THÈSE PRÉSENTÉE
POUR OBTENIR LE GRADE DE

DOCTEUR DE L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE MATHÉMATIQUES ET INFORMATIQUE

SPÉCIALITÉ : INFORMATIQUE

par **Edern LE BOT**

Méthodes d'apprentissage profond pour la parcellisation fine du cerveau à partir d'IRM FLAIR

Deep Learning Methods for Fine-grained Parcellation of the Brain from FLAIR MRIs

Sous la direction de :
Pierrick Coupé

Soutenue le 17 décembre 2025

Devant la comission d'examen composée de :

M. Jean-Philippe Domenger	Professeur des universités, Université de Bordeaux	Président
M. Pierrick Coupé	Directeur de Recherche, CNRS	Directeur
Mme Victoria Bourgeais	Maîtresse de conférences, Université de Bordeaux	Examinatrice
Mme Caroline Petitjean	Professeure des Universités, Université de Rouen Normandie (Saint-Étienne-du-Rouvray)	Examinatrice
M. Pierre-Henri Conze	Maître de conférences, IMT Atlantique (Brest)	Rapporteur
M. Marc-Emmanuel Bellemare	Maître de conférences, Université d'Aix-Marseille, Polytech' Marseille	Rapporteur
M. Rémi Giraud	Maître de conférences, Bordeaux INP	Invité
M. Boris Mansencal	Ingénieur de Recherche, Bordeaux INP	Invité

Table of Contents

Introduction **10**

1 MRI Fundamentals and Clinical Context **16**
1 Magnetic Resonance Imaging 17
1.A Clinical Context 18
1.B Image Acquisition 18
1.C Clinical MRI Sequences 19
1.D Sources of Variability 20
2 Neurological Conditions 22
3 MRI Acquisition Protocols and Guidelines 25
4 Scientific Challenges 26

2 Literature Review of Brain Segmentation Methods **29**
1 Common Regions of Interests 30
2 Semantic Segmentation Methods for Brain MRI 32
2.A Traditional Brain Segmentation 32
2.B Deep Learning Segmentation Methods 33
2.C Modality Agnostic Methods 38
3 Evaluation Metrics and Training Objectives 39
3.A Segmentation Metrics 39
3.B Training Objectives and Loss Functions 41
4 Current Limitations and Research Gaps 41

3 FLAIR-ICE: Intracranial Cavity Extraction from FLAIR MRI **45**
1 Introduction 46
2 Materials 48
2.A Augmentations 51
2.B Training configuration 53
3 Results 54
3.A Quantitative Evaluation 54
3.B Qualitative analysis 56
4 Conclusion 57

4 FLAIRBrainSeg: Fine-Grained Brain Segmentation Using FLAIR MRI Only **59**
1 Introduction 60
2 Materials & Methods 61

2.A Datasets . . . 61
2.B Preprocessing . . . 62
2.C T1w-based segmentation . . . 62
2.D Quality control . . . 63
2.E Deep learning model description . . . 65
2.F Validation framework . . . 66
3 Results . . . 66
4 Conclusion . . . 69

5 FLAIRBrainSeg^{+}: Generalized Anatomical Segmentation on Pathological FLAIR MRI 72
1 Introduction . . . 73
2 Methods . . . 76
2.A Synthetic Data Generation . . . 76
2.B Training Configuration: . . . 79
2.C Validation Framework . . . 79
3 Materials . . . 81
3.A Datasets . . . 83
4 Results . . . 85
5 Conclusion . . . 92

6 Conclusion and Perspectives 95
1 Summary of Findings . . . 96
2 Perspectives . . . 97

Appendixes 101
1 Appendix for Chapter 4 . . . 101
2 Appendix for Chapter 5 . . . 103

Bibliography 107

List of Figures 118

List of Tables 123

L'imagerie par résonance magnétique est aujourd'hui un outil incontournable pour l'étude du cerveau et le diagnostic de nombreuses maladies neurologiques. Elle permet d'observer la structure cérébrale et les anomalies de manière non invasive. Parmi les différentes séquences d'acquisition, la séquence FLAIR joue un rôle essentiel pour la détection des lésions cérébrales. Ces images sont largement utilisées dans la pratique clinique, notamment pour des pathologies telles que la sclérose en plaques, les tumeurs cérébrales ou les accidents vasculaires cérébraux. Cependant, leur analyse automatique demeure difficile en raison du faible contraste entre les tissus, du rapport signal/bruit limité et de la variabilité entre appareils d'imagerie.

En neuroimagerie, l'analyse quantitative du cerveau repose de plus en plus sur des outils de segmentation automatique, capables d'identifier les structures anatomiques et les régions pathologiques. La majorité de ces méthodes a été conçue pour des images T1w, considérées comme la référence pour la segmentation anatomique. Or, cette séquence n'est pas systématiquement acquise dans les protocoles cliniques, ce qui limite l'utilisation des outils existants. Dans de nombreux cas, la séquence FLAIR est la seule disponible, mais les approches classiques, conçues pour la T1w, échouent à fournir des résultats fiables.

Cette thèse s'inscrit dans ce contexte et explore la possibilité de réaliser une segmentation fine du cerveau à partir d'images FLAIR uniquement. L'objectif est de développer des méthodes robustes, capables d'extraire des informations anatomiques et pathologiques pertinentes à partir de cette séquence, tout en s'adaptant à la variabilité des acquisitions et à la présence de lésions. Les travaux menés ont conduit au développement d'outils spécifiquement adaptés à la FLAIR, destinés à améliorer les performances des étapes de prétraitement et de segmentation dans un cadre clinique réaliste.

Les résultats obtenus montrent qu'il est possible d'obtenir des segmentations cérébrales fines et cohérentes à partir de FLAIR seule, ouvrant ainsi la voie à l'exploitation de vastes bases de données cliniques jusque-là sous-utilisées. Ces approches démontrent également l'intérêt de concevoir des modèles spécialisés par modalité plutôt que des méthodes génériques indépendantes de la séquence. À terme, les outils développés pourraient être intégrés à la plateforme en ligne *volBrain* (`https://volbrain.net`) afin de faciliter leur diffusion auprès de la communauté scientifique et clinique.

Mots-Clés : IRM FLAIR, Segmentation cérébrale, Apprentissage profond, Modèle spécifique à la modalité

MÉTHODES D'APPRENTISSAGE PROFOND POUR LA PARCELLISATION FINE DU CERVEAU À PARTIR D'IRM FLAIR

Magnetic Resonance Imaging (MRI) is a key tool for studying the brain and diagnosing numerous neurological diseases. It enables non-invasive observation of brain structure and abnormalities. Among MRI sequences, the Fluid-Attenuated Inversion Recovery (FLAIR) sequence plays a crucial role in detecting brain lesions and is widely used in clinical practice, particularly for disorders such as multiple sclerosis, brain tumors, and stroke. However, automatic analysis of FLAIR images remains challenging due to low tissue contrast, reduced signal-to-noise ratio, and variations between scanners and acquisition protocols.

In neuroimaging, quantitative brain analysis increasingly relies on automated segmentation tools that can identify anatomical structures and pathological regions. Most existing segmentation methods have been developed for T1w MRI, which is considered the standard reference for anatomical segmentation. However, T1w images are not always acquired in clinical protocols, which limits the applicability of existing tools. In many cases, FLAIR is the only available sequence, but conventional T1w-based approaches fail to produce reliable segmentations under such conditions.

This thesis addresses this limitation by investigating the feasibility of performing fine-grained brain segmentation using FLAIR MRI alone. The goal is to design robust methods capable of extracting meaningful anatomical and pathological information from FLAIR data, while remaining resilient to acquisition variability and the presence of lesions. The proposed work led to the development of tools specifically adapted to FLAIR imaging, improving the robustness of both preprocessing and segmentation steps in realistic clinical settings.

The results demonstrate that fine-grained and reliable brain segmentations can be obtained from FLAIR MRI alone, enabling the use of large clinical datasets that are often underexploited. These findings also highlight the advantages of modality-specific approaches over generic, modality-agnostic methods. In the long term, the tools developed in this work could be integrated into the volBrain online platform (https://volbrain.net) to facilitate their dissemination to the scientific and clinical communities.

KeyWords : FLAIR MRI, Brain Segmentation, Deep Learning, Modality-Specific Imaging, Modality-Specific Models

---

DEEP LEARNING METHODS FOR FINE-GRAINED PARCELLATION OF THE BRAIN FROM FLAIR MRIs

Resumé étendu en français

L'imagerie par résonance magnétique est aujourd'hui un outil essentiel pour l'étude du cerveau et la détection de nombreuses maladies neurologiques. Elle permet d'observer la structure cérébrale et les anomalies de manière non invasive. L'augmentation du volume de données issues de l'IRM rend nécessaire le développement de méthodes automatiques pour en faciliter l'analyse.

En neuroimagerie, l'analyse quantitative de la structure cérébrale et des maladies repose de plus en plus sur des outils de segmentation automatique. Ces outils permettent d'extraire des informations à partir d'images obtenues par Imagerie par Résonance Magnétique sur l'anatomie, notamment sur la taille et la localisation des différentes structures du cerveau, mais aussi sur certaines maladies, caractérisées par l'apparition de lésions ou de cellules cancéreuses, que les outils de segmentation permettent d'identifier.

L'application des outils de segmentation automatique à de larges bases de données d'images permet de réaliser des études pour mieux comprendre les mécanismes de certaines maladies neurologiques. Néanmoins, la plupart des méthodes de segmentation existantes ont été conçues pour traiter des images T1w volumétriques, un certain type d'acquisition spécifique. Or, les pratiques cliniques peuvent ne pas toujours inclure systématiquement cette séquence. Dans ce cas, les outils existants ne peuvent pas être appliqués. De fait, en l'absence de disponibilité de ces images T1w, il n'est pas possible d'extraire des informations volumétriques sur ces sujets. La situation est particulièrement manifeste dans des contextes tels que le suivi de la sclérose en plaques où, dans le suivi des patients, si l'image T1w est acquise, elle peut l'être de manière non volumétrique, à partir de laquelle il est difficile de définir des délinéations entre les structures avec une plus faible résolution spatiale.

Dans certains de ces contextes, où l'image T1w n'est pas acquise, d'autres séquences et types d'images peuvent être disponibles et préférés par les cliniciens. Une séquence en particulier, la FLAIR (Fluid-Attenuated Inversion Recovery), est utilisée pour l'identification de lésions cérébrales dans certains de ces cas. Les images FLAIR permettent de mettre en évidence les lésions cérébrales, visible sous forme d'hyperintensité dans l'image. Cependant, malgré son importance clinique, la segmentation anatomique fine à partir de ces images reste très peu explorée, en raison du faible contraste entre les tissus, du rapport signal/bruit réduit et de la présence plus fréquente d'artefacts par rapport aux autres séquences.

Additionnellement, les outils existants de segmentation anatomique sont souvent conçus à partir de bases de données contenant uniquement des sujets sains. En l'absence de variations telles que l'apparition de lésions de sclérose en plaques ou de tissus cancéreux dans ces bases de données, ces outils peuvent aussi produire

des erreurs dans l'interprétation des images. Ces limitations sont importantes, car la majorité des acquisitions d'IRM sont faites sous suspicion de la présence d'une maladie neurologique.

Cette thèse vise à étudier la possibilité de développer des méthodes de segmentation fine du cerveau à partir d'images FLAIR uniquement, en se concentrant sur les situations et les scénarios où les images T1w volumétriques peuvent potentiellement ne pas être acquises. Dans ce cas, les méthodes doivent être développées pour prendre en compte la présence d'hyperintensités sur l'image, liées à la présence de la maladie.

Extraction robuste de la cavité intracrânienne à partir de données FLAIR

L'application d'outils d'analyse d'image tels que la segmentation est souvent précédée par des étapes de prétraitement de l'image, destinées à faciliter l'extraction d'informations en améliorant la qualité de l'image, notamment en présence de bruit. Une de ces étapes, l'extraction du cerveau (ou *skull-stripping*), vise à extraire et différencier le volume cérébral des tissus non cérébraux. Cette étape est essentielle, car elle permet de restreindre l'analyse aux régions d'intérêt et d'éviter que la présence des tissus non cérébraux impacte les autres étapes de traitement. En particulier, la qualité des résultats des outils d'analyse peut être dégradée dans le cas où l'extraction du cerveau est inexacte.

Les outils existants d'extraction du cerveau sont souvent conçus en premier lieu pour le traitement d'images T1w. Ces outils peuvent ne pas être particulièrement adaptés à certains contextes, notamment en présence d'hyperintensités ou de tissus cancéreux dans l'image, qui peuvent conduire à des confusions dans l'étape d'extraction. Un autre contexte où ces outils peuvent se révéler inadaptés est celui du traitement d'images prétraitées par des procédures d'anonymisation, telles que le *defacing* ou un *skull-stripping*. Ces étapes, parfois appliquées avant le partage des données pour protéger l'identité des patients, peuvent conduire à d'autres erreurs dans l'extraction du cerveau.

La première contribution de ce manuscrit porte sur le développement d'une méthode adaptée aux images FLAIR, appelée FLAIR-ICE, conçue pour maintenir des performances robustes en présence d'hyperintensités mais aussi de potentiels prétraitements d'anonymisation appliqués aux images. Nous comparons les performances de cet outil à SynthStrip, une méthode d'extraction du cerveau adaptée pour le traitement de plusieurs types de modalités. Nos résultats montrent que FLAIR-ICE permet d'obtenir une extraction du cerveau fiable à partir d'images FLAIR dans plusieurs contextes, en évaluant la qualité de l'extraction sur plusieurs bases de données différentes et dans différents contextes. FLAIR-ICE permet de construire un *pipeline* de prétraitement adapté aux images FLAIR.

Segmentation anatomique fine du cerveau à partir de FLAIR uniquement

La seconde contribution de ce manuscrit introduit une méthode de segmentation fine du cerveau à partir de l'IRM FLAIR. En l'absence de disponibilité d'images

FLAIR annotées, la méthode proposée, baptisée FLAIRBrainSeg, est conçue pour reproduire une segmentation anatomique fine obtenue à partir de séquences T1w à l'aide d'outils automatiques, mais en utilisant l'image FLAIR seulement. La capacité de FLAIRBrainSeg à reproduire l'information extraite de la T1w est évaluée en comparaison avec SynthSeg, une méthode de segmentation "agnostique à la modalité". Nos résultats montrent que FLAIRBrainSeg est en mesure de produire des segmentations anatomiques fidèles à celles obtenues à partir de la T1w, y compris en présence d'hyperintensités liées à la maladie de la sclérose en plaques.

Segmentation jointe de l'anatomie et des anomalies pathologiques dans les IRM FLAIR

L'utilisation des outils de segmentation anatomique et de segmentation d'hyperintensités ou de tissus cancéreux est souvent traitée séparément. Cette séparation est justifiée par une préférence pour l'utilisation de la T1w dans la segmentation anatomique et pour d'autres modalités, comprenant la FLAIR, pour la détection d'hyperintensités. Les outils sont utilisés dans des objectifs différents, visant soit à identifier des régions d'atrophie par des analyses volumétriques sur les structures, soit dans un but de détection et d'identification d'hyperintensités dans l'image liées à la présence de maladie.

Cette séparation ne permet pas de produire une vision complète de l'impact d'une maladie ou d'une affliction, car ces maladies sont souvent caractérisées par ces deux aspects, comme par exemple dans le cas de la sclérose en plaques. Aussi, la localisation des régions hyperintenses dans le cerveau, relative aux régions fonctionnelles, est intéressante pour attribuer et comprendre les fonctions cérébrales impactées par la maladie.

La troisième contribution de ce manuscrit propose de concevoir un modèle unifié pour la segmentation simultanée de l'anatomie et des hyperintensités, marqueurs de certaines maladies. La méthode proposée, baptisée FLAIRBrainSeg+, est entraînée à partir d'images augmentées, à l'aide d'un processus d'insertion de lésions provenant de différentes images annotées dans des images saines préalablement segmentées. Ce processus permet de générer des exemples annotés contenant des lésions variées avec une segmentation anatomique correspondante. Les performances de FLAIRBrainSeg+ sont comparées à celles de SynthSeg-WMH, une méthode de segmentation "agnostique à la modalité" visant à améliorer SynthSeg en permettant la segmentation de certaines hyperintensités. FLAIRBrainSeg+ permet d'obtenir une segmentation de l'anatomie et d'hyperintensités fiable dans différents contextes, y compris en présence de lésions de sclérose en plaques et de tissus cancéreux.

En conclusion, cette thèse propose un ensemble de méthodes visant à exploiter les séquences IRM FLAIR pour la production de segmentations fines du cerveau. Les résultats obtenus semblent mettre en évidence l'intérêt de développer des outils et des méthodes spécifiques à la modalité, en raison des performances supérieures obtenues

dans ce manuscrit en comparaison aux méthodes "indépendantes de la modalité". Ces applications ouvrent de nouvelles perspectives quant à l'utilisation de différents types d'images pour des études volumétriques. À terme, les outils développés dans le cadre de cette thèse pourraient être intégrés à la plateforme en ligne *volBrain* (`https://volbrain.net`) afin de faciliter leur diffusion auprès de la communauté scientifique et clinique.

Des perspectives supplémentaires d'amélioration et d'extension des méthodes restent ouvertes, notamment pour améliorer la robustesse des modèles à des acquisitions provenant de différents centres ou en présence de différents types de maladies affectant le cerveau, les strategies d'augmentation et de génération d'exemple introduite dans ce manuscrit n'étant pas limitées aux maladies étudiées.

Extended Abstract

Magnetic Resonance Imaging (MRI) is an essential tool for studying the brain and detecting numerous neurological diseases. It enables non-invasive observation of brain structure and abnormalities. The growing volume of MRI data highlights the need for automated methods capable of facilitating quantitative analysis.

In neuroimaging, quantitative analysis of brain structure and pathology increasingly relies on automated segmentation tools. These tools extract anatomical and pathological information from MRI scans, such as the size and location of brain regions or the presence of lesions and tumors. Applying such tools to large imaging databases enables the study of neurological mechanisms on a population scale. However, most existing segmentation methods have been designed for volumetric T1w MRI, a specific type of acquisition that is not always included in clinical protocols. When T1w images are unavailable, these tools cannot be applied, preventing the extraction of volumetric information for many patients. This situation is particularly common in disorders such as multiple sclerosis, where T1w images are often missing or acquired in non-volumetric form, limiting their use for precise structural delineation.

In these contexts, other MRI sequences may be available and are often preferred by clinicians. Among them, the Fluid-Attenuated Inversion Recovery (FLAIR) sequence is widely used for visualizing brain lesions. FLAIR imaging highlights pathological regions as hyperintense signals, but fine-grained anatomical segmentation from FLAIR remains poorly explored due to its low tissue contrast, reduced signal-to-noise ratio, and frequent imaging artifacts. Moreover, most existing anatomical segmentation tools are trained on datasets containing only healthy subjects. When applied to images showing lesions or tumors, they can produce errors or misinterpretations. These limitations are important, as the majority of MRI scans acquired in clinical settings involve suspected neurological disease.

This thesis investigates the feasibility of performing fine-grained brain segmentation using FLAIR MRI alone, focusing on scenarios in which volumetric T1w images are unavailable. The goal is to develop approaches that can account for the presence of disease-related hyperintensities and still provide reliable anatomical information.

Robust Intracranial Cavity Extraction from FLAIR MRI

Image analysis tasks such as segmentation are often preceded by preprocessing steps designed to improve image quality and facilitate feature extraction. One such step, brain extraction (or skull stripping), isolates the brain from surrounding non-brain tissues. This operation is critical, as errors at this stage can propagate to later processing steps.

Most existing brain extraction tools were developed for T1w MRI. As such, these tools may fail and produce errors in cases with strong hyperintensities. Another possible causes for failure can be the processing of pre-processed images for anonymization (defaced or pre-stripped) images. These steps, sometimes applied before data sharing

to protect patient identity, can further complicate brain extraction.

The first contribution of this thesis introduces a FLAIR-specific brain extraction method, FLAIR-ICE, designed to maintain robust performance across heterogeneous data, including images affected by pathology or prior anonymization. We compare its performance to state-of-the-art public tools and demonstrate reliable extraction quality across multiple datasets. FLAIR-ICE establishes a consistent preprocessing pipeline for FLAIR-based segmentation.

Fine-Grained Anatomical Brain Segmentation Using FLAIR MRI Only

The second contribution of this work introduces FLAIRBrainSeg, a method for fine-grained brain segmentation using only FLAIR MRI. In the absence of annotated FLAIR data, the method is trained to reproduce anatomical segmentations derived from T1w MRI, but using FLAIR images as input. FLAIRBrainSeg is compared to SynthSeg, a public modality-agnostic segmentation method, and achieves superior performance, demonstrating that accurate fine-grained anatomical segmentation is feasible from FLAIR alone, even in the presence of disease-related hyperintensities. To our knowledge, this represents the first approach specifically designed for fine-grained anatomical segmentation from FLAIR MRI.

Unified Segmentation of Brain Anatomy and Pathology in FLAIR MRI

The use of anatomical and pathological segmentation tools is often treated separately. This separation arises from the conventional use of T1w MRI for anatomical segmentation and other modalities, including FLAIR, for lesion detection. Each type of segmentation is typically used for a distinct purpose, either to identify structural atrophy through volumetric analyses or to detect and characterize hyperintensities associated with neurological disease.

However, this separation prevents the generation of a comprehensive view of the overall impact of disease on the brain, since many neurological conditions combine both structural and pathological alterations. Multiple sclerosis, for example, involves both focal lesions and diffuse atrophy. Moreover, the spatial relationship between hyperintense regions and functional brain areas provides valuable information for understanding how disease affects specific neural functions.

The third contribution of this thesis proposes a unified model for the simultaneous segmentation of brain anatomy and hyperintensities, which serve as markers of various neurological diseases. The proposed method, FLAIRBrainSeg+, is trained using an augmentation process that inserts annotated lesions from multiple sources into previously segmented healthy images. This approach generates diverse and realistic examples containing both lesions and corresponding anatomical segmentations. The performance of FLAIRBrainSeg+ is compared with SynthSeg-WMH, a modality-agnostic segmentation method extending SynthSeg to include white matter hyperintensity segmentation. FLAIRBrainSeg+ achieves accurate and reliable

segmentation of both anatomy and hyperintensities across different datasets and conditions, including multiple sclerosis and brain tumors.

To conclude, this thesis presents a set of methods designed to exploit FLAIR MRI sequences for fine-grained brain segmentation. The results highlight the benefit of developing modality-specific methods, which outperform generic, modality-agnostic approaches. These tools open new perspectives for the volumetric analysis of MRI data acquired in clinical settings. Furthermore, the methods developed in this work could be integrated into the **volBrain** online platform (`https://volbrain.net`)

Additional perspectives for improving and extending these methods remain open, particularly to enhance the robustness of the models to data acquired from different centers or in the presence of various types of brain diseases. The augmentation and example generation strategies introduced in this thesis are not limited to the pathologies studied.

Introduction

Medical imaging plays a central role in modern healthcare by enabling non-invasive visualization of internal structures for both diagnosis and treatment planning. Among available modalities, Magnetic Resonance Imaging (MRI) is particularly valuable, as it provides high-resolution, three-dimensional images without exposing patients to ionizing radiation. By adjusting acquisition parameters, radiologists can obtain different sequences that emphasize specific tissue properties and reveal distinct biological or pathological features. In clinical practice, this flexibility allows MRI to serve a wide range of diagnostic and monitoring purposes.

Over the past five decades, MRI has become an essential clinical tool, leading to the generation of vast amounts of imaging data. To process these data, numerous image processing methods have been developed to transform raw MRI scans into standardized, interpretable representations. These tools aim to transform raw MRI scans into more interpretable forms, useful for both clinical and research applications.

For this purpose, Artificial Intelligence (AI) and Deep Learning (DL) have been increasingly adopted for medical image analysis, including neuroimaging. Such tools assist radiologists and clinicians in diverse tasks such as detection of abnormal brain changes (*e.g.*, brain tumors, Multiple Sclerosis (MS), or stroke lesions), disease progression monitoring (*e.g.*, cortical thickness estimation for Alzheimer's Disease (AD) or cognitive impairment), and surgical planning. A key task among these applications is *segmentation*, which refers to the process of identifying and delineating brain regions within an image.

Segmentation provides a foundation for most quantitative neuroimaging analyses. By defining boundaries between tissues, segmentation enables volumetric and morphometric measurements such as brain tissue volumes, lesion counts, and cortical thickness estimates. However, current AI-based segmentation models face significant challenges in clinical use. Most deep learning segmentation approaches rely on supervised learning, where models are trained on curated datasets of images and corresponding expert annotations. Producing such manual annotations is extremely time-consuming, especially for volumetric data, as it requires radiologists to assign labels to each voxel of MRI volumes. While tools for the production of semi-automatic annotations exist, they still require manual validation, resulting in similar time consumption. This results in small quantities of annotated data being available for training. As such, the data may not fully represent the diversity of scanner types, acquisition settings, and possible disease-related appearance changes. Under these conditions, supervised DL models can yield substantial errors when applied to unseen data, called "hallucinations" that compromise downstream analyses.

For this reason, most existing segmentation tools are restricted to a single MRI

modality, typically the T1w sequence. This sequence offers high contrast between gray matter, white matter, and cerebrospinal fluid, making it ideal for detailed anatomical segmentation. However, in clinical diagnosis and follow-up, radiologists are often more interested in detecting anomalies related to disease rather than normal anatomy. As a result, clinical protocols frequently include sequences designed to highlight pathological changes. Among these, the Fluid-Attenuated Inversion Recovery (FLAIR) sequence is one of the most widely used, particularly in the assessment of Multiple Sclerosis, stroke, brain tumors, and traumatic brain injury.

Despite its clinical importance, segmentation tools developed for T1w MRI cannot be directly applied to FLAIR data due to differences in tissue contrast, intensity distributions, and image appearances. Furthermore, because most anatomical segmentation datasets contain only healthy T1w images, existing DL models trained on this data struggle with three main limitations:

(1) Models generalize poorly to the presence of hyperintensities related to pathologies not present in healthy populations.

(2) Models struggle with changes in tissue contrast and image appearances specific to FLAIR.

(3) Models are more vulnerable to noise and artifacts found in clinical FLAIR scans.

The growing clinical reliance on FLAIR has led to increasing interest in tools capable of robust FLAIR-based segmentation, as illustrated by recent commercial releases such as QP-Brain. The work presented in this thesis contributes to this direction by providing scientifically transparent, reproducible, and publicly accessible segmentation tools through the volBrain[1] platform.

Problem Statement

This thesis addresses the challenge of performing accurate brain segmentation from FLAIR MRI alone, with a particular emphasis on clinical images that contain pathological abnormalities and lack other structural contrasts.

FLAIR presents unique segmentation challenges: reduced gray-white matter contrast, high variability across acquisition settings, and artifact-related signal distortions. Most critically, clinical FLAIR scans often contain pathological abnormalities such as lesions, since MRI is typically performed under suspicion of an underlying neurological condition. In this context, it is essential not only to detect lesions but also to accurately segment normal brain structures to enable reliable volumetric analysis, as changes in tissue volumes (e.g., atrophy or ventricular enlargement) play a key role in diagnosis, monitoring, and treatment planning across many neurological diseases.

However, as mentioned above, most existing segmentation tools for volumetry are only applicable to volumetric T1w images, the preferred modality for semantic

[1] `https://volbrain.net`

segmentation, as it offers clearer anatomical contrast. As a result, these tools are unusable in scenarios where only FLAIR is available.

The central hypothesis of this work is that segmentation performance on FLAIR can be significantly improved beyond modality-agnostic approaches by tailoring model design, training strategies, and data augmentation specifically to FLAIR characteristics, rather than relying on generic or modality-agnostic frameworks.

Contributions

This thesis focuses on the development of methods for the segmentation of FLAIR MRI, targeting scenarios in which FLAIR is the only available volumetric anatomical sequence. The main contributions are summarized below:

- **C1. FLAIR-ICE:** A preprocessing pipeline specifically designed for FLAIR MRI, providing robust brain extraction and intensity normalization across heterogeneous clinical acquisitions. It ensures reliable preparation of data containing artifacts, lesions, and variable coverage, forming a stable foundation for downstream segmentation.
- **C2. FLAIRBrainSeg:** A fine-grained segmentation framework that delineates anatomical structures and pathological regions from FLAIR alone. The model incorporates lesion-aware augmentation and FLAIR-specific training strategies, demonstrating that anatomical segmentation comparable to T1w methods can be achieved from FLAIR only.
- **C3. FLAIRBrainSeg$^+$:** An integrated, end-to-end clinical pipeline capable of simultaneously segmenting brain anatomy and FLAIR hyperintensities (lesions and tumors). The method is fully containerized and publicly deployed through the volBrain[2] platform, promoting accessibility and reproducibility in clinical neuroimaging.

Publications and Software

The following outputs have resulted from the work conducted in this thesis. *Entries shown in gray correspond to work currently under review or in submission.*

Journal

Edern Le Bot, Rémi Giraud, Boris Mansencal, Thomas Tourdias, José V. Manjón, and Pierrick Coupé (2025). *FLAIRBrainSeg$^+$: Unified Fine-Grained Brain and Lesion Segmentation from FLAIR MRI Only.* Submitted to *Journal of Machine Learning for Biomedical Imaging (MELBA).*

[2] `https://volbrain.net`

Conference

Edern Le Bot, Rémi Giraud, Boris Mansencal, Thomas Tourdias, José V. Manjón, and Pierrick Coupé (2025). *FLAIRBrainSeg: Fine-Grained Brain Segmentation Using FLAIR MRI Only.* In *Proceedings of the Medical Imaging with Deep Learning (MIDL 2025)*, Salt Lake City, USA, July 2025. *Oral presentation.*

National Conference

Edern Le Bot, Rémi Giraud, Boris Mansencal, Thomas Tourdias, José V. Manjón, and Pierrick Coupé (2025). *FLAIRBrainSeg: Fine-Grained Brain Segmentation Using FLAIR MRI Only.* In *Colloque Français d'Intelligence Artificielle en Imagerie Biomédicale*, Nice, France, Mars 2025. *Poster presentation.*

Software

Edern Le Bot, Rémi Giraud, Boris Mansencal, Thomas Tourdias, José V. Manjón, and Pierrick Coupé (2025). *FLAIRBrainSeg^{+}. Under deposit.*

Other Publications

Arthur Longuefosse, Edern Le Bot, Baudouin Denis de Senneville, Rémi Giraud, Boris Mansencal, Pierrick Coupé, Pascal Desbarats, and Fabien Baldacci (2024). *Adapted nnU-Net: A Robust Baseline for Cross-Modality Synthesis and Medical Image Inpainting.* In *International Workshop on Simulation and Synthesis in Medical Imaging (SASHIMI 2024)*, MICCAI 2024 Workshop, Springer.

José V. Manjón, Sergio Morell-Ortega, Marina Ruiz-Perez, Boris Mansencal, Edern Le Bot, Marien Gadea, Enrique Lanuza, Gwenaelle Catheline, Thomas Tourdias, Vincent Planche, *et al.* (2025). *Ultra-high Resolution Multimodal MRI Dense Labelled Holistic Brain Atlas.* Submitted to *Scientific Reports.*

Impact Statement

This thesis focuses on developing segmentation methods designed specifically for FLAIR MRI. The main goal is to make deep learning tools for brain analysis usable in clinical situations where only FLAIR scans are available and other MRI sequences are missing. The work presents a complete set of methods for FLAIR MRI segmentation, including preprocessing, brain extraction, and detailed structural segmentation. It also highlights the need for methods that remain reliable on clinical data, where disease-related changes are often present. This is important because the diagnosis and follow-up of many neurological conditions require information about both lesion location and size, as well as measurements of normal brain structures. To address this need, we also propose a unified model that can segment both anatomical structures and pathological regions directly from FLAIR images.

Thesis Outline

This thesis is organized into six chapters, following the conceptual and methodological trajectory of the work, from the motivation for FLAIR-based brain analysis to the development of dedicated segmentation frameworks and their integration into clinical pipelines.

Chapter 1
: introduces the physical principles of Magnetic Resonance Imaging (MRI) and discusses the clinical relevance of the FLAIR sequence in neuroimaging. It details the sources of image variability across scanners and acquisition protocols, emphasizing how these factors affect image contrast and signal consistency. The chapter outlines the challenges of performing quantitative brain analysis when only FLAIR is available and motivates the need for segmentation frameworks capable of operating under clinical constraints where other contrasts (including volumetric T1w) can be missing.

Chapter 2
: reviews the state of the art in brain segmentation, covering traditional atlas-based and Bayesian probabilistic frameworks as well as modern deep learning approaches. It analyzes how these methods typically rely on the structural contrast provided by T1w images and the difficulties that arise when such models are applied directly to FLAIR data. The review also examines previous attempts at lesion and pathology segmentation, establishing the methodological foundations for the approaches proposed in this thesis.

Chapter 3
: presents a brain-extraction pipeline specifically designed for FLAIR MRI. The method addresses typical clinical challenges such as inhomogeneous intensity profiles, acquisition artifacts, and the presence of large lesions. This component, named **FLAIR-ICE (contribution C1)**, is a pre-processing step for subsequent segmentation by ensuring accurate skull removal and intensity standardization across diverse clinical acquisitions.

Chapter 4
: presents the first core contribution of this work: a fine-grained anatomical segmentation model operating exclusively on FLAIR images. The proposed approach, **FLAIRBrainSeg (contribution C2)**, delineates detailed tissue classes and substructures, including cortical and subcortical regions, ventricles. This chapter investigates how a model explicitly optimized for FLAIR characteristics compares with modality-agnostic segmentation frameworks. It further demonstrates that the anatomical delineations typically achievable from T1w scans can be faithfully reproduced from FLAIR alone, provided that the model architecture and training are adapted to this modality's specific properties.

Chapter 5
: extends the segmentation framework to include pathological regions alongside

normal anatomy. This chapter introduces **FLAIRBrainSeg$^+$ (contribution C3)**, a unified model for the simultaneous segmentation of FLAIR hyperintensities (including white matter lesions and tumorous abnormalities) and healthy brain structures. The approach leverages lesion-aware data augmentation and multi-scale representations to handle diverse pathological appearances. This joint segmentation paradigm enables integrated volumetric analyses that combine anatomical and pathological information within a single FLAIR-only workflow.

Chapter 6

concludes the manuscript by summarizing the main findings and contributions. It discusses the implications of enabling accurate brain segmentation from FLAIR alone, including improved accessibility of automated volumetric tools for clinical datasets. The chapter also outlines current limitations and proposes directions for improvements and future research, such as handling other pathologies and improving cross-scanner robustness.

Chapter 1

MRI Fundamentals and Clinical Context

1 Magnetic Resonance Imaging 17
1.A Clinical Context 18
1.B Image Acquisition 18
1.C Clinical MRI Sequences 19
1.D Sources of Variability 20
2 Neurological Conditions 22
3 MRI Acquisition Protocols and Guidelines 25
4 Scientific Challenges 26

1 Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI) enables non-invasive visualization of brain anatomy and pathology by using diverse contrast mechanisms to highlight different tissue properties. Among these, the FLAIR sequence plays a critical role in clinical neuroimaging due to its sensitivity to white matter abnormalities such as lesions associated with Multiple Sclerosis (MS), stroke, tumors, and Traumatic Brain Injury (TBI). FLAIR's ability to suppress cerebrospinal fluid (CSF) signal enhances the visibility of periventricular and cortical lesions, making it a standard component in clinical protocols for disease detection and monitoring.

Segmentation refers to the delineation of anatomical and pathological structures within MRI scans and is a foundational task in neuroimaging. Segmentation is required for numerous downstream applications, including volumetric analysis, lesion quantification, disease progression tracking, and surgical planning. Accurate segmentation supports evidence-based medical decision-making and enables population-level research. Traditionally, segmentation pipelines have relied on T1w images due to their high contrast between gray and white matter, clear anatomical boundaries, and widespread availability in research datasets.

However, as clinical practices evolve toward faster, pathology-focused protocols, many scenarios (particularly in MS monitoring or stroke management) omit T1w sequences to reduce scan duration or because FLAIR alone provides sufficient diagnostic information. Consequently, large volumes of clinical and retrospective data lack the T1w inputs assumed by most existing segmentation methods. This mismatch limits the real-world applicability of current tools and creates a pressing need for segmentation approaches that operate directly and robustly on FLAIR images.

Existing cross-modality methods such as SynthSeg [Billot et al., 2023a], while effective in controlled settings, often struggle to reproduce the pathological variations present in real-world FLAIR images acquired in clinical environments. Efforts to overcome these limitations are generally restricted to lesions with "common" intensity profiles [Laso et al., 2024]. This gap underscores the limitations of current pipelines, emphasizing the importance of modality-specific, pathology-aware approaches. It also motivates the content of this thesis.

This chapter introduces the physical principles, clinical protocols, and sequence-specific characteristics of MRI relevant to neuroimaging, with a focus on the FLAIR sequence. Despite FLAIR's widespread clinical use, particularly in aging, MS, stroke, and neurovascular disorders, robust segmentation methods tailored to this modality remain limited. Existing tools, even those designed for cross-modality generalization, are not well-suited for FLAIR and frequently exhibit failure cases, especially in the conditions highlighted above. By introducing the fundamental concepts underlying this work, this chapter provides the context for the segmentation approaches developed in later chapters.

1.A Clinical Context

MRI is a non-invasive technique that generates detailed images of internal body structures. It provides high soft-tissue contrast by exploiting the magnetic properties of hydrogen nuclei (^{1}H) in water and lipid molecules. Unlike computed tomography (CT) and X-ray imaging, MRI uses non-ionizing radiofrequency (RF) energy instead of potentially harmful radiation, making it safer for pediatric patients, pregnant women, and cases requiring repeated imaging [ICNIRP, 2017]. Many MRI studies can be performed without contrast agents, as the technique inherently differentiates tissues based on their magnetic properties [Le Bihan, 2012]. These advantages have made MRI a primary tool for many clinical applications, despite its higher cost and longer acquisition times compared to CT and conventional radiography [Dill, 2008].

MRI is essential in neuroscience for visualizing brain anatomy and assessing neurological conditions [Brown et al., 2014]. Common applications include the detection of ischemic stroke [Lansberg et al., 2000], brain tumors [Ellingson et al., 2015], demyelinating diseases [Filippi and Rocca, 2011], and neurodegenerative disorders [Jack Jr. et al., 2008]. In oncology, MRI supports initial detection, treatment planning, response monitoring, and follow-up care [Ellingson et al., 2015]. By adjusting acquisition parameters, MRI produces various tissue contrasts that enhance diagnostic capabilities across both clinical and research applications in brain imaging [Haacke et al., 2014].

1.B Image Acquisition

MRI images are generated by applying a strong magnetic field B_0, typically ranging from 1.5–3.0 T in clinical settings and up to 7–11.5 T in research environments. Under these conditions, hydrogen nuclei within the body align either parallel or anti-parallel to B_0, with a slight excess favoring the lower-energy parallel orientation. This alignment creates a net longitudinal magnetization [Brown et al., 2014].

An RF pulse disrupts this alignment, rotating nuclei away from their original orientation by 90° or 180° [Brown et al., 2014]. When the pulse ceases, nuclei realign with B_0, emitting an RF signal known as Free Induction Decay (FID). Receiver coils capture this magnetic resonance signal as raw data for image reconstruction. The MRI signal populates k-space, a frequency-domain representation that is transformed into the final image via the inverse Fourier transform $\mathcal{F}^{-1}$ [Haacke et al., 2014]. Different tissues generate distinct signals based on their relaxation properties, creating contrast that reveals anatomical structures and pathological changes.

In practice, a directional gradient magnetic field $B_x(x)$ is applied to distinguish signals from different spatial locations. These gradients create position-dependent frequency shifts in proton precession, expressed as $\omega_x(x) = \lambda(B_0 + B_x(x))$, where $\omega_x(x)$ is the precession frequency at position x. With protons at different positions resonating at unique frequencies, the MRI system can identify the spatial origin of each signal. By applying RF pulses at specific frequencies, the system excites

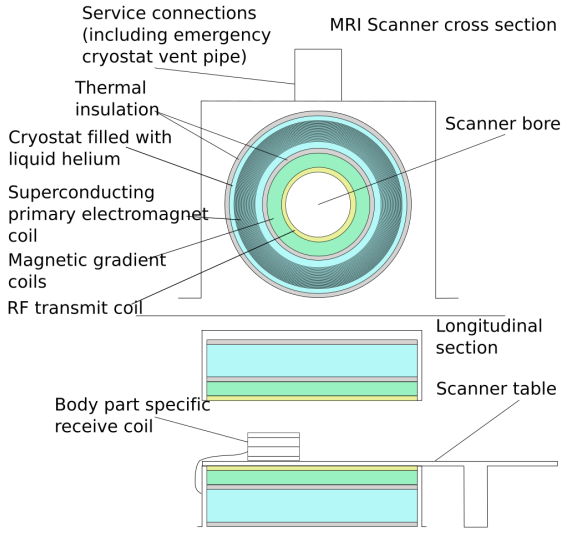


Figure 1.1: Schematic of a cylindrical superconducting MRI scanner. Top: cross-section of the cylinder showing primary, gradient, and RF transmit coils. Bottom: longitudinal section showing the same coils and the RF receive coil. Source: `https://www.wikipedia.fr`.

selected protons while the gradient field maps frequencies to physical positions. This spatial encoding enables the creation of detailed images that depict internal tissue structures.

MRI image appearance depends on acquisition configuration. Parameters such as timing, duration, frequency, and amplitude of RF pulses, as well as gradient field strength and orientation, can be adjusted to encode tissue properties differently. These settings define the pulse sequence, determining how tissue relaxation characteristics translate into signal intensities.

1.C Clinical MRI Sequences

MRI sequences generate images with different contrasts by exploiting distinct tissue properties, primarily tissue relaxation characteristics or water diffusion behavior. These contrasts determine how brain structures and pathologies appear, which directly affects diagnostic interpretation and segmentation.

Relaxation-based sequences manipulate acquisition parameters to emphasize differences in how tissues return to equilibrium after RF excitation. Key pulse sequence parameters include Repetition Time (TR), Echo Time (TE), and Inversion Time (TI), which shape the timing of RF pulses and signal measurement. Adjusting these parameters produces distinct tissue contrasts:

T1w
: images use short echo (TE) and repetition (TR) times to highlight differences in longitudinal relaxation ($T1$) properties, resulting in bright fat and dark fluid signals. This contrast provides clear gray–white matter delineation, making

T1w the preferred sequence for structural brain mapping, tissue segmentation, and cortical surface reconstruction. As such, T1w images form the foundation of most automated segmentation pipelines and large-scale neuroimaging datasets.

T2w

images use long TE and TR to emphasize transverse relaxation ($T2$) properties, where fluids such as cerebrospinal fluid (CSF) and edema appear bright. T2w imaging is valuable for identifying pathology associated with increased water content, including edema, cysts, and certain tumor components. It is often paired with T1w for differential diagnosis and multimodal tissue characterization.

FLAIR

sequences are T2-weighted acquisitions with an inversion pulse that suppresses free fluid signals, particularly from CSF. This suppression enhances the visibility of periventricular and cortical lesions that would otherwise be obscured by bright fluid in standard T2w scans. FLAIR is used in clinical follow-up of multiple sclerosis (MS), stroke, and traumatic brain injury (TBI), often serving as the key sequence for lesion quantification and disease progression monitoring.

Diffusion-weighted imaging (DWI) differs from relaxation-based sequences by applying diffusion-sensitizing gradients that encode the random motion of water molecules into signal attenuation. These contrasts provide unique sensitivity to tissue microstructure and integrity, complementing relaxation-based modalities.

DWI

highlights regions with restricted water diffusion as hyperintense areas. It is highly sensitive to acute ischemic stroke, where cytotoxic edema limits water mobility within minutes of onset [Nagaraja, 2021]. DWI is also used in oncology and infection imaging for detecting hypercellular lesions.

Apparent Diffusion Coefficient (ADC)

maps are quantitative parametric images derived from DWI through voxel-wise fitting. They separate true diffusion effects from T2 signal response, allowing for more accurate characterization. True diffusion restriction manifests as high signal on DWI and corresponding low signal on ADC. ADC metrics are used for tumor grading, stroke staging, and assessment of treatment response.

These sequence types and parameter variations produce a broad spectrum of image appearances tailored to clinical needs. Even within a single sequence, parameter adjustments can alter contrast, signal-to-noise ratio, and artifact susceptibility, all of which influence segmentation accuracy and reliability.

1.D Sources of Variability

MRI scans can vary significantly across sessions, scanners, and institutions, even when using the same sequence and imaging the same anatomy. These differences affect intensity distributions, contrast, and spatial resolution, introducing variability that complicates automated analysis. Below are common sources of such variation.

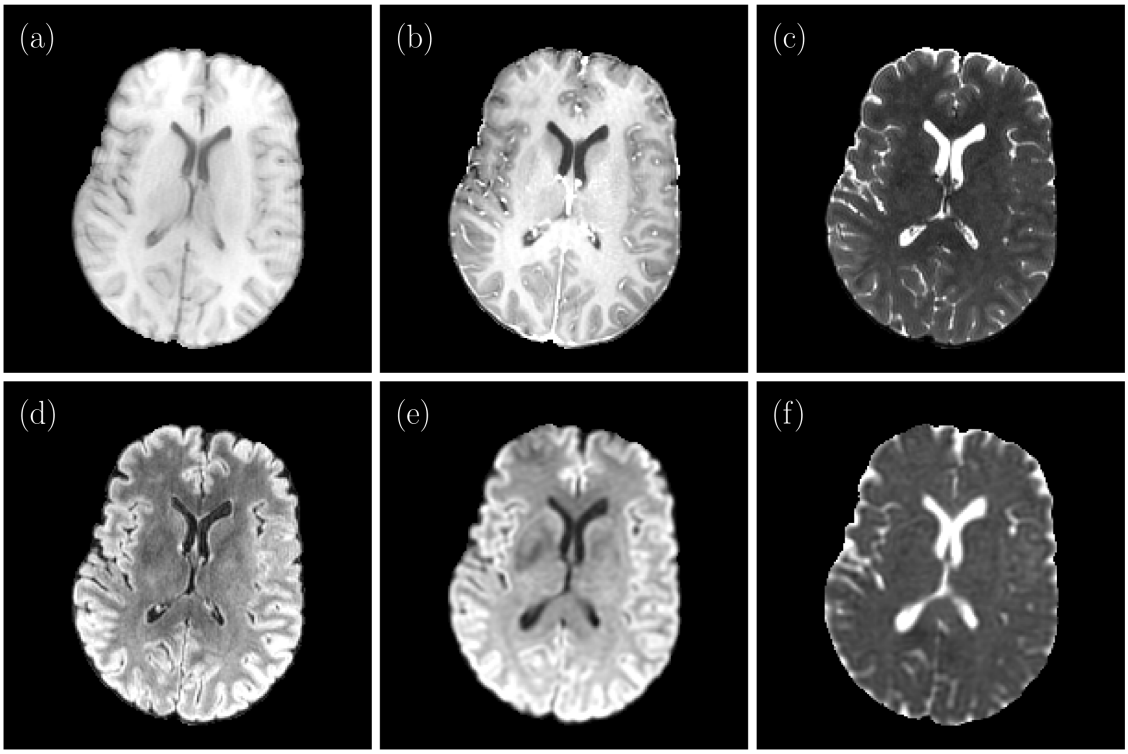


Figure 1.2: Examples of axial slices across common MRI sequences (native, Right-Anterior-Superior orientation) acquired from a single UCSF-PDGM subject. Panels show: (a) T1-weighted, (b) T1c, (c) T2, (d) FLAIR, (e) DWI, and (f) ADC. Contrast variations across modalities highlight the complementarity of MRI sequences in characterizing tissue properties. *T1c corresponds to T1w acquisitions with prior contrast agent administration to the patient. This enhances the visibility of brain regions, but is not always used due to potential reactions.*

- *Scanner field strength:* affects signal-to-noise ratio, tissue contrast, and artifact susceptibility. Scans from 1.5 T and 3 T systems can differ in intensity profiles and lesion appearance.
- *Bias field artifacts:* appear as smooth, spatially varying intensity shifts across the image, stemming from RF coil inhomogeneities and magnetic field distortions. These are more pronounced at higher field strengths and disrupt intensity-based segmentation.
- *Anisotropic resolution:* occurs when through-plane resolution (slice thickness) is lower than in-plane resolution, causing partial volume effects and blurring along the slice axis.
- *Image reconstruction artifacts:* such as Gibbs ringing arise from k-space truncation and appear as ripples near sharp edges, obscuring anatomical boundaries.
- *Low-resolution imaging:* increases partial volume effects, reducing boundary clarity and hampering fine-grained analysis, especially for FLAIR images.

Other artifacts (illustrated in Figure 1.3) include motion artifacts, limited field

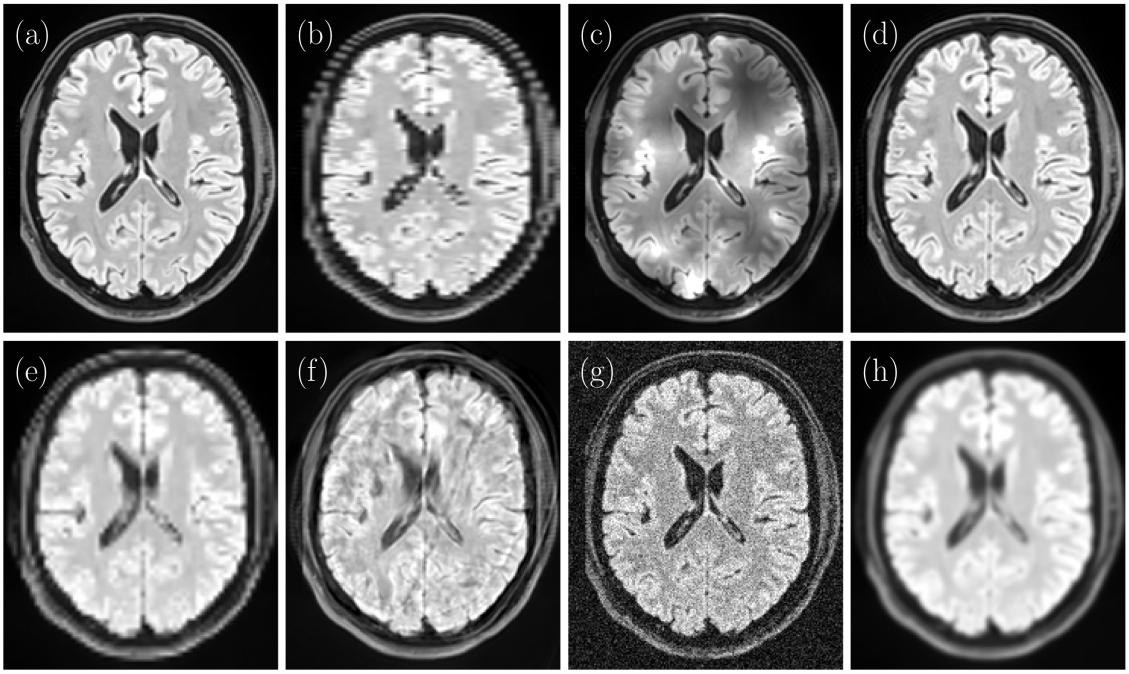


Figure 1.3: Illustration of simulated MRI artifacts applied to an axial FLAIR slice. (a) Original FLAIR MRI, (b) anisotropy, (c) bias field, (d) Gibbs ringing, (e) low-resolution sampling, (f) motion-induced artifacts, (g) Rician noise, and (h) partial-volume effect. All panels share the same windowing and geometry. Augmentation parameters were chosen for high visual contrast.

of view (FOV), RF interference, and ghosting. These artifacts introduce noise and spatial inconsistencies that degrade image quality and complicate segmentation. For instance, bias field inhomogeneities shift the intensity distribution, creating artificially high-intensity regions that may be confused with hyperintense pathological structures such as tumors or gliomas. Other artifacts induce local variations in contrast and gradient patterns, which can be misinterpreted by segmentation algorithms as changes in tissue properties.

The combined effect of these factors creates a challenging environment for image analysis algorithms. Methods developed and validated on data from a single scanner or protocol often generalize poorly to images with different acquisition characteristics. While standardizing imaging protocols could reduce variability, this is often impractical in clinical settings. Therefore, robust preprocessing and segmentation methods that handle diverse image appearances are essential.

2 Neurological Conditions

Neurological diseases are conditions that affect the brain and nervous system. They can be caused by immune dysfunction, blood flow anomalies, tumors, injury, infections,

or other factors. These conditions often lead to memory issues, impaired movement or speech, and other neurological dysfunctions [Organization, 2006]. MRI is commonly used in these cases because it provides detailed visualization of brain structure and disease-related changes. Neurological diseases in which MRI, and specifically FLAIR imaging, plays a crucial role includes:

Stroke
: is an acute neurological event caused by interrupted blood flow to brain tissue, leading to ischemia and cell death. It is a leading cause of death and disability worldwide, with over 12 million new cases annually [Steinmetz et al., 2024]. This results in infarcts, localized areas of tissue necrosis and damage. Treatment aims to restore perfusion and minimize injury. On FLAIR, ischemic infarcts appear as focal hyperintensities, often accompanied by edema and mass effect, following the affected vascular territory. DWI is the gold standard for acute detection, but FLAIR remains essential for delineating infarct evolution and guiding therapy. In chronic phases, stroke can lead to encephalomalacia, sulcal widening, and ventricular enlargement. Monitoring stroke progression with MRI is crucial for assessing tissue viability, guiding therapy, and predicting functional outcomes [Stulberg et al., 2023].

Multiple Sclerosis
: is a chronic autoimmune demyelinating disease of the central nervous system that affects approximately 2.8 million people worldwide [Steinmetz et al., 2024]. It is a leading cause of neurological disability in young adults, often following a relapsing–remitting course that may progress to chronic disability. On FLAIR, MS lesions appear as small, ovoid hyperintensities, most characteristically located in periventricular and juxtacortical regions, but also involving infratentorial and spinal structures. Their number, size, and distribution evolve over time, and in advanced stages, lesions may become confluent, obscuring normal tissue boundaries. Beyond focal lesions, MS is associated with neurodegeneration and global brain atrophy, producing macroscopic anatomical changes.

Brain Tumors
: are abnormal proliferations of neural or glial cells affecting roughly 300,000 people annually [Steinmetz et al., 2024]. They include gliomas, glioblastomas, and meningiomas. High-grade gliomas, in particular, are among the most aggressive neurological diseases, with median survival measured in months despite aggressive treatment. On FLAIR, tumors appear as heterogeneous hyperintense lesions, often surrounded by vasogenic edema. MRI is indispensable for detection, characterization, surgical planning, and treatment monitoring. By displacing and compressing adjacent tissue, tumors significantly distort brain anatomy.

Traumatic Brain Injury (TBI)
: results from external mechanical forces that cause brain damage, including contusions, hemorrhages, and diffuse axonal injury. Globally, TBI incidence is

estimated at 69 million cases annually [Steinmetz et al., 2024]. The injury leads to neuronal and vascular disruption, inflammation, and edema. On FLAIR, acute contusions appear as cortical–subcortical hyperintensities, while diffuse axonal injury is reflected by multifocal white matter lesions in characteristic locations such as the corpus callosum. Chronic stages often show diffuse atrophy and gliosis. These structural alterations manifest as tissue loss and asymmetric ventricular dilation. MRI is critical for identifying injury patterns, such as microbleeds on susceptibility-weighted imaging (SWI) and edema on FLAIR [Lee and Newberg, 2005].

Conditions mentioned above produce not only focal lesions (seen as white matter hyperintensities on FLAIR) but also broader anatomical alterations, including atrophy, ventricular enlargement, and displacement of brain structures. White matter hyperintensities on FLAIR are often non-specific, as similar patterns may reflect demyelination, ischemia, or trauma. Therefore, a comprehensive understanding of the disease or condition affecting the brain requires consideration of both focal changes (lesions or local abnormalities) and global alterations (such as atrophy or ventricular enlargement).

Mapping cortical and white matter lesions relative to functionally meaningful regions (e.g., motor, language, or visual systems) and major tracts is crucial for understanding symptoms and recovery potential. This contextualization explains why lesions of similar size can have very different clinical effects depending on their location. Combining parcellation with lesion maps yields detailed impact representations for treatment planning and rehabilitation.

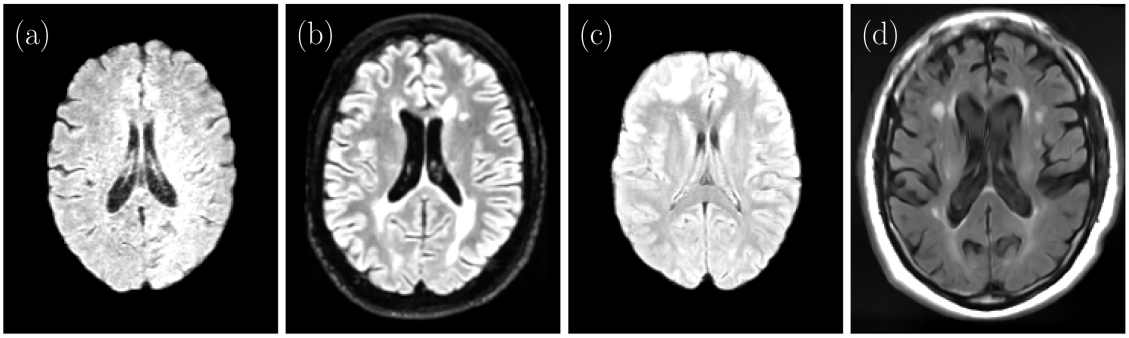


Figure 1.4: Representative axial FLAIR slices (MNI, RAS) showing disease-specific lesion appearances. (a) Ischemic stroke with a focal territorial hyperintensity in the left hemisphere. (b) Multiple sclerosis presenting with small, ovoid periventricular hyperintensities. (c) Brain tumor with heterogeneous signal intensity. (d) Traumatic brain injury with patchy white matter hyperintensities of irregular intensity. These examples illustrate both lesion distribution variability and the diversity of FLAIR intensity patterns across pathologies. Intensity scaling differs slightly across panels and reflects acquisition conditions. *MNI refers to the Montreal Neurological Institute space, a standardized brain template commonly used in neuroimaging, using templates is primordial to be able to compare images from different subjects using the same orientation and size.*

This thesis focuses on MS and tumors because they represent major conditions where FLAIR imaging is especially relevant clinically. Since FLAIR MRI is routinely acquired in their evaluation, these diseases constitute scenarios where T1w images are not the preferred modality and may not be acquired according to guidelines. In all these use cases, FLAIR imaging is standard due to its sensitivity to pathology, while T1w imaging, though valuable for structural segmentation, is often omitted to reduce scan time, especially in urgent or resource-constrained settings.

These neurological diseases alter MR image appearance through features such as lesions, infarcts, tumors, and diffuse atrophy. Such pathological changes can compromise the reliability of automatic image processing tools, including brain tissue segmentation and anatomical localization. Robust analysis pipelines must account for these abnormalities to ensure accurate interpretation and reliable downstream clinical or research use.

3 MRI Acquisition Protocols and Guidelines

MRI sequence selection depends on the clinical context, with radiologists choosing contrasts and parameters that best highlight disease-specific tissue properties. Because MRI sequence selection varies with clinical objectives and disease-specific needs, radiology consortia have established standardized acquisition protocols to harmonize imaging practices across sites. Radiology consortia produce clinical protocols to formalize these choices into standardized guidelines defining which sequences to acquire, along with parameters such as resolution, slice thickness, and scan timing. These standards promote consistency across time points and institutions, supporting reliable diagnosis and longitudinal follow-up, particularly in chronic neurological diseases. Protocol development is guided by expert consensus and national or international initiatives aimed at harmonizing imaging practices across clinical sites, providing a representative overview of routine clinical acquisition strategies. Within these standardized protocols, FLAIR imaging is of particular importance for certain neurological conditions, as it enhances lesion visibility near fluid spaces. Its diagnostic utility across diverse pathologies has made it one of the most consistently acquired MRI sequences in both clinical and research settings. Even more so, in particular, certain scenarios put more importance on the acquisition of FLAIR images.

Brain Trauma Foundation

[Carney et al., 2017] recommends MRI for assessing TBI. FLAIR sequences are advised for detecting cortical contusions, microhemorrhages, and edema following brain injury. T1w sequences are also recommended for anatomical reference and structural damage assessment, but FLAIR is often prioritized in acute settings due to its sensitivity to pathology.

American Stroke Association

[Jauch et al., 2013] provides MRI recommendations following stroke. Because intervention is time-sensitive, imaging typically begins with CT to rapidly

exclude hemorrhage. FLAIR MRI, in combination with diffusion imaging, is used for follow-up and longitudinal monitoring to detect ischemic and subacute lesions.

MAGNIMS–CMSC–NAIMS

[Goodkin et al., 2021] offers guidelines for MRI in MS. FLAIR sequences are used to highlight white matter lesions, while T1w-Gd sequences detect active inflammation. When T1w-Gd is contraindicated (e.g., due to contrast agent allergy), FLAIR serves as the primary anatomical sequence for monitoring MS progression.

These scenarios are the primary motivations behind the focus on FLAIR imaging in this thesis, where in certain cases, volumetric (*e.g.*, 3D) T1w images are sometimes not acquired at full-resolution. Some of the guidelines above even mentions the possibility of ommitting it entirely in certain cases.

4 Scientific Challenges

Although FLAIR imaging plays a central role in clinical neurology, particularly for detecting white matter abnormalities, most automated segmentation pipelines still assume T1w input with strong tissue contrast and clear cortical boundaries. As discussed in Section 3, certain clinical protocols may deprioritize volumetric T1w imaging entirely, in favor of other sequences, more relevant. However, in these scenarios, high-resolution FLAIR is especially relevant, and seems to be systematically acquired. While FLAIR might not provide the same anatomical detail as T1w, it still contains discernable tissue contrast. Additionally, it is valuable thanks to its sensitivity to disease-related changes. Therefore, to be able to process images acquired in these scenarios, providing a method operating directly on FLAIR images would allow to include them in volumetric studies. Something that is exceedingly important, due to volumetric analysis being able to point out atrophy patterns, a pattern that is relevant in the above mentioned diseases.

FLAIR imaging presents distinct challenges for automated segmentation compared to T1w sequences. Whereas T1w images provide consistent gray–white matter contrast and anatomical clarity, FLAIR prioritizes pathology visualization, often at the expense of structural definition. This trade-off reduces contrast between normal brain structures, particularly in subcortical regions, making fine-grained anatomical parcellation based only on FLAIR more difficult. Moreover, FLAIR contrast is highly sensitive to parameters such as inversion time, leading to pronounced variability across scanners and acquisition settings. Variability in CSF suppression and susceptibility to motion and flow artifacts further degrade image quality and may produce false hyperintensities or obscure anatomical boundaries. These specific characteristics of FLAIR images makes the development of DL based segmentation methods challenging, due to the additional variability and artifacts mentioned above.

Pathology further complicates the problem. FLAIR Hyper Intensities (FHI) are

intrinsically non-specific, as similar signal changes may result from demyelination, ischemia, or trauma. Therefore, intensity patterns alone can be insufficient for reliable interpretation, and as such obtaining a complete understanding of a specific condition affecting the brain requires additional biomarkers. While volumetric structure analysis could provide additional context, as mentioned in 3, due to volumetric T1w images not always being acquired, existing T1w-based tools, trained with healthy data are not applicable in these scenarios. Tools designed for Clinically Normal (CN) brains additionally often fail in presence of FHI.

As an important addition, since the clinical impact of FHI depends on their location; the ability to obtain or relate structure segmentation to FHI location provides functional and anatomical context (e.g., motor, language, or visual systems) could provide crucial insights into symptoms and recovery potential. This could provide further insight in the condition at hand, either for treatment planning or rehabilitation.

This thesis addresses the gap created by the lack of tools to process images in such scenarios, by developing segmentation approaches that operate regardless of the presence of FHI, on FLAIR alone. These methods target both brain extraction and fine-grained anatomical segmentation, with a particular focus on robustness to pathological variation. By enabling detailed analysis from single-contrast FLAIR scans, this work expands the reach of automated neuroimaging to a broader range of clinical data.

Chapter 2

Literature Review of Brain Segmentation Methods

1 Common Regions of Interests 30
2 Semantic Segmentation Methods for Brain MRI 32
2.A Traditional Brain Segmentation 32
2.B Deep Learning Segmentation Methods 33
2.C Modality Agnostic Methods 38
3 Evaluation Metrics and Training Objectives 39
3.A Segmentation Metrics 39
3.B Training Objectives and Loss Functions 41
4 Current Limitations and Research Gaps 41

1 Common Regions of Interests

Segmentation in neuroimaging serves two principal purposes. First, the delineation of anatomical or functional brain structures enables volumetric analysis, which supports a range of clinical and research applications such as tracking neurodegenerative atrophy or studying structural brain variability. Second, the segmentation of pathological regions, such as tumors or lesions in multiple sclerosis (MS), provides critical biomarkers (e.g., lesion load or tumor volume), which are essential for diagnosis, disease monitoring, and treatment planning.

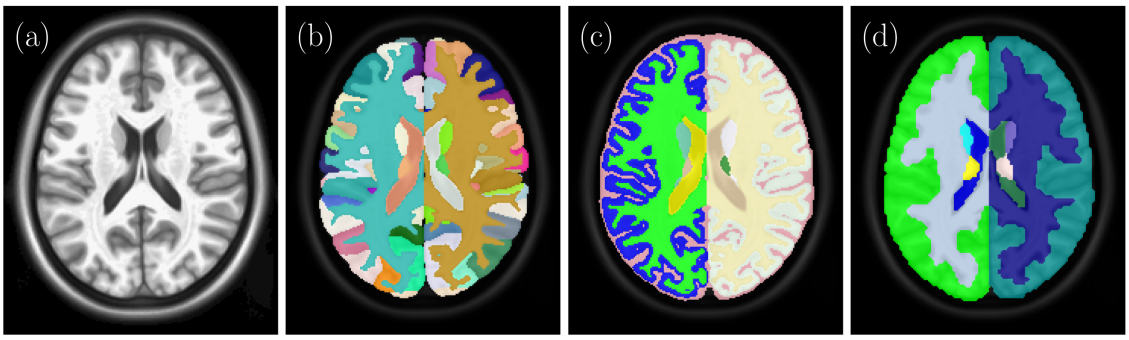


Figure 2.1: Examples of common brain segmentation protocols. *b.* NeuroMorphometrics (138 regions), *c.* SynthSeg (63 regions), *d.* Harvard-Oxford (48 cortical + 21 subcortical regions). *Brain segmentation protocols vary by both their number of structures and definitions.*

Segmentation protocols provide standardized frameworks for labeling and delineating brain regions, ensuring reproducibility and enabling cross-study comparisons. These protocols vary not only in the number of segmented structures, but also in their conceptual foundations: some rely on macroscopic anatomical landmarks (*e.g.*, sulci boundaries, see Fig.2.1), while others incorporate structural or functional connectivity to define coherent brain regions. In what follows, we categorize widely used segmentation protocols into coarse-grained and fine-grained anatomical segmentation schemes:

Desikan-Killiany:
A surface-based cortical parcellation protocol comprising 68 regions (34 per hemisphere), defined on gyral anatomy (e.g., ridges or sulci dividing the cerebral cortex). It is the default atlas in FreeSurfer, a suite of tools for the processing of MRI files (DICOM/NIFTI) [Desikan et al., 2006].

DKT (Desikan-Killiany-Tourville) Atlas:
An improved version of the Desikan-Killiany atlas, designed for greater anatomical consistency by refining sulcal and gyral definitions. It is used by Mindboggle [Klein and Tourville, 2012].

Harvard-Oxford Atlas:
A volumetric probabilistic atlas including 69 regions (48 cortical + 21 sub-

cortical), derived from structural MRI of healthy individuals. It is widely used in the FMRIB Software Library (FSL) and CONN [Makris et al., 2006, Desikan et al., 2006, Frazier et al., 2005, Goldstein et al., 2007][1].

Fine-grained protocol commonly refer to protocols defining over 100 regions, enabling more detailed morphometric and connectomic analyses.

NeuroMorphometrics Atlas:
A volumetric atlas with approximately 138 labeled structures, based on manual expert segmentation of T1w images (e.g., BrainCOLOR). It is used in Statistical Parametric Mapping (SPM) via CAT12 and volBrain [NeuroMorphometrics Inc., 2024].

Brainnetome Atlas:
A volumetric atlas with 246 regions (210 cortical + 36 subcortical), based on both functional and structural connectivity. It is aligned to MNI152 space [Fan et al., 2016].

HCP-MMP1.0 Atlas:
A multimodal surface-based parcellation with 180 regions per hemisphere, incorporating T1, T2, myelin maps, and connectivity profiles. It is used in the Human Connectome Project (HCP) [Glasser et al., 2016].

Beyond traditional atlases, recent work introduces high-resolution parcellations that aim to capture even finer anatomical detail or accommodate specialized segmentation tasks. This often includes the delineation of smaller anatomical or functional structures—such as fine sulci, deep nuclei, or small vessels—that are not visible at standard 1 mm isotropic resolution. These atlases support advanced modeling but often require higher-resolution data 0.125–0.5 mm, because finer structures are only resolvable at submillimetric scales.

NextBrain Atlas:
A next-generation probabilistic atlas built from 3D reconstructions of serial histology in five human hemispheres. It defines 333 cortical and subcortical regions based on dense cytoarchitectural labeling and enables Bayesian segmentation of in vivo and ex vivo MRI at submillimetric resolution [Casamitjana et al., 2024].

HoliBrain Atlas:
A structural, multi-modal MRI-based atlas defining 350 substructures, built from high-resolution (0.125 mm^3) T1w, T2-weighted (T2w) scans of 75 healthy subjects. Labels are derived from a fusion of multiple segmentation protocols and manually refined to ensure anatomical consistency across scales [Manjón et al., 2025].

[1] references as provided on FSL documentation

2 Semantic Segmentation Methods for Brain MRI

2.A Traditional Brain Segmentation

While manual segmentation remains the benchmark for accuracy, automated methods allow large-scale studies by drastically reducing processing time and inter-rater variability.

Atlas-Based Segmentation: Traditional brain segmentation typically relies on image registration, where pre-labeled volumetric atlases (see Section 1) are aligned to unlabeled target images. Registration involves the estimation of spatial transformation parameters, such as rotation angles and scaling factors for affine models, or dense deformation fields, by optimizing similarity metrics such as mutual information or cross-correlation [Avants et al., 2008]. Once aligned, anatomical labels from the atlas can be transferred to the target image [Collins et al., 1995]. Atlases can represent either individual subjects or population averages to reduce inter-subject variability. However, single-atlas approaches can struggle to generalize across populations because of anatomical and demographic variability (e.g., sex, ethnicity, acquisition protocols).

Multi-Atlas Segmentation: To address these limitations, Rohlfing et al. propose template selection strategies that identify the most anatomically similar atlas for each subject [Rohlfing et al., 2003]. A more robust alternative is multi-template segmentation (MAS in the literature), in which multiple annotated atlases are registered to each target image [Klein et al., 2005]. Label fusion techniques are then used to combine the resulting segmentations. While effective, MAS is computationally expensive: registering multiple high-resolution 3D atlases can require up to 8 hours per subject on standard workstations (*e.g.*, Intel Core i7, 32 GB RAM), depending on the number of templates and registration complexity.

Recent advances include the use of neural networks to estimate registration parameters, both affine (e.g., rotation, scaling) and non-linear (e.g., dense deformation fields) [Balakrishnan et al., 2019], allowing for significantly faster learning-based registration pipelines.

The accuracy of MAS depends on precise registration and on the quality and diversity of the available atlases. MAS methods require the availability of diverse atlas libraries to avoid inducing population bias related to available atlases. However it is important to not that DL methods are also subject to similar considerations with regards to the composition of the training database.

Surface-based segmentation: Surface-based segmentation, as used in the Desikan-Killiany and DKT atlases, constructs a cortical mesh from structural MRI and assigns anatomical labels to mesh vertices. After extracting the cortical surface,

the mesh is inflated and mapped onto a sphere. Each vertex on this sphere encodes geometric features (e.g., sulcal depth), and surface registration adjusts vertex positions so that the subject's folding patterns align with a spherical template. Once aligned, anatomical labels from the template are transferred to the subject via vertex correspondence [Fischl, 2004].

Patch-based segmentation: Patch-based segmentation methods are introduced to mitigate the reliance on precise registration in traditional label fusion. Instead of depending on voxel-wise alignment, these methods compare local image neighborhoods (patches) following coarse alignment (e.g., affine registration to MNI space). Labels are propagated by identifying similar patches using intensity-based similarity metrics. Non-local patch-based strategies, first introduced by Coupé et al., extend the search beyond immediate spatial neighborhoods, allowing for more reliable matches [Coupé et al., 2010]. These methods achieve high segmentation accuracy by leveraging local context, non-local redundancy, and robust label aggregation, while being less susceptible to non-linear registration errors. Importantly, patch-based methods can be adapted to different modalities by using modality-specific similarity metrics. Improvements on non-local patch-based methods include the integration of machine learning techniques to enhance patch similarity assessment and label fusion [Xie et al., 2019].

2.B Deep Learning Segmentation Methods

Convolutional Neural Networks

Convolutional Neural Networks (CNNs) become the foundation for nearly all computer vision tasks, including medical image segmentation. They replace handcrafted features with learned representations by exploiting spatial structure through convolutions that aggregate local information. This design draws inspiration from early biological vision, where neurons respond to input from neighboring photoreceptors to extract spatial patterns.

CNNs can be seen as an extension of Multi-Layer Perceptrons (MLP) by introducing crucial spatial inductive biases. Unlike fully connected MLPs that treat each input feature independently and ignore spatial relationships, convolutional layers apply small filters across spatial dimensions. This allows CNNs to learn translation-invariant features with far fewer parameters. Instead of connecting each output pixel to all input pixels, outputs depend only on local neighborhoods defined by the filter size. This architectural shift improves scalability and computational efficiency, making CNNs especially well suited for image data where spatial locality is essential for capturing complex patterns required for accurate segmentation.

The U-Net architecture [Ronneberger et al., 2015, Çiçek et al., 2016] (see Figure 2.2) becomes the dominant framework for medical image segmentation due to key advantages suited to medical data. Its symmetric encoder-decoder design uses

skip connections to preserve fine spatial details while learning features at multiple scales. The encoder progressively downsampled features, efficiently expanding the receptive field to capture broader context without needing very deep networks.

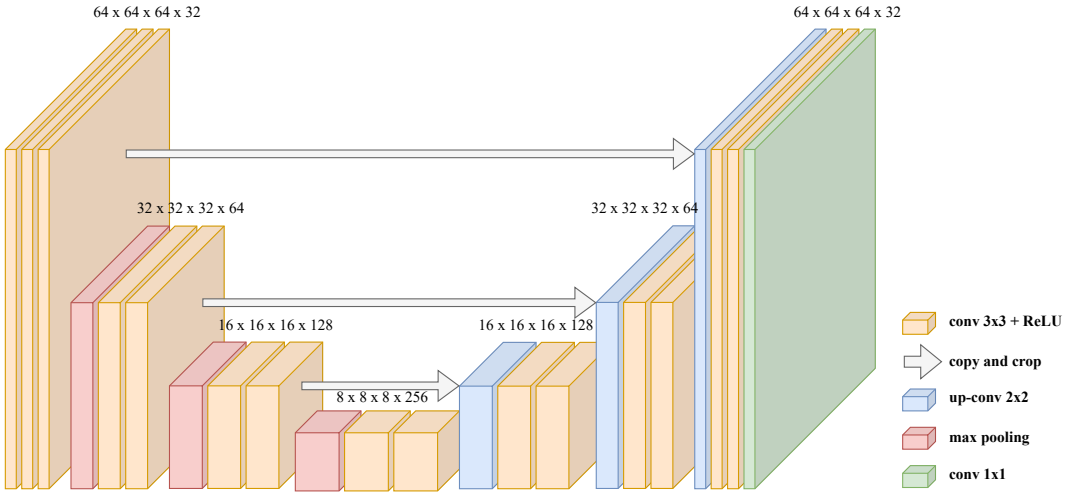


Figure 2.2: Architecture of a 3D U-Net model for volumetric segmentation. The network followed an encoder–decoder structure with skip connections. The contracting path (left) progressively downsampled the input through repeated applications of $3 \times 3 \times 3$ convolutions with ReLU activation (yellow) and max pooling (red), increasing feature channels while reducing spatial resolution. The bottleneck layer (center) encoded the most compressed representation. The expansive path (right) restored spatial resolution using up-convolutions (blue), concatenating feature maps from the contracting path (copy and crop, grey arrows) to preserve localization information. The final $1 \times 1 \times 1$ convolution (green) mapped the feature maps to the output segmentation classes.

This efficiency is especially important in medical imaging, where training data is often limited compared to natural images. U-Net variants use far fewer parameters than typical computer vision models while delivering competitive segmentation performance. The encoder-decoder structure with skip connections also promotes stable training, which is crucial for handling the complex anatomy and class imbalances common in medical segmentation.

U-Net remains highly prevalent in medical imaging, even as more complex architectures dominate natural image tasks. One possible reason is that structural MRI modalities, such as T1w, are captured under standardized protocols, resulting in limited intensity variation and highly consistent anatomical layouts across subjects. Compared to natural images, which feature broad variation in object appearances, medical scans mostly involve segmenting subtle deviations from a relatively fixed anatomical template. This narrower variability space means that U-Net's simpler encoder-decoder structure with limited parameter count is sufficient, and often preferable, for capturing the consistent features necessary for robust segmentation.

Training data limitations and annotation quality are major bottlenecks in medical image segmentation [Maier-Hein et al., 2018]. Expert segmentations show significant inter- and intra-rater variability, introducing noise in ground truth labels [Maier-Hein et al., 2018]. Small datasets and inconsistent annotation standards further complicate training. In this context, compact architectures like U-Net strike a better balance between model capacity and trainability than more complex models. U-Net's strong spatial inductive biases align well with anatomical data, and its simplicity aids optimization under difficult training conditions.

One thing of note is the introduction of fully automated segmentation frameworks like nnU-Net [Isensee et al., 2021] (see Fig 2.3) and Auto3D-Seg (included in MONAI [Cardoso et al., 2022]). These frameworks automate the entire pipeline, from preprocessing and architecture selection to training and postprocessing, removing the need for time consuming manual-tuning of training hyperparameters. Both frameworks consistently win major medical imaging segmentation challenges by delivering robust, high-performance results across diverse anatomical targets and imaging modalities. They also adopt different strategies with regards to their self configurations: while nnUNet use rules based heuristics extracted from the training data to adapt the architecture and training pipeline, Auto3D-Seg trains multiple candidate configurations and select the best performing models. The success of these frameworks stems from combining U-Net's strong spatial inductive biases with automated configuration and training, making them highly adaptable to the variability, limited data, and heterogeneity characteristic of medical imaging. This automation significantly accelerates deployment and enhances reproducibility in clinical and research settings.

While traditional multi-atlas segmentation (MAS) is computationally expensive and dependent on registration accuracy (often requiring hours per subject), CNN-based segmentation shifts the field towards fast, end-to-end pipelines that complete inference in seconds to minutes. Hybrid approaches, however, complicate this dichotomy. Methods such as Deep Label Fusion (DLF) combine MAS principles with CNNs, achieving segmentation accuracy comparable to state-of-the-art U-Net variants while offering improved generalizability on unseen domains [Xie et al., 2019]. This suggests that while CNNs offer a clear advantage in efficiency and reduced error propagation, hybrid strategies remain relevant where robustness across heterogeneous datasets is critical.

Despite strong performance in controlled settings, CNN-based segmentation faces key limitations affecting clinical use. Models trained on healthy brains often struggle with pathological cases involving lesions, tumors, or other abnormalities. Acquisition artifacts like motion blur, intensity inhomogeneity, or low signal-to-noise ratio can cause systematic errors that go unnoticed. Population bias in training data can reduce accuracy for underrepresented groups, and age-related anatomical changes in pediatric or elderly populations are poorly captured when training relies mainly on healthy adults. Uncertainty quantification approaches begin to address this limitation by providing confidence estimates alongside segmentation results, but

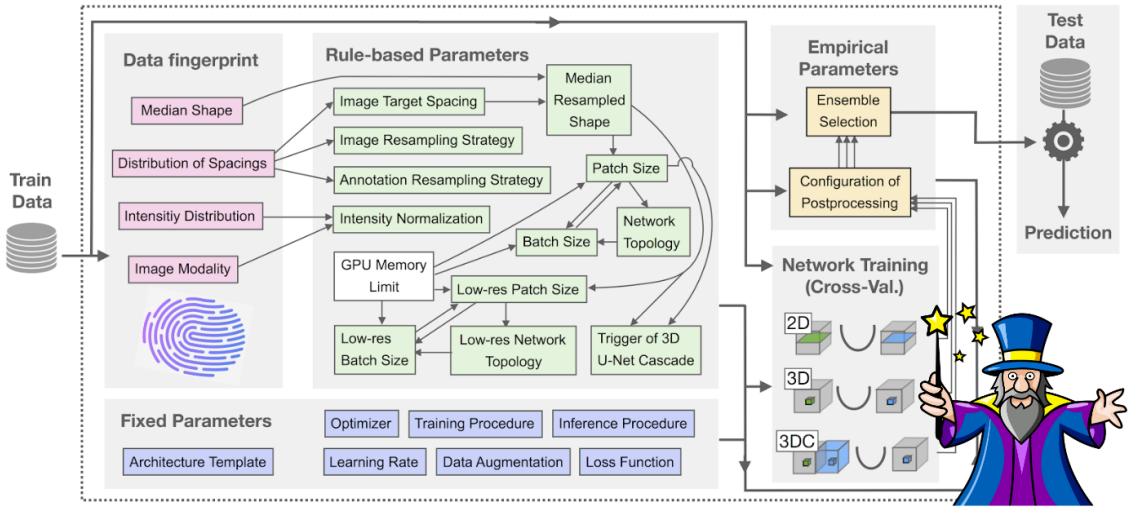


Figure 2.3: Overview of the nnU-Net framework, proposed as a "self-configuring wizard" for deep learning–based medical image segmentation. nnU-Net automatically adapted all key design choices to a given dataset, including image preprocessing, network architecture and hyperparameters, data augmentation strategies, training schedules, and postprocessing heuristics. By standardizing these components, nnU-Net provided a strong and reproducible baseline across a wide range of medical segmentation tasks. Source: [Isensee et al., 2021]

robust uncertainty estimation remains an active area of research.

Ensembling 2D models improves segmentation robustness by combining predictions from networks analyzing slices in multiple orientations. FastSurfer exemplifies this by using an ensemble of 2D CNNs processing axial slices from different directions [Henschel et al., 2020, Henschel et al., 2022]. This multi-view approach addresses anisotropic voxel spacing and low slice resolution by leveraging complementary information from orthogonal axes, reducing slice-specific errors while maintaining 2D CNNs' computational efficiency.

Ensembling 3D models enhances accuracy in complex anatomy by leveraging spatial context and regional specialization. SLANT27 and AssemblyNet use ensembles of 3D CNNs trained on spatially localized patches within MNI space [Huo et al., 2018, Coupé et al., 2020]. This patch-based strategy manages GPU memory efficiently while enabling detailed modeling of heterogeneous structures. Combining region-specific predictions mitigates local errors and improves reliability, which is critical for fine-grained neuroanatomical parcellation.

3D CNNs is shown to outperform 2D models in certain contexts for medical image segmentation tasks [Avesta et al., 2023] and in cases where 3d convolutions allows leveraging of full volumetric context. By processing the entire 3D volume or volumetric patches, 3D CNNs capture spatial dependencies along all three anatomical axes simultaneously, enabling more coherent and anatomically consistent segmentations. This advantage is particularly important in neuroimaging, where complex structures extend across slices and subtle spatial relationships are critical.

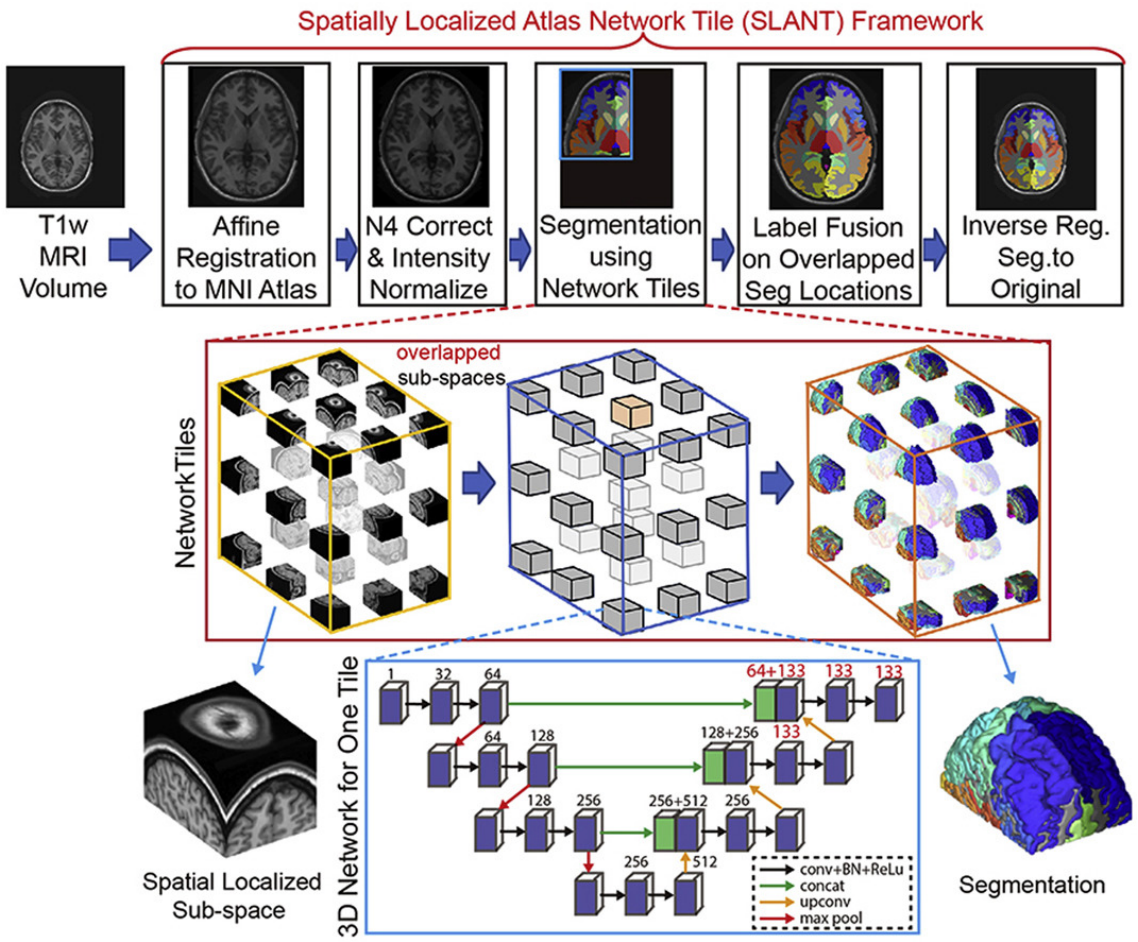


Figure 2.4: Regional ensemble approach used by SLANT and AssemblyNet. Multiple 3D CNNs were trained on local regions of the MNI patches, and the final prediction was obtained by aggregating all predictions. Source [Huo et al., 2018]

Studies consistently show that 3D models achieve higher accuracy and robustness compared to their 2D counterparts, despite increased computational and memory demands [Avesta et al., 2023]. It is important to note that depending on image resolution, slice spacing, and dataset characteristics however, 2d models are still used in certain cases, and included in major medical segmentation projects [Isensee et al., 2021]. They also allows allows for smaller memory footprints and computing requirements [Henschel et al., 2022].

Visual Transformer Architectures

Transformer architectures, originally developed for natural language processing, attract attention in medical imaging for their ability to model global context through self-attention. Unlike CNNs where the receptive field grows only gradually with depth and typically requires many stacked layers (or pooling, striding, or dilation) to capture

long-range dependencies, the self-attention mechanism used in vision transformers (ViTs) processes images by dividing them into patches, typically 16×16 to 32×32 pixels, and captures long-range dependencies [Dosovitskiy et al., 2020]. However, their quadratic computational cost limits direct application to full-resolution 3D brain volumes. To address this, patch-based and hierarchical attention methods are developed, such as the Swin Transformer, which uses windowed attention to balance efficiency and global modeling [Liu et al., 2021].

In medical imaging, a common strategy combines the strengths of vision transformers with the proven U-Net architecture. For example, SwinUNetr replaces U-Net's encoder convolutions with hierarchical Swin Transformer blocks to preserve multi-scale feature extraction and skip connections while adding efficient global context modeling [He et al., 2023, Hatamizadeh et al., 2021]. This design achieves accurate segmentation with lower computational demands than naïve transformer models.

Top-performing segmentation methods often employ ensembles that combine SwinUNetr with classic U-Net variants; for example, the 2023 BraTS-GLI segmentation challenge winning solution leverages the complementary benefits of transformer-based global attention and convolutional spatial biases [Ferreira et al., 2024]. These ensembles consistently deliver state-of-the-art accuracy and robustness across various imaging tasks and modalities.

2.C Modality Agnostic Methods

Modality-agnostic methods aim to segment brain structures across varying intensity distributions and imaging modalities without retraining or fine-tuning for specific acquisition settings. This flexibility is especially valuable in clinical practice, where MRI sequences differ widely in contrast, resolution, and scanner protocols. Traditional CNNs typically need modality-specific training, which limits their use on images outside the training distribution.

SynthSeg exemplifies a successful modality-agnostic approach [Billot et al., 2023a, Billot et al., 2023b] (see Figure 2.5). SynthSeg relies on synthetic data with random contrast generated from the target segmentation labels, with extensive randomization of contrast, resolution, morphology, artifacts, and noise. This forces the network to focus on domain-independent anatomical features rather than modality-specific intensities. Training requires only anatomical label maps, not real images, allowing the incorporation of diverse populations and pathological cases.

However, SynthSeg's reliance on synthetic training data introduces key limitations. Because the synthetic generation is based on healthy anatomical priors, it does not capture the full complexity of real-world cases, especially when lesions or pathologies significantly alter normal anatomy. Generated images contain only as much detail as is present in the provided segmentation maps. Consequently, the trained model struggles when presented with novel intensity patterns or anatomical distortions absent from the synthetic training samples; this can occur when examples containing

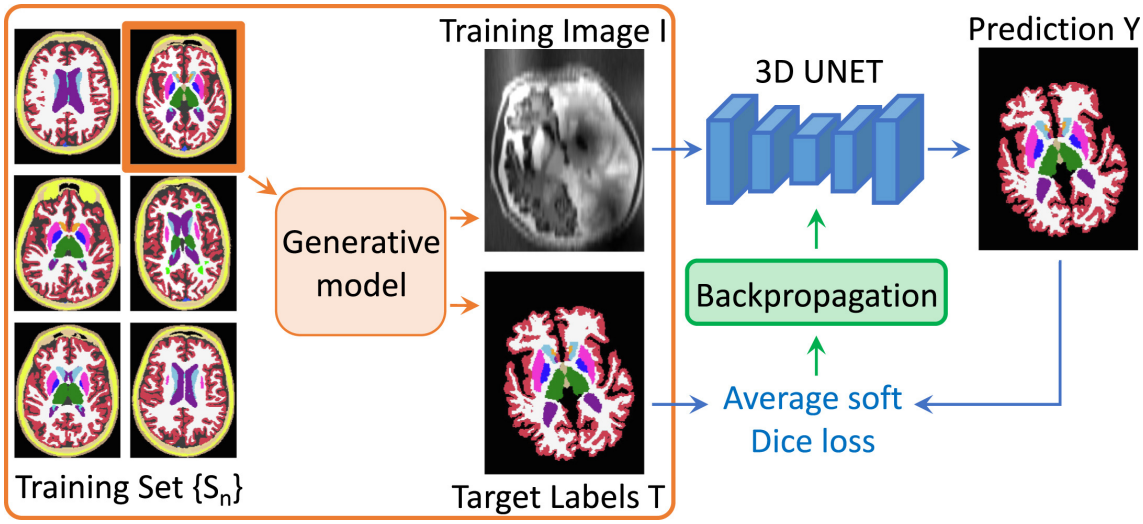


Figure 2.5: Illustration of the training sample generation process described in SynthSeg. Source: [Billot et al., 2023a]. At each step, a synthetic training image was sampled from a generative model based on the prior/available segmentation maps. The same segmentation map was used as a target for training the SynthSeg network.

tumors or hyperintensities are presented to the model. To account for this, SynthSeg-WMH propose integrating WMH into the synthetic generation pipeline by modulating their appearance based on surrounding white matter intensities (WMH), simulating either "T2-like" or "T1-like" lesion appearances [Laso et al., 2024]. While this improves the model's ability to recognize WMH, this solution remains limited: it depends on simplified models of lesion appearance. This limits the generalizability of SynthSeg-WMH when applied to other disease cases, such as tumors.

3 Evaluation Metrics and Training Objectives

3.A Segmentation Metrics

Evaluation metrics quantify the agreement between predicted and reference segmentations and are essential to assess the accuracy, robustness, and consistency of a model. In medical image segmentation, metrics are typically divided into two families: overlap-based measures, which evaluate volumetric agreement, and boundary-based measures, which assess spatial precision along structure surfaces. In this work, we describe the Dice similarity coefficient (DSC), the Average Symmetric Surface Distance (ASSD), and the 95th percentile Hausdorff distance (HD95), as they jointly capture region- and boundary-level accuracy. Other metrics such as the Jaccard index or volumetric overlap error are sometimes used, but they provide equivalent or redundant information with respect to DSC.

Dice Similarity Coefficient (DSC)

$$\mathrm{DSC} = \frac{2\,|X \cap Y|}{|X| + |Y|}$$

where X and Y denote the sets of voxels belonging to the predicted and reference segmentations, and $|\cdot|$ is the voxel count. The DSC measures volumetric overlap between two segmentations and ranges from 0 (no overlap) to 1 (perfect match). It is equivalent to the voxel-wise F1 score in binary classification. DSC is relatively robust to moderate class imbalance but tends to under-penalize boundary errors on large structures. In practice, it provides a good overall indicator of segmentation quality and is the most commonly reported metric in medical imaging challenges.

While overlap-based metrics such as DSC provide a global estimate of segmentation accuracy, they do not reflect local geometric discrepancies along structure boundaries. For instance, two segmentations can yield similar DSC values despite differing substantially in boundary smoothness or spatial alignment. To address this limitation, surface-based metrics such as the Average Symmetric Surface Distance (ASSD) and the Hausdorff distance (HD95) are often employed to quantify boundary precision and outlier deviations.

Average Symmetric Surface Distance (ASSD)

$$\mathrm{ASSD} = \tfrac{1}{2}\left(\frac{1}{|S_P|} \sum_{x \in S_P} d(x, S_G) + \frac{1}{|S_G|} \sum_{y \in S_G} d(y, S_P) \right)$$

where S_P and S_G denote the predicted and reference surfaces, and $d(x, S_G)$ is the minimum Euclidean distance from a point x to the reference surface S_G. ASSD measures the mean distance between the two surfaces and reflects systematic boundary discrepancies. Unlike DSC, it is sensitive to shape alignment and surface smoothness, providing complementary information on spatial accuracy.

Hausdorff Distance (HD95)

$$\mathrm{HD} = \max\Big\{ \sup_{x \in S_P} d(x, S_G),\ \sup_{y \in S_G} d(y, S_P) \Big\}$$

The Hausdorff distance quantifies the largest boundary deviation between two segmentations. Because it is highly sensitive to isolated outliers, the 95th percentile of the distance distribution (denoted HD95) is commonly used as a more robust alternative. HD95 captures local mis-segmentations and outlier regions that might not strongly affect DSC or ASSD, thus offering a complementary perspective on segmentation reliability.

3.B Training Objectives and Loss Functions

The training objective defines how the model parameters are updated to minimize discrepancies between the predicted and ground-truth segmentations. In medical imaging, loss functions have to handle severe class imbalance, irregular object shapes, and the need for probabilistic calibration. Commonly used loss functions include the cross-entropy, Dice, and focal losses, each offering complementary properties depending on the segmentation task.

Cross-Entropy (CE) Loss: CE remains the standard choice for multi-class segmentation because it provides well-behaved gradients and calibrated probabilistic outputs. During early training, when predictions are noisy and overlap is low, CE stabilizes convergence by rewarding voxel-wise accuracy rather than global region agreement. However, in FLAIR-based lesion segmentation where lesions represent a minute fraction of voxels, CE tends to bias toward the background unless balanced through class weighting or patch sampling.

Dice Loss: Dice loss directly optimizes a differentiable form of the DSC metric. It compensates for class imbalance by emphasizing relative overlap rather than absolute voxel counts. This makes it highly effective for small or irregular structures. Yet, when used alone, Dice loss can produce over-smoothed contours or poorly calibrated probabilities, issues that affect both uncertainty estimation and boundary-based metrics such as HD95.

Hybrid Dice + Cross-Entropy Loss. In practice, most segmentation frameworks combine complementary loss terms to balance overlap accuracy and probabilistic calibration. The hybrid loss function used in this work is defined as:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{Dice}}\,\mathcal{L}_{\text{Dice}} + \lambda_{\text{CE}}\,\mathcal{L}_{\text{CE}},$$

where λ_{Dice} and λ_{CE} control the relative contribution of each term. Combining Dice and cross-entropy losses leverages their complementary properties. During early training, the cross-entropy term dominates, providing stable gradients and calibrated probabilistic outputs. As the segmentation becomes more localized, the Dice term gains influence, refining overlap accuracy and improving the handling of rare voxels. This combination aligns optimization objectives with the evaluation metrics (Dice and HD95) while preserving output calibration, consistent with the metric-aware principles advocated by Metrics Reloaded.

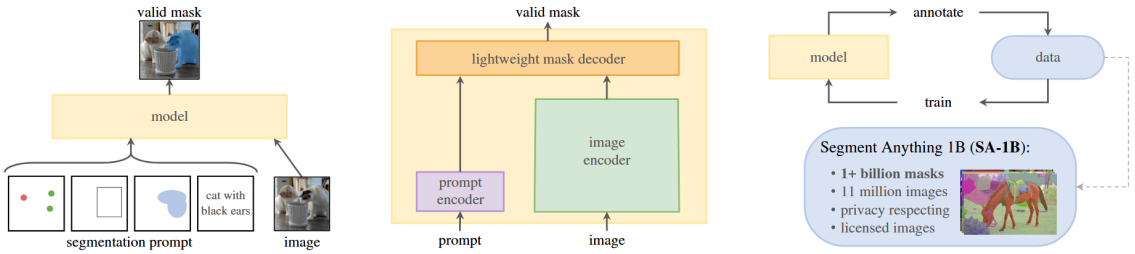
4 Current Limitations and Research Gaps

The review highlights critical gaps in current segmentation methods for FLAIR-only analysis in pathological cases:

- *Modality Mismatch:* The vast majority of segmentation models are trained and validated on T1w images. T1w images show high tissue contrast between structures. FLAIR image acquisition results in fundamentally different intensity distributions with less distinct tissue boundaries. The differences between FLAIR and T1w alter both global appearance and local texture, leading to poor feature correspondence between modalities. As a result, models trained on T1w data cannot be reliably applied to FLAIR images. While modality-agnostic approaches are proposed to enable cross-modality generalization, they typically are trained to rely on contrast learned from synthetic training examples rather than directly extracting features from real-world images. As a result, such approaches can be less accurate than a supervised learning approach, as shown in baseline comparisons. For example [Billot et al., 2023a] show that the proposed modality agnostic methods achieved lower Dice scores than T1w-trained supervised baseline, where SynthSeg achieved an average Dice score of 0.88 vs 0.91 for the supervised baseline. Another strategy is modality synthesis [Morell-Ortega et al., 2025], where missing contrasts (e.g., T1w) are generated from available acquisitions such as FLAIR and subsequently used with established segmentation pipelines. However, these methods introduce an additional layer of uncertainty, because segmentation accuracy depends on the quality of the synthesized images. In this thesis, we do not pursue modality synthesis and instead focus on developing segmentation methods that operate directly on FLAIR images.
- *Pathological Robustness:* Deep learning segmentation models are typically trained on datasets composed largely of clinically normal subjects. As a result, they learn to expect anatomical structures with regular shapes, consistent spatial locations, and typical intensity patterns. In clinical scenarios, however, MRIs are acquired under suspicion of neurological conditions, and the presence of lesions, edema, or mass effect can significantly alter both the appearance and spatial configuration of brain structures. This mismatch between the training distribution and the clinical imaging domain often leads to degraded performance, particularly in areas affected by pathology.
- *Limited Training Data:* To the best of our knowledge, there are no publicly available FLAIR datasets that include detailed anatomical annotations beyond lesion labels. While FLAIR is widely used in clinical practice, existing datasets typically focus on lesion segmentation (e.g., white matter hyperintensities, tumors) rather than full brain parcellation. The absence of manually labeled anatomical structures on FLAIR images severely limits the development and benchmarking of supervised segmentation methods tailored to this modality.

An additional note concerns the recent introduction of foundation models such as the Segment Anything Model (SAM) and MedSAM, which aim to provide generalizable segmentation capabilities across modalities and tasks. These models operate in a prompt-based manner, requiring either a point or bounding box to guide the segmentation. Independent assessments using manually drawn 3D box prompts on real-world

clinical data reveal substantial performance gaps compared to classical U-Net models. For example, Fu et al. find that MedSAM performs significantly worse than a supervised nnU-Net on the public glioma dataset MSD-BRATS [Fu et al., 2025]. This gap might be partly explained by the current architecture of MedSAM, which is limited to 2D slice-wise inference and expects RGB input, thus constraining its ability to fully exploit 3D context and native medical image contrast. To our knowledge, the performance of foundation models for anatomical structure segmentation has not yet been systematically evaluated. In this thesis, however, we focus on segmentation methods specifically designed for FLAIR images rather than on foundation model approaches.



(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**) (c) **Data**: data engine (top) & dataset (bottom)

Figure 2.6: SAM is a prompt-based segmentation model requiring both a color-based image and a user-provided prompt to obtain a segmentation. Source: [Kirillov et al., 2023]. Attempts at using SAM for medical image segmentation are achieved by fine-tuning the model on medical images (MedSAM) [Ma et al., 2024]. However, these models are limited to 2D slice-wise inference and expect RGB input, thus constraining their ability to fully exploit 3D context and native medical image contrast.

These limitations highlights the need for segmentation methods specifically designed for FLAIR images, *i.e.*, approaches that can handle pathological variability and operate without relying on additional modalities such as T1w MRI. The following chapters present novel methodologies tailored to the contrast properties of FLAIR and the challenges of clinical neuroimaging.

Chapter 3

FLAIR-ICE: Intracranial Cavity Extraction from FLAIR MRI

1 Introduction 46
2 Materials 48
2.A Augmentations 51
2.B Training configuration 53
3 Results 54
3.A Quantitative Evaluation 54
3.B Qualitative analysis 56
4 Conclusion 57

Chapter Overview. This chapter introduces *FLAIR-ICE*, a FLAIR-specific brain extraction method designed to ensure robust performance across heterogeneous MRI data. The approach addresses the variability of image provenance, including anonymization or prior preprocessing procedures, and remains reliable in the presence of pathology-related changes that alter image appearance. It allows us to establish a preprocessing pipeline for subsequent FLAIR-based segmentation tasks.

---

1 Introduction

Brain extraction refers to the process of separating brain tissue from non-brain structures (e.g., skull, scalp, and dura) in MRI scans. It is a critical preprocessing step in most neuroimaging pipelines and is required for downstream analyses such as bias field correction, spatial normalization, and tissue segmentation. Early methods for brain extraction typically relied on image intensity heuristics (usually applied to T1w images due to their excellent tissue contrast), deformable surface models [Smith, 2002]. Atlas-based anatomical priors also contributed to early solutions [Ashburner and Friston, 2005]. These classical approaches established the foundation for automated segmentation, but they often required manual parameter tuning and showed limited robustness when applied to heterogeneous data, including the presence of lesions, tumors, or acquisition-related artifacts. Notable traditional brain extraction methods include:

BET A fast, surface-based method that segments brain tissue by evolving a deformable mesh initialized near the brain center. BET is a popular choice thanks to its simplicity and integration in the FSL software suite, but its performance can be sensitive to parameter settings and image quality [Smith, 2002].

BeaST A patch-based approach leveraging a library of pre-segmented, pre-registered images. By fusing multi-scale patch predictions, BeaST improves robustness and accuracy across heterogeneous datasets [Eskildsen et al., 2012].

NICE Extends the patch-based framework of BeaST and includes external cerebrospinal fluid (CSF), yielding anatomically complete intracranial masks. NICE is particularly suited for total intracranial volume estimation and is widely used in large-scale neuroimaging studies [Manjón et al., 2014].

While traditional methods remain effective under controlled conditions, they are typically time-consuming and sensitive to acquisition variability. Recent advances in deep learning have transformed brain extraction, enabling near real-time segmentation and improved generalization to unseen data, and brain extraction should no longer represent a computational bottleneck.

Several deep learning-based approaches have emerged to address these limitations:

DeepICE Employs a 3D U-Net architecture to segment the intracranial cavity, mirroring NICE's anatomical definition but leveraging deep learning for improved speed and automation. Trained on large, heterogeneous datasets, DeepICE achieves robust performance across diverse pathologies and scanners while drastically reducing computation time [Manjón et al., 2020b].

SynthStrip Utilizes a U-Net trained exclusively on synthetic MRIs generated from anatomical label maps, following the SynthSeg framework. By simulating broad anatomical and intensity variability, SynthStrip generalizes well across modalities and scanner types, achieving robust skull stripping even on previously unseen data [Hoopes et al., 2022].

HD-BET A 3D U-Net trained on a large, multi-site, multi-modal dataset encompassing diverse MRI sequences and clinical pathologies. Its broad training distribution enables reliable performance across different acquisition protocols and neurological conditions, making it suitable for both research and clinical applications [Isensee et al., 2019].

	FLAIR-ICE	**DeepICE**	**SynthStrip**	**HD-BET**
Modality	FLAIR	T1w	any*	any*
Cerebral & cerebellar WM	✓	✓	✓	✓
Cerebral & cerebellar GM	✓	✓	✓	✓
Ventricular CSF	✓	✓	✓	✓
Surface CSF	✓	✓	✓	×
Brainstem (pons, medulla)	✓	✓	✓	✓
Internal brain vessels	✓	✓	✓	✓
External CSF	✓	✓	✓	×

Table 3.1: Comparison of anatomical content included in the output masks of NICE, DeepICE, SynthStrip (default FreeSurfer implementation), and HD-BET. WM: white matter; GM: gray matter; surface CSF: cerebrospinal fluid in deep sulci and the subarachnoid space.

Despite their strong performance, deep learning-based methods remain sensitive to data heterogeneity. A key challenge arises from variations in preprocessing and anonymization, particularly in public or user-submitted datasets such as those hosted on VOLBRAIN.NET. Images may arrive partially skull-stripped, facially masked, or normalized using undocumented intensity transformations. Such inconsistencies introduce distribution shifts that can degrade model reliability when applied to real-world data, and segmentation errors at this stage can propagate through subsequent

pipelines, leading to biased tissue quantification or distorted volumetric estimates of atrophy and lesion load.

Another important distinction across methods lies in the definition of the "intracranial cavity." For instance, NICE, DeepICE, and SynthStrip include external CSF within their intracranial masks, whereas HD-BET excludes it (see Table 3.1). This discrepancy complicates cross-method comparisons and highlights the need for consistent anatomical definitions when integrating brain extraction outputs into larger neuroimaging frameworks.

Finally, the generalization capacity of CNN-based methods remains strongly linked to the diversity of their training data. Models trained primarily on healthy subjects or restricted pathologies may fail when applied to populations with complex neurological conditions, such as multiple sclerosis, traumatic brain injury, or brain tumors, where lesions and structural deformations alter tissue appearance and boundary contrast.

Problem Statement: Despite the widespread clinical use of FLAIR MRI, particularly in aging, multiple sclerosis, stroke, and neurovascular disorders, robust brain extraction tools tailored to this modality remain limited. Existing approaches, even those designed for cross-modality generalization, frequently underperform on FLAIR images due to their high inter-scanner variability and frequent presence of pathology, which challenge both classical and deep learning-based methods.

Contribution: In this work, we present **FLAIR-ICE**, a robust brain extraction method specifically designed for FLAIR MRI. Our approach is engineered to handle the heterogeneity of FLAIR images and to remain resilient to upstream preprocessing artifacts such as anonymization-induced masking or skull-stripping errors. By focusing solely on FLAIR, FLAIR-ICE enables accurate brain extraction even when T1w scans are unavailable, corrupted, or incomplete, broadening the applicability of automated neuroimaging pipelines in both clinical and retrospective studies. We benchmark FLAIR-ICE against existing state-of-the-art tools, including SynthStrip and HD-BET, and demonstrate superior performance across diverse and challenging imaging scenarios.

2 Materials

Overview: This section describes the FLAIR MRI databases and preprocessing procedures used for training and evaluating the proposed FLAIR-based intracranial extraction method. The databases includes clinically normal subjects as well as patients with various neurological conditions, ensuring diversity across both acquisition parameters and anatomical variability. A complete overview of all training and testing databases used in this chapter is presented in Table 3.2.

Training Databases: Two primary databases provided training data, with additional augmentation to improve model robustness to pathological and acquisition-related variability.

- *Multiple Sclerosis* (OFSEP) – A large multi-site clinical cohort of patients with multiple sclerosis (MS) acquired across 36 sites in France, using a wide range of scanner models and protocols [Vukusic et al., 2020]. The cohort includes subjects with varying disease severity (EDSS 0–9, median 2.5). For this work, we used only T1w and FLAIR images. Brain masks were obtained using the NICE segmentation method [Manjón et al., 2014]. We randomly selected 600 images from the full cohort after quality control using RegQCNet [Denis de Senneville et al., 2020]. We retained 100 images for testing after a train-test split. Of the remaining 500 images, 250 images were left as is, and 250 were used for synthetic tumor augmentation (see Section 2.A). It is important to note that train-test split was performed at the subject level, ensuring that no subject included in the test set was used during training.
- *Glioma*(UCSF-PDGM) – 501 patients with grade II–IV gliomas contributed by the University of California, San Francisco, as part of the UCSF-PDGM dataset. [Calabrese et al., 2022] The dataset includes volumetric T1w, T1c, T2w, and FLAIR sequences. We used only FLAIR images in this work. The dataset was distributed skull-stripped as part of anonymization, so we generated brain masks via simple intensity thresholding. Tumor masks were available and divided into subregions (enhancing tumor, peritumoral edema, necrotic core); we used only the union of all subregions. Of the 501 subjects, 100 images were retained for testing after a train-test split. The remaining images contributed to synthetic image augmentation; in total, we randomly selected 250 images for generation of synthetic training data (see Section 2.A). Similarly to the previous dataset, the train-test split was performed at the subject level, ensuring that no subject included in the test set was used during training.

Evaluation Databases: Model evaluation used a diverse collection of FLAIR MRI databases designed to assess both in-domain and out-of-distribution generalization. The databases encompass healthy controls, multiple neurological conditions, and heterogeneous acquisition protocols. We considered two main evaluation groups: (1) real and synthetic clinical datasets for assessing generalization across pathologies, and (2) augmented datasets simulating acquisition-induced variability.

- *Clinically Normal* (UKB) – 100 FLAIR MRIs selected from the UK Biobank (UKB) database. [Bycroft et al., 2018] The subjects are clinically normal adults with no reported neurological or psychiatric disorders. The dataset benefits from harmonized acquisition protocols and extensive quality control, providing high-quality reference data for in-domain evaluation. The cohort primarily includes middle-aged to older adults (mean age ∼65 years).
- *Multiple Sclerosis* (OFSEP) – 100 FLAIR MRIs from the multicentric OFSEP cohort, including patients with varying MS lesion loads (EDSS range 0–9). Images were acquired at 36 French clinical centers using heterogeneous scanners and acquisition parameters. We used this dataset to evaluate robustness under clinically realistic variability in both pathology and acquisition.

- *Traumatic Brain Injury* (ADNIDOD) – 100 FLAIR MRIs from the DOD–ADNI study, comprising subjects with mild to moderate traumatic brain injury. The dataset introduces anatomical and intensity heterogeneity distinct from demyelinating lesions, allowing evaluation of cross-pathology generalization.
- *Glioblastoma* (UCSF-PDGM) – 100 FLAIR MRIs from the UCSF-PDGM glioma cohort, including patients with WHO grade II–IV gliomas. All images were skull-stripped during anonymization. The dataset provides expert tumor annotations and represents a strong out-of-distribution test set due to the presence of large, heterogeneous tumor masses.
- *Unknown* (VOLBRAIN.NET) – 100 FLAIR MRIs submitted by users to volbrain.net. The dataset includes anonymized scans from diverse scanners, institutions, and preprocessing pipelines (e.g., skull-stripped or masked images). We used it to assess robustness to uncontrolled preprocessing and acquisition conditions.

To further evaluate resilience to acquisition-related artifacts, we generated augmented variants of the OFSEP dataset that simulate common sources of intensity and geometric variation:

- OFSEP* – Augmented with random intensity and resolution perturbations to emulate inter-scanner or inter-protocol variability.
- OFSEP† – Modified through elastic deformation and partial skull-stripping errors to simulate anonymization- or preprocessing-induced distortions.
- OFSEP^{+} – Enhanced with "perfect" skull stripping based on ground-truth intracranial masks obtained using NICE segmentation, serving as an upper-bound reference for brain extraction accuracy.

Synthetic Training Dataset: The training data for FLAIR-ICE consisted of 500 FLAIR MRIs synthetically generated by lesion insertion onto normal FLAIR images. Specifically, 250 clinically normal OFSEP FLAIR scans were augmented with tumors derived from the UCSF-PDGM datasets. Half of the original OFSEP FLAIR images were left unaltered (containing native WMH), while the other half were augmented with synthetic tumors. This design exposed the network to a balanced distribution of healthy, lesion-bearing, and tumor-augmented cases, improving generalization to unseen anomalies.

- OFSEP (250 images): Multi-site FLAIR MRIs from patients with multiple sclerosis, containing a wide range of native white matter hyperintensities (WMH) and lesion loads. These images provide real pathological variability and reflect clinical acquisition heterogeneity.
- OFSEP/UCSF-PDGM (250 images): Synthetic FLAIR MRIs generated by inserting glioma lesions (from UCSF-PDGM) into OFSEP FLAIR scans. This augmentation simulates tumor pathology superimposed on MS backgrounds, increasing the diversity of lesion types and anatomical presentations encountered during training.

Dataset	Subjects	Description
		Training Images
OFSEP	250	Multiple sclerosis cohort (real FLAIR; WMH)
OFSEP/UCSF-PDGM	250	MS FLAIR with synthetic tumors (glioma augmentation)
		Testing Images
UKB	100	Clinically normal (in-domain baseline)
OFSEP	100	Multiple sclerosis (WMH)
ADNIDOD	100	Traumatic brain injury (structural variability)
UCSF-PDGM	100	Glioblastoma (tumor pathology)
VOLBRAIN.NET	100	Heterogeneous preprocessing (user-submitted)
		Augmented Testing Images
OFSEP*	100	Intensity/resolution perturbations (scanner simulation)
OFSEP†	100	Simulated bad skull-stripping (elastic deformation)
OFSEP^{+}	100	Perfect skull-stripping (NICE ground truth)

Table 3.2: Overview of FLAIR MRI datasets used for training and evaluation of the proposed intracranial extraction method.

Preprocessing: We preprocessed all training and testing images using the DeepLesionBrain pipeline [Kamraoui et al., 2022]. The following steps were applied uniformly across datasets:

(*1*) **Denoising:** We denoised images using the adaptive non-local means algorithm. [Manjón et al., 2010b] This reduces random noise while preserving anatomical detail.

(*2*) **Bias Field Correction:** We corrected intensity non-uniformities with N4ITK. [Tustison et al., 2010] This improves tissue contrast and segmentation accuracy.

(*3*) **Co-registration:** We linearly registered FLAIR images to their corresponding T1w images using ANTs to ensure anatomical alignment across modalities. [Avants et al., 2008]

(*4*) **Normalization:** We nonlinearly normalized T1w images to MNI152 space, and then applied the resulting transformation to FLAIR images for cross-subject spatial consistency. [Fonov et al., 2011]

(*5*) **Brain Extraction:** We generated intracranial masks from T1w images using NICE, providing consistent brain boundaries and enabling computation of total intracranial volume (TIV) for normalization and quality control. [Coupé et al., 2010]

2.A Augmentations

To enhance the robustness of our model to acquisition variability and preprocessing artifacts, we employed a series of augmentation strategies during training. These augmentations reflect realistic scenarios encountered in clinical data, particularly those represented in our modified test datasets (OFSEP*, OFSEP†, and VOLBRAIN.NET).

Resolution Augmentations: To simulate the variability introduced by acquisition protocols, particularly 2D FLAIR sequences with low through-plane resolution, we applied random downsampling along each spatial axis, with factors up to $4\times$. We applied additional random downsampling specifically along the axial (z) axis to emulate thick-slice, anisotropic acquisitions. These transformations reflect common characteristics of clinical FLAIR images and introduce resolution-induced artifacts such as partial volume effects. We also used this type of augmentation to generate the OFSEP* test set.

Simulated Preprocessing Artifacts: To improve robustness to artifacts introduced by anonymization procedures such as defacing or skull stripping, we simulated anonymization errors by applying smooth elastic deformations to the ground-truth brain mask. Specifically, we generated a random displacement vector field followed by Gaussian smoothing, and then performed skull stripping using the inaccurate deformed mask. We trained the model to segment the brain within the intersection of the original and deformed brain masks, effectively teaching it to ignore erroneously removed or distorted regions. We also used this transformation to construct the OFSEP† test set, allowing for controlled evaluation under such distortions.

Augmented Synthetic Lesions Generation: To enhance the model's sensitivity to pathological hyperintensities and mitigate overfitting to the training set lesion distribution, we implemented a lesion augmentation strategy. We extracted hyperintense lesions from publicly available annotated datasets (listed in Section 2) and then spatially aligned them to healthy FLAIR images from our training set. For each lesion-containing image we first registered it to the target FLAIR image using linear registration, and we transformed the corresponding lesion mask accordingly. We obtained the final augmented image using Poisson image editing [Pérez et al., 2003], a gradient-based blending technique that seamlessly integrates the tumor region by solving a Poisson partial differential equation. This method minimizes intensity discontinuities at lesion boundaries by reconstructing pixel values based on gradient information from both the source lesion and target image, ensuring smooth intensity transitions. The Poisson blending approach preserves realistic intensity profiles and spatial characteristics, making the inserted lesions appear naturally integrated into the target image while maintaining the underlying anatomical context. We then used the augmented images to train the model, allowing it to learn to segment both healthy and pathological hyperintensities effectively.

These augmentations ensure that the model is exposed to a broad spectrum of lesion appearances and intensity profiles, not limited to WMH hyperintensities caused by MS lesions. This exposure encourages better generalization to real pathological cases, including lesions and tumors, even when trained on healthy or limited pathology datasets. However, this approach depends on the availability of well-defined lesion masks, since the augmentation process requires explicit delineation of the region to

be transplanted. Such high-quality masks are not systematically available across all neurological conditions; for example, some stroke datasets derive segmentation from low-resolution diffusion-weighted images, resulting in unclear boundaries after alignment with FLAIR images.

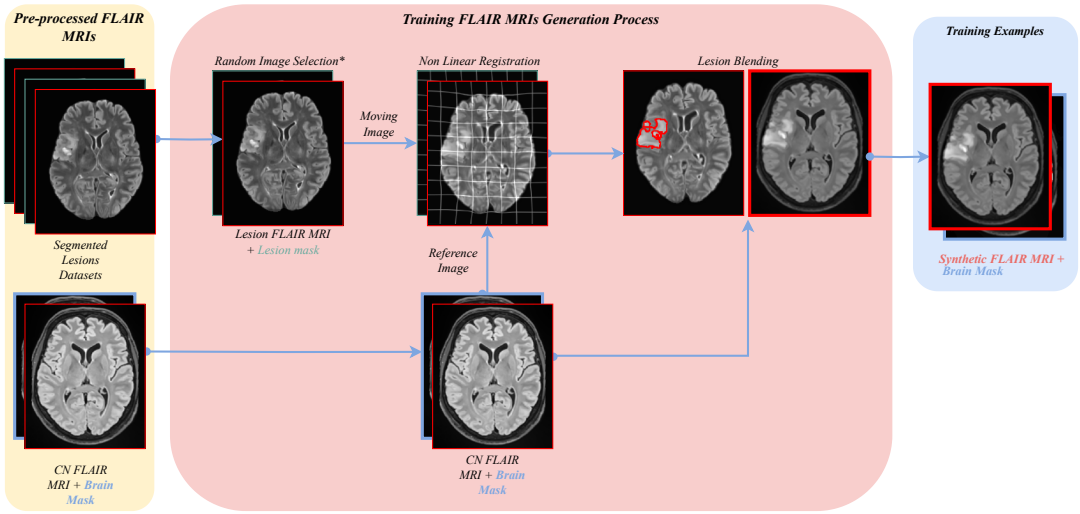


Figure 3.1: Illustration of the augmented training sample generation process. A randomly selected subject with a segmented lesion or tumor is registered to a randomly chosen healthy training image. The lesion or tumor is then inserted into the target image using a blending method to produce a realistic augmented example.

2.B Training configuration

We used a 3D U-Net based on the 3D full-resolution configuration of nnU-Net: a reliable and well-established segmentation framework with demonstrated robustness across a wide range of medical image segmentation challenges. [Isensee et al., 2021] Adopting nnU-Net as a baseline enables reproducible training and consistent evaluation without relying on highly customized, task-specific architectures. We trained networks using stochastic gradient descent (SGD) with Nesterov momentum, $\alpha = 0.99$, an initial learning rate of 0.1, and a polynomial learning rate scheduler defined as $lr(t) = (1 - t/t_{\max})^{0.9}$. We trained for 1000 epochs, each consisting of 250 iterations, with a batch size of 4.

During training, we randomly sampled patches of size $112 \times 128 \times 112$ voxels from preprocessed volumes. We applied standard nnU-Net augmentations on-the-fly, including spatial transforms (rotation, scaling) and intensity modifications (Gaussian noise, Gaussian blur, brightness/contrast adjustment, gamma correction). Unlike the default configuration, we performed simulated low-resolution resampling in two steps, as described in Section 2.A. We optimized the segmentation objective using a combination of Dice and cross-entropy loss functions. At inference, we segmented

full volumes using a sliding window approach with overlapping patches, following the default nnU-Net strategy, with $\alpha = 0.5$ denoting the overlap between patches.

3 Results

3.A Quantitative Evaluation

We evaluated the accuracy of FLAIR-ICE in comparison to publicly available segmentation methods applicable to FLAIR images. Since most brain extraction methods do not share the same anatomical definition or are not optimized for FLAIR, we restricted the comparison to SynthStrip, which is compatible with FLAIR and includes a comparable definition. We assessed performance using two standard metrics: Dice Similarity Coefficient (DSC) and 95th percentile Hausdorff distance (HD95), where higher DSC and lower HD95 indicate better agreement with the reference.

	DSC ↑		**HD95 [mm] ↓**		**ASD [mm] ↓**		**CC ↑**	
Dataset	SynthStrip	FLAIR-ICE	SynthStrip	FLAIR-ICE	SynthStrip	FLAIR-ICE	SynthStrip	FLAIR-ICE
ADNIDOD ($N = 100$)	94.61 ± 0.72	97.55 ± 0.39	7.38 ± 10.02	2.80 ± 0.51	3.00 ± 5.02	1.08 ± 0.16	0.92 ± 0.09	0.97 ± 0.01
OFSEP ($N = 100$)	95.58 ± 0.50	98.85 ± 0.25	5.09 ± 0.42	1.50 ± 0.36	1.93 ± 0.21	0.56 ± 0.10	0.94 ± 0.01	0.98 ± 0.00
Volbrain.net ($N = 100$)	96.05 ± 0.55	98.53 ± 0.71	4.73 ± 0.51	1.92 ± 0.56	1.73 ± 0.22	0.69 ± 0.15	0.95 ± 0.01	0.98 ± 0.01
UCSF-PDGM ($N = 100$)	97.33 ± 0.80	98.93 ± 0.30	3.04 ± 1.01	1.22 ± 0.34	1.02 ± 0.23	0.52 ± 0.13	0.97 ± 0.01	0.99 ± 0.00
UKB ($N = 100$)	95.67 ± 0.48	98.81 ± 0.17	5.13 ± 0.44	1.50 ± 0.26	1.90 ± 0.20	0.59 ± 0.07	0.94 ± 0.01	0.98 ± 0.00
Augmented Testing Images								
OFSEP* ($N = 100$)	95.71 ± 0.57	98.67 ± 0.34	5.08 ± 0.45	1.67 ± 0.43	1.87 ± 0.24	0.64 ± 0.14	0.94 ± 0.01	0.98 ± 0.00
OFSEP† ($N = 100$)	97.73 ± 0.58	98.85 ± 0.71	3.29 ± 0.74	2.66 ± 1.53	1.19 ± 0.31	0.98 ± 0.61	0.96 ± 0.01	0.97 ± 0.02
OFSEP^{+} ($N = 100$)	97.88 ± 0.31	99.77 ± 0.07	2.86 ± 0.41	1.00 ± 0.00	0.97 ± 0.12	0.12 ± 0.04	0.97 ± 0.00	1.00 ± 0.00

Table 3.3: Comparison of brain extraction performance between SynthStrip and FLAIR-ICE across multiple test sets. Metrics: Dice similarity coefficient (DSC), 95th percentile Hausdorff distance (HD95), average surface distance (ASD), and cross-correlation (CC). Higher DSC and CC, and lower HD95 and ASD indicate better performance.

Across all datasets, FLAIR-ICE achieved equal or superior performance compared to SynthStrip. Improvements were particularly notable in pathological datasets (OFSEP, UCSF-PDGM), where competing methods exhibited greater variability. For example, in OFSEP, FLAIR-ICE obtained higher DSC and lower HD95 than SynthStrip (see Table 3.3). These consistent gains demonstrate better robustness to disease-related intensity variations.

See Figure 3.2 for a broader view of performance; the figure shows the distribution of DSC scores across all datasets and methods. FLAIR-ICE (Panel B) displays consistently higher median values than SynthStrip (Panel A), with narrower variability. This suggests overall better accuracy and reliability, in the selected test conditions.

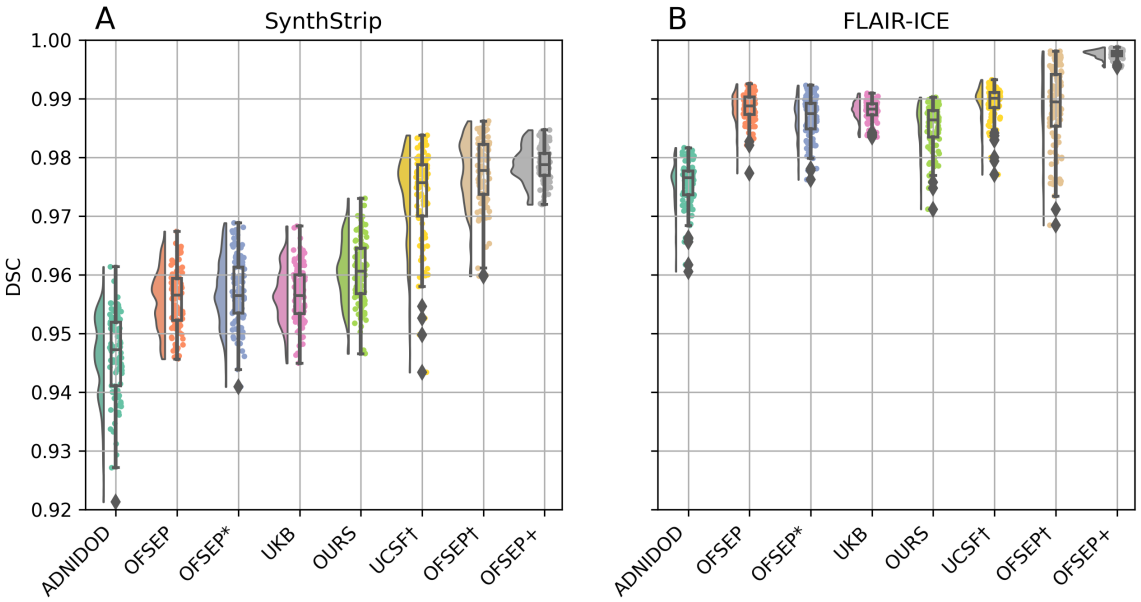


Figure 3.2: Raincloud plots showing the distribution of Dice similarity coefficient (DSC) scores for SynthStrip and FLAIR-ICE across all evaluation datasets. Each point represents a subject; violin plots illustrate score distributions. FLAIR-ICE consistently achieves higher median DSC and reduced variability compared to SynthStrip. *UCSF† corresponds to the UCSF-PDGM dataset, and is labelled this way due to images being skull-stripped as per BraTS sharing policies. Ours denotes the testing set composed of images from* VOLBRAIN.NET.

3.B Qualitative analysis

Figure 3.3 illustrates representative results for challenging cases. Even in its lowest-scoring examples (approximately 0.93 DSC on OFSEP), FLAIR-ICE produced anatomically coherent masks that preserved cortical outlines, whereas SynthStrip often generated fragmented or incomplete segmentations. Furthermore, FLAIR-ICE successfully handled skull-stripped and anonymized inputs, while SynthStrip systematically overestimated the intracranial cavity. This highlights a practical advantage, since many clinical and research pipelines rely on anonymized data.

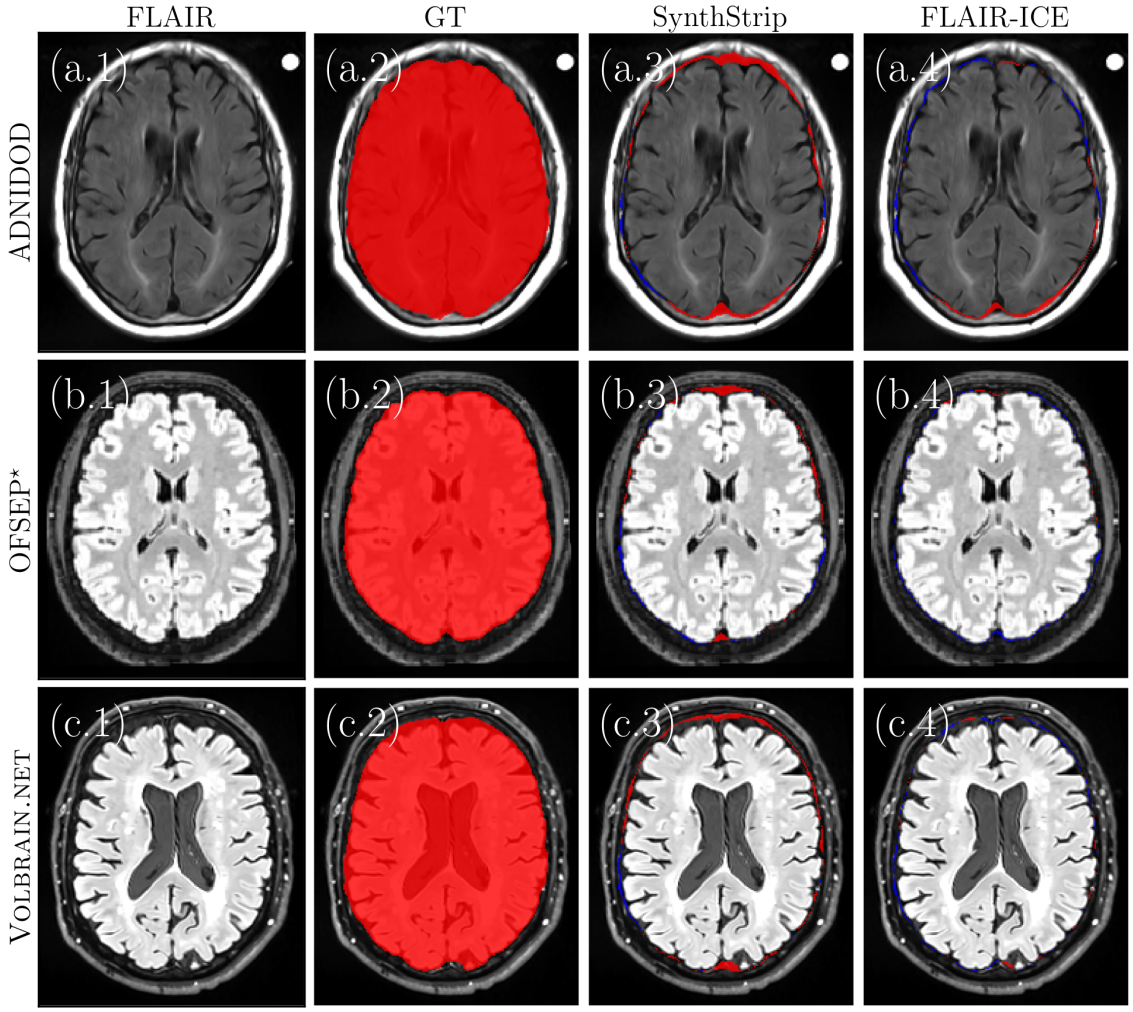


Figure 3.3: Qualitative examples of brain extraction on evaluation datasets. Columns (a) and (b) show the FLAIR image and the ground truth segmentation overlaid in red. Columns (c) and (d) display segmentation errors of SynthStrip and FLAIR-ICE, respectively, overlaid on the FLAIR image. False positives are shown in blue and false negatives in red. While SynthStrip often fragments or overestimates the intracranial cavity in challenging cases, FLAIR-ICE better preserves anatomical fidelity.

4 Conclusion

We introduced FLAIR-ICE, a brain extraction method tailored to FLAIR MRI. Unlike most existing tools, which are typically optimized for T1w images or assume full-head inputs, FLAIR-ICE operates directly on FLAIR scans and accommodates a wider range of preprocessing conditions, including images that have already undergone skull stripping or facial masking. Such cases are increasingly common due to data sharing and anonymization policies, which often apply de-identification procedures prior to public release or platform submission. Our results, evaluated across several datasets with varied pathologies and acquisition characteristics, show that FLAIR-ICE performs consistently well and compares favorably to other publicly available methods. The approach also demonstrates robustness to common artifacts such as resolution variability and anonymization-induced distortions, which can affect segmentation quality in practice.

By focusing specifically on FLAIR, this method may help improve brain extraction quality in settings where T1w scans are not available or usable. The use of synthetic lesion augmentation during training further contributes to its ability to generalize to a range of pathological cases. Overall, FLAIR-ICE offers a practical solution for brain extraction in clinical and retrospective studies, especially in contexts where modality constraints and prior preprocessing steps are unavoidable.

Chapter 4

FLAIRBrainSeg: Fine-Grained Brain Segmentation Using FLAIR MRI Only

1 Introduction 60
2 Materials & Methods 61
2.A Datasets 61
2.B Preprocessing 62
2.C T1w-based segmentation 62
2.D Quality control 63
2.E Deep learning model description 65
2.F Validation framework 66
3 Results 66
4 Conclusion 69

Chapter Overview: This chapter investigates the feasibility of performing fine-grained anatomical brain segmentation using only FLAIR MRI. It introduces FLAIRBrainSeg, a deep learning–based method specifically designed for FLAIR images and trained to reproduce the anatomical information typically obtained from T1w MRI. The method is evaluated against modality-agnostic alternatives, demonstrating that accurate anatomical segmentation is achievable from FLAIR alone, even in the presence of disease-related changes.

The methods discussed in this chapter correspond to the following publication:

Edern Le Bot, Rémi Giraud, Boris Mansencal, Thomas Tourdias, José V. Manjón, and Pierrick Coupé (2025). *FLAIRBrainSeg: Fine-Grained Brain Segmentation Using FLAIR MRI Only.* In *Proceedings of the Medical Imaging with Deep Learning (MIDL 2025)*, Salt Lake City, USA, July 2025. *Oral presentation.*

---

1 Introduction

Problem Statement: Accurate and detailed brain segmentation is essential for various clinical applications, including diagnosis, treatment planning, and monitoring of neurological diseases. Traditionally, T1w MRI has been the standard modality for brain segmentation. However, a significant limitation of relying solely on T1w MRIs is that they are not always available, particularly in certain clinical settings. For example, in the management of acute stroke patients, T1w images are not obtained as part of the initial imaging protocol [Powers et al., 2019]. Moreover, guidelines for diagnosing MS consider T1w acquisition optional, recommending only contrast-enhanced T1w images. If lesion segmentation and brain volumetry can be effectively performed using only FLAIR images, the acquisition of T1w sequences could become redundant [Goodkin et al., 2021], allowing for a reduction in protocol acquisition times.

Whole brain segmentation based on T1w MRIs: A large number of currently available segmentation methods rely on T1w MRIs for generating anatomical segmentations. Due to the challenges of applying deep learning (DL) methods in the context of limited training data, many existing methods prioritize coarse-grained segmentation (*e.g.*, fewer than 50 anatomical structures) [Henschel et al., 2020, Roy et al., 2019]. In contrast, fine-grained segmentation, which delineates over 100 structures, provides a more detailed understanding of brain anatomy, including cortical parcellation. The increased complexity of fine-grained segmentation, compared to coarse-grained approaches, necessitates the adoption of specialized strategies such as patch-based methods and model ensembling [Coupé et al., 2020, Huo et al., 2019],

or architecture improvements [Hatamizadeh et al., 2021] to achieve meaningful and accurate results.

Modality-agnostic whole brain segmentation: Modality-agnostic methods, make an attempt at resolving domain-shift differences in contrast and appearances in between different acquisition modalities. SynthSeg [Billot et al., 2023a], achieves this by introducing a domain randomization strategy to randomize the appearance of anatomical structures during training. In SynthSeg$^+$ [Billot et al., 2023b], the authors improve upon this, in order to obtain more detailed segmentations, where the initial segmentation obtained by SynthSeg is refined using a second "denoising" network. A final network is used to perform cortical parcellation, and allows for the obtention of a final fine-grained segmentation, at the cost of using multiple deep-learning networks. It is important to note that unsupervised (although limited to coarse-grained) modality-agnostic segmentation also exists. Notably, SAMSEG [Puonti et al., 2016], uses a Bayesian framework, relying on an intensity clustering algorithm and a probabilistic brain atlas for the production of coarse-grained segmentations.

To address the challenges of brain segmentation in the absence of T1w images, we propose a novel fine-grained brain segmentation based on FLAIR MRI only. We compare the accuracy of a domain-randomization strategy with randomized tissue intensities, as used by SynthSeg, to more traditional and direct "supervised learning" approaches based on labeled FLAIR data. Specifically, we highlight differences in the segmentations obtained in the presence of unexpected anomalies, such as multiple sclerosis lesions, which can significantly alter the appearance of MRI images and challenge the robustness of segmentation methods.

2 Materials & Methods

2.A Datasets

- Training dataset: 3577 pairs of T1w and FLAIR MRIs selected after the QC steps introduced below from the OFSEP [Vukusic et al., 2020] database were used for training data. The OFSEP database contains images from 36 different sites (with multiple different scanners) in France, from clinically diagnosed patients with multiple sclerosis, at multiple stages of advancement. T1w and FLAIR images selected from the OFSEP database had a wide range of resolutions ranging from 0.25x0.25x0.5mm to 2x2x3mm. However most available images had a 1mm isotropic sampling.
- Testing dataset:
  A stratified train-test split of all available images (prior to QC) was used to create the testing sets, ensuring balance across age groups and genders.
  - In-domain testing set: 100 pairs of images, not included in the training set, from the OFSEP database, were used for validation.

– Out-of-domain testing set: to perform the comparison on out-of-domain images, a selection of 100 pairs of images, from the UKBiobank [Bycroft et al., 2018] database, contains clinically normal patients from 3 different sites in the United Kingdom.

	Dataset	Number of subjects	Average age (years)	Sex distribution	EDSS ratings
Train Set	OFSEP	3577	41.77 ± 11.56	2613F/964M	$[0, 9.0]$
Test Set	OFSEP	100	41.93 ± 12.09	73F/27M	$[0, 6.0]$
	UKB	100	64.80 ± 11.01	50F/50M	n/a

Table 4.1: Dataset overview. Please note that the number of subjects and dataset statistics are reported *after all the quality control steps.* We include the range of Expanded Disability Status Scale (EDSS) ratings to provide an idea of the stage of progression of MS in the in-domain dataset (i.e., OFSEP).

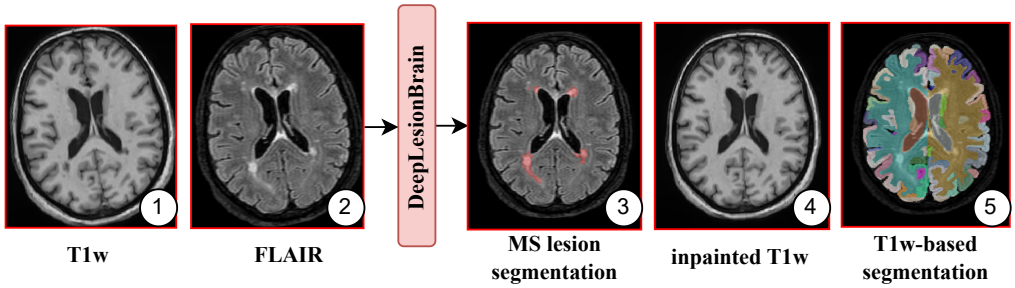
2.B Preprocessing

All the used pairs of FLAIR and T1w MRIs from the OFSEP database were processed using the DeepLesionBrain pipeline [Kamraoui et al., 2022]. The following steps were applied during the preprocessing : 1) all images were denoised using [Manjón et al., 2010b] 2) images were corrected using N4ITK bias-field correction [Tustison et al., 2010] 3) FLAIR images were linearly registered to the corresponding T1w MRI images, using ANTS [Avants et al., 2011] 4) FLAIR and T1w images were transformed to the MNI space. This last step was achieved by first registering the T1w images to the McGill MNI template [Fonov et al., 2011]. The resulting T1w-to-MNI transformation was then applied to the FLAIR images to ensure consistent alignment in the MNI space.

2.C T1w-based segmentation

The segmentation of lesions using FLAIR MRI and fine-grained brain segmentation using T1w were performed in a sequential pipeline. First, DeepLesionBrain [Kamraoui et al., 2022] was employed to segment MS lesions. The goal of producing lesion segmentation was to enable masking of MS lesions in the T1w image, thereby facilitating more accurate anatomical segmentation by AssemblyNet [Coupé et al., 2020]. Next, the lesion map was used to perform T1w inpainting [Manjón et al., 2020a]. Finally, AssemblyNet was applied to the inpainted T1w MRI (see Fig. 4.1), resulting in a fine-grained segmentation of 132 structures following the Neuromorphometrics protocol (see `https://neuromorphometrics.com/`). An overview of this training data generation process is available in Fig. 4.2 and Fig. 4.1.

Figure 4.1: Schematic overview of the T1w-based segmentation construction. From left to right, the images are : the (1) T1w and (2) FLAIR images, the (3) lesion maps, (4) inpainted T1w and (5) the obtained fine-grained segmentation.



AssemblyNet was selected for its proven test-retest reliability across multiple acquisitions in clinical settings [Coupé et al., 2020], ensuring robustness in varied imaging conditions. Additionally, it was specifically trained for fine-grained segmentation, aligning with the study's objective of achieving detailed anatomical delineation. After verification for possible misalignment between T1w and FLAIR MRIs, the T1w-based segmentation can be used as target labels for FLAIR images.

2.D Quality control

MNI misalignment: First, we assessed possible misalignment of the T1w-based segmentation and the FLAIR *training image* by first controlling registration errors between the FLAIR image and the MNI template, using RegQCNet [Denis de Senneville et al., 2020], a robust deep-learning-based method designed to estimate registration errors to the MNI space. In addition to detecting misalignment during registration, RegQCNet facilitated the exclusion of images with excessively poor quality, ensuring a more reliable dataset for analysis.

Segmentation misalignment: To assess possible misalignment between the generated segmentation and the FLAIR image, we generated a synthetic FLAIR-like image (see Fig. 4.3 at left), from the segmentation as follow : 1) For each anatomical structure in the segmentation, we extracted all corresponding voxel intensities from the FLAIR image, then we computed the local median intensity within each region to obtain an estimate of the typical intensity for that structure. 2) We generated the synthetic image by replacing every voxel within a labeled structure with its corresponding median intensity value observed above. 3) To simulate the effect of partial volume averaging, we applied a uniform filter over the synthetic image. Finally, we computed the correlation between the real FLAIR image and the synthetic FLAIR-like image based on T1w segmentation. The computed score reflects the alignment of the segmentation (*i.e.,* T1w MRI) and the FLAIR image, where higher correlation

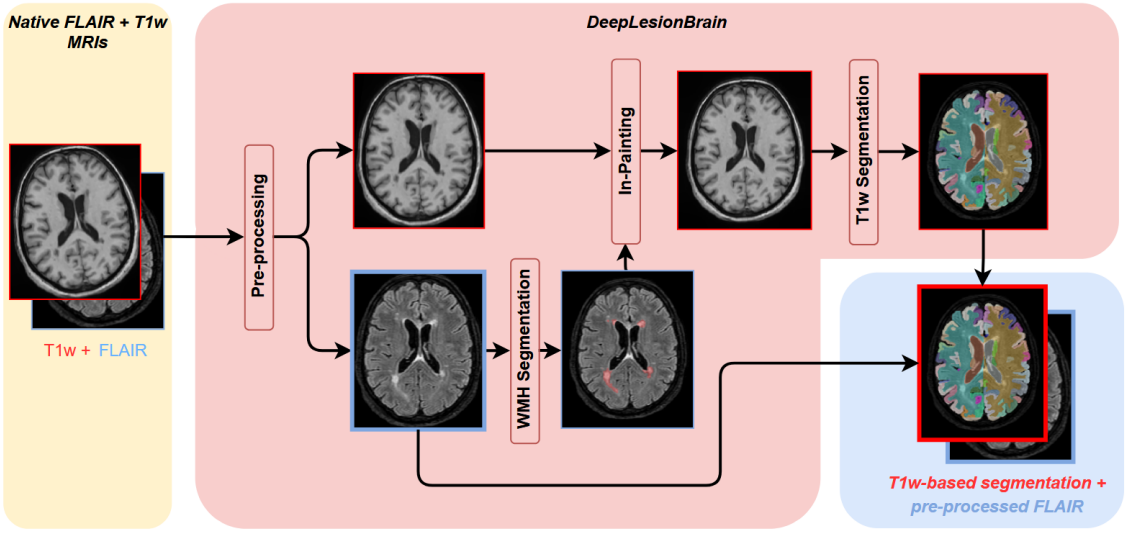
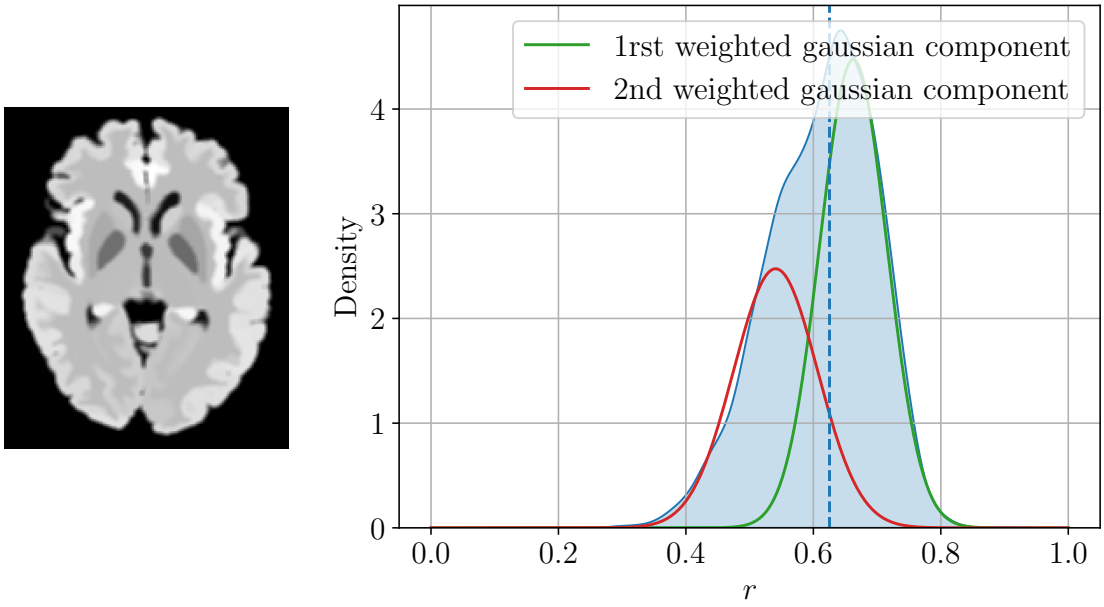


Figure 4.2: Illustration of the training generation process. Pairs of T1w and FLAIR images are pre-processed and processed by (1) segmenting the WMH visible in the FLAIR image, (2) performing an in-painting operation to replace WMH intensities with intensities of the surounding tissues (3) performing the fine-grained segmentation of the in-painted T1w image.

values indicate better spatial alignment. To identify and reject possibly misaligned images, we defined a rejection threshold based on the computed correlation metrics: given $S = \{s_1, s_2, \ldots, s_n\}$, where s_i is the computed correlation scores between the synthetic and the FLAIR image, by fitting a two-component (*i.e.* $K = 2$) Gaussian Mixture Model (GMM) $p(S) = \sum_{i=1}^{K} \phi_i \mathcal{N}(\mu_i, \Sigma_i)$ on the computed correlation scores, and selecting threshold $= \mu_0 + z^{0.9}\sigma_0$, where μ_0 and σ_0 are the mean and standard deviation of the lower Gaussian component corresponding to the misaligned FLAIR and segmentation, and $z^{0.9}$ is the 90th percentile of a standard normal distribution. The decision to fit a two-component GMM was based on the hypothesis that the correlation scores for failed and successful registrations would naturally cluster into two distinct distributions. Specifically, successful registrations are expected to yield high correlation scores, forming one Gaussian component, while failed registrations, with lower correlation scores, form a separate component (see Fig. 4.3 at right).

Visual assessment of testing images: Lastly, we conducted a human visual inspection of the registration process for all the *testing images*. The authors systematically reviewed the images for potential misalignment, specifically assessing discrepancies between the segmentation and the underlying anatomical structures. Any images exhibiting misalignment or visible defects were excluded from the testing dataset. For the in-domain dataset, 13 images were removed following this visual assessment, while none of the selected test images in the out-of-domain dataset required removal. To replace the removed images, a new test split to select replacement test images from the pool of available images was performed.

Figure 4.3: **Left**: Illustration of a synthetic FLAIR-like image created from T1w-based segmentation as part of the quality control process to address registration discrepancies. **Right**: Distribution of correlation scores on the training set. The vertical dotted line corresponds to the selected threshold.



2.E Deep learning model description

We used a 3D U-Net derived from the 3D configuration of the nnU-Net [Isensee et al., 2021]. The networks were trained using SGD with momentum ($\alpha = 0.99$), with a batch size of 4. A polynomial learning rate (lr) scheduling policy was used, where $lr(t) = (1 - t/t_{max})^{0.9}$, with an initial learning rate of 0.1. Similarly to nnU-Net, a combination of Dice and Cross-Entropy loss was used during training.

At each iteration, batch samples were randomly selected from the training dataset, before applying nnU-Net data augmentations, and randomly sampling a patch of size $112 \times 128 \times 112$ voxels. Augmentations introduced this way included linear spatial transforms (scaling, rotation), intensity transforms (Gaussian noise, Gaussian blur, brightness, contrast, and gamma transforms), and low-resolution resampling (up to 4mm). Unlike the default nnU-Net configuration, augmentations did not include image flipping, due to neuromorphometrics labels being lateralized. The segmentation network was trained during 1000 epochs, where each epoch was made of 250 iterations. While alternative methods, such as patch-based approaches [Coupé et al., 2020, Huo et al., 2019, Henschel et al., 2020], are often employed to address the complexities of fine-grained whole-brain segmentation, nnU-Net was selected for its robustness, ease of reproducibility, and established performance across multiple medical imaging segmentation challenges [Isensee et al., 2024]. Its adoption as a baseline allows for consistent comparisons while avoiding the variability

introduced by more complex, task-specific strategies.

2.F Validation framework

During our experiments, we compared FLAIRBrainSeg with SynthSeg [Billot et al., 2023a] and SynthSeg$^+$ [Billot et al., 2023b]. As previously described, SynthSeg relies on a single network to perform coarse-grained segmentation (*i.e.*, 34 structures) while SynthSeg$^+$ uses a cascade of four networks to produce fine-grained segmentation (*i.e.*, 99 structures). To ensure a fair comparison, we performed two experiments.

- **Experiment 1:** We retrained SynthSeg (using the code available at: `https://github.com/BBillot/SynthSeg`) using our T1w-based segmentations. Therefore, we obtained SynthSeg-132, a version of SynthSeg segmenting 132 structures according to the Neuromorphometrics protocol. Parameters proposed by [Billot et al., 2023a] were re-used for training, except for the number of iterations, where we used an increased number of 500 000 steps instead of 300 000 steps to account for the increased number of structures to segment. For evaluation, since the Neuromorphometrics protocol does not include non-brain tissues, we skull-stripped the test images with the foreground labels of the ground-truth segmentations. This can advantage SynthSeg-132, since other methods might estimate structures outside the brain mask. However, we wanted to make a fair comparison and be as close as possible to the original SynthSeg training, where every part of the image had corresponding labels.
- **Experiment 2:** For SynthSeg$^+$, we used the binary provided by FreeSurfer (7.4.1) to segment our testing datasets. To compare SynthSeg$^+$, SynthSeg-132 and FLAIRBrainSeg which follow different protocols, we manually selected the 35 corresponding structures between FreeSurfer protocol of SynthSeg$^+$ and Neuromorphometrics protocol of SynthSeg-132 and FLAIRBrainSeg. We emphasize that we acknowledge that this approach is not perfect, as differences may exist not only in the identified structures but also in their borders across protocols.

3 Results

The quantitative results based on the average Dice similarity coefficient (DSC) are presented in Fig. 4.5 and in Fig. 4.4, alongside the 95th percentile of the surface distance (SD95). A comparison of SynthSeg-132 and FLAIRBrainSeg by structure groups is detailed in Appendix 1. We assessed the statistical significance of our results by using a Wilcoxon signed-rank test. All p-values are inferior to 10^{-13} except for in-domain Experiment 2 between SynthSeg-132 and SynthSeg$^+$ ($p = 0.017$). Results from Experiment 1 show that on 132 structures, FLAIRBrainSeg outperforms SynthSeg-132 method by a substantial margin (respectively 7pp and 10pp higher

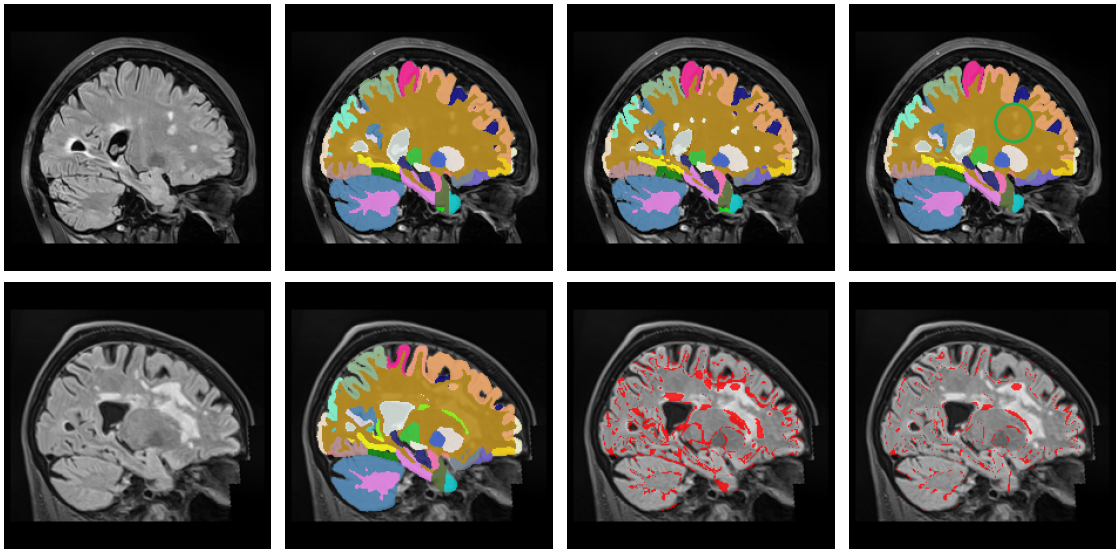


Figure 4.4: Examples of segmentation of 132 structures obtained with FLAIRBrainSeg and SynthSeg-132 in Experiment 1 are shown on different slices and images, for the in-domain (top) and out-of-domain (bottom) testing sets. The images were selected from cases with median mean DSC scores for SynthSeg-132 to ensure a fair representation of its performance. The top row highlights the accuracy of FLAIRBrainSeg, even in the presence of MS lesions, which are incorrectly labeled as ventricle or cortex by SynthSeg. The bottom row illustrates the differences between the segmentations obtained and the ground-truth labels on another example, further emphasizing the advantages of FLAIRBrainSeg in handling modality-specific challenges.

DSC) on the in-domain and out-of-domain datasets.

Additionally, results from Experiment 2 shows that FLAIRBrainSeg also outperforms SynthSeg-132 and SynthSeg$^+$ on the selected 35 structures used for comparison. SynthSeg-132 shows a statistically significant but minimal improvement over SynthSeg$^+$ (0.5pp higher DSC, $p = 0.017$) on the in-domain testing set; while SynthSeg$^+$ shows a statistically significant improvement in accuracy compared to SynthSeg-132 on the out-of-domain testing set (1pp higher DSC, $p < e^{13}$). Additionally, the variability and spread of accuracy distribution are reduced on the out-of-domain testing set compared to the in-domain set. This is consistent with the experiment design, as this second testing set contains higher quality images from clinically normal patients, less challenging to the segmentation models than the OFSEP clinical data of MS patients used as in-domain.

Examples of segmentation results from the in-domain and out-of-domain test sets are shown at the top of Fig. 4.4 and in Appendix 1. The inaccuracies observed in SynthSeg-132 highlight that its training scheme, designed to induce resilience across multiple contrasts and modalities, may reduce its performance when compared to a modality-specific model. Additionally, due to synthetic training images used

Figure 4.5: Average DSC for Experiment 1 and Experiment 2. **Top:** results on the in-domain dataset (i.e., OFSEP), **Bottom:** Results on the out-of-domain dataset (i.e., UKB), **Left:** Comparison of SynthSeg-132 and FLAIRBrainSeg over the 132 structures of the Neuromorphometrics protocol (Experiment 1), **Right:** Comparison of SynthSeg-132, SynthSeg$^+$, and FLAIRBrainSeg on the 35 common structures between FreeSurfer and Neuromorphometrics protocols (Experiment 2).

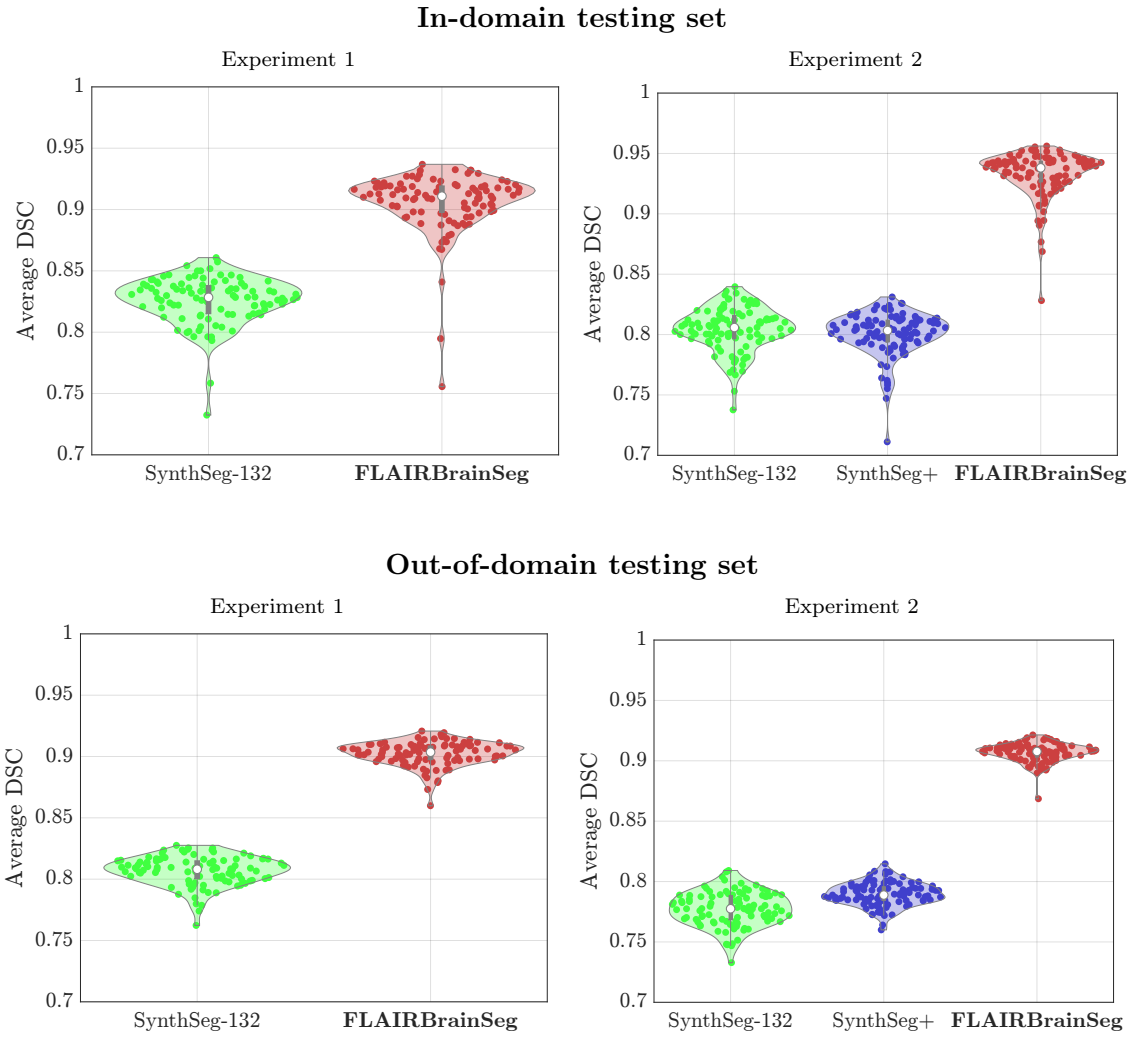


Table 4.2: Table report of Mean Dice Score and voxel-wise 95th Surface Distance (SD95) for each testing dataset on Experiment 1

			Mean	**Min**	**50%**	**Max**
In-Domain	FLAIRBrainSeg	Dice	0.90 ± 0.02	0.75	0.91	0.93
		SD95	0.56 ± 0.13	0.35	0.53	1.15
	SynthSeg	Dice	0.83 ± 0.01	0.73	0.83	0.86
		SD95	2.31 ± 0.86	1.15	2.11	5.70
Out-of-domain	FLAIRBrainSeg	Dice	0.91 ± 0.01	0.86	0.91	0.93
		SD95	0.56 ± 0.07	0.40	0.54	0.73
	SynthSeg	Dice	0.81 ± 0.01	0.76	0.81	0.83
		SD95	1.48 ± 0.20	1.34	1.45	2.27

by SynthSeg being generated from the segmentation, and since the segmentation does not include a label for the lesion, white matter hyperintensities from MS lesions are never seen in training. As a result, they often end up misclassified (for example, in Fig. 4.4, MS lesions are incorrectly classified as cortical and ventricles structures in the highlighted area for SynthSeg). While not surprising, this is not the case for FLAIRBrainSeg, as the structures are accurately described in the output segmentation. Therefore, in light of those comparisons, while SynthSeg and SynthSeg$^+$ offer the advantage of adaptability across multiple modalities, specialized models (such as FLAIRBrainSeg) might be preferable when targeting a specific modality.

While SynthSeg and FLAIRBrainSeg both employ single models for their respective segmentation tasks, SynthSeg$^+$ makes use of four different CNN models. This difference in model architecture directly impacts computational resource requirements and runtime. The inclusion of a post-processing step during SynthSeg and SynthSeg$^+$ inference also adds additional complexity, extending the overall runtime compared to FLAIRBrainSeg. For instance, when processed on GPU, the processing of 10 cases (without the inclusion of the pre-processing time) requires 29.1 seconds for FLAIRBrainSeg, 160.6 seconds for SynthSeg$^+$, and 120.8 seconds for SynthSeg.

4 Conclusion

In this work, we introduce FLAIRBrainSeg, a novel framework designed for fine-grained segmentation of FLAIR MRIs. The framework enables the production of detailed anatomical segmentations typically obtained from reliable T1w MRI-based methods, but solely using FLAIR MRIs. Our results demonstrate the effectiveness of this modality-specific approach, particularly in addressing challenges posed by abnormalities such as white matter hyperintensities. Furthermore, we highlight

the limitations of domain randomization techniques, particularly when dealing with pathological conditions in FLAIR MRIs, underlining the need for tailored methods for modality-specific segmentation tasks. A current limitation of our work is the generalization of FLAIRBrainSeg to low-resolution FLAIR images although our method was trained using down-sampling as data augmentation. In clinical practice, FLAIR scans are often acquired at a coarse resolution or as 2D slices with limited coverage along the z-axis. This can affect the image appearance after the linear resampling to the 1 mm template resolution, potentially decreasing segmentation accuracy. Future improvements could leverage recent advancements in super-resolution generative models [Morell-Ortega et al., 2024] to enhance the quality of low-resolution FLAIR images before segmentation. Future work will also focus on enabling the processing of FLAIR images with various types of lesions and disease-related anomalies, in particular incorporating patients with strokes and tumors. This will enhance the applicability of FLAIRBrainSeg and ensure its relevance in clinical settings. In the future, we plan to release this tool for free on `https://volbrain.net/`.

Chapter 5

FLAIRBrainSeg^{+}: Generalized Anatomical Segmentation on Pathological FLAIR MRI

1 Introduction 73
2 Methods 76
2.A Synthetic Data Generation 76
2.B Training Configuration: 79
2.C Validation Framework 79
3 Materials 81
3.A Datasets 83
4 Results 85
5 Conclusion 92

Chapter Overview: This chapter extends the FLAIR-based segmentation framework to achieve the joint segmentation of brain anatomy and pathology. It introduces FLAIRBrainSeg+, a unified model trained using a synthetic lesion generation process that inserts annotated lesions into previously segmented healthy images. This strategy increases the diversity of pathological appearances encountered during training and aims to improve the model's robustness and generalization when segmenting both anatomical structures and disease-related hyperintensities.

The methods discussed in this chapter correspond to the following publication:

Edern Le Bot, Rémi Giraud, Boris Mansencal, Thomas Tourdias, José V. Manjón, and Pierrick Coupé (2025). *FLAIRBrainSeg^{+}: Unified Fine-Grained Brain and Lesion Segmentation from FLAIR MRI Only.* Submitted to *Journal of Machine Learning for Biomedical Imaging (MELBA).*

---

1 Introduction

Problem Statement: A major challenge in neuroimaging is the joint segmentation of brain anatomy and pathological lesions. Fine-grained anatomical segmentation is crucial for tasks such as structural volumetry and cortical parcellation, while lesion segmentation provides complementary information about disease burden in conditions such as multiple sclerosis (MS), stroke, and glioma. The presence of pathologies introduces local signal alterations, intensity heterogeneity, and mass effects that can disrupt the assumptions underlying standard anatomical segmentation methods. At the same time, most lesion segmentation approaches are developed independently of anatomical frameworks, prioritizing the detection of abnormal tissues without explicitly incorporating information about the surrounding brain structures. This separation limits the clinical utility of automated tools for two main reasons:

(1) Accurate assessment of disease progression and severity often requires consideration of both anatomical and pathological information.

(2) In certain clinical settings where only single-sequence acquisitions are available (*e.g.*, FLAIR in stroke or MS), existing methods are not applicable due to required modalities or sequences not being available. For example, most anatomical segmentation methods rely on T1w, while for example in MS or stroke, the acquisition of volumetric T1w MRI is not systematically performed for follow-up imaging visits [Wattjes et al., 2021, Schmierer et al., 2019].

Addressing these challenges requires the development of unified segmentation frameworks that can robustly delineate both normal anatomy and pathological lesions from single-sequence MRI data, while accounting for the complex interactions between tissue types and disease processes.

Anatomical Brain Segmentation: Numerous methods have been developed for anatomical brain segmentation, with protocols and levels of granularity that vary considerably. Fine-grained approaches, which may delineate more than 100 structures, are less common than coarse-grained methods that typically identify fewer than 20 structures without cortical parcellation. Examples of widely used coarse-grained segmentation tools include FreeSurfer [Fischl, 2012], FSL [Jenkinson et al., 2012], and SPM [Ashburner and Friston, 2005]. While these approaches are valuable, fine-grained methods provide more detailed anatomical maps. Examples of fine-grained segmentation methods include AssemblyNet [Coupé et al., 2020] and SLANT [Huo et al., 2018], which can segment up to 133 structures, as well as FastSurfer [Henschel et al., 2020].

These fine-grained segmentation tools are designed for clinically normal T1-weighted MRIs, although some methods attempt to generalize to other modalities. For instance, SynthSeg [Billot et al., 2023a, Billot et al., 2023b] introduces a modality-agnostic segmentation method (for 33 to 68 structures) by training on synthetic contrasts generated using a generative model.

However, like other approaches, these methods are not inherently designed to handle pathological cases such as tumors, white matter hyperintensities, or stroke lesions, and can therefore produce inaccurate segmentations in such scenarios, as further illustrated in Chapter 4 and in our previous work [Le Bot et al., 2025].

Segmentation of tumors and White Matter Hyperintensites (WMH): In parallel, a large body of work has focused on lesion segmentation, with major advances catalyzed by public challenges such as BraTS for brain tumors and the WMH challenges (ICPR-MSLesSeg, ISBI, MICCAI). The primary motivation for tumor segmentation is to monitor tumor volume and assess treatment response, whereas WMH segmentation enables quantification of lesion load. Despite their clinical relevance, most FLAIR Hyper Intensity (FHI) segmentation frameworks are optimized independently of anatomical segmentation pipelines and typically rely on multi-modal MRI acquisitions. As a result, they provide limited anatomical context and often lack robustness when processing single-sequence data, such as FLAIR. Recent challenge leaderboards nonetheless highlight top-performing methods, including [LaBella et al., 2025] in BraTS-MEN (2023) a segmentation challenge for the segmentation of meningioma tumors, [Ferreira et al., 2024] in BraTS-GLI (2023) a segmentation challenge for glioma tumors, and [Rondinella et al., 2024] in MSLesSeg/ICPR (2024) a segmentation challenge for MS lesions [1]. These methods predominantly utilize deep learning architectures, often based on 3D U-Nets [Ronneberger et al., 2015], and incorporate various attention mechanisms and ensemble strategies to enhance performance. However, they are typically trained on multi-modal datasets and may not generalize well to single-sequence inputs. One

[1]These specific editions are mentioned because they are the most recent challenges aligned with the problem settings described here. Subsequent editions have shifted toward distinct clinical scenarios, *e.g.*, postoperative imaging or low-resolution data.

thing of note is that none of the methods mentioned above target multiple pathologies, that is, they are specifically designed for either tumor or a specific type of WMH.

Joint anatomical and WMH segmentation: Only a small number of methods have been proposed to support joint anatomical and lesion segmentation. [Dorent et al., 2021, Himmetoglu et al., 2025] introduced frameworks for learning joint coarse-grained anatomical and, respectively, SVD-related hyperintensities and tumor masks. More recently, SynthSeg-WMH [Laso et al., 2024], was introduced to segment both brain anatomy and WMH from FLAIR images. SynthSeg-WMH extends the SynthSeg [Billot et al., 2023a] framework by incorporating WMH into the training process, enabling simultaneous segmentation of 33 anatomical structures and WMH. However, this method is limited to WMH and does not address other lesion types such as tumors or stroke lesions. Additionally, SynthSeg-WMH relies on synthetic training data generated from a random generative model using label maps as prior for the image generation. This model is trained on specifics sets of limited of MS WMH cases, which may restrict its ability to generalize to diverse pathological conditions beyond MS.

Synthetic Data for Segmentation Methods mentioned above rely on supervised deep learning, in order to reach state-of-art performance. Due to supervised learning requiring a large diversity of examples to ensure proper generalisation, segmentation methods rely on data augmentation, to increase the variety of examples shown. In addition to usual classical intensity (gaussian noise, bias field simulation, *etc* ...) and spatial data augmentations (rotation, flip, *etc* ...) widely used for training both anatomical and abnormality segmentation models, recent developments have introduced new strategies to increase the diversity of images seen during training. In the case of brain abnormalities, [Ferreira et al., 2024] won the 2023 Adult Glioma Segmentation Challenge by combining traditional augmentations with images synthesized via Generative Adversarial Network (GAN) [Ferreira et al., 2024] to generate additional tumors. This approach increases the diversity of abnormalities encountered during training, thereby enhancing model robustness on the test set. Similar attempts have been made using more complex architectures for image synthesis; for example, [Lawry Aguila et al., 2025] employed a Latent Denoising Diffusion Probabilistic Model (L-DDPM) to generate additional lesions from augmented tumor masks. However, it is important to note that, due to the nature of 3D data, the use of diffusion models for 3D image synthesis remains limited by high computational costs. For example in the Healthy Tissue Synthesis Challenge [Kofler et al., 2024], few entries rely on the use of DDPM. Another kind of approach discards the effort for realistic generation of tumors. LF-SynthSeg [Xu et al., 2024] generates new tumors in clinically normal images by generating synthetic ellipsoids, with intensities derived from the white matter. It is important to note that as reported in [Lambert, 2024], experiments comparing the use of synthesis for the generation of synthetic MS lesions for the training of a subsequent segmentation model showed no clear evidence

that generating higher-quality synthetic lesions improves the accuracy of supervised segmentation models.

Contribution In this chapter, we present a novel method, FLAIRBrainSeg^{+}, for unified fine-grained anatomical and FHI segmentation. More particularly, we propose to train a 3D CNN by leveraging synthetic training data. This enables the creation of realistic variations of lesion-related FHI on clinically normal FLAIR MRIs. This strategy explicitly integrates anatomical and pathological information within a single label map. We demonstrate that incorporating such synthetic examples allows the training of a segmentation model capable of accurately delineating both FHI and anatomical structures. The method is evaluated on multiple public lesion datasets, and its performance is compared against SynthSeg-WMH. We perform this comparison using both the publicly available models distributed on FreeSurfer, and a retrained version of SynthSeg-WMH using our training data. This evaluation is performed on several in-domain and out-of-domain databases. Finally, we assess the reliability of the proposed approach using a scan–rescan dataset acquired across multiple sites.

2 Methods

2.A Synthetic Data Generation

In order to train our supervised segmentation model, on a large and diverse set of images, we introduce the following synthetic data generation process. To increase the diversity of training images, we propose to emulate the presence of FHI on real clinically normal FLAIR images. More specifically, the simulated FHI are generated on the FLAIR image by inserting "real" FHI from other images containing either WMH or tumors.

Pre-processing steps: Prior to the generation of synthetic training data, we preprocess all images according to the following scheme. The preprocessing process is illustrated in Figure 5.1. The pipeline includes denoising using a non-local means filter [Manjón et al., 2010a], rigid registration to a common template space using ANTs [Avants et al., 2011], bias field correction with N4ITK [Tustison et al., 2010], and intensity normalization using NICE [Manjón et al., 2014].

Description of the generation process: The generation process consists of four main steps and is illustrated in Figure 5.2. For each available FLAIR image, the process is as follows:

1. We preprocess all the selected target Clinically Normal (CN) image for the extraction of WMH and anatomical segmentation. For this purpose, we use DeepLesionBrain [Kamraoui et al., 2022], that relies on the availability of the

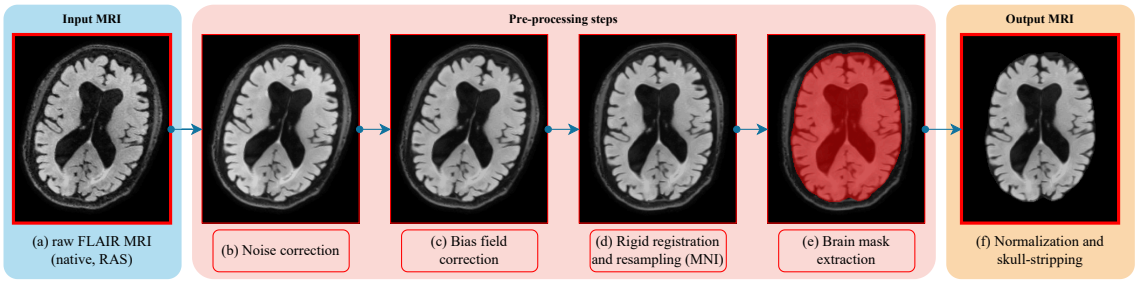


Figure 5.1: Illustration of the preprocessing steps applied to FLAIR MRI data: (a) raw MRI acquisition; (b) noise correction; (c) intensity inhomogeneity (bias-field) correction; (d) spatial normalization via rigid registration to the MNI152 template; (e) brain-mask extraction; (f) intensity normalization and skull stripping. Please note that in this instance, registration to the MNI was performed using the associated T1w images.

FLAIR image to extract WMH location, and the same subject T1w image for the extraction of anatomical information. We emphasize the fact that this extraction process is only required for the training image generation. As such, while the current generation process requires both modalities, the final segmentation model will not require it.

2. We randomly select a FLAIR MRI image from a database containing labelled WMH and tumor masks. The selected FLAIR MRIs image are preprocessed using the same pre-processing steps as above.
3. The selected FLAIR MRI image is non-linearly registered to the clinically normal (CN) FLAIR image. The computed deformation field is stored and applied to FHI mask.
4. The registered FHI mask is then compared with the target CN anatomical label map to ensure valid anatomical placement. The only constraint introduced at this stage is to prevent lesion placement outside the brain or within the ventricles. In cases where the FHI mask intersects with invalid regions, the overlapping areas are removed entirely from the mask.
5. A gradient-based blending method [Pérez et al., 2003] is used to merge the new hyperintense region into the CN FLAIR image. The label map is then updated with the validly placed FHI mask to generate the final target label map for training.

Alternative methods for FHI synthesis are explored in Appendix 2. We retained the gradient-based blending method because it offered a good compromise between simplicity, computational cost, and the visual quality of the generated images.

An overview of the training image generation process is provided in Figure 5.2. By using clinically normal (CN) images as targets for the alignment and in-painting processes, we ensure the integration of anatomical segmentation from a clinically normal image (the standard use case for most anatomical segmentation methods) with the synthesized FHI information derived from the post-registration FHI masks. As a result, the generated training images contain both anatomical and synthetic

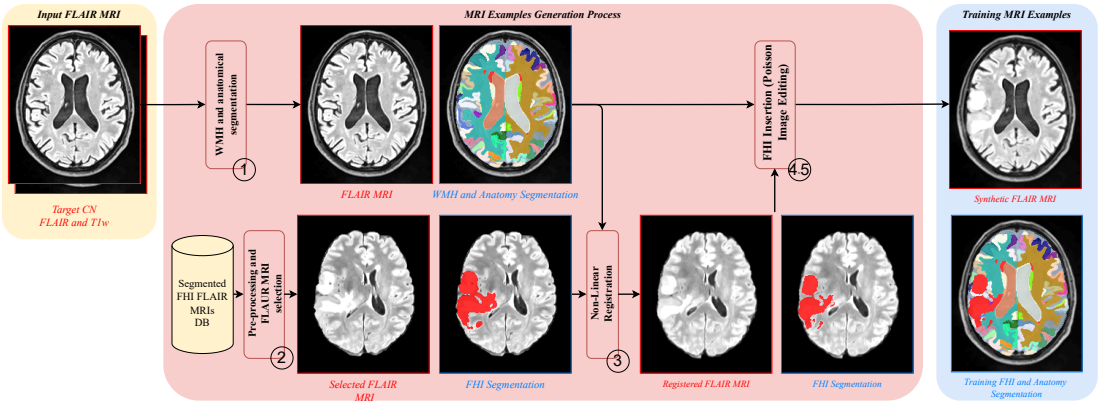


Figure 5.2: Overview of the training image generation process. Each pre-processed FLAIR image from a cognitively normal (CN) subject is first segmented for WMH, as CN cases may still present WMH due to incidental findings or age-related small vessel disease. The CN image is then randomly paired with another FLAIR image from the FHI dataset, containing WMH or tumors. The selected pathology image is non-linearly registered to the CN image, and a gradient-based image processing method is applied to merge the pathological regions onto it.

FHI information.

Limitations and drawbacks: A limitation of the augmented image generation process is the requirement on the availability of accurate pathology tissue mask. Specifically if the pathology mask includes healthy tissues, the generated label map will wrongly include labelling of the inserted healthy tissues as pathology. In that case we observe a discrepancy between the generated FLAIR image and the label map.

However, lesion delineation on the FLAIR modality is not always accurate. In case labelling is performed on different modality, at different resolutions or spatial spacing, when resampled and aligned, the quality and precision of FHI masks can vary considerably. For example, some stroke datasets provide segmentations derived from diffusion-weighted images at variable resolutions, resulting in imprecise FHI boundaries after alignment with FLAIR. Additionally, either in the case of manual or automatic segmentation, biases are introduced leading to poor mask quality. For example, with manual segmentation, inter- and intra- rater variation lead to differences in the segmentation mask [Coupé et al., 2020]. With automatic segmentation, the label maps can also contain segmentation errors.

Because of this, the images used for the training generation process were carefully selected to ensure good lesion mask quality. However, this selection process inherently limits the diversity of lesions used for training to a few FHI types (that is, tumors and MS WMH) and a restricted number of datasets (see Table 5.1). These limitations are the main reasons why we did not include stroke lesions in the generation process,

due to the databases available to us not including good quality stroke lesion masks, despite stroke being one of the main use case where FLAIR images are available over volumetric T1w images.

Synthetic testing smages: To evaluate the generalization capabilities of the proposed method, we compiled a comprehensive set of testing images from multiple sources, including both synthetic and real images with lesions. The datasets used for the generation of the synthetic testing images are described in Section 3.

2.B Training Configuration:

The segmentation network was based on a 3D U-Net architecture based on the 3D configuration of nnU-Net [Isensee et al., 2021]. The network was trained using a combination of Dice loss and cross-entropy loss to optimize segmentation accuracy. Training was performed using stochastic gradient descent (SGD) with a polynomial learning rate (lr) scheduling policy defined as $lr(t) = (1 - t/t_{max})^{0.9}$, with an initial learning rate of 1, which was reduced upon plateauing of the validation loss. At each iteration, batch samples were randomly selected from the generated synthetic training images (UKB/UCSF-PDGM, UKB/In-House, UKB/BraTS-MEN), from which patches of size $128 \times 128 \times 128$ voxels were extracted. The batch size was set to 2, constrained by GPU memory limitations. Data augmentation techniques, including random rotations and scaling, were applied during training to further enhance model robustness. The model was trained for 1,000 epochs. While alternative training configurations or architectures could have been used, the focus of this work is to evaluate the feasibility of using the proposed synthetic data generation process to train a segmentation network capable of delineating both anatomical structures and FHI. Moreover, nnU-Net–like architectures have consistently achieved state-of-the-art performance across multiple medical image segmentation tasks [Isensee et al., 2021]; therefore, this architecture represents a reasonable and well-validated choice for the present study.

2.C Validation Framework

Comparison settings: The comparison was performed using SynthSeg-WMH, a segmentation method based on the SynthSeg framework but adapted to additionally segment white matter hyperintensities (WMH). To the best of our knowledge, SynthSeg-WMH is one of the few existing methods designed to simultaneously segment both WMH and brain anatomy from FLAIR images. Two configurations were evaluated. First, we used the published SynthSeg-WMH method shared with FreeSurfer 7.4.1 and evaluated the accuracy of the WMH segmentation only. Second, we used the official SynthSeg codebase available on GitHub `https://github.com/BBillot/SynthSeg` to implement the modifications to the Gaussian mixture model (GMM) intensity sampling described in [Laso et al., 2024].

In summary, these modifications introduce an additional constraint on the sampling of WMH intensities, requiring the WMH class to be either brighter or darker than the sampled white matter intensity. This approach is introduced to an approach designed to emulate WMH appearance on T1- and T2-weighted contrasts, respectively. We used this modified implementation to train a new model using our synthetic training data label maps, containing both WMH and tumors, and also 132 anatomical structures. This retrained version is referred to as SynthSeg-FHI-133[2].

This setup allowed us to evaluate the performance of SynthSeg-FHI-133 for both anatomical and FHI segmentation. The model was trained for 150 epochs, each consisting of 5,000 iterations. Following the training parameters described in the official repository and documentation, the network was trained with a fixed learning rate of 1×10^{-4} and used the same augmentation strategies as the original SynthSeg model [Billot et al., 2023a]. These augmentations included deformation of label maps prior to synthetic image generation, as well as scaling, rotation, shearing, and a variety of intensity-based transformations to generate contrast variability. *Figure 3 illustrates representative training examples generated by SynthSeg-WMH, highlighting the diversity and limitations of the GMM-based image synthesis approach.*

Validation metrics were computed as follows. For the WMH datasets (In-House, MSLesSeg, and UKB), we measured the Dice Similarity Coefficient (DSC) and the 95th percentile Hausdorff distance (HD95) between the ground-truth and predicted WMH masks. For the tumor datasets (UCSF-PDGM, Brats-MEN), where only tumor annotations were available, the predicted FHI masks were processed as follows: (1) each connected component in the predicted mask was assigned a unique label; (2) for each component, we computed its center of mass and measured its distance to the center of mass of the ground-truth tumor mask; (3) the predicted component closest to a ground-truth tumor was retained for the computation of DSC and HD95. This procedure accounts for the presence of unrelated (that is, non-tumor) FHI in the image, which the model may segment by design. By selecting the component closest to the ground-truth tumor, the evaluation focuses on the most relevant prediction.

Ablation studies: To assess the impact of different training configurations on the generalization ability of FLAIRBrainSeg$^+$, we trained several model variants incorporating alternative data augmentation schemes, loss functions emphasizing FHI segmentation, and pre-training on larger and more diverse datasets. The different configurations are described as follows:

1. **Baseline:** The baseline configuration described above, trained on synthetic images generated from the UKB and pathology datasets.
2. **Baseline + Data Augmentations:** A variant trained using the same configuration as the baseline but with extended data augmentation strategies, including additional spatial and intensity transformations.

[2]The name SynthSeg-FHI-133 reflects the inclusion of 132 anatomical structures and one additional label for FLAIR hyperintensities (FHIs). Tumors are not strictly considered WMH, although they also appear hyperintense on FLAIR.

3. **Pre-training (VOLBRAIN.NET):** A variant pre-trained on a large dataset of FLAIR images from VOLBRAIN.NET before fine-tuning on the synthetic training set.
4. **Pre-training (MultiTalent):** A variant initialized with the pre-trained weights provided in the [Ulrich et al., 2025] method, inspired by the [Ulrich et al., 2023] approach introduced to provide a multi-dataset approach to medical image segmentation.
   For both fine tuning experiments, the pre-trained weights were fine-tuned on the synthetic training set using a linear learning rate increase for warmup up to 0.0001 over 50 epochs, followed by a default schedule.

Scan–rescan reliability: Scan–rescan reliability is an important property of automatic segmentation methods, reflecting the consistency of the segmentation across repeated acquisitions of the same subject under varying scanner and site conditions. To assess the scan–rescan reliability of the proposed method, we used FLAIR images from the SIMON [Duchesne et al., 2019] dataset. All scans with the highest QC rating according to RegQCNet [Denis de Senneville et al., 2020] were selected, resulting in five FLAIR acquisitions of the same subject across five different sites. Each scan was segmented using FLAIRBrainSeg^{+} and SynthSeg-FHI-133. We then computed the Dice similarity coefficient (DSC) between all pairs of segmentations across sites, yielding ten pairwise comparisons per method. The mean DSC across these pairs was used as the reliability measure.

3 Materials

Databases: The FLAIR MRI databases used in our experiments are detailed below. Specifically, we employed a large set of clinically normal (CN) FLAIR images described below:

- Clinically Normal (UKB) – 500 images, selected from the UK Biobank database [Bycroft et al., 2018], a large-scale biomedical database and research resource containing in-depth genetic and health information from half a million UK participants. We selected subjects with no reported neurological or psychiatric conditions, ensuring a clinically normal cohort. The selected subjects had both FLAIR and T1-weighted images available, which were used for preprocessing and anatomical segmentation. The dataset primarily consists of middle-aged to older adults (mean age 64.8 years, standard deviation 11.0 years). The dataset benefits from extensive preprocessing and quality-control pipelines, resulting in high-quality images with minimal artifacts. While subjects were screened through medical records, it is important to consider that some participants may still present incidental findings. Only MRI and demographic data were used in this study, despite the UKB containing a wide range of health-related information. Demographic details were used to ensure balanced train and test

splits.

FHI insertion databases: To generate synthetic FHI on the selected CN FLAIR images, we incorporated FLAIR MRIs with corresponding tumor or white matter hyperintensity (WMH) segmentations from multiple available databases:

- Glioblastoma (UCSF-PDGM) – 500 images, from the UCSF-PDGM publicly shared and included the BraTS-GLI challenge, contains a number of subjects with grade II to IV gliomas contributed from the University of Califorina, San Francisco, USA. The datasets includes volumetric T1w, T1c, T2w and FLAIR images. Only FLAIR images were used in this work. A tumor segmentation is provided for each subject, divided into the enhancing tumor, the peritumoral edema and the necrotic core. For this work, only the whole tumor segmentation was used. UCSF-PDGM are provided without non-brain tissues, as skull-stripping was performed as an anonymisation procedure.
- Multiple Sclerosis (In-House) – 49 images, acquired in Bordeaux, with 3 different scanners and at multiple resolutions. The FHI masks were delineated by a human expert. Since the cohort only included 49 subject, we used all available cases rather than sampling 500 as for the other datasets. Datasets includes T1w and FLAIR. Only FLAIR images were used in this work.
- Meningioma (BraTS-MEN) – 500 images, publicly available from the BraTS-MEN challenge, including T1w, T2w, FLAIR, T1c images from patients with confirmed meningioma diagnosis (grade I to III) from 6 different institutions in the United States. Only the FLAIR images are used in this work. Similar to UCSF-PDGM A tumor segmentation is provided for each subject, divided in tumor subregions. For this work, only the whole tumor segmentation was used.

Out-of-Domain testing databases: To evaluate the robustness of our models, we also retained additional image databases for testing in out-of-domain (OOD) settings. These databases include:

- Multiple Sclerosis (OFSEP) is a multi-site clinical cohort of MS patients acquired in 36 different sites in France, using a variety of scanner models and acquisition protocols. The dataset includes T1w and FLAIR images. Images were pre-processed using the same pipeline as UKB images, and a structure segmentation and WMH lesion segmentation were obtained using DeepLesionBrain [Kamraoui et al., 2022]. OFSEP is mainly comprised of middle age adults (average age 45.6 years old, standard deviation 11.2 years old).
- Multiple Sclerosis (MSLesSeg) is a multi-site challenge dataset, including 89 FLAIR images from multiple sclerosis patients acquired in a study lead by an Italian hospital with expert WMH segmentations. The dataset includes T1w and FLAIR images. Images were pre-processed using the same pipeline as UKB images.

Unfortunately, we were not able to include other OOD databases for the OOD evaluation of tumor segmentation.

Additional pre-training data: To explore possible improvements to FLAIR-BrainSeg, we explored pre-training on a larger and more diverse set of FLAIR images. We used the following database for this purpose:

- PreTraining (VOLBRAIN.NET) – 1,000 images, from the VOLBRAIN.NET. `https://volbrain.net` is a platform with a set of publically available online tools for the processing of brain MRIs. When submitting image on the platform, users are given the option to share their images for research purposes. Over time, this resulted in a large set of images with a large variety of acquisition settings and scanners. We used this database in the training of additional variants of FLAIRBrainSeg^{+} (see Section 4). The dataset includes T1w and FLAIR images. Images were pre-processed using the same pipeline as UKB images.

Scan-rescan experiment database: Finally, we retained images for a scan-rescan experiment. We used the following databases:

- Clinically Normal (SIMON) [Duchesne et al., 2019] – 5 images, Images were pre-processed using the same DeepLesionBrain pipeline. The dataset contains repeated acquisitions of a single healthy subject, scanned at multiple different sites.

3.A Datasets

An overview of all datasets used for training and testing is provided in Table 5.1.

Training images For training, we used synthetic images generated from 500 clinically normal subjects from the UK Biobank (UKB), combined with lesion masks derived from three pathology databases: UCSF-PDGM (glioblastoma), IN-HOUSE (WMH), and BRATS-MEN (meningioma). For each UKB image and each pathology dataset, a random FLAIR image from the annotated dataset was selected for the generation and in-painting process. Consequently, each synthetic subset comprised 500 images, yielding a total of 1,500 synthetic training images. Each image was associated with a corresponding label map containing 132 anatomical structures and one additional label for both lesions or tumors.

Testing images For testing, we used both synthetic and real images. The synthetic testing images were generated using the same procedure as for training but with a separate held-out set of 100 UKB subjects.

The synthetic testing images were generated using the same procedure as for training but with a separate held-out set of 100 UKB subjects and held-out images from each pathology dataset (UCSF-PDGM, IN-HOUSE, BRATS). The real in-domain testing images were obtained from the same pathology datasets used for training generation, but from held-out subjects: UCSF-PDGM, 9 images from IN-HOUSE, and 100 images from BRATS-MEN. Additionally, we included an external dataset, MSLESSEG, which contains 89 FLAIR images from multiple sclerosis patients

Dataset	Subjects	Description
		Synthetic Training Images
UKB/UCSF-PDGM	500	Glioblastoma (synthetic, derived from UCSF-PDGM)
UKB/InHouse	500	MS WMH (synthetic, derived from InHouse)
UKB/BraTS-MEN	500	Meningioma (synthetic, derived from BraTS-MEN)
		Ablation Study Pre-training images
volbrain.net	1000	Unknown (mixed images; automatic annotations)
		Testing Images
BraTS-MEN	100	Meningioma (public challenge; expert annotations)
InHouse	9	MS WMH (local clinical cohort; expert annotations)
UCSF-PDGM	100	Glioblastoma (public challenge; expert annotations)
UKB	100	Clinically Normal (population cohort; automatic annotations)
UKB/UCSF-PDGM	100	Glioblastoma (synthetic, derived from UCSF-PDGM)
UKB/InHouse	100	MS WMH (synthetic, derived from InHouse)
UKB/BraTS-MEN	100	Meningioma (synthetic, derived from BraTS-MEN)
MSLesSeg	89	MS WMH (multi-site challenge; expert annotations)
OFSEP	100	MS WMH (multi-site clinical MS cohort; automatic annotations)
		Scan-rescan testing images
Simon	5	Multi-site same subject acquisition

Table 5.1: Overview of FLAIR MRI datasets used for training, testing, and scan–rescan experiments. The table reports the number of subjects and a concise description of dataset characteristics.

acquired at the University of North Carolina, USA, with expert WMH segmentations. An overview of all datasets used for training and testing is provided in Table 5.1.

4 Results

Quantitative Analysis of FHI Segmentation: Results from the quantitative evaluation are summarized in Table 5.2. We report the Dice Similarity Coefficient (DSC) and 95th percentile Hausdorff distance (HD95) for FHI across selected testing sets.

	DSC ↑			HD95 [mm] ↓		
Dataset	SynthSeg-WMH	SynthSeg-FHI-133	FLAIRBrainSeg$^+$	SynthSeg-WMH	SynthSeg-FHI-133	FLAIRBrainSeg$^+$
"FHI" segmentation metrics						
BraTS-MEN ($N = 100$)	0.001 ± 0.005	0.240 ± 0.295	$\mathbf{0.514 \pm 0.314}^*$	53.39 ± 17.69	34.61 ± 18.19	$\mathbf{31.63 \pm 25.83}^*$
InHouse ($N = 9$)	0.332 ± 0.125	0.519 ± 0.117	$\mathbf{0.714 \pm 0.188}^*$	18.57 ± 8.76	18.73 ± 5.46	$\mathbf{14.73 \pm 9.61}^*$
UCSF-PDGM ($N = 100$)	0.009 ± 0.034	0.418 ± 0.272	$\mathbf{0.852 \pm 0.127}^*$	47.93 ± 11.46	29.21 ± 14.27	$\mathbf{12.09 \pm 14.42}^*$
UKB ($N = 100$)	0.212 ± 0.203	0.261 ± 0.218	$\mathbf{0.828 \pm 0.109}^*$	31.54 ± 21.29	33.74 ± 21.22	$\mathbf{9.49 \pm 17.83}^*$
UKB/BraTS-MEN ($N = 100$)	0.108 ± 0.107	0.445 ± 0.169	$\mathbf{0.950 \pm 0.031}^*$	32.94 ± 11.64	18.40 ± 4.69	$\mathbf{1.63 \pm 0.86}^*$
UKB/InHouse ($N = 100$)	0.359 ± 0.156	0.489 ± 0.132	$\mathbf{0.874 \pm 0.038}^*$	21.37 ± 10.49	17.18 ± 6.50	$\mathbf{1.09 \pm 0.18}^*$
UKB/UCSF-PDGM ($N = 100$)	0.105 ± 0.106	0.453 ± 0.196	$\mathbf{0.953 \pm 0.028}^*$	33.45 ± 11.53	27.82 ± 13.70	$\mathbf{1.47 \pm 0.68}^*$
MSLesSeg ($N = 89$)	0.319 ± 0.204	0.292 ± 0.206	$\mathbf{0.522 \pm 0.226}^*$	33.60 ± 12.74	33.37 ± 12.92	$\mathbf{32.02 \pm 19.28}^*$
OFSEP ($N = 100$)	0.022 ± 0.011	0.284 ± 0.233	$\mathbf{0.501 \pm 0.240}^*$	48.50 ± 4.89	35.26 ± 24.78	$\mathbf{33.64 \pm 17.98}^*$

Table 5.2: Comparison of WMH and tumor segmentation performance between methods across the selected test sets. Metrics: Dice similarity coefficient (DSC) and 95th percentile Hausdorff distance (HD95). Higher DSC and lower HD95 indicate better performance. Statistical significance between FLAIRBrainSeg$^+$ and other methods was assessed using the Wilcoxon signed-rank test with $p < 0.05$. * indicates statistically significant improvement over both SynthSeg-WMH and SynthSeg-FHI-133.

Results from the SynthSeg-WMH model evaluated using the FreeSurfer binaries indicate that the model fails to segment tumors, as evidenced by the very low Dice Similarity Coefficient (DSC $\simeq 0.01$) obtained on the tumor datasets. This outcome is consistent with expectations, given that the model was trained specifically for WMH segmentation. On the MS datasets, SynthSeg-WMH achieved poor but somewhat higher performance (DSC $\simeq 0.3$) on the selected testing set from In-House. The cause of this suboptimal performance remains unclear, despite the model's specialization for WMH. Qualitatively, we observed that the segmentations produced by SynthSeg-WMH were consistently smaller than the corresponding ground-truth masks, suggesting possible discrepancies in lesion size, boundary definition, or anatomical distribution in the data used to train the publicly released model.

The retrained SynthSeg-FHI-133 model, trained using the label maps from the FLAIRBrainSeg$^+$ training pairs, showed significantly improved performance on WMH segmentation compared to SynthSeg-WMH (DSC $\simeq 0.519$, $p < 0.05$). However, its performance remained substantially lower than that of FLAIRBrainSeg$^+$ (DSC $\simeq 0.714$, $p < 0.05$). We hypothesize that this difference arises because the GMM-based domain randomization model used in SynthSeg-FHI-133 may not generate sufficiently realistic image contrasts for the network to learn features relevant to FHI

segmentation in the test datasets. Additional factors related to model complexity or training configuration could also contribute, as SynthSeg-FHI-133 was trained strictly following the default parameters and procedures provided in the official repository.

FLAIRBrainSeg$^+$ significantly outperformed both SynthSeg-WMH and SynthSeg-FHI-133 across all evaluated test sets, achieving consistent accuracy on the In-House and UCSF-PDGM datasets. However, its performance on BraTS-MEN was notably lower, with several failure cases in which the model failed to accurately localize tumors. These failures are attributed to the lower image quality and resolution of the BraTS-MEN dataset. The elevated HD95 values observed on MSLesSeg, OFSEP, and BraTS-MEN primarily reflect distant false-positive detections of FHIs. These limitations are further discussed below in relation to identified failure cases.

Quantitative Analysis of Structure Segmentation: For brain structure segmentation, metrics were computed between the obtained segmentations and those derived from the T1w images processed with AssemblyNet from the DLB pipeline. This comparison was necessary due to the lack of existing fine-grained anatomical segmentation datasets containing FLAIR images. SynthSeg-FHI-133 achieved reasonable overlap scores (DSC $\simeq 0.8$) but was significantly outperformed by FLAIRBrainSeg$^+$ (DSC $\simeq 0.93$, $p < 0.05$). This finding is consistent with the results shown in Figure 4.5, where a similar comparison between SynthSeg and a modality-specific U-Net trained on FLAIR images was performed for fine-grained segmentation.

We attribute this difference to FLAIRBrainSeg$^+$ being able to extract more relevant features for structural delineation by learning from real image contrasts, as opposed to the synthetic contrasts generated by the GMM-based domain randomization model. FLAIRBrainSeg$^+$ consistently achieved high accuracy in reproducing the T1w-based segmentation targets across all test sets. These results support our previous findings in Chapter 4 and [Le Bot et al., 2025], confirming that FLAIR images contain sufficient information to produce fine-grained anatomical segmentations with accuracy comparable to those derived from T1w images.

Contextual comparison with state-of-the-art methods. For context, we report the accuracy of FLAIRBrainSeg$^+$ relative to top-performing methods from the segmentation challenges corresponding to the pathology datasets used for testing (Table 5.4). Results from competing methods are presented as reported in the original publications and are *not directly comparable, since the challenges employed different data splits and, in some cases, targeted slightly different tasks (that is, using additional modalities,* etc*).* Notably, challenge participants were allowed to use all available MRI modalities for segmentation (T1w, T1c, T2w, and FLAIR), whereas our method relies solely on FLAIR images. Competing methods were identified from the official challenge leaderboards and the corresponding winning-team publications.

For BraTS-MEN, we report the winning solution from the BraTS-MEN23 challenge, as later editions focus on post-operative segmentation tasks using different data. For UCSF-PDGM, we report the accuracy of the winning solution from

Dataset	Mean DSC ↑ SynthSeg-FHI-133	FLAIRBrainSeg$^+$
Structure segmentation metrics		
UKB ($N = 100$)	0.798 ± 0.010	$\mathbf{0.927 \pm 0.021}^*$
UKB/BraTS-MEN ($N = 100$)	0.723 ± 0.051	$\mathbf{0.937 \pm 0.020}^*$
UKB/InHouse ($N = 100$)	0.801 ± 0.017	$\mathbf{0.953 \pm 0.007}^*$
UKB/UCSF-PDGM ($N = 100$)	0.772 ± 0.033	$\mathbf{0.937 \pm 0.019}^*$
OFSEP ($N = 100$)	0.782 ± 0.0195	$\mathbf{0.879 \pm 0.029}^*$

Table 5.3: Comparison of brain structure segmentation performance between methods across the selected test sets. Metric: Dice similarity coefficient (DSC). Higher DSC indicates better overlap with the reference segmentation. Statistical significance between FLAIRBrainSeg$^+$ and SynthSeg-FHI-133 was assessed using the Wilcoxon signed-rank test with $p < 0.05$. * indicates statistically significant improvement of FLAIRBrainSeg$^+$ over SynthSeg-FHI-133. *HD95 values were not computed for structure segmentation due to the large number of regions and limited interpretability of boundary-based metrics in this context.*

the BraTS-GLI23 challenge for consistency. Finally, for In-House, we include the results of [Kamraoui et al., 2022], which, to the best of our knowledge, is the only method evaluated on this dataset. Incidentally, this same method was the top-performing approach in the MSSeg-2 multiple sclerosis lesion segmentation challenge [Commowick et al., 2021].

Despite relying exclusively on FLAIR images, FLAIRBrainSeg$^+$ achieved FHI segmentation performance approaching that of multi-modal state-of-the-art methods on UCSF-PDGM and performed competitively on In-House. Performance was lower on BraTS-MEN, likely due to the lower resolution and overall quality of that dataset (see Failure Cases). HD95 values were not reported in some original references and are therefore omitted in the table.

Dataset	Method	Modalities	DSC ↑	HD95 [mm] ↓
FHI segmentation metrics				
UCSF-PDGM	*Faking-it* [1rst place BraTS-GLI23]	T1+T1c+T2+FLAIR	**0.910 ± 0.021**	**11.1 ± –**
	FLAIRBrainSeg$^+$	FLAIR	0.852 ± 0.127	12.09 ± 14.42
InHouse	*DeepLesionBrain* [1rst place MSSeg-2]	T1+FLAIR	0.696 ± –	–
	FLAIRBrainSeg$^+$	FLAIR	**0.714 ± 0.188**	**14.73 ± 9.61**
BraTS-MEN	*NVAUTO* [1rst place BraTS-MEN23]	T1+T1c+T2+FLAIR	**0.930 ± 0.018**	**9.0 ± 2.3**
	FLAIRBrainSeg$^+$	FLAIR	0.514 ± 0.314	31.63 ± 25.83

Table 5.4: Comparison of FLAIRBrainSeg$^+$ against reported State of the Art (SoTA) methods. *Results from competing methods are reported as published in the challenge reports and **are not directly comparable**, since SoTA approaches use **multiple modalities and different data splits**. These values are shown for contextual reference only. Empty entries indicate that the information was not reported in the corresponding reference.*

Ablation study. Results from the ablation study are summarized in Table 5.5. The experiments were designed to evaluate whether pre-training or stronger data augmentations could improve the generalization capabilities of FLAIRBrainSeg$^+$. This investigation was primarily motivated by the observation that the UKB images used for synthetic training data generation are of substantially higher quality than typical clinical scans or test images from the pathology datasets, such as BraTS-MEN. We therefore hypothesized that pre-training on a larger variety of images, or applying more aggressive data augmentations, could enhance the model's ability to generalize across different imaging conditions.

	Datasets			
Method Variant	BraTS-MEN	InHouse	UCSF-PDGM	UKB
Baseline	$\mathbf{0.514 \pm 0.314}$	0.714 ± 0.188	0.852 ± 0.127	$\mathbf{0.927 \pm 0.021}$
Baseline + data augmentations	0.511 ± 0.353	0.718 ± 0.118	0.859 ± 0.130	0.921 ± 0.015
Pre-training (volbrain.Net)	$0.464 \pm 0.318^\star$	$\mathbf{0.731 \pm 0.128}$	0.850 ± 0.131	0.917 ± 0.012
Pre-training (MultiTalent)	0.506 ± 0.356	0.717 ± 0.105	$\mathbf{0.863 \pm 0.111}$	0.917 ± 0.012

Table 5.5: Comparison of segmentation performance between different training configurations of FLAIRBrainSeg$^+$ across multiple selected test sets. Metrics: Dice similarity coefficient (DSC). *For BraTS-MEN, InHouse and UCSF-PDGM, DSC is computed with regards to the ground-truth lesion label. For UKB, DSC is computed with regards to the T1w-based anatomical segmentation.* Statistical significance between the baseline and other variants was assessed using the Wilcoxon signed-rank test with $p < 0.05$. Statistically significant decrease in performance compared to the baseline is indicated by *.

Results from the ablation study indicate that neither pre-training nor the inclusion of stronger data augmentations improved the generalization capabilities of FLAIRBrainSeg$^+$. This finding held true for both types of pre-training tested, namely using either publicly available pretrained weights or our own pre-training dataset. The baseline configuration consistently outperformed all other variants, suggesting that the information learned during pre-training did not help the model converge to a better minimum during fine-tuning. Furthermore, the use of aggressive data augmentations, intended to better match the quality and appearance of pathological testing datasets, negatively impacted final performance by disrupting the learning of optimal features for segmentation.

Qualitative analysis. Qualitative results for FHI and brain structure segmentation on real testing datasets are shown in Figure 5.3. The figure illustrates representative examples of segmentations obtained with FLAIRBrainSeg$^+$, SynthSeg-WMH, and SynthSeg-FHI-133 across the three real testing datasets.

Both versions of SynthSeg are highly sensitive to the presence of hyperintensities and tend to produce over-segmentations. As seen in Figure 5.3 panel (b.4), the retrained SynthSeg-FHI-133 model additionally misclassifies darker brain regions as FHI, leading to a higher number of false positives. These errors are likely related

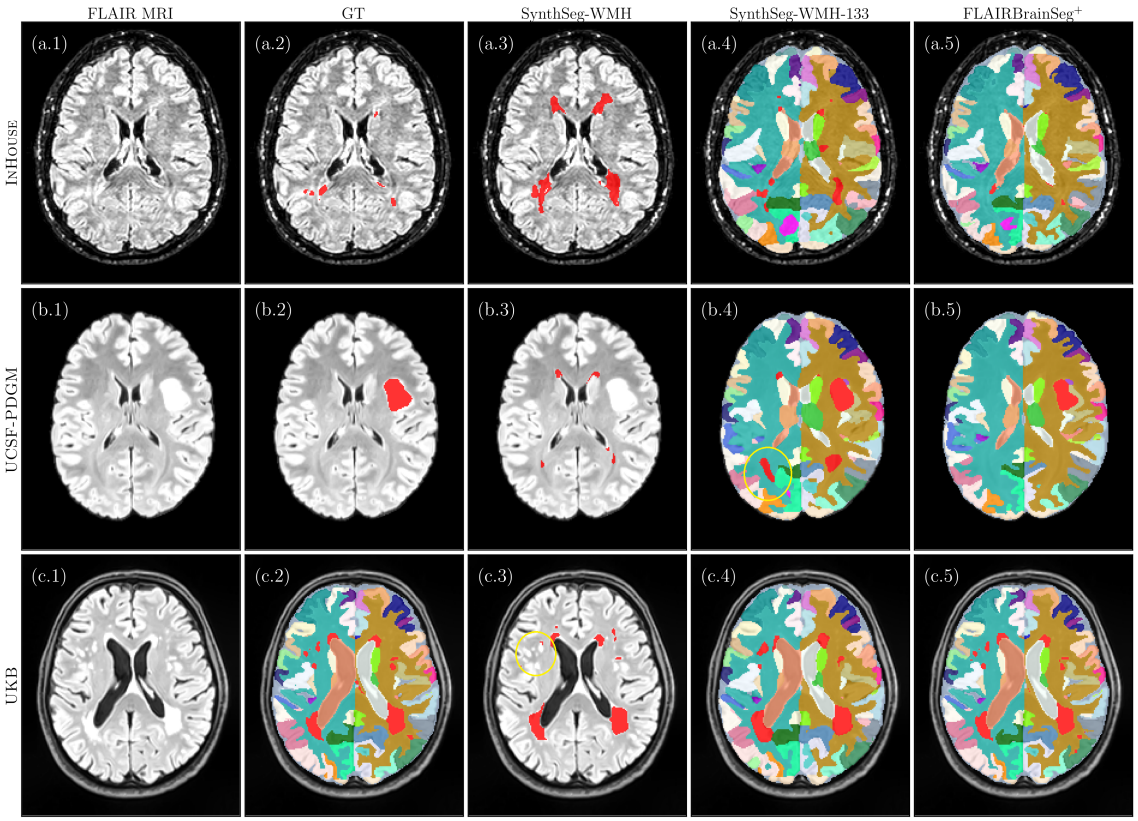


Figure 5.3: Examples of FLAIR-based segmentation results across the three real testing datasets. Each row corresponds to a different dataset (InHouse, UCSF-PDGM, and UKB). Columns show, from left to right: FLAIR MRI, ground truth (GT), FreeSurfer SynthSeg-WMH segmentation, SynthSeg-FHI-133, and the segmentation obtained with the proposed method (FLAIRBrainSeg^{+}). In case of InHouse and UCSF-PDGM, GT is the corresponding segmentation provided in the dataset. In case of UKB, the GT is the T1w-based anatomical segmentation merged with the WMH identified by DLB.

to the training process of SynthSeg-WMH: the GMM constraints introduced during training aim to model FHI appearance in both T1w (where FHI appear darker than white matter) and T2w-type contrasts (where FHI appear as white matter hyperintensities). While SynthSeg-WMH can delineate small WMH boundaries, it fails to detect larger tumors, as shown in Fig. 5.3 panel (a.3).

Additionally, Fig. 5.3 (c.5) demonstrates that FLAIRBrainSeg$^+$ nearly perfectly reproduces the anatomy obtained from the T1w image. In contrast to the previous models, although its FHI delineations may appear slightly smaller than the ground truth (*e.g.*, Fig. 5.3 panel (a.5)), the method achieves high DSC scores and consistently localizes both WMH and tumors.

These results illustrate the superior performance of FLAIRBrainSeg$^+$ in accurately delineating both lesions and anatomical structures across diverse pathological conditions. In contrast, SynthSeg-WMH struggles to identify lesions, particularly tumor, while SynthSeg-FHI-133 shows improved lesion detection but still lacks anatomical precision compared to FLAIRBrainSeg$^+$. These qualitative examples underscore the effectiveness of our synthetic training data generation strategy in enabling robust and accurate segmentation from FLAIR images.

Scan–rescan reliability. Results from the scan–rescan experiment are summarized in Table 5.6. Both methods produced consistent results across the selected scans, that consists in repeated acquisition of the same subject. FLAIRBrainSeg$^+$ segmentations were more consistent by 0.03 DSC points compared to SynthSeg-FHI-133, as reflected by higher mean DSC values. This indicates that FLAIRBrainSeg$^+$ is more robust to variations in image acquisition protocols and scanner models. Such robustness is particularly important in multi-site studies and longitudinal analyses, where differences in scanner hardware or acquisition parameters are common. High scan–rescan reliability ensures that observed differences in segmentation reflect true anatomical or pathological changes rather than acquisition-related variability.

	Mean DSC ↑	
Dataset	SynthSeg-FHI-133	FLAIRBrainSeg$^+$
SIMON	0.835 ± 0.029	$\mathbf{0.860 \pm 0.027}$

Table 5.6: Scan-rescan reliability results, measured as the mean Dice similarity coefficient (DSC) between all pairs of segmentations from the same subject acquired at different sites. Higher DSC indicates better reliability. *Given the limited number of samples in the scan-rescan dataset, no statistical tests were performed.*

Failure cases and limitations. While FLAIRBrainSeg$^+$ generally performed well across all test sets, some failure cases were observed, particularly for lesion segmentation in the BRATS-MEN and MSLESSEG datasets. In these instances, the

model occasionally failed to accurately localize or fully delineate lesions. To contextualize these errors, Figure 5.4 compares FLAIRBrainSeg$^+$ with SynthSeg-WMH and SynthSeg-FHI-133 on the same challenging subjects. All methods exhibited degraded performance, though not to the same extent or for the same reasons. Specifically, failures were associated with low-resolution or noisy FLAIR scans, pronounced bias-field inhomogeneities, and lesions with atypical or weakly hyperintense signal profiles. Despite these challenging conditions, FLAIRBrainSeg$^+$ maintained higher structure consistency than the competing methods.

Figure 4 presents a comparison of segmentation error maps across the different methods on the UKB test set. This visualization highlights the spatial distribution and magnitude of segmentation errors, allowing for a direct assessment of each method's accuracy and consistency relative to the T1w-based anatomical reference.

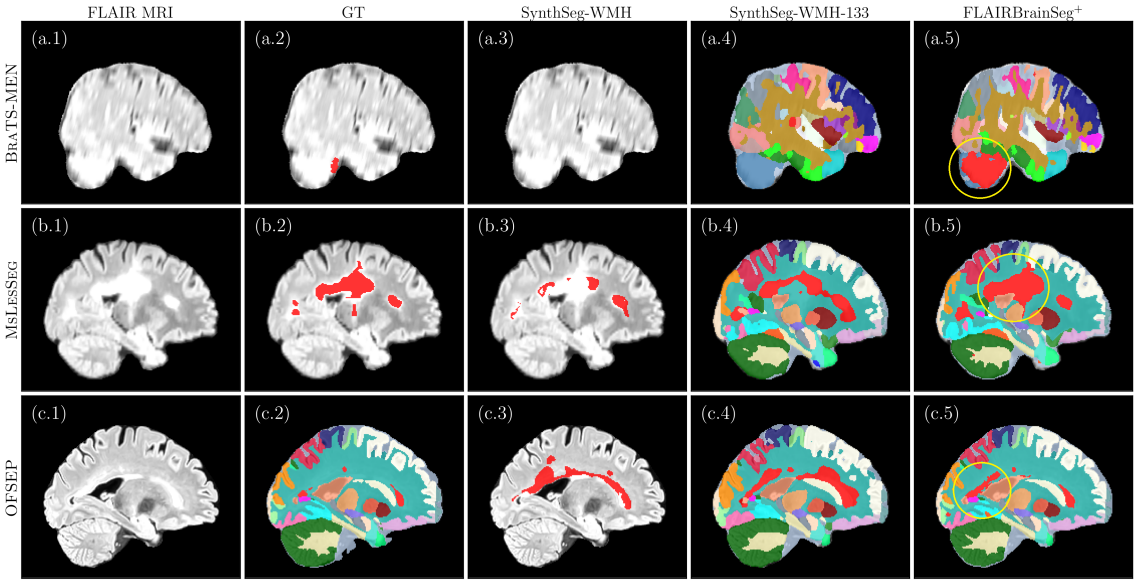


Figure 5.4: Representative failure cases across selected OOD datasets. Each row shows a subject from BraTS-MEN, MSLesSeg, and OFSEP. Sagittal slice is shown to highlight differences in FLAIR resolution and image quality across datasets: anisotropic and noisy scans in BraTS-MEN, moderate variability in MSLesSeg, and high-quality isotropic scans in OFSEP. *Segmentation errors made by FLAIRBrainSeg$^+$ are circled in yellow.* Columns show, from left to right: FLAIR MRI, ground truth (GT), FreeSurfer SynthSeg-WMH segmentation, SynthSeg-FHI-133, and the segmentation obtained with the proposed method (FLAIRBrainSeg$^+$).

On out-of-distribution (OOD) datasets, failure cases led to a performance drop, particularly in cohorts with heterogeneous acquisition parameters (*e.g.*, scanner vendor, spatial resolution, patient preparation). This highlights an important limitation of the proposed approach: synthetic data augmentation alone does not fully compensate for the variability encountered across clinical sites. These findings are consistent with prior studies reporting similar degradations under severe domain

shifts, even in multi-modal frameworks. Future work should therefore investigate explicit domain adaptation or contrast harmonization strategies to further enhance generalization in FLAIR-based segmentation.

5 Conclusion

In this work, we introduced FLAIRBrainSeg^{+}, a deep learning–based framework for fine-grained segmentation of both brain structures and FHI directly from FLAIR MRI. This dual-capability approach addresses an important yet largely unexplored problem, as most existing segmentation tools are optimized for either structural or lesion segmentation, but not both. The method leverages synthetic data generation by combining clinically normal FLAIR scans with lesion masks from multiple pathology datasets, enabling the network to learn from a broad spectrum of anatomical and pathological variability.

Experimental results demonstrated that FLAIRBrainSeg^{+} consistently outperforms existing baselines, including SynthSeg-WMH and the retrained SynthSeg variant (SynthSeg-FHI-133), across multiple independent test sets. The model achieved robust anatomical segmentation and reliable lesion detection, even in challenging pathological cases. However, out-of-distribution evaluations revealed that performance remains sensitive to acquisition variability, particularly in datasets with heterogeneous spatial resolution or pronounced bias-field inhomogeneities. These findings highlight the persistent challenge of domain generalization [Wen et al., 2023], a recurring issue in medical image analysis where variations in scanner hardware, acquisition protocols, or patient populations often lead to significant performance degradation across sites and datasets.

Future work will focus on addressing this limitation by integrating super-resolution [Morell-Ortega et al., 2024] reconstruction within the processing pipeline. Enhancing the effective resolution of anisotropic or low-quality FLAIR scans prior to analysis may improve robustness and anatomical fidelity across diverse clinical acquisitions. In parallel, incorporating domain adaptation and contrast harmonization strategies could further reduce performance variability across sites and scanner types, moving toward a clinically reliable, FLAIR-only segmentation framework.

Another limitation of our proposed approach is that the generation process does not account for mass effect, which is a common feature of many brain pathologies, particularly tumors. Mass effect refers to the displacement of brain structures due to the presence of a tumor, which can significantly alter the surrounding anatomy. In our current generation process, lesions are inserted into clinically normal anatomies without simulating the associated mass effect. This simplification may limit the realism of the synthetic images and potentially affect the model's ability to generalize to real pathological cases where mass effect is present. However, despite this limitation, our method still achieves robust performance on real pathological data, suggesting that the anatomical information learned from augmented clinically normal images

provides sufficient structural priors for accurate segmentation. Future work could explore methods to simulate mass effect during the insertion process, such as using biomechanical models or advanced image synthesis techniques.

Chapter 6

Conclusion and Perspectives

1 Summary of Findings

This thesis investigated the problem of fine-grained brain segmentation using FLAIR MRI, a non-standard modality for structure segmentation, with the goal not to explore novel architectures; but to investigate how data design can enhance the generalisation capacities of segmentation models. In general, segmentation is performed using other modalities, such as T1w MRI, which provide better anatomical contrast. However, in some cases described in this thesis, volumetric T1w images may not be available or of sufficient quality. Incidentally, FLAIR is a recurrent modality (even though not often volumetric), thanks to its sensitivity to lesions, it is used in clinical protocols of a few of the most prevalent neurological diseases (*e.g.*, TBI, MS, cancerous tissues). Yet it is rarely used as the primary input for fine-grained segmentation due to its non-uniform intensity profile, limited tissue contrast, and susceptibility to artifacts. The work presented here investigated the feasibility of using this single modality to perform fine-grained brain segmentation.

The first contribution of this work focused on enabling reliable FLAIR-based analysis through accurate brain extraction. A robust extraction method was developed to separate brain tissue from non-brain structures even in the presence of large or irregular lesions. The resulting improvement in brain delineation was shown to be able to perform brain extraction reliably in a wide variety of conditions. The method, called FLAIR-ICE, allowed us to build a preprocessing pipeline tailored towards FLAIR MRI.

Building on this foundation, we introduced FLAIRBrainSeg, a new segmentation framework to perform fine-grained anatomical parcellation directly from FLAIR. The proposed method was built by attempting to reproduce segmentations obtained from T1w MRI. We showed that FLAIRBrainSeg was able to reliably reproduce these segmentations very accurately (with a DSC > 0.9). These results were specifically compared with "modality-agnostic" methods, specifically SynthSeg, designed to work with any input modality. We showed that FLAIRBrainSeg outperformed these methods by a significant margin. These results seem to suggest that while the generative process adopted by SynthSeg is effective to generalize to multiple modalities, training with "real" FLAIR images, can lead to better results. The results also seem to suggest that FLAIR MRI contains sufficient information to perform fine-grained segmentation.

The final contribution of this thesis focused on trying to produce a model to jointly segment both anatomy and FLAIR hyperintensities. We introduced FLAIRBrainSeg$^+$, a model trained with synthetic data generated by inserting simulated lesions into healthy FLAIR images. This allowed us to create a diverse training set without requiring extensive manual annotations. The results demonstrated that FLAIRBrainSeg$^+$ was able to accurately segment both anatomical structures and lesions, achieving competitive performance compared to state-of-the-art methods that rely on multiple modalities.

Taken together, the findings demonstrate that FLAIR MRI, when processed and

modeled according to its intrinsic characteristics, can serve as a standalone modality for comprehensive brain segmentation. Segmentation accuracy and interpretability both improve markedly when model design, training, and preprocessing are explicitly tailored to the imaging physics and contrast behavior of FLAIR.

Despite these advances, important challenges remain. The proposed methods achieved strong results on diverse datasets but still exhibited sensitivity to inter-scanner and inter-population variability. Differences in acquisition parameters, field strength, and patient demographics continue to affect model robustness and limit the clinical deployment of such systems. Addressing these generalization challenges through harmonization, self-supervised learning, and domain adaptation represents a critical next step toward routine clinical adoption of FLAIR-based segmentation tools. The contributions in this thesis also contain important limitations. Unlike original segmentation methods trained on T1w MRI in a supervised manner with manual expert annotations, our tools merely attempted to reproduce the outputs of existing segmentation methods applied to T1w images. One possibility is that the errors or bias of the tools used for the obtention of these "pseudo-labels" were reproduced by our proposed methods. As such, we emphasized on using well-established and validated segmentation tools for the label generation labels. While, future work could investigate the possibility of training FLAIR-based segmentation models directly from manual expert annotations, it is very unlikely that datasets containing annotations for structures on FLAIR MRI will ever be made available.

2 Perspectives

The methods developed in this thesis have demonstrated that accurate anatomical and lesion segmentation can be achieved directly from FLAIR MRI. However, several challenges remain, and these open directions for future work that naturally extend the current contributions.

Pre-training While the models developed in this work achieved strong results on the available datasets, their robustness across scanners, institutions, and populations can vary, as shown in Fig 5.4. Preliminary experiments on large-scale pretraining, using extensive anatomical databases to initialize FLAIR-specific segmentation networks, did not lead to measurable improvements in downstream performance. However, further investigations might be warranted, as intuitively, exposing the model to different settings (scanners, resolutions, *etc* ...) would help to improve generalisation. Possible areas of exploration include self-supervised learning approaches, such as contrastive learning, which have shown promise in other medical imaging contexts.

Synthetic Image Generation In chapter 5, we introduced a training example generation process to create synthetic training data, that allowed us to insert and

combine simulated lesions with real anatomical context. This approach enabled the training of FLAIRBrainSeg^{+} to jointly segment anatomy and pathology from FLAIR MRI. The proposed generation strategy merely focused on inserting lesions into healthy images without modeling the associated mass effect (where a displacement of brain structures is observed, due to the appearance of cancerous tissue for example). As such a potential improvement could be to make an attempt to model these deformations, in the case of tumors.

Additionally, the insertion process itself remains to be actually explored. A quick exploration of alternative insertion strategies did not seem to lead to large improvements in segmentation accuracy. This is counter intuitive, as one would ideally like the training examples to be as close as possible to real "test" images. However, these results seem to be supported by the findings in [Lambert, 2024], where the authors showed that striving for realism does not necessarily lead to better segmentation accuracy. Possible improvement to the insertion process could leverage more complex generative models, such as latent diffusion models, which have seen limited use in medical imaging, due to the specifics of the domain (high-computational cost, limited data availability, *etc* ...).

The proposed lesion generation strategy is not specific to any conditions or disease. While we restricted our experiments to MS and brain tumors, further consideration should be given to extending this approach to other context, such as TBI or stroke for example. While initially planned to be included in this work, this was ultimately not done due to the lack of available "good" annotated datasets. In the case of stroke for example, labeling is often done using ADC, with limited spatial resolution. This makes it difficult to align the expert segmentation mask with the FLAIR image, required for the insertion process. However, one should understand that as we add more conditions, it is reasonable to expect the model to lose specificity to other conditions. The decision to use a single label for all types of lesions was made due to hyper-intensities on FLAIR images being notoriously difficult to differentiate or attribute to specific diseases.

Harmonization and super-resolution A different direction for future work is to investigate the possible use of "advanced" pre-processing techniques, such as harmonization to nullify scan-related variances, or super-resolution to enhance the quality of low-resolution images. In such pipelines, missing information would be inferred from the low-resolution or low-quality examples using a set of pre-trained priors, to provide "enhanced" good quality inputs to the segmentation model. In the first case, (harmonisation techniques), an attempt is made to reduce the variability induced by different scanners or acquisition protocols. In the second case (super-resolution techniques), an attempt is made to recover high-frequency details not acquired or lost during the acquisition process. While these approaches could improve performance, they rely on "implicit" or "generated" missing information to recover semantic information, which is often frowned upon by clinicians.

Beyond immediate research directions, a general goal of the development of segmentation tools is the idea of "clinical deployments", where a GPU available on a computer near the scanner could be used to provide instant insight to radiologists. Despite the introduction of CNNs in 2015 (using UNet as a reference) more than a decade ago, and despite the large increase in the number of AI-based tools approved by FDA or EC, roughly after COVID-19, current clinical practice does not include the use of segmentation models as a standard tool. While this can be partly attributed to the lack of available GPUs in clinical environments, the main challenge remains the robustness of AI models to the diverse range of data acquired in practice, and possibly, the difference in alignment between the developed models and clinical needs [Gupta et al., 2024]. As such, leveraging AI tools requires that models are aligned with clinical practices, such as the acquired modalities and sometimes limited quality of images, mainly due to time-constraints of radiologists, linked to the high patient load typical of clinical routines.

Recent developments have made it easier than ever to deploy AI models in clinical environments. Platforms such as MONAI Deploy `https://github.com/Project-MONAI/monai-deploy`, facilitates the packaging and distribution of medical AI applications in clinical production. These advances largely remove the technical barriers that previously limited clinical integration. The remaining challenge is therefore no longer computational, but one of trust, by ensuring that proposed models are reliable, interpretable, and safe to use. Future research should therefore focus on explainability and uncertainty estimation, combined with rigorous clinical validation.

1 Appendix for Chapter 4

Figure 1: Comparison on Experiment 1 of SynthSeg-132 and FLAIRBrainSeg across structure groups (e.g., tissue, cortical, subcortical, etc.) for the 132 structures in the Neuromorphometrics protocol on the in-domain test set.

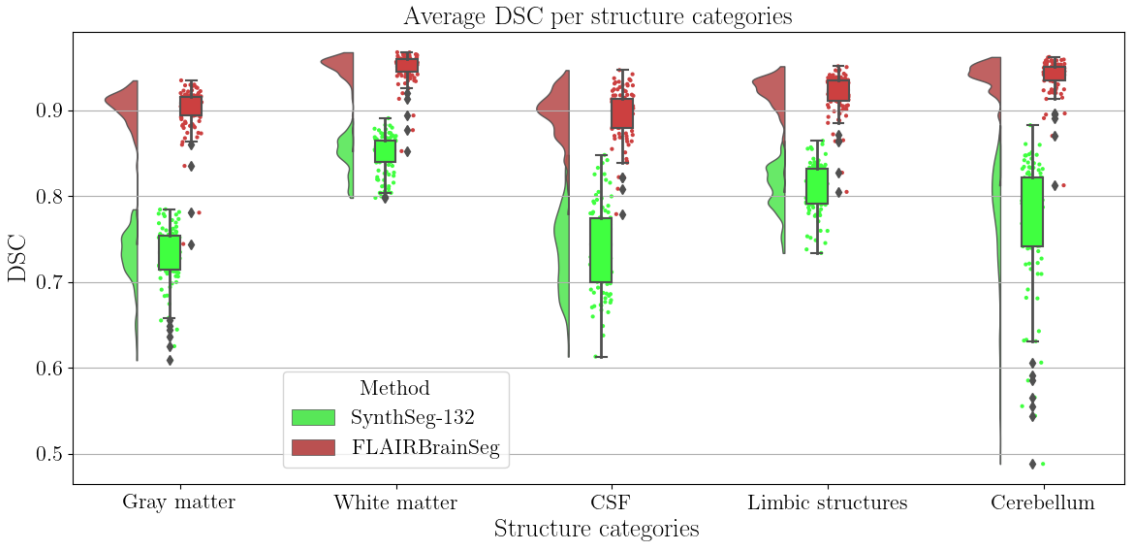
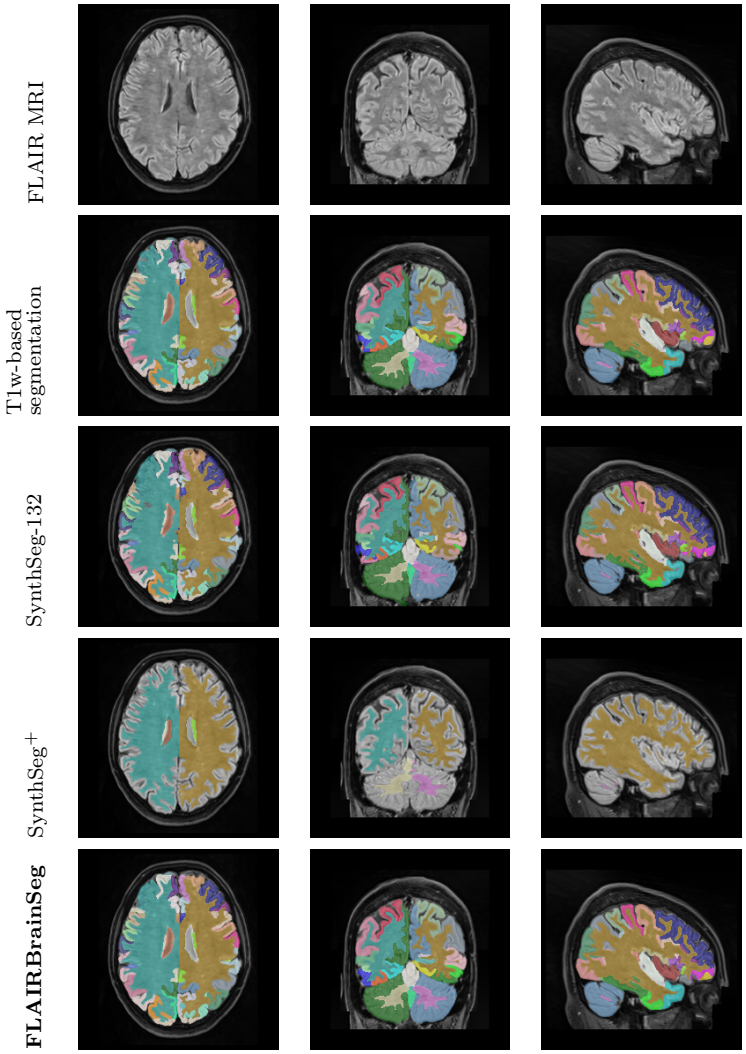


Figure 2: Examples of obtained segmentations with FLAIRBrainSeg, SynthSeg-132, and SynthSeg^{+}. Please note that the SynthSeg^{+} segmentation includes only the 35 selected structures for comparison which do not include cortical structures and cerebellum grey matter.



2 Appendix for Chapter 5

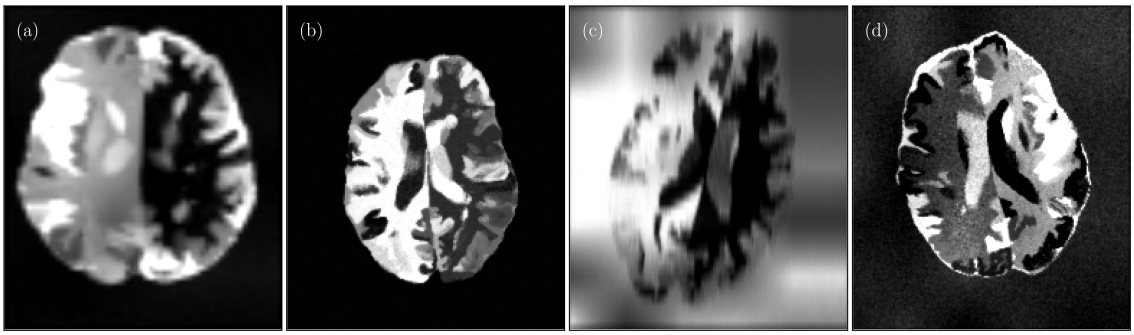


Figure 3: Examples of training examples generated by SynthSeg-WMH. Each row shows a different example. SynthSeg-WMH generates images by sampling from a Gaussian Mixture Model (GMM) constrained by the provided label map. The GMM parameters are randomly sampled at each training iteration, resulting in a wide variety of image contrasts. However, the generated images often lack realistic anatomical details and may contain artifacts due to the simplistic nature of the GMM-based synthesis process. This can lead to suboptimal feature learning for segmentation tasks, particularly when applied to real-world FLAIR images with complex pathologies.

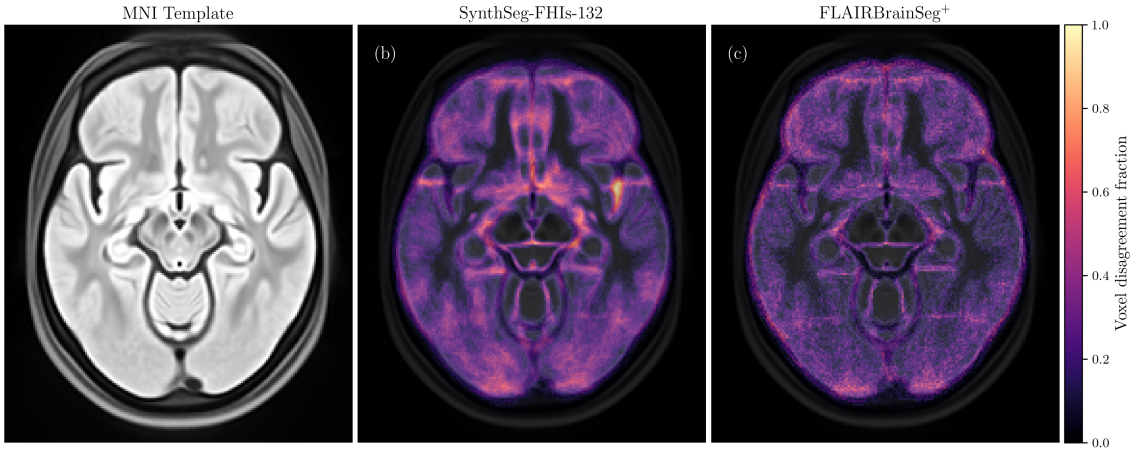


Figure 4: Comparison of segmentation errors across methods with 133 labels on UKB test set. Columns show, from left to right: FLAIR MRI, FreeSurfer SynthSeg-WMH segmentation error map, SynthSeg-FHI-133 error map, and the segmentation error map obtained with the proposed method (FLAIRBrainSeg^{+}). Error maps are computed as the absolute difference between the predicted segmentation and the T1w-based anatomical segmentation merged with the WMH identified by DLB, summed over all images in the UKB test set and projected onto the MNI152 FLAIR template for visualization.

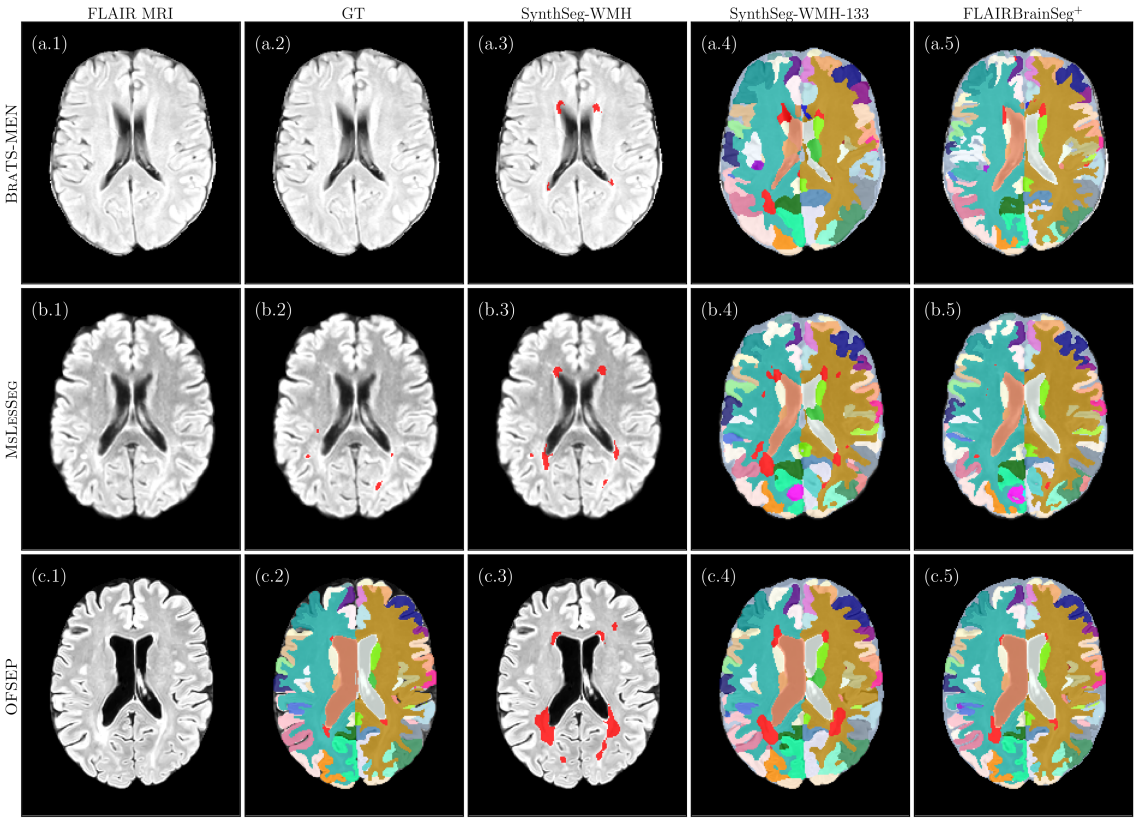


Figure 5: Axial view of representative failure cases across selected OOD datasets. Each row shows a subject from BraTS-MEN, MSLesSeg, and OFSEP. The datasets differ in FLAIR resolution and image quality: anisotropic and noisy scans in BraTS-MEN, moderate variability in MSLesSeg, and high-quality isotropic scans in OFSEP. These differences are the main contributors to the observed variations in segmentation performance.

Comparison of insertion strategies. To validate the choice of FHI insertion strategy for the generation of synthetic training images, we compared three methods for merging FHI masks onto clinically normal (CN) FLAIR MRIs. For each method, we generated synthetic training images by inserting FHI masks from the UCSF-PDGM training set into CN FLAIR images from the UKB training set. Each synthetic dataset was composed of 50% real images from UCSF-PDGM and 50% synthetic tumor images created with the listed blending method. The three blending methods compared were:

Poisson Image Editing
: [Pérez et al., 2003] is a classical image blending technique that seamlessly merges the FHI mask and its corresponding intensities into the target FLAIR image by solving a Poisson equation to minimize intensity discontinuities at the boundaries.

Copy-Paste
: is a simple copy-paste approach, where the FHI mask and its corresponding intensities are directly inserted into the target FLAIR image without any intensity blending or harmonization.

nnUNet-translation
: [Longuefosse et al., 2024] is a nnU-Net–based model trained to locally refine FHI boundaries. The model was trained to recreate lesions from voided FHI regions in UCSF FLAIR images, using the UCSF-PDGM training set.

	Dice Similarity Coefficient (DSC) ↑			
Blending Method	**Mean ± SD**	**Min**	**Median**	**Max**
Poisson Image Editing	0.8346 ± 0.1230	**0.5033**	0.8814	**0.9584**
nnUNet translation	**0.8587 ± 0.1290**	0.3636	**0.9093**	0.9578
Copy–Paste	0.8185 ± 0.1545	0.3291	0.8833	0.9454

Table 1: Comparison of FHI insertion strategies for synthetic training data generation. For each synthetic dataset, a 3D nnU-Net was trained with identical parameters and data splits, and evaluated on the held-out UCSF-PDGM test set. Each model used 50% real images from UCSF-PDGM and 50% synthetic tumor images created with the listed blending method. Given the trade-off between segmentation performance, stability, and efficiency, Poisson blending was selected as the default FHI-merging strategy for all subsequent experiments.

Bibliography

[Ashburner and Friston, 2005] Ashburner, J. and Friston, K. J. (2005). Unified segmentation. *NeuroImage*, 26(3):839–851. 46, 74

[Avants et al., 2008] Avants, B. B., Epstein, C. L., Grossman, M., and Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Med. Image Anal.*, 12(1):26–41. 32, 51

[Avants et al., 2011] Avants, B. B., Tustison, N. J., Song, G., Cook, P. A., Klein, A., and Gee, J. C. (2011). A reproducible evaluation of ANTs similarity metric performance in brain image registration. *NeuroImage*, 54(3):2033–2044. 62, 76

[Avesta et al., 2023] Avesta, A., Hossain, S., Lin, M., Aboian, M., Krumholz, H. M., and Aneja, S. (2023). Comparing 3D, 2.5D, and 2D Approaches to Brain Image Auto-Segmentation. *Bioengineering (Basel, Switzerland)*, 10(2):181. 36, 37

[Balakrishnan et al., 2019] Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J., and Dalca, A. V. (2019). Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8):1788–1800. 32

[Billot et al., 2023a] Billot, B., Greve, D. N., Puonti, O., Thielscher, A., Van Leemput, K., Fischl, B., Dalca, A. V., Iglesias, J. E., et al. (2023a). Synthseg: Segmentation of brain MRI scans of any contrast and resolution without retraining. *Medical image analysis*, 86:102789. 17, 38, 39, 42, 61, 66, 74, 75, 80, 119

[Billot et al., 2023b] Billot, B., Magdamo, C., Cheng, Y., Arnold, S. E., Das, S., and Iglesias, J. E. (2023b). Robust machine learning segmentation for large-scale analysis of heterogeneous clinical brain MRI datasets. *Proceedings of the National Academy of Sciences*, 120(9):e2216399120. 38, 61, 66, 74

[Brown et al., 2014] Brown, R. W., Cheng, Y.-C. N., Haacke, E. M., Thompson, M. R., and Venkatesan, R. (2014). *Signal, Contrast, and Noise*, pages 325–373. John Wiley & Sons, Ltd. 18

[Bycroft et al., 2018] Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209. 49, 62, 81

[Calabrese et al., 2022] Calabrese, E., Villanueva-Meyer, J., Rudie, J., Rauschecker, A., Baid, U., Bakas, S., Cha, S., Mongan, J., and Hess, C. (2022). The University of California San Francisco preoperative diffuse glioma MRI (UCSF-PDGM). Dataset. 49

[Cardoso et al., 2022] Cardoso, M. J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., et al. (2022). Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*. 35

[Carney et al., 2017] Carney, N., Totten, A. M., O'Reilly, C., Ullman, J. S., Hawryluk, G. W., Bell, M. J., Bratton, S. L., Chesnut, R., Harris, O. A., Kissoon, N., et al. (2017). Guidelines for the management of severe traumatic brain injury. *Neurosurgery*, 80(1):6–15. 25

[Casamitjana et al., 2024] Casamitjana, A., Mancini, M., Robinson, E., Peter, L., Annunziata, R., Althonayan, J., Crampsie, S., Blackburn, E., Billot, B., Atzeni, A., et al. (2024). A next-generation, histological atlas of the human brain and its application to automated brain MRI segmentation. *bioRxiv*. 31

[Çiçek et al., 2016] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer. 33

[Collins et al., 1995] Collins, D. L., Holmes, C. J., Peters, T. M., and Evans, A. C. (1995). Automatic 3-d model-based neuroanatomical segmentation. *Human Brain Mapping*, 3(3):190–208. 32

[Commowick et al., 2021] Commowick, O., Cervenansky, F., Cotton, F., and Dojat, M. (2021). Msseg-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure. In *MICCAI 2021-24th international conference on medical image computing and computer assisted intervention*, page 126. 87

[Coupé et al., 2020] Coupé, P., Mansencal, B., Clément, M., Giraud, R., de Senneville, B. D., Ta, V.-T., Lepetit, V., and Manjon, J. V. (2020). Assemblynet: A large ensemble of cnns for 3d whole brain MRI segmentation. *NeuroImage*, 219:117026. 36, 60, 62, 63, 65, 74, 78

[Coupé et al., 2010] Coupé, P., Manjón, J. V., Fonov, V., Pruessner, J., Robles, M., and Collins, D. L. (2010). Nonlocal patch-based label fusion for hippocampus segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2010*, volume 6363 of *Lecture Notes in Computer Science*, pages 129–136. Springer. 33, 51

[Denis de Senneville et al., 2020] Denis de Senneville, B., Manjon, J. V., and Coupé, P. (2020). RegQCNET: Deep quality control for image-to-template brain MRI affine registration. *Physics in Medicine & Biology*, 65(22):225022. 49, 63, 81

[Desikan et al., 2006] Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., and Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, 31(3):968–980. 30, 31

[Dill, 2008] Dill, T. (2008). Contraindications to magnetic resonance imaging: non-invasive imaging. *Heart*, 94(7):943–948. 18

[Dorent et al., 2021] Dorent, R., Booth, T., Li, W., Sudre, C. H., Kafiabadi, S., Cardoso, J., Ourselin, S., and Vercauteren, T. (2021). Learning joint segmentation of tissues and brain lesions from task-specific hetero-modal domain-shifted datasets. *Medical image analysis*, 67:101862. 75

[Dosovitskiy et al., 2020] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*. 38

[Duchesne et al., 2019] Duchesne, S., Chouinard, I., Potvin, O., Fonov, V. S., Khademi, A., Bartha, R., Bellec, P., Collins, D. L., Descoteaux, M., Hoge, R., McCreary, C. R., Ramirez, J., Scott, C. J., Smith, E. E., Strother, S. C., Black, S. E., and for the CIMA-Q group and the CCNA group (2019). The canadian dementia imaging protocol: Harmonizing national cohorts. *Journal of Magnetic Resonance Imaging*, 49(2):456–465. 81, 83

[Ellingson et al., 2015] Ellingson, B. M. et al. (2015). Radiographic response criteria for brain tumors in clinical trials: practical aspects and limitations. *Radiology*, 275(2):346–364. 18

[Eskildsen et al., 2012] Eskildsen, S. F., Coupé, P., Fonov, V., Manjón, J. V., Leung, K. K., Guizard, N., Wassef, S. N., Østergaard, L. R., and Collins, D. L. (2012). Beast: Brain extraction based on nonlocal segmentation technique. *NeuroImage*, 59(3):2362–2373. 46

[Fan et al., 2016] Fan, L., Li, H., Zhuo, J., Zhang, Y., Wang, J., Chen, L., Yang, Z., Chu, C., Xie, S., Laird, A. R., Fox, P. T., Eickhoff, S. B., Yu, C., and Jiang, T. (2016). The human brainnetome atlas: A new brain atlas based on connectional architecture. *Cerebral Cortex*, 26(8):3508–3526. 31

[Ferreira et al., 2024] Ferreira, A., Solak, N., Li, J., Dammann, P., Kleesiek, J., Alves, V., and Egger, J. (2024). How we won brats 2023 adult glioma challenge? just faking it! enhanced synthetic data augmentation and model ensemble for brain tumour segmentation. *arXiv preprint arXiv:2402.17317*. 38, 74, 75

[Filippi and Rocca, 2011] Filippi, M. and Rocca, M. A. (2011). Magnetic resonance imaging in neuroscience: the contribution of clinical and experimental studies. *Journal of Magnetic Resonance Imaging*, 33(4):649–661. 18

[Fischl, 2004] Fischl, B. (2004). Automatically parcellating the human cerebral cortex. *Cerebral Cortex*, 14(1):11–22. 33

[Fischl, 2012] Fischl, B. (2012). Freesurfer. *NeuroImage*, 62(2):774–781. 74

[Fonov et al., 2011] Fonov, V., Evans, A. C., Botteron, K., Almli, C. R., McKinstry, R. C., and Collins, D. L. (2011). Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage*, 54(1):313–327. 51, 62

[Frazier et al., 2005] Frazier, J. A., Chiu, S., Breeze, J. L., Makris, N., Lange, N., Kennedy, D. N., Herbert, M. R., Bent, E. K., Koneru, V. K., Dieterich, M. E., Hodge, S. M., Rauch, S. L., Grant, P. E., Cohen, B. M., Seidman, L. J., Caviness, V. S., and Biederman, J. (2005). Structural brain magnetic resonance imaging of limbic and thalamic volumes in pediatric bipolar disorder. *American Journal of Psychiatry*, 162(7):1256–1265. 31

[Fu et al., 2025] Fu, G., Nichelli, L., Herran, D., Valabregue, R., Alentorn, A., Hoang-Xuan, K., Houillier, C., Dormont, D., Lehéricy, S., and Colliot, O. (2025). Comparing foundation models and nnu-net for segmentation of primary brain lymphoma on clinical routine post-contrast t1-weighted mri. In *Medical Imaging 2025: Clinical and Biomedical Imaging*, volume 13410. SPIE. 43

[Glasser et al., 2016] Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., Smith, S. M., and Van Essen, D. C. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178. Epub 2016 Jul 20. 31

[Goldstein et al., 2007] Goldstein, J. M., Seidman, L. J., Makris, N., Ahern, T., O'Brien, L. M., Caviness, V. S. J., Kennedy, D. N., Faraone, S. V., and Tsuang, M. T. (2007). Hypothalamic abnormalities in schizophrenia: sex effects and genetic vulnerability. *Biological Psychiatry*, 61(8):935–945. 31

[Goodkin et al., 2021] Goodkin, O., Prados, F., Vos, S. B., Pemberton, H., Collorone, S., Hagens, M. H. J., Cardoso, M. J., Yousry, T. A., Thornton, J. S., Sudre, C. H., Barkhof, F., and MAGNIMS study group (2021). FLAIR-only joint volumetric analysis of brain lesions and atrophy in clinically isolated syndrome (CIS) suggestive of multiple sclerosis. *NeuroImage Clin.*, 29(102542):102542. 26, 60

[Gupta et al., 2024] Gupta, V., Erdal, B., Ramirez, C., Floca, R., Genereaux, B., Bryson, S., Bridge, C., Kleesiek, J., Nensa, F., Braren, R., et al. (2024). Current state of community-driven radiological ai deployment in medical imaging. *JMIR AI*, 3(1):e55833. 99

[Haacke et al., 2014] Haacke, E. M., Brown, R. W., Thompson, M. R., and Venkatesan, R. (2014). *Magnetic Resonance Imaging: Physical Principles and Sequence Design*. Wiley-Blackwell, Hoboken, NJ, 2nd edition. 18

[Hatamizadeh et al., 2021] Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H. R., and Xu, D. (2021). Swin unetr: Swin transformers for semantic segmentation of brain tumors in MRI images. In *International MICCAI brainlesion workshop*, pages 272–284. Springer. 38, 61

[He et al., 2023] He, Y., Nath, V., Yang, D., Tang, Y., Myronenko, A., and Xu, D. (2023). Swinunetr-v2: Stronger swin transformers with stagewise convolutions for 3D medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 416–426. Springer. 38

[Henschel et al., 2020] Henschel, L., Conjeti, S., Estrada, S., Diers, K., Fischl, B., and Reuter, M. (2020). Fastsurfer-a fast and accurate deep learning based neuroimaging pipeline. *NeuroImage*, 219:117012. 36, 60, 65, 74

[Henschel et al., 2022] Henschel, L., Kügler, D., and Reuter, M. (2022). FastSurferVINN: Building resolution-independence into deep learning segmentation methods—A solution for HighRes brain MRI. *NeuroImage*, 251:118933. 36, 37

[Himmetoglu et al., 2025] Himmetoglu, M., Ciernik, I., and Konukoglu, E. (2025). Learning to segment anatomy and lesions from disparately labeled sources in brain mri. *CoRR*, abs/2503.18840. 75

[Hoopes et al., 2022] Hoopes, A., Mora, J. S., Dalca, A. V., Fischl, B., and Hoffmann, M. (2022). SynthStrip: skull-stripping for any brain image. *NeuroImage*, 260:119474. 47

[Huo et al., 2018] Huo, Y., Xu, Z., Aboud, K., Parvathaneni, P., Bao, S., Bermudez, C., Resnick, S. M., Cutting, L. E., and Landman, B. A. (2018). Spatially localized atlas network tiles enables 3d whole brain segmentation from limited data. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part III 11*, pages 698–705. Springer. 36, 37, 74, 119

[Huo et al., 2019] Huo, Y., Xu, Z., Xiong, Y., Aboud, K., Parvathaneni, P., Bao, S., Bermudez, C., Resnick, S. M., Cutting, L. E., and Landman, B. A. (2019). 3D whole brain segmentation using spatially localized atlas network tiles. *NeuroImage*, 194:105–119. 60, 65

[ICNIRP, 2017] ICNIRP (2017). Icnirp statement on diagnostic devices using non-ionizing radiation: Existing regulations and potential health risks. *Health Physics*, 112(3):305–321. 18

[Isensee et al., 2021] Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., and Maier-Hein, K. H. (2021). nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211. 35, 36, 37, 53, 65, 79, 119

[Isensee et al., 2019] 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. 47

[Isensee et al., 2024] Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., and Jäger, P. F. (2024). nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 488–498, Cham. Springer Nature Switzerland. 65

[Jack Jr. et al., 2008] Jack Jr., C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P. J., L. Whitwell, J., Ward, C., Dale, A. M., Felmlee, J. P., Gunter, J. L., Hill, D. L., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C. S., Krueger, G., Ward, H. A., Metzger, G. J., Scott, K. T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J. P., Fleisher, A. S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J., and Weiner, M. W. (2008). The alzheimer's disease neuroimaging initiative (adni): MRI methods. *Journal of Magnetic Resonance Imaging*, 27(4):685–691. 18

[Jauch et al., 2013] Jauch, E. C., Saver, J. L., Adams, H. P., Bruno, A., Connors, J. B., Demaerschalk, B. M., Khatri, P., McMullan, P. W., Qureshi, A. I., Rosenfield, K., Scott, P. A., Summers, D. R., Wang, D. Z., Wintermark, M., and Yonas, H. (2013). Guidelines for the early management of patients with acute ischemic stroke. *Stroke*, 44(3):870–947. 25

[Jenkinson et al., 2012] Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., and Smith, S. M. (2012). Fsl. *NeuroImage*, 62(2):782–790. 74

[Kamraoui et al., 2022] Kamraoui, R. A., Ta, V.-T., Tourdias, T., Mansencal, B., Manjon, J. V., and Coupé, P. (2022). DeepLesionBrain: Towards a broader deep-learning generalization for multiple sclerosis lesion segmentation. *Medical Image Analysis*, 76:102312. 51, 62, 76, 82, 87

[Kirillov et al., 2023] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. (2023). Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026. 43, 119

[Klein et al., 2005] Klein, A., Mensh, B., Ghosh, S., Tourville, J., and Hirsch, J. (2005). Mindboggle: automated brain labeling with multiple atlases. *BMC medical imaging*, 5:1–14. 32

[Klein and Tourville, 2012] Klein, A. and Tourville, J. (2012). 101 labeled brain images and a consistent human cortical labeling protocol. *Frontiers in neuroscience*, 6:171. 30

[Kofler et al., 2024] Kofler, F., Meissen, F., Steinbauer, F., Graf, R., Ehrlich, S. K., Reinke, A., Oswald, E., Waldmannstetter, D., Hoelzl, F., Horvath, I., Turgut, O., Shit, S., Bukas, C., Yang, K., Paetzold, J. C., de da Rosa, E., Mekki, I., Vinayahalingam, S., Kassem, H., Zhang, J., Chen, K., Weng, Y., Durrer, A., Cattin, P. C., Wolleb, J., Sadique, M. S., Rahman, M. M., Farzana, W., Temtam, A., Iftekharuddin, K. M., Adewole, M., Anwar, S. M., Baid, U., Janas, A., Kazerooni, A. F., LaBella, D., Li, H. B., Moawad, A. W., Conte, G.-M., Farahani, K., Eddy, J., Sheller, M., Pati, S., Karagyris, A., Aristizabal, A., Bergquist, T., Chung, V., Shinohara, R. T., Dako, F., Wiggins, W., Reitman, Z., Wang, C., Liu, X., Jiang, Z., Johanson, E., Meier, Z., Familiar, A., Davatzikos, C., Freymann, J., Kirby, J., Bilello, M., Fathallah-Shaykh, H. M., Wiest, R., Kirschke, J., Colen, R. R., Kotrotsou, A., Lamontagne, P., Marcus, D., Milchenko, M., Nazeri, A., Weber, M.-A., Mahajan, A., Mohan, S., Mongan, J., Hess, C., Cha, S., Villanueva-Meyer, J., Colak, E., Crivellaro, P., Jakab, A., Fatade, A., Omidiji, O., Lagos, R. A., Olatunji, O. O., Khanna, G., Kirkpatrick, J., Alonso-Basanta, M., Rashid, A., Bornhorst, M., Nabavizadeh, A., Lepore, N., Palmer, J., Porras, A., Albrecht, J., Anazodo, U., Aboian, M., Calabrese, E., Rudie, J. D., Linguraru, M. G., Iglesias, J. E., Leemput, K. V., Bakas, S., Wiestler, B., Ezhov, I., Piraud, M., and Menze, B. H. (2024). The brain tumor segmentation (BraTS) challenge: Local synthesis of healthy brain tissue via inpainting. 75

[LaBella et al., 2025] LaBella, D., Baid, U., Khanna, O., McBurney-Lin, S., McLean, R., Nedelec, P., Rashid, A., Tahon, N. H., Altes, T., Bhalerao, R., et al. (2025). Analysis of the brats 2023 intracranial meningioma segmentation challenge. *Machine Learning for Biomedical Imaging*, 3(March 2025 issue):38–58. 74

[Lambert, 2024] Lambert, B. (2024). *Quantification et caractérisation de l'incertitude de segmentation d'images médicales par des réseaux profonds.* PhD thesis, Université Grenoble Alpes. 75, 98

[Lansberg et al., 2000] Lansberg, M. G., Thijs, V. N., Bammer, R., et al. (2000). Mri for the assessment of patients with acute stroke. *The Lancet Neurology*, 1(11):700–708. 18

[Laso et al., 2024] Laso, P., Cerri, S., Sorby-Adams, A., Guo, J., Mateen, F., Goebl, P., Wu, J., Liu, P., Li, H. B., Young, S. I., et al. (2024). Quantifying white matter hyperintensity and brain volumes in heterogeneous clinical and low-field portable mri. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE. 17, 39, 75, 79

[Lawry Aguila et al., 2025] Lawry Aguila, A., Liu, P., Puonti, O., and Iglesias, J. E. (2025). Conditional diffusion models for guided anomaly detection in brain MRI using fluid-driven anomaly randomization. In *MICCAI Workshop on Deep Generative Models*, pages 67–77. Springer. 75

[Le Bihan, 2012] Le Bihan, D. (2012). Diffusion mri: what water tells us about biological tissues. *PLoS biology*, 10(7):e1001366. 18

[Le Bot et al., 2025] Le Bot, E., Giraud, R., Mansencal, B., Tourdias, T., Manjon, J. V., and Coupe, P. (2025). FLAIRBrainseg: Fine-grained brain segmentation using FLAIR MRI only. In *Medical Imaging with Deep Learning*. 74, 86

[Lee and Newberg, 2005] Lee, B. and Newberg, A. (2005). Neuroimaging in traumatic brain imaging. *NeuroRx*, 2(2):372–383. 24

[Liu et al., 2021] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022. 38

[Longuefosse et al., 2024] Longuefosse, A., Bot, E. L., De Senneville, B. D., Giraud, R., Mansencal, B., Coupé, P., Desbarats, P., and Baldacci, F. (2024). Adapted nnu-net: A robust baseline for cross-modality synthesis and medical image inpainting. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 24–33. Springer. 106

[Ma et al., 2024] Ma, J., He, Y., Li, F., Han, L., You, C., and Wang, B. (2024). Segment anything in medical images. *Nature Communications*, 15(1):654. 43, 119

[Maier-Hein et al., 2018] Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A. P., Carass, A., Feldmann, C., Frangi, A. F., Full, P. M., van Ginneken, B., Hanbury, A., Honauer, K., Kozubek, M., Landman, B. A., März, K., Maier, O., Maier-Hein, K., Menze, B. H., Müller, H., Neher, P. F., Niessen, W., Rajpoot, N., Sharp, G. C., Sirinukunwattana, K., Speidel, S., Stock, C., Stoyanov, D., Taha, A. A., van der Sommen, F., Wang, C.-W., Weber, M.-A., Zheng, G., Jannin, P., and Kopp-Schneider, A. (2018). Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature Communications*, 9(1):5217. 35

[Makris et al., 2006] Makris, N., Goldstein, J. M., Kennedy, D., Hodge, S. M., Caviness, V. S., Faraone, S. V., Tsuang, M. T., and Seidman, L. J. (2006). Decreased volume of left and total anterior insular lobule in schizophrenia. *Schizophrenia Research*, 83(2-3):155–171. 31

[Manjón et al., 2010a] Manjón, J. V., Coupé, P., Martí-Bonmatí, L., Collins, D. L., and Robles, M. (2010a). Adaptive non-local means denoising of mr images with spatially varying noise levels. *Journal of Magnetic Resonance Imaging*, 31(1):192–203. 76

[Manjón et al., 2014] Manjón, J. V., Eskildsen, S. F., Coupé, P., Romero, J. E., Collins, D. L., and Robles, M. (2014). Nonlocal intracranial cavity extraction. *International Journal of Biomedical Imaging*, 2014:820205. 46, 49, 76

[Manjón et al., 2020a] Manjón, J. V., Romero, J. E., Vivo-Hernando, R., Rubio, G., Aparici, F., de la Iglesia-Vaya, M., Tourdias, T., and Coupé, P. (2020a). Blind MRI brain lesion inpainting using deep learning. In *Simulation and Synthesis in Medical Imaging*, pages 41–49. 62

[Manjón et al., 2020b] Manjón, J. V., Romero, J. E., Vivo-Hernando, R., Rubio-Navarro, G., De la Iglesia-Vaya, M., Aparici-Robles, F., and Coupé, P. (2020b). Deep ice: A deep learning approach for mri intracranial cavity extraction. *arXiv preprint arXiv:2001.05720*. 47

[Manjón et al., 2010b] Manjón, J. V., Tohka, J., and Robles, M. (2010b). Improved estimates of partial volume coefficients from noisy brain MRI using spatial context. *NeuroImage*, 53(2):480–490. 51, 62

[Manjón et al., 2025] Manjón, J. V., Morell-Ortega, S., Ruiz-Perez, M., Mansencal, B., Bot, E. L., Gadea, M., Lanuza, E., Catheline, G., Tourdias, T., Planche, V., Giraud, R., Rivière, D., Mangin, J.-F., Labra-Avila, N., Vivo-Hernando, R., Rubio, G., Aparici, F., de la Iglesia-Vaya, M., and Coupé, P. (2025). Ultra-high resolution multimodal MRI densely labelled holistic structural brain atlas. 31

[Morell-Ortega et al., 2024] Morell-Ortega, S., Perez, M. R., Coupe, P., and Manjon, J. V. (2024). REMIX: Resolution enhancement through mixture of experts. In *Medical Imaging with Deep Learning*. 70, 92

[Morell-Ortega et al., 2025] Morell-Ortega, S., Ruiz-Perez, M., Gadea, M., Vivo-Hernando, R., Rubio, G., Aparici, F., de la Iglesia-Vaya, M., Tourdias, T., Mansencal, B., Coupé, P., et al. (2025). Robust deep MRI contrast synthesis using a prior-based and task-oriented 3d network. *Imaging Neuroscience*, 3:IMAG–a. 42

[Nagaraja, 2021] Nagaraja, N. (2021). Diffusion weighted imaging in acute ischemic stroke: a review of its interpretation pitfalls and advanced diffusion imaging application. *Journal of the neurological sciences*, 425:117435. 20

[NeuroMorphometrics Inc., 2024] NeuroMorphometrics Inc. (2024). Neuromorphometrics: Experts in human brain labeling. Accessed: 2025-06-26. 31

[Organization, 2006] Organization, W. H. (2006). *Neurological disorders: public health challenges*. World Health Organization. 23

[Pérez et al., 2003] Pérez, P., Gangnet, M., and Blake, A. (2003). Poisson image editing. In *ACM SIGGRAPH 2003 Papers*, pages 313–318. Association for Computing Machinery. 52, 77, 106

[Powers et al., 2019] Powers, W. J., Rabinstein, A. A., Ackerson, T., Adeoye, O. M., Bambakidis, N. C., Becker, K., Biller, J., Brown, M., Demaerschalk, B. M., Hoh, B., Jauch, E. C., Kidwell, C. S., Leslie-Mazwi, T. M., Ovbiagele, B., Scott, P. A., Sheth, K. N., Southerland, A. M., Summers, D. V., and Tirschwell, D. L. (2019). Guidelines for the early management of patients with acute ischemic stroke: 2019 update to the 2018 guidelines for the early management of acute ischemic stroke: A guideline for healthcare professionals from the american heart association/american stroke association. *Stroke*, 50(12):e344–e418. 60

[Puonti et al., 2016] Puonti, O., Iglesias, J. E., and Van Leemput, K. (2016). Fast and sequence-adaptive whole-brain segmentation using parametric bayesian modeling. *NeuroImage*, 143:235–249. 61

[Rohlfing et al., 2003] Rohlfing, T., Brandt, R., Menzel, R., and Maurer Jr, C. R. (2003). Segmentation of three-dimensional images using non-rigid registration: Methods and validation with application to confocal microscopy images of bee brains. In *Medical Imaging 2003: Image Processing*, volume 5032, pages 363–374. SPIE. 32

[Rondinella et al., 2024] Rondinella, A., Guarnera, F., Crispino, E., Russo, G., Di Lorenzo, C., Maimone, D., Pappalardo, F., and Battiato, S. (2024). Icpr 2024 competition on multiple sclerosis lesion segmentation—methods and results. In *International Conference on Pattern Recognition*, pages 1–16. Springer. 74

[Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer. 33, 74

[Roy et al., 2019] Roy, A. G., Conjeti, S., Navab, N., Wachinger, C., Initiative, A. D. N., et al. (2019). QuickNAT: A fully convolutional network for quick and accurate segmentation of neuroanatomy. *NeuroImage*, 186:713–727. 60

[Schmierer et al., 2019] Schmierer, K., Campion, T., Sinclair, A., van Hecke, W., Matthews, P. M., and Wattjes, M. P. (2019). Towards a standard MRI protocol for multiple sclerosis across the UK. *Br. J. Radiol.*, 92(1101):20180926. 73

[Smith, 2002] Smith, S. M. (2002). Fast robust automated brain extraction. *Human Brain Mapping*, 17(3):143–155. 46

[Steinmetz et al., 2024] Steinmetz, J. D., Seeher, K. M., Schiess, N., Nichols, E., Cao, B., Servili, C., Cavallera, V., Cousin, E., Hagins, H., Moberg, M. E., Mehlman, M. L., et al. (2024). Global, regional, and national burden of disorders affecting the nervous system,

1990–2021: a systematic analysis for the global burden of disease study 2021. *The Lancet Neurology*, 23(4):344–381. 23, 24

[Stulberg et al., 2023] Stulberg, E. L., Sachdev, P. S., Murray, A. M., Cramer, S. C., Sorond, F. A., Lakshminarayan, K., and Sabayan, B. (2023). Post-stroke brain health monitoring and optimization: A narrative review. *Journal of Clinical Medicine*, 12(23):7413. 23

[Tustison et al., 2010] Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., and Gee, J. C. (2010). N4ITK: Improved N3 bias correction. *IEEE Transactions on Medical Imaging*, 29(6):1310–1320. 51, 62, 76

[Ulrich et al., 2023] Ulrich, C., Isensee, F., Wald, T., Zenk, M., Baumgartner, M., and Maier-Hein, K. H. (2023). Multitalent: A multi-dataset approach to medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 648–658. Springer. 81

[Ulrich et al., 2025] Ulrich, C., Wald, T., Isensee, F., and Maier-Hein, K. H. (2025). Large scale supervised pretraining for traumatic brain injury segmentation. *arXiv e-prints*, pages arXiv–2504. 81

[Vukusic et al., 2020] Vukusic, S., Casey, R., Rollot, F., Brochet, B., Pelletier, J., Laplaud, D.-A., De Sèze, J., Cotton, F., Moreau, T., Stankoff, B., et al. (2020). Observatoire français de la sclérose en plaques (OFSEP) : A unique multimodal nationwide MS registry in France. *Multiple Sclerosis Journal*, 26(1):118–122. 49, 61

[Wattjes et al., 2021] Wattjes, M. P., Ciccarelli, O., Reich, D. S., Banwell, B., de Stefano, N., Enzinger, C., Fazekas, F., Filippi, M., Frederiksen, J., Gasperini, C., Hacohen, Y., Kappos, L., Li, D. K. B., Mankad, K., Montalban, X., Newsome, S. D., Oh, J., Palace, J., Rocca, M. A., Sastre-Garriga, J., Tintoré, M., Traboulsee, A., Vrenken, H., Yousry, T., Barkhof, F., Rovira, À., Magnetic Resonance Imaging in Multiple Sclerosis study group, Consortium of Multiple Sclerosis Centres, and North American Imaging in Multiple Sclerosis Cooperative MRI guidelines working group (2021). 2021 MAGNIMS-CMSC-NAIMS consensus recommendations on the use of MRI in patients with multiple sclerosis. *Lancet Neurol*, 20(8):653–670. 73

[Wen et al., 2023] Wen, G., Shim, V., Holdsworth, S. J., Fernandez, J., Qiao, M., Kasabov, N., and Wang, A. (2023). Machine learning for brain MRI data harmonisation: a systematic review. *Bioengineering*, 10(4):397. 92

[Xie et al., 2019] Xie, L., Wang, J., Dong, M., Wolk, D. A., and Yushkevich, P. A. (2019). Improving multi-atlas segmentation by convolutional neural network based patch error estimation. In Shen, D., Liu, T., Peters, T. M., Staib, L. H., Essert, C., Zhou, S., Yap, P.-T., and Khan, A., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 347–355, Cham. Springer International Publishing. 33, 35

[Xu et al., 2024] Xu, P., Lyu, J., Lin, L., Cheng, P., and Tang, X. (2024). Lf-synthseg: Label-free brain tissue-assisted tumor synthesis and segmentation. *IEEE Journal of Biomedical and Health Informatics*. 75

List of Figures

1.1 Schematic of a cylindrical superconducting MRI scanner. Top: cross-section of the cylinder showing primary, gradient, and RF transmit coils. Bottom: longitudinal section showing the same coils and the RF receive coil. Source: `https://www.wikipedia.fr`. 19

1.2 Examples of axial slices across common MRI sequences (native, Right-Anterior-Superior orientation) acquired from a single UCSF-PDGM subject. Panels show: (a) T1-weighted, (b) T1c, (c) T2, (d) FLAIR, (e) DWI, and (f) ADC. Contrast variations across modalities highlight the complementarity of MRI sequences in characterizing tissue properties. *T1c corresponds to T1w acquisitions with prior contrast agent administration to the patient. This enhances the visibility of brain regions, but is not always used due to potential reactions.* 21

1.3 Illustration of simulated MRI artifacts applied to an axial FLAIR slice. (a) Original FLAIR MRI, (b) anisotropy, (c) bias field, (d) Gibbs ringing, (e) low-resolution sampling, (f) motion-induced artifacts, (g) Rician noise, and (h) partial-volume effect. All panels share the same windowing and geometry. Augmentation parameters were chosen for high visual contrast. 22

1.4 Representative axial FLAIR slices (MNI, RAS) showing disease-specific lesion appearances. (a) Ischemic stroke with a focal territorial hyperintensity in the left hemisphere. (b) Multiple sclerosis presenting with small, ovoid periventricular hyperintensities. (c) Brain tumor with heterogeneous signal intensity. (d) Traumatic brain injury with patchy white matter hyperintensities of irregular intensity. These examples illustrate both lesion distribution variability and the diversity of FLAIR intensity patterns across pathologies. Intensity scaling differs slightly across panels and reflects acquisition conditions. *MNI refers to the Montreal Neurological Institute space, a standardized brain template commonly used in neuroimaging, using templates is primordial to be able to compare images from different subjects using the same orientation and size.* 24

2.1 Examples of common brain segmentation protocols. *b.* NeuroMorphometrics (138 regions), *c.* SynthSeg (63 regions), *d.* Harvard-Oxford (48 cortical + 21 subcortical regions). *Brain segmentation protocols vary by both their number of structures and definitions.* 30

2.2 Architecture of a 3D U-Net model for volumetric segmentation. The network followed an encoder–decoder structure with skip connections. The contracting path (left) progressively downsampled the input through repeated applications of $3 \times 3 \times 3$ convolutions with ReLU activation (yellow) and max pooling (red), increasing feature channels while reducing spatial resolution. The bottleneck layer (center) encoded the most compressed representation. The expansive path (right) restored spatial resolution using up-convolutions (blue), concatenating feature maps from the contracting path (copy and crop, grey arrows) to preserve localization information. The final $1 \times 1 \times 1$ convolution (green) mapped the feature maps to the output segmentation classes. 34
2.3 Overview of the nnU-Net framework, proposed as a "self-configuring wizard" for deep learning–based medical image segmentation. nnU-Net automatically adapted all key design choices to a given dataset, including image preprocessing, network architecture and hyperparameters, data augmentation strategies, training schedules, and postprocessing heuristics. By standardizing these components, nnU-Net provided a strong and reproducible baseline across a wide range of medical segmentation tasks. Source: [Isensee et al., 2021] . . . 36
2.4 Regional ensemble approach used by SLANT and AssemblyNet. Multiple 3D CNNs were trained on local regions of the MNI patches, and the final prediction was obtained by aggregating all predictions. Source [Huo et al., 2018] . 37
2.5 Illustration of the training sample generation process described in SynthSeg. Source: [Billot et al., 2023a]. At each step, a synthetic training image was sampled from a generative model based on the prior/available segmentation maps. The same segmentation map was used as a target for training the SynthSeg network. 39
2.6 SAM is a prompt-based segmentation model requiring both a color-based image and a user-provided prompt to obtain a segmentation. Source: [Kirillov et al., 2023]. Attempts at using SAM for medical image segmentation are achieved by fine-tuning the model on medical images (MedSAM) [Ma et al., 2024]. However, these models are limited to 2D slice-wise inference and expect RGB input, thus constraining their ability to fully exploit 3D context and native medical image contrast. 43

3.1 Illustration of the augmented training sample generation process. A randomly selected subject with a segmented lesion or tumor is registered to a randomly chosen healthy training image. The lesion or tumor is then inserted into the target image using a blending method to produce a realistic augmented example. 53
3.2 Raincloud plots showing the distribution of Dice similarity coefficient (DSC) scores for SynthStrip and FLAIR-ICE across all evaluation datasets. Each point represents a subject; violin plots illustrate score distributions. FLAIR-ICE consistently achieves higher median DSC and reduced variability compared to SynthStrip. *UCSF† corresponds to the UCSF-PDGM dataset, and is labelled this way due to images being skull-stripped as per BraTS sharing policies. Ours denotes the testing set composed of images from* VOLBRAIN.NET. 55

3.3 Qualitative examples of brain extraction on evaluation datasets. Columns (a) and (b) show the FLAIR image and the ground truth segmentation overlaid in red. Columns (c) and (d) display segmentation errors of SynthStrip and FLAIR-ICE, respectively, overlaid on the FLAIR image. False positives are shown in blue and false negatives in red. While SynthStrip often fragments or overestimates the intracranial cavity in challenging cases, FLAIR-ICE better preserves anatomical fidelity. 56

4.1 Schematic overview of the T1w-based segmentation construction. From left to right, the images are : the (1) T1w and (2) FLAIR images, the (3) lesion maps, (4) inpainted T1w and (5) the obtained fine-grained segmentation. . . 63

4.2 Illustration of the training generation process. Pairs of T1w and FLAIR images are pre-processed and processed by (1) segmenting the WMH visible in the FLAIR image, (2) performing an in-painting operation to replace WMH intensities with intensities of the surounding tissues (3) performing the fine-grained segmentation of the in-painted T1w image. 64

4.3 **Left**: Illustration of a synthetic FLAIR-like image created from T1w-based segmentation as part of the quality control process to address registration discrepancies. **Right**: Distribution of correlation scores on the training set. The vertical dotted line corresponds to the selected threshold. 65

4.4 Examples of segmentation of 132 structures obtained with FLAIRBrainSeg and SynthSeg-132 in Experiment 1 are shown on different slices and images, for the in-domain (top) and out-of-domain (bottom) testing sets. The images were selected from cases with median mean DSC scores for SynthSeg-132 to ensure a fair representation of its performance. The top row highlights the accuracy of FLAIRBrainSeg, even in the presence of MS lesions, which are incorrectly labeled as ventricle or cortex by SynthSeg. The bottom row illustrates the differences between the segmentations obtained and the ground-truth labels on another example, further emphasizing the advantages of FLAIRBrainSeg in handling modality-specific challenges. 67

4.5 Average DSC for Experiment 1 and Experiment 2. **Top:** results on the in-domain dataset (i.e., OFSEP), **Bottom:** Results on the out-of-domain dataset (i.e., UKB), **Left:** Comparison of SynthSeg-132 and FLAIRBrainSeg over the 132 structures of the Neuromorphometrics protocol (Experiment 1), **Right:** Comparison of SynthSeg-132, SynthSeg$^+$, and FLAIRBrainSeg on the 35 common structures between FreeSurfer and Neuromorphometrics protocols (Experiment 2). 68

5.1 Illustration of the preprocessing steps applied to FLAIR MRI data: (a) raw MRI acquisition; (b) noise correction; (c) intensity inhomogeneity (bias-field) correction; (d) spatial normalization via rigid registration to the MNI152 template; (e) brain-mask extraction; (f) intensity normalization and skull stripping. Please note that in this instance, registration to the MNI was performed using the associated T1w images. 77

5.2 Overview of the training image generation process. Each pre-processed FLAIR image from a cognitively normal (CN) subject is first segmented for WMH, as CN cases may still present WMH due to incidental findings or age-related small vessel disease. The CN image is then randomly paired with another FLAIR image from the FHI dataset, containing WMH or tumors. The selected pathology image is non-linearly registered to the CN image, and a gradient-based image processing method is applied to merge the pathological regions onto it. 78

5.3 Examples of FLAIR-based segmentation results across the three real testing datasets. Each row corresponds to a different dataset (InHouse, UCSF-PDGM, and UKB). Columns show, from left to right: FLAIR MRI, ground truth (GT), FreeSurfer SynthSeg-WMH segmentation, SynthSeg-FHI-133, and the segmentation obtained with the proposed method (FLAIRBrainSeg^{+}). In case of InHouse and UCSF-PDGM, GT is the corresponding segmentation provided in the dataset. In case of UKB, the GT is the T1w-based anatomical segmentation merged with the WMH identified by DLB. 89

5.4 Representative failure cases across selected OOD datasets. Each row shows a subject from BraTS-MEN, MSLesSeg, and OFSEP. Sagittal slice is shown to highlight differences in FLAIR resolution and image quality across datasets: anisotropic and noisy scans in BraTS-MEN, moderate variability in MSLesSeg, and high-quality isotropic scans in OFSEP. *Segmentation errors made by FLAIRBrainSeg^{+} are circled in yellow.* Columns show, from left to right: FLAIR MRI, ground truth (GT), FreeSurfer SynthSeg-WMH segmentation, SynthSeg-FHI-133, and the segmentation obtained with the proposed method (FLAIRBrainSeg^{+}). 91

1 Comparison on Experiment 1 of SynthSeg-132 and FLAIRBrainSeg across structure groups (e.g., tissue, cortical, subcortical, etc.) for the 132 structures in the Neuromorphometrics protocol on the in-domain test set. 101

2 Examples of obtained segmentations with FLAIRBrainSeg, SynthSeg-132, and SynthSeg^{+}. Please note that the SynthSeg^{+} segmentation includes only the 35 selected structures for comparison which do not include cortical structures and cerebellum grey matter. 102

3 Examples of training examples generated by SynthSeg-WMH. Each row shows a different example. SynthSeg-WMH generates images by sampling from a Gaussian Mixture Model (GMM) constrained by the provided label map. The GMM parameters are randomly sampled at each training iteration, resulting in a wide variety of image contrasts. However, the generated images often lack realistic anatomical details and may contain artifacts due to the simplistic nature of the GMM-based synthesis process. This can lead to suboptimal feature learning for segmentation tasks, particularly when applied to real-world FLAIR images with complex pathologies. 103

4 Comparison of segmentation errors across methods with 133 labels on UKB test set. Columns show, from left to right: FLAIR MRI, FreeSurfer SynthSeg-WMH segmentation error map, SynthSeg-FHI-133 error map, and the segmentation error map obtained with the proposed method (FLAIRBrainSeg^{+}). Error maps are computed as the absolute difference between the predicted segmentation and the T1w-based anatomical segmentation merged with the WMH identified by DLB, summed over all images in the UKB test set and projected onto the MNI152 FLAIR template for visualization. 104
5 Axial view of representative failure cases across selected OOD datasets. Each row shows a subject from BraTS-MEN, MSLesSeg, and OFSEP. The datasets differ in FLAIR resolution and image quality: anisotropic and noisy scans in BraTS-MEN, moderate variability in MSLesSeg, and high-quality isotropic scans in OFSEP. These differences are the main contributors to the observed variations in segmentation performance. 105

List of Tables

3.1 Comparison of anatomical content included in the output masks of NICE, DeepICE, SynthStrip (default FreeSurfer implementation), and HD-BET. WM: white matter; GM: gray matter; surface CSF: cerebrospinal fluid in deep sulci and the subarachnoid space. 47

3.2 Overview of FLAIR MRI datasets used for training and evaluation of the proposed intracranial extraction method. 51

3.3 Comparison of brain extraction performance between SynthStrip and FLAIR-ICE across multiple test sets. Metrics: Dice similarity coefficient (DSC), 95th percentile Hausdorff distance (HD95), average surface distance (ASD), and cross-correlation (CC). Higher DSC and CC, and lower HD95 and ASD indicate better performance. 54

4.1 Dataset overview. Please note that the number of subjects and dataset statistics are reported *after all the quality control steps.* We include the range of Expanded Disability Status Scale (EDSS) ratings to provide an idea of the stage of progression of MS in the in-domain dataset (i.e., OFSEP). . 62

4.2 Table report of Mean Dice Score and voxel-wise 95th Surface Distance (SD95) for each testing dataset on Experiment 1 69

5.1 Overview of FLAIR MRI datasets used for training, testing, and scan–rescan experiments. The table reports the number of subjects and a concise description of dataset characteristics. 84

5.2 Comparison of WMH and tumor segmentation performance between methods across the selected test sets. Metrics: Dice similarity coefficient (DSC) and 95th percentile Hausdorff distance (HD95). Higher DSC and lower HD95 indicate better performance. Statistical significance between FLAIRBrainSeg^{+} and other methods was assessed using the Wilcoxon signed-rank test with $p < 0.05$. * indicates statistically significant improvement over both SynthSeg-WMH and SynthSeg-FHI-133. 85

5.3 Comparison of brain structure segmentation performance between methods across the selected test sets. Metric: Dice similarity coefficient (DSC). Higher DSC indicates better overlap with the reference segmentation. Statistical significance between FLAIRBrainSeg^{+} and SynthSeg-FHI-133 was assessed using the Wilcoxon signed-rank test with $p < 0.05$. * indicates statistically significant improvement of FLAIRBrainSeg^{+} over SynthSeg-FHI-133. *HD95 values were not computed for structure segmentation due to the large number of regions and limited interpretability of boundary-based metrics in this context.* 87

5.4 Comparison of FLAIRBrainSeg^{+} against reported State of the Art (SoTA) methods. *Results from competing methods are reported as published in the challenge reports and* ***are not directly comparable****, since SoTA approaches use* ***multiple modalities and different data splits****. These values are shown for contextual reference only. Empty entries indicate that the information was not reported in the corresponding reference.* 87
5.5 Comparison of segmentation performance between different training configurations of FLAIRBrainSeg^{+} across multiple selected test sets. Metrics: Dice similarity coefficient (DSC). *For BraTS-MEN, InHouse and UCSF-PDGM, DSC is computed with regards to the ground-truth lesion label. For UKB, DSC is computed with regards to the T1w-based anatomical segmentation.* Statistical significance between the baseline and other variants was assessed using the Wilcoxon signed-rank test with $p < 0.05$. Statistically significant decrease in performance compared to the baseline is indicated by *. 88
5.6 Scan-rescan reliability results, measured as the mean Dice similarity coefficient (DSC) between all pairs of segmentations from the same subject acquired at different sites. Higher DSC indicates better reliability. *Given the limited number of samples in the scan-rescan dataset, no statistical tests were performed.* 90

1 Comparison of FHI insertion strategies for synthetic training data generation. For each synthetic dataset, a 3D nnU-Net was trained with identical parameters and data splits, and evaluated on the held-out UCSF-PDGM test set. Each model used 50% real images from UCSF-PDGM and 50% synthetic tumor images created with the listed blending method. Given the trade-off between segmentation performance, stability, and efficiency, Poisson blending was selected as the default FHI-merging strategy for all subsequent experiments. 106