FACULTÉ DE SCIENCES ET INGÉNIERIE

---

Master Degree's Internship Report

Master Données, Apprentissage, Connaissances (DAC)

# *Artificial intelligence assisted evaluation of post-hemorrhagic ventricular dilatation on brain MRI of preterm infants*

---

*Autor:* Garance Lucas

*Supervised by:* Isabelle Bloch & Sarah Stricker

*Laboratories:* LIP6 - Sorbonne Université
&
IMAG2 - Imagine Institute - Necker Hospital

March - August 2024

# Table of contents


**1 Introduction** **3**

**2 Medical context** **3**
2.1 Cerebral Ventricular System and Hydrocephalus . . . 3
2.2 Post-Hemorrhagic Ventricular Dilation in Premature Infants . . . 4
2.3 Neurosurgical Management . . . 4
2.4 Traditional and Computational Neuroimaging of the Premature Brain . . . 4
2.5 Image caracteristics . . . 5

**3 Related works** **6**

**4 Data** **8**
4.1 Necker Hospital dataset . . . 8
4.2 FeTA challenge dataset . . . 10
4.2.1 University Children's Hospital Zurich (Kispi) dataset . . . 10
4.2.2 General Hospital Vienna/Medical University of Vienna dataset . . . 11

**5 Segmentation methods** **12**
5.1 Data preprocessing . . . 13
5.2 U-Net based methods . . . 14
5.2.1 Architectures 2D U-Net / 3D U-Net . . . 14
5.2.2 Training caracteristics for U-Net models . . . 14
5.3 SynthSeg . . . 16
5.3.1 SynthSeg - Gererative network for synthetic training images . . . 16
5.3.2 SynthSeg - Segmentation network for brain structures segmentation . . . 18
5.3.3 Training characteristics . . . 19

**6 Results and discussion** **19**
6.1 Results of original SynthSeg on fetal and preborn MRI . . . 19
6.2 Results with FeTA 2022: best 2D U-Net model searching . . . 20
6.2.1 2D U-Net model with 1 dataloader on random MRI vs non-empty MRI . . . 20
6.2.2 2D U-Net model with 2 dataloaders vs 1 dataloader . . . 23
6.3 Results with FeTA 2024: best 3D U-Net model searching . . . 25
6.4 Results with FeTA 2024: best SynthSeg model searching . . . 27
6.5 Results with FeTA 2024 - 7 labels . . . 29
6.6 Results with FeTA 2024 - 1 label . . . 30
6.7 Results with FeTA 2024 and Necker - 1 label . . . 31

**7 Conclusion and perspectives** **32**

**8 Appendix** **34**

List of abbreviations:

CSF - Cerebrospinal fluid
MRI - Magnetic resonance imaging
PHVD - Post-hemorrhagic ventricular dilatation
IVH - Intraventricular hemorrhage
VSG - Ventriculo-subgaleal shunt
GA - Gestational age
TR - Repetition time
TE - Echo time

# 1 Introduction

Severe intraventricular hemorrhage (IVH) is a comorbidity of prematurity and can lead to post-hemorrhagic ventricular dilation (PHVD), which often necessitates ventricular shunting to mitigate secondary sequelae. The operative risks are elevated, and in addition to known clinical risk factors, radiological factors can be assessed through precise interpretation of initial and secondary cerebral lesions on magnetic resonance imaging (MRI).

The objective of this work is to achieve automatic segmentation of the ventricles in pre and post-operative MRIs of premature infants with PHVD who have undergone a ventriculo-subgaleal shunt (VSG).

The first phase has involved the creation of a MRI images with annotations. Out of 100 high-resolution MRIs from a cohort of 77 premature infants who underwent VSG for PHVD at Necker Enfants Malades Hospital between 2013 and 2024, the ventricles are manually segmented by the neurosurgeon Dr. Sarah Stricker.
The second phase involves developing three automatic segmentation models. The algorithms are trained and tested on both the publicly available fetal MRI FeTA dataset [33, 34] and the PHVD data from the Necker Enfants Malades Hospital cohort.

This work demonstrates the feasibility of automatic segmentation of the ventricles on MRI scans of premature infants with PHVD. The best model in the study achieved a Dice score of 0.844 on the ventricles segmentation with a standard deviation of 0.105. The segmentation is useful for estimating ventricular volumes. Ideally, this contributes to stratifying risk and determining surgical timing based on ventricular volume.

This project has been held under the supervision of Professor Isabelle Bloch (LIP6 - Sorbonne Université), Dr. Sarah Stricker (Necker Hospital - Paris) and the engineer team of the IMAG2 laboratory of the Imagine Institute Giammarco La Bardera, Enzo Bonnot and Thomas Isla.

# 2 Medical context

This work focuses on rare subjects with unique characteristics, specifically neonatal infants with hydrocephalus, potentially complicated by intraventricular hemorrhage. These conditions impart specific features to the images being studied. Understanding the medical aspects of these cases is crucial to fully grasp the objectives and stakes of this study, as the clinical implications of accurate segmentation and analysis of these images directly impact patient care and surgical decision-making.

## 2.1 Cerebral Ventricular System and Hydrocephalus

The cerebral ventricular system comprises two lateral ventricles, a third ventricle, and a fourth ventricle, as illustrated in Figure 41 in the appendix (8). Hydrocephalus is a condition characterized by the abnormal accumulation of cerebrospinal fluid (CSF) within the brain ventricles, leading to ventricular dilation as an indicator of increased intracranial pressure. In neonates, hydrocephalus can be caused by infection, congenital malformations detected prenatally, or hemorrhage.

## 2.2 Post-Hemorrhagic Ventricular Dilation in Premature Infants

Prematurity leads to various comorbidities, including germinal matrix bleeding, which can result in cerebral palsy and, consequently, multiple disabilities. In premature neonates, the paraventricular germinal matrix is an area with immature blood vessels, rendering it susceptible to hemorrhage. The severity of germinal matrix bleeding, classified according to Papile, depends on the extent of bleeding into the cerebral parenchyma and ventricles [32] (Figure 41 in the appendix (8)). In the case of intraventricular hemorrhage, this can result in secondary hydrocephalus, which leads to additional cerebral damage. In developed countries, approximately 38 premature neonates per 100,000 live births [10] will develop post-hemorrhagic ventricular dilation, with 50% of high-grade IVHs (grades 3 and 4) leading to PHVD [6]. The etiology of PHVD is multifactorial, including obstruction of CSF circulation due to blood and limited CSF absorption due to post-hemorrhagic inflammatory reaction [13, 37].

## 2.3 Neurosurgical Management

The surgical treatment of secondary hydrocephalus aims to mitigate further cerebral damage. Standard surgical methods involve the implantation of permanent or temporary ventricular drainage devices to evacuate excess CSF, such as a VSG (Figures 42 and 43 in the appendix (8)). Advances in neurosurgical, obstetrical, and intensive care technologies have increased the number of patients and therapeutic options [1, 11, 38, 25, 2, 45, 46]. However, surgery remains challenging due to the high risk of complications associated with low body weight, intraventricular hemorrhage, and neonatal comorbidities. While early interventions to promote brain development in premature infants are crucial for improving long-term outcomes, rigorous patient selection is essential [35, 13, 29, 23]. In addition to clinical evaluation, magnetic resonance imaging is employed to assess the extent of associated cerebral lesions and to guide patient selection.

## 2.4 Traditional and Computational Neuroimaging of the Premature Brain

Neuroimaging of the premature brain includes ultrasound for the diagnosis and monitoring of PHVD [13, 18]. Magnetic resonance imaging is employed to detect brain lesions associated with PHVD and assess the impact on brain development [18], considering that the brain at this age is not yet myelinated and maturation is incomplete. Prenatal brain atlases with annotated structures illustrate stages of brain growth and organization according to gestational age (GA) [9, 15, 43, 16, 12]. These atlases are used to evaluate developmental brain pathologies on MRI in premature infants, as prematurity can affect developmental processes. Atlases also provide a reference for brain structuring on MRI, varying with gestational age, for the annotation of structures in MRI scans.

Magnetic resonance imaging uses strong magnetic fields and radio waves to generate detailed images of the body internal structures, leveraging the abundance of hydrogen atoms in water, which makes up a significant portion of the human body. When exposed to a strong magnetic field, these hydrogen atoms align with the field. A brief pulse of radio waves then disrupts this alignment. After the pulse stops, the hydrogen atoms gradually realign, releasing energy that the MRI machine detects to create images of the body tissues. Different imaging types, such as T1 and T2, arise from measuring the hydrogen atoms' return to alignment in distinct ways. T1 measures the time it takes for hydrogen atoms to realign with the magnetic field, producing images where fat tissues are more visible. T2 measures the time for hydrogen atoms to lose the energy from the radio wave pulse and cease interacting, highlighting fluids in the images.

## 2.5 Image caracteristics

The medical images are constituted of voxels ($mm^3$). In the context of medical imaging, particularly in MRI of the brain, a voxel, or volumetric pixel, represents a discrete, three-dimensional unit of spatial resolution within the scanned volume. Each voxel corresponds to a specific tissue volume, characterized by its MRI signal intensity values, which reflect the relaxation characteristics of hydrogen protons in a tissue within the external magnetic field (typical field strength are 1.5T and 3T). The intensity value is defined as the mean value within the voxel, lies between 0 and the maximum defined during the acquisition protocol.

By adjusting specific MRI acquisition parameters, such as repetition time (TR) and echo time (TE), different signal responses can be obtained, resulting in various imaging sequences. The most commonly used sequences are T1-weighted (see Figure 1 left side) and T2-weighted (see Figure 1 right side) sequences. T1-weighted MRI sequences are used to demonstrate the morphological structuring of the brain, while T2-weighted MRI sequences are employed to visualize CSF pathways.

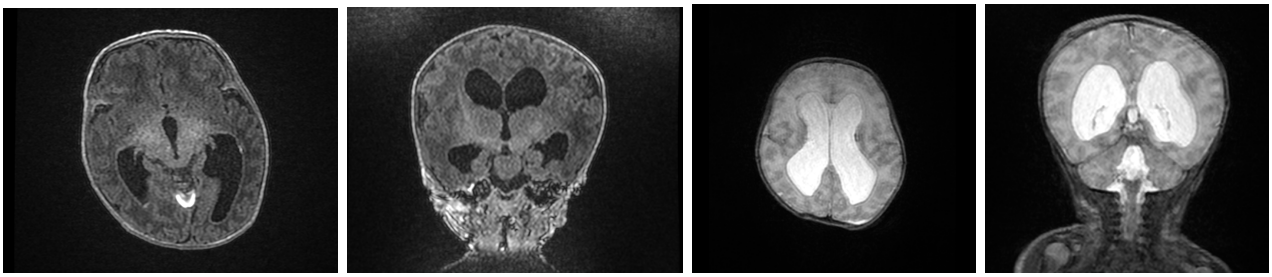

Figure 1: Left side: T1-weighted MRI - axial and coronal planes
Right side: T2-weighted MRI - axial and coronal planes

The spatial dimensions are defined by the MRI machine maker and the acquisition protocol.

The pixel matrix (mm) is the in-plan resolution of the image slice. The resolution of the third dimension is given by the slice thickness (mm) and slice spacing (mm).

Slice spacing refers to the distance between two consecutive image slices along the body. This distance is measured from the center of one slice to the center of the next slice. The finer the slice spacing is, the higher the resolution of the resulting images will be, allowing for more detailed and accurate visualization of anatomical structures. Conversely, larger slice spacing can result in faster image acquisition but may compromise the level of detail and potentially miss small lesions or subtle abnormalities. Therefore, the choice of slice spacing involves a balance between the need for image detail and the practical considerations of scanning time and patient comfort.
The multiplanar reconstruction is possible with methods such as interpolation that fill the gaps caused by slice spacing. The voxel size of the 3D image results from this reconstruction.

Voxel size refers to the three-dimensional unit of measurement within a reconstructed MRI image, determined by the product of the pixel dimensions and the slice thickness. Smaller voxel sizes enable higher spatial resolution, allowing for the detection of finer details in the scanned tissues. However, reducing the voxel size typically increases the scan time and the amount of data to process.

An MRI acquired in three dimensions is characterized by three planes: coronal, sagittal, and axial. The coronal plane divides the body into front (anterior) and back (posterior) sections, providing a view from the front to the back. The sagittal plane divides the body into left and right sections, offering a side view from either the left or right. The axial plane, also known as the transverse plane, divides the body into upper (superior) and lower (inferior) sections, giving a horizontal view from top to bottom.

A notable characteristic of fetal brain MRI is the significant heterogeneity within the dataset. This variability arises from the rapid development of brain structures in newborns, leading to considerable differences in the images as it can be seen in Figure 2 of spatiotemporal fetal brain MRI atlas [16]. Additionally, the natural movement of the fetus during image acquisition can introduce artifacts, which further complicate the analysis and increase the variability in image quality. Consequently, depending on the acquisition protocol and examination conditions, the resulting MRIs can vary widely in their 'quality,' posing additional challenges for accurate segmentation and analysis (see Figure 2).

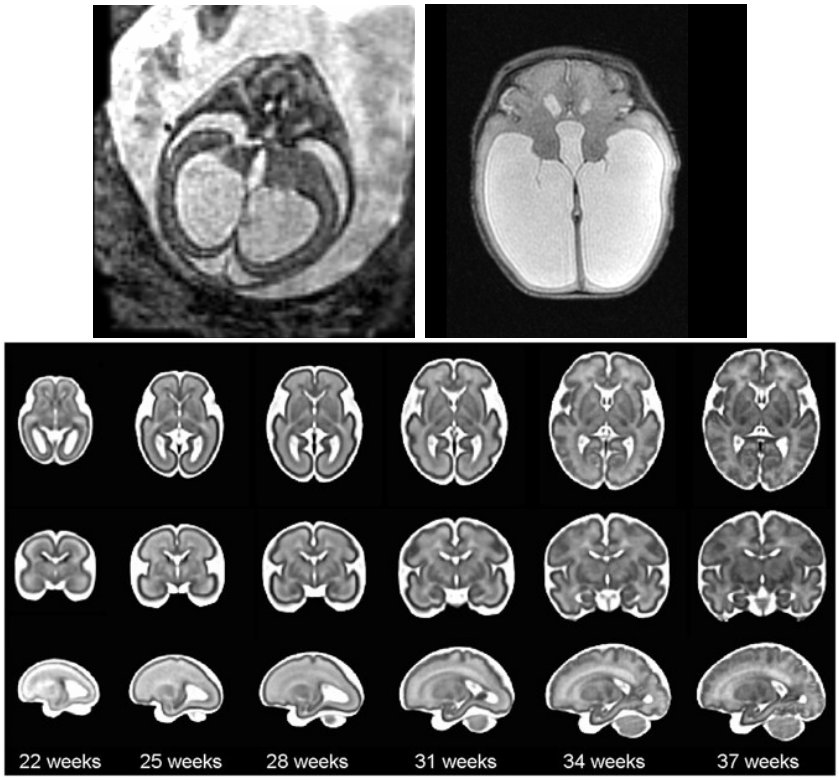


Figure 2: Above: "Bad quality" MRI and "good quality" MRI
Below: Spatiotemporal fetal brain MRI atlas (CRL fetal brain atlas) at six representative gestational ages: 22, 25, 28, 31, 34, and 37 weeks [16]

# 3 Related works

The automated segmentation of neonatal or fetal brain MRI is a challenging task due to significant changes in the overall brain structure and considerable variations in image intensity, which reflect the rapid tissue maturation that occurs around birth [22]. To address these challenges, several segmentation techniques have been developed, such as multi-atlas label fusion methods, specifically designed to enhance robustness against these factors [3, 15, 24, 26, 27]. These methods have been applied to large open datasets, including the developing Human Connectome Project (dHCP) [28].

Recently, supervised deep-learning techniques have emerged as the next generation of segmentation methods for medical images, demonstrating higher performance and reduced computation time compared to earlier approaches. The U-Net architecture, in particular, has surpassed previous methods in numerous challenges [19, 36]. However, a well-recognized drawback of supervised learning methods is their significant drop in performance when applied to unseen data, a problem known as the "domain gap" [21]. This issue has been identified as a

major obstacle in the field [31, 48].
One common strategy to address this problem involves augmenting the training set with synthetic perturbations that explicitly manage deviations from the original training data [17]. This approach has the clear advantage of avoiding the costly process of acquiring additional training data. In the context of brain development, data augmentation can be divided into two key aspects that correspond to the major challenges mentioned earlier: 1) spatial augmentation, which accounts for variations in the spatial arrangement of different tissues and the shape of specific anatomical structures (such as the increase in cortical folding with age); and 2) style (or appearance) augmentation, which compensates for changes in tissue contrast that may result from differences in acquisition settings, scanner variations, or brain maturation.
The main disadvantage of these methods is the need to train a new model for each new domain.

Recent advancements in domain adaptation techniques have concentrated on developing unsupervised learning methods that generate realistic synthetic training sets without the need for manually labeled data in the target external domain. Some of these techniques involve learning a shared latent space between the original domain, where ground truth labels are available, and the target external domain [14, 20, 40]. Others focus on direct image-to-image translation [47]. Chen et al. (2019) [8] combine these two approaches in their work. Additionally, the use of adversarial generative models for domain adaptation has been explored by Chartsias et al. (2018) [7].

One of the most promising synthetic approaches is the recent work by Billot et al., who introduced a method called SynthSeg [4, 5] that eliminates the need for real MRI data during the training process. The key innovation in this approach is the avoidance of potential bias towards the training set domain by employing a framework that enables model training without any real imaging data. Instead, a fully synthetic training dataset is generated from a set of real label maps. Assuming tissue homogeneity, the image signal for each tissue is sampled from a Gaussian distribution with distinct mean and variance, reflecting the fact that MRI data typically consist of Gaussian intensity mixtures. This generated signal is then enhanced with additional common random transformations, such as bias field distortions, Gaussian noise, and spatial deformations. Billot et al.'s approach [4, 5] is rooted in the concept of domain randomization [39, 42], which posits that the variations observed in real data across different domains should be captured within the distribution of the generated synthetic data.
In Billot et al. [5], the authors validated their approach on highly heterogeneous data obtained from adults using various clinical settings, showcasing remarkable robustness to challenging variations in image contrast and resolution. They reported superior segmentation accuracy and resilience compared to other domain adaptation methods. Furthermore, the study explored the impact of training set size on SynthSeg performance, finding that only a small number of training examples were needed for the model to achieve its maximum accuracy on an adult population.

Valabregue et al. (2023) [44] recently compared the performance of a standard U-Net model trained on T2-weighted images from 583 infants. The study population was born at term equivalent age (37–44 weeks post-menstrual age) without any known pregnancy or neonatal problems and are regarded as healthy. They also evaluated the synthetic learning approach using the SynthSeg model by Billot et al. [5] and confirmed its robustness to variations in image contrast. However, they observed that the SynthSeg model performance on neonatal brain segmentation was not as strong as expected, particularly noting a clear influence of the infant's age on the model predictions.

From this discussion, while U-Net and SynthSeg approaches are promising, no study has dealt with new borns with pathologies. This raises specific problem, that we want to address. This work aims to compare different models for segmenting neonatal brains from preborn babies with hydrocephalus, with a particular focus on segmenting the ventricles. Accurate segmentation of the ventricles is crucial, as their volume is a key factor in the medical decision-making.
The study compares three models: 2D U-Net, 3D U-Net [36], and SynthSeg [5]. The U-Net architecture is still considered a state-of-the-art method in medical image segmentation [19, 36]. On the other hand, SynthSeg represents a very recent and promising approach, offering the ability to segment various types of sequence scan images, in particular T1 and T2 MRI. Various implementations of each model are tested, with the comparison based on the performance of the best version of each. The models are evaluated on three different datasets of fetal or neonatal brain MRI from three hospitals, with only one dataset containing images of subjects with severe hydrocephalus, potentially complicated by stroke.

# 4 Data

This work utilizes three distinct datasets. One of these datasets comprises images of preborn infants from Necker Hospital, which is central to this project. The other two datasets are publicly available and contain images of fetal MRI. They are incorporated to enhance the learning process of the different models under study and to provide more robust test sets for evaluating their performances.

## 4.1 Necker Hospital dataset

The data collected from Necker Hospital were acquired between 2013 and 2024 and are derived from a cohort of preborn infants with ventriculo-subgaleal shunts, which are the particularities of the Necker dataset: all the subjects are preborn infant and have hydrocephalus pathology, sometimes with hemorragy. The Necker dataset includes 40 subjects. Some subjects have undergone a surgical operation and thus have pre-operative MRI and post-operative MRI. Due to the significant differences between pre- and post-operation images for the same patient (as illustrated in Figure 3 - pre- and post-operation T1 of a patient on the left and pre- and post-operation T2 of another patient on the right), these images are treated as representing distinct subjects in this analysis. Therefore, the dataset effectively includes 60 'relative patients'.

The mean gestational age at birth of the subjects is 31.5 weeks and a standard deviation of 5 weeks, with a minimum of 24 weeks and 5 days, and a maximum of 41 weeks and 5 days.

The images have been acquired with a pixel size between $0.35mm \times 0.35mm$ and $0.66mm \times 0.66mm$, a slice thickness from 1 to 1.2mm and a slice spacing from 0.45mm to 1.2mm. The volume size of the images varies.

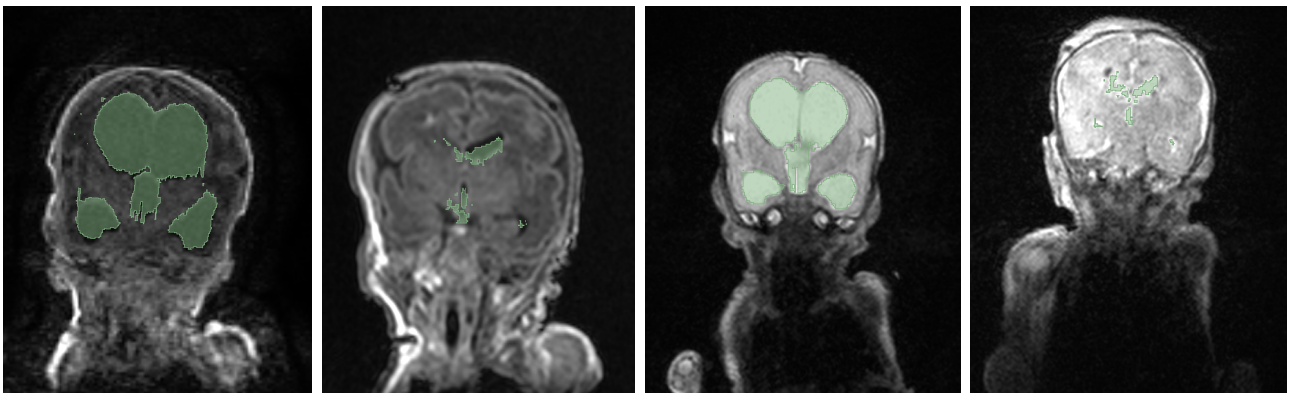

Figure 3: pre- and post-operation T1-weighted and T2-weighted MRI from Necker Hospital dataset with segmentation - coronal plane

The distribution of the type of sequence and the moment of the image acquisition (pre-operation or post-operation) are presented in Figure 4.

| Necker dataset | Pre-surgery only | Post-surgery only | Pre & post-surgery | Total |
|---|---|---|---|---|
| **T1 only** | 12 | 3 | 0 | 15 |
| **T2 only** | 5 | 19 | 6 (12) | 36 |
| **T1 & T2** | 6 (12) | 6 (12) | 6 (24) | 48 |
| **Total** | 29 | 34 | 36 | 99 |

Figure 4: Table of MRI distribution in the Necker dataset

The Necker dataset comprises a total of 99 MRI scans, with 47 pre-operation and 52 post-operation images. Among these, there are 39 T1-weighted images and 60 T2-weighted images, all manually segmented by the doctor Sarah Stricker.

The dataset includes 6 subjects with both pre- and post-operation exams for T1 and T2 sequences. Given that pre- and post-operation images differ significantly, they are treated as representing 12 distinct subjects (6 subjects $\times$ 2 imaging sessions). Those MRI are used for testing the volumetry and the segmentation performances on the Necker data.

The T2-only, T2 in pre-surgery only (of T1&T2) and T2 in post-surgery only (of T1&T2) give a total of 48 T2-weighted images and are used as train and validation set.

T1 in pre-operation only (of T1&T2) and the T1 in post-operation only (of T1&T2) give a total of 12 T1-weighted images. Those images could be used together with the previous 12 complementary T2, to fine-tune a model which takes as input both sequences of an image. This could be useful in order to compare the result with the SynthSeg model (presented below), trained with artificially generated images in order to be able to segment all kinds of sequences. This approach enables a comparison with the SynthSeg model, potentially providing insights into whether it is advantageous to consistently acquire both T1 and T2 sequences during imaging.

T1-only images are not directly utilized in this study because the networks under review are trained exclusively with T2 sequences. Additionally, there is a sufficient number of images with both T1 and T2 sequences to evaluate network performance. Nonetheless, the presence of T1-only scans in the dataset highlights the need for models capable of segmenting T1 sequences, as T1-only acquisitions can occasionally occur. However, since 2018, in France, MRI exams are usually conducted using both T1 and T2 acquisitions, rather than relying on just one sequence.

In the Necker dataset, the manual segmentations associated with the MRI scans include two labels: ventricles and intraventricular hemorrhage. Some subjects exhibit intraventricular hemorrhage, as shown in Figure 5 (left images), while others do not, as illustrated in the last

two images of the same figure (right images).

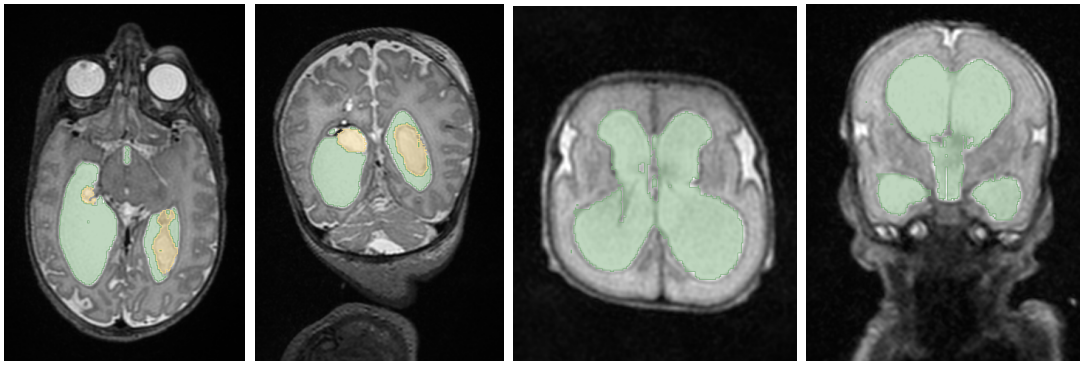

Figure 5: T2-weighted MRI with and without stroke from Necker Hospital dataset with segmentation - axial and coronal planes

Initially, the goal was to extend segmentation beyond just the ventricles to include other brain structures. This approach aimed to facilitate the training or at least testing of models, particularly SynthSeg, for segmenting a broader range of structures. The objective was to determine whether learning to segment structures surrounding the ventricles could enhance the accuracy of ventricles segmentation, especially on Necker images, and if so, quantify the extent of this improvement. However, due to the time-consuming and challenging nature of segmenting these structures manually according to the medical doctors (see Figure 6 with T1 and T2 slices examples), just ventricles and intraventricular hemorrage has been segmented to all the patients.

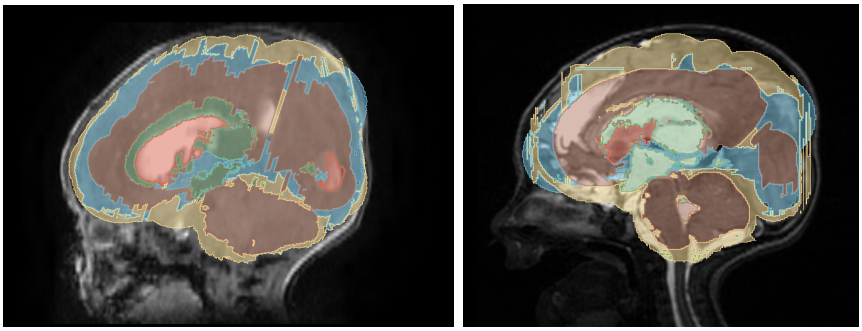

Figure 6: T1 and T2-weighted MRI from Necker Hospital dataset with five labels segmentation - sagittal plane

It is important to note that, because segmenting the images was highly time-consuming for the doctor, and additionally, due to the need for extensive discussions and coordination with medical collaborators, the definitive Necker dataset was finalized and acquired on August 23, 2024, which have been a real constrain in the study.

## 4.2 FeTA challenge dataset

### 4.2.1 University Children's Hospital Zurich (Kispi) dataset

**Data acquisition:**

The data collected from the University Children's Hospital utilized 1.5T and 3T clinical GE whole-body scanners. T2-weighted single shot Fast Spin Echo sequences were captured, featuring an in-plane resolution of $0.5mm \times 0.5mm$ and slice thickness ranging from 3 to 5 mm. The sequence parameters included TR values between 2000-3500ms, a minimum TE of 120ms, a flip angle of 90°, and a sampling percentage of 55%. The imaging plane was oriented relative to the fetal brain, and axial, coronal, and sagittal images were obtained.

**Post-processing:**

For each subject, the acquired fetal brain images were manually reviewed for quality to compile a stack of images. Each stack included at least one brain scan per orientation, with more scans incorporated when available, resulting in 3 to 13 scans per stack. For subjects sub-001 to sub-040, each image in the stack was reoriented to a standard plane, and a mask of the fetal brain was created using a semi-automated atlas-based custom MeVisLab (MeVis Medical Solutions AG, Bremen, Germany) [41]. A super-resolution reconstruction algorithm was then applied to each subject's stack of images and brain masks, producing a 3D super-resolution volume of brain morphology with an isotropic size of $0.5mm^3$. Each image underwent histogram matching using Slicer and was zero-padded to dimensions of $256 \times 256 \times 256$ voxels.

**Data descriptions:**

The University Children's Hospital Zurich (Kispi) dataset comprises imaging data from 80 subjects, each with a single 3D T2-weighted brain MRI scan. These MRI scans have a size of $256 \times 256 \times 256$ voxels. Although the voxel spacing varies between subjects, each MRI ensures isotropic voxel dimensions (e.g., $0.5mm \times 0.5mm \times 0.5mm$).

For every subject, the dataset includes both the brain MRI and a corresponding segmentation file. The segmentation file serves as a mask that aligns with the associated MRI, delineating seven distinct regions within the brain, as shown in the Figures 7 and 8:

1. External Cerebrospinal Fluid (pale green label on the images of Figures 7 and 8)
2. Grey Matter (yellow label on the images of Figures 7 and 8)
3. White Matter (brown label on the images of Figures 7 and 8)
4. Ventricles (blue label on the images of Figures 7 and 8)
5. Cerebellum (neon red label on the images of Figures 7 and 8)
6. Deep Grey Matter (red label on the images of Figures 7 and 8)
7. Brainstem (neon green label on the images of Figures 7 and 8)

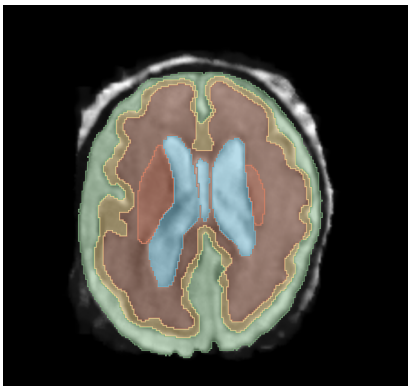

Figure 7: T2-weighted MRI from Zurich Hospital dataset with segmentation - axial plane

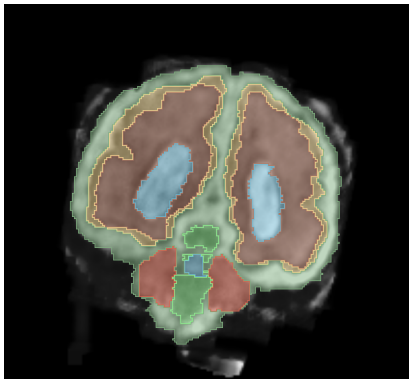

Figure 8: T2-weighted MRI from Zurich Hospital dataset with segmentation - coronal plane

### 4.2.2 General Hospital Vienna/Medical University of Vienna dataset

**Data acquisition:**

The data collected from the University of Vienna included 40 cases, acquired using 1.5 T and 3T magnets without any maternal or fetal sedation. For each case, a minimum of three T2-weighted single-shot fast spin echo (ssFSE) sequences (TE=80-140ms) were obtained in three orthogonal planes (axial, coronal, sagittal) aligned with the fetal brain stem axis and/or the axis of the corpus callosum, using a 1.5 Tesla Philips Intera MR scanner. The slice thickness ranged from 3mm to 5mm with a gap of 0.3-1mm, pixel size varied between 0.65mm and

1.17mm, and acquisition times ranged from 13.46 to 41.19 seconds.

**Post-processing:**

The post-processing pipeline involves several steps, starting with data denoising, followed by in-plane super resolution, and automatic brain masking. This process culminates in a single 0.5mm isotropic slice-wise motion correction and volumetric super-resolution reconstruction. The final volumes are then rigidly aligned to a common reference space.

**Data descriptions:**

The General Hospital Vienna/Medical University of Vienna dataset comprises imaging data from 40 subjects, each with a single 3D T2-weighted brain MRI scan. The images of the dataset have the same properties as the University Children's Hospital Zurich dataset, and the same segmentation labels.

One of the main differences with the MRI from the University Children's Hospital and the ones from Necker Hospital is that the MRI of the Medical University of Vienna are fetal MRI. This means that the imaging captures the infant while still in the womb (see Figure 9).

For the Zurich Hospital dataset, post-processing was done to isolate the baby's brain in the images. However, in the initial Vienna dataset, no such post-processing was performed, so the images also included the mother's organs (as shown in Figure 9 - images 1 and 3). To address this, a mask was applied to exclude voxels not associated with a segmentation label, thereby isolating the fetal brain in the images (see Figure 9 - images 2 and 4).

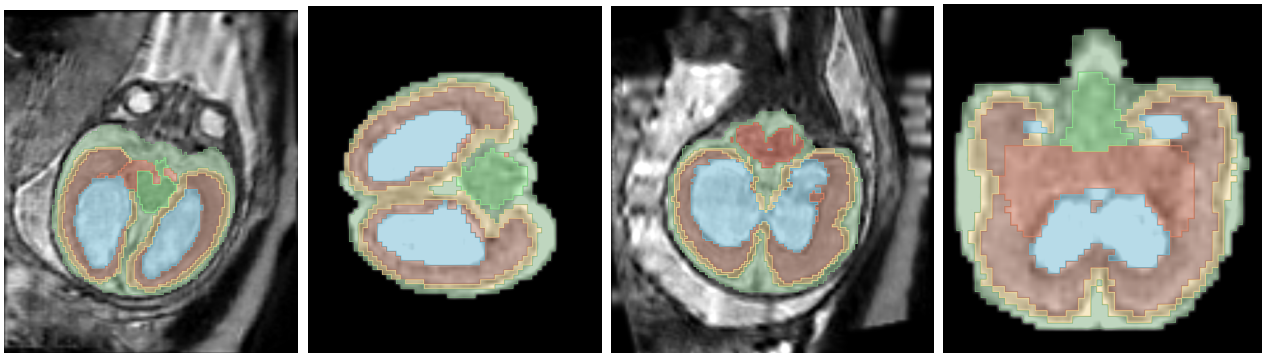

Figure 9: On the left: T2-weighted MRI from the initial Vienna dataset with segmentation and from the modified Vienna dataset (not the same subject) - axial plane
On the right: T2-weighted MRI from the initial Vienna dataset with segmentation and from the modified Vienna dataset (not the same subject) - coronal plane

It is important to note that due to administrative procedures for data acquisition, the Hospital Vienna dataset was obtained later in the study, specifically on July 23, 2024.

# 5 Segmentation methods

In order to perform the ventricles segmentation task over the cohort of Necker patients, various models were examined. Three distinct models were selected: the U-Net model [36], chosen for its widespread popularity for medical image segmentation, and the SynthSeg model [5], selected for its promising approach. These models were trained and tested across three different datasets to compare their performance. The calculation of the ventricles volume was derived directly from the segmentation results.

## 5.1 Data preprocessing

A pre-processing step is applied consistently across all datasets, with the same procedure used for both FeTA (Zurich and Vienna datasets) and Necker data. Two pre-processing are defined, as for the 3D models (3D U-Net and SynthSeg) the training input data need to be standardized to the same volume size.

**Data preprocessing for 3D models** The data preprocessing for 3D models is the following:

1. **Noise reduction:** A median filter was applied in order to reduce the noise in the image (due to MRI acquisition or patient's movements). The median filter was applied by moving a window of a specified size across the image and replacing each voxel value with the median value of the voxel within the window. The median value is the middle value when all the voxel values in the window are sorted in ascending order, which means it is less affected by extreme values or outliers compared to the mean. This process effectively removes salt-and-pepper noise, which appears as randomly distributed white and black voxels, without blurring the edges of objects in the image.
   The size of the filter is an hyperparameter. After testing different values for the filter size, the size of $3 \times 3 \times 3$ (voxels) has been used.

2. **Cropping:** A cropping of the 3D image was applied in order to reduce the background and focus better on the structures. The cropping has been done retrieving first the indices of voxels higher than a defined threshold (after few tests, the value of the threshold took on was 200). In this list of indices, the max and min indices for each dimension were retrieved. Then, the MRI was cropped based on the min and max indices. The associated segmentation is cropped on the same indices in the same dimension as the MRI.

3. **Normalisation:** The MRI is normalised in a range between 0 and 1. That helps handling the fact that the voxels values range differs in function of the acquisition machine.

4. **Padding:** After cropping the MRI, a padding is realised in order to get all the three dimensions of the same size. The dimensions with the smallest sizes are padded (from both sides in order to have a centered image) to the size of the other dimension. This operation is useful for the resampling.

5. **Resampling:** The MRI scans are resampled to a standard volume size of $256 \times 256 \times 256$ by applying a zoom. This zoom uses linear interpolation for the images and nearest neighbor interpolation for the segmentations, ensuring that the segmentation labels remain as integers. Both interpolation methods are reflection-based, meaning they reflect the image data at its borders. This approach minimizes artifacts and produces smoother results by considering mirrored data points around the edges.

In this work, voxel size is not standardized across the dataset. Given that the images are from infants of varying ages and sizes, enforcing a uniform voxel size would not be beneficial. Instead, it is preferable to maximize the brain occupancy within the $256 \times 256$ matrix by adjusting the zoom factor accordingly—greater than one for smaller brains and less than one for larger brains. This approach should ensure that the anatomical structures occupy a similar number of voxels, aside from differences attributable to the pathology.
The range 0-1 obtained with the normalization is kept.

**Data preprocessing for 2D models** The data preprocessing for 2D models is the following:

1. **Noise reduction:** The noise reduction operation was the same as the one for 3D models.
2. **Cropping:** As for the cropping operation for the 3D models, for training the 2D models the MRI was cropped based on min and max indices, except the third dimension (axial axis) that has not been modified in order to not loose information in the acquisition direction and due to the variable number of slices in the Necker dataset.
   The segmentation corresponding to the MRI is cropped in the same dimension as the MRI (except on the third dimension).
3. **Normalisation:** The normalisation operation was the same as the one for 3D models.
4. **Padding:** The padding operation was the same as the one for 3D models. The third dimension is not padded.
5. **Resampling:** The resampling operation was the same as the one for 3D models. The third dimension is not resampled.
6. **3D to 2D slices:** In order to get 2D data, the 3D MRI have been cut into slices. For each MRI, an iteration is done through each axial slice in the 3D volume as the majority of the images has been acquired along this axis, and each 2D slices are extracted from both MRI and the associated segmentation. For training and validation sets, only one 2D slice every three slices is kept, in order to avoid having too similar slices in the datasets.

## 5.2 U-Net based methods

### 5.2.1 Architectures 2D U-Net / 3D U-Net

The U-Net architecture [36] is tailored for medical image segmentation, employing a symmetric encoder-decoder structure with skip connections. The encoder serves to capture contextual information by progressively downsampling the input and extracting feature representations, and the decoder to reconstruct the spatial resolution of the image by upsampling and refining these features to produce a segmented output. The skip connection allows the direct transfer of high-resolution feature maps from the encoder to the decoder, preserving spatial information. For both 3D and 2D U-Net models, the encoder path consists of five levels, each beginning with a convolutional block. At each level, a convolutional block applies two consecutive $3 \times 3$ convolutions with ReLU activations and batch normalization, progressively increasing the number of feature channels from 64 to 1024. After each convolutional block, a max-pooling operation reduces the spatial resolution by a factor of 2.
In the decoder path, the model utilizes upsampling layers followed by convolutional blocks to progressively restore the spatial resolution of the feature maps. Each upsampling operation is paired with a concatenation of the corresponding feature maps from the encoder path, followed by a convolutional block that refines the concatenated features. This process is repeated across four levels, reducing the number of feature channels from 1024 to 64. The final layer consists of a $1 \times 1$ convolution, which maps the feature channels to the desired output channels, followed by a softmax activation function to produce the final probabilistic segmentation map.

### 5.2.2 Training caracteristics for U-Net models

In order to be able to compare as exact as possible the three models, the learning rate, optimizer, loss function and evaluation measures used were shared by the three models. The values of the

hyperparameters were set mostly according to the ones recommended for SynthSeg (presented in 5.3), which is composed by a generative model and a 3D U-Net. SynthSeg hyperparameters have been optimised by the authors, and can be considered as reference values.

The learning rate was set at $10^{-4}$ and the optimizer is the Adam optimiser.

In the 2D and 3D U-Net models, the loss function is a combination of CrossEntropy and Dice index. The CrossEntropy quantifies the difference between the predicted probability distribution and the true class distribution for each voxel. The Dice index measures the overlap between predicted and true segmentations, providing a metric that reflects the model's performance on the global structure of the segmented regions.
The CrossEntropy belongs to the voxel-wise family of loss functions, which means they evaluate the segmentation quality at the individual voxel level. This loss function is effective in guiding the model towards voxel-level accuracy. By contrast, the Dice index provides a more holistic assessment by evaluating the overall agreement between predicted and true segmentations.
Combining these loss functions, particularly at the beginning of the training process, offers several advantages as demonstrated by Isensee and al. [19]. The CrossEntropy loss serves as regularizer by focusing on voxel-wise accuracy, which is particularly beneficial for background regions (easily distinguishable). This regularization helps the model to stabilize and converge more quickly. By constraining the search area for the Dice index, this voxel-wise loss ensure that the model does not overfit to the details of the background, allowing the Dice loss function to focus on more challenging areas and thereby enhancing the model's ability to learn and generalize effectively.
The use of a combined loss function in the U-Net models differs from the training approach used in the SynthSeg model. In SynthSeg, two different losses are applied sequentially—one during pre-training (weighted L2 loss function as voxel-wise regularizer) and another during training (Dice loss function as global-wise one)—rather than being combined. Conversely, in the U-Net models, both losses are integrated and applied together throughout the entire training process.

To be able to evaluate the models and track the training evolution, the Dice score has been used.

$$\text{Dice Score} = \frac{2 \times |A \cap B|}{|A| + |B|}$$

Where: $A$ is the set of elements in the predicted segmentation, $B$ is the set of elements in the ground truth segmentation, $|A \cap B|$ is the number of elements common to both sets (the intersection). The Dice loss function used is defined as: Dice Loss = 1 - Dice Score.

**Training caracteristics for 3D U-Net model** For the 3D U-Net model, patches were extracted from the 3D images for training. A patch size of $(128, 128, 128)$ was selected to ensure consistency with the patch sizes used in the SynthSeg model. A stride of 32 was employed to balance the number of patches and memory constraints. During inference, a stride of 64 was used, making the training stride of 32 beneficial for improving the model robustness to image translation.

From the training dataset, 125 patches per image were extracted, resulting in a total of $100 \times 125 = 12,500$ patches for training on FeTA and $148 \times 125 = 18,500$ patches for training on FeTA and Necker datasets. Each epoch utilized 500 randomly selected patches, ensuring that the model did not see the same patches in every epoch and thereby enhancing its robustness.

## 5.3 SynthSeg

Proposed by B. Billot et al. [5], SynthSeg is a segmentation-task model that is enable to segment different brain structures on MRI scans, and that pretends to be robust against changes in contrast and resolution. The main specificity of this approach is that SynthSeg is trained on synthetic data sampled from a generative model conditioned on segmentations. This method allows the guarantees of the model.
SynthSeg relies on two networks: a generation part and a segmentation part.

The image dimensions for training the model need to be cubic and divisible by 2, but the dimensions for inference do not need to match the training images, as cropping or resizing is applied within the model implementation in inference.

### 5.3.1 SynthSeg - Gererative network for synthetic training images

The goal of the generative part of SynthSeg is to generate synthetic MRI from segmentation, in order to train a segmentation network (the segmentation part of SynthSeg).
Here is a description of the generative model, which is illustrated in Figure 10.

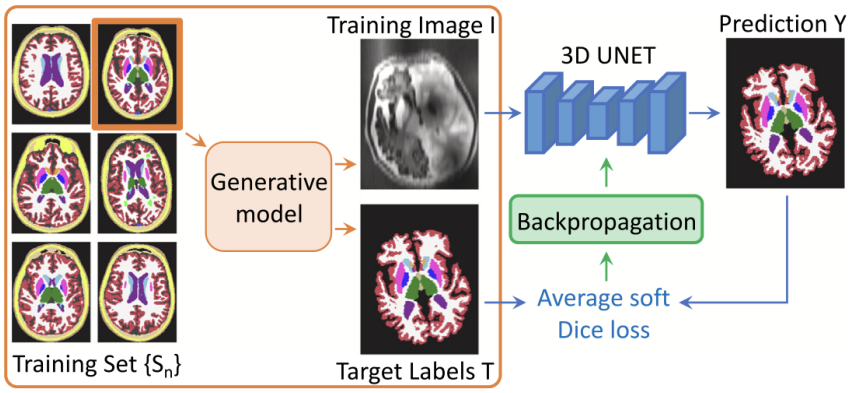


Figure 10: Overview of a training step of the SynthSeg model

**Label map selection and spatial augmentation** The generative model assumes the existence of $N$ training label maps $S_{n\,n=1}^{N}$ defined over discrete spatial coordinates $(x, y, z)$. Each label map $(x, y, z)$ is assumed to take values from a set of $K$ labels, i.e., $S_n(x, y, z) \in 1, \ldots, K$.

The generative process initiates by selecting a segmentation $S_i$ from the training dataset at random. To introduce greater variability among the segmentations, $S_i$ is subjected to a random spatial transformation $\phi$, which is the composition of an affine and non-linear transformations.

The affine transformation $\phi_{\text{aff}}$ is the composition of three rotations $(\theta_x, \theta_y, \theta_z)$, three scalings $(s_x, s_y, s_z)$, three shearings $(sh_x, sh_y, sh_z)$, and three translations $(t_x, t_y, t_z)$. The parameters for these transformations are sampled from uniform distributions:

$$\begin{aligned}
\theta_x, \theta_y, \theta_z &\sim \mathcal{U}(a_{rot}, b_{rot}),\\
s_x, s_y, s_z &\sim \mathcal{U}(a_{sc}, b_{sc}),\\
s_h, s_h, s_h &\sim \mathcal{U}(a, b),\\
t_x, t_y, t_z &\sim \mathcal{U}(a_{tr}, b_{tr}),\\
\phi_{\text{aff}} &= \text{Aff}(\theta_x, \theta_y, \theta_z, s_x, s_y, s_z, s_{hx}, s_{hy}, s_{hz}, t_x, t_y, t_z),
\end{aligned}$$

where, $a_{\text{rot}}, b_{\text{rot}}, a_{\text{sc}}, b_{\text{sc}}, a_{sh}, b_{sh}, a_{tr}$, and $b_{tr}$ denote the predefined bounds for the uniform distributions from which the parameters are sampled. The notation $\text{Aff}(\cdot)$ represents the composition of the aforementioned affine transformations.

The non-linear component $\phi_{\text{nonlin}}$ is a diffeomorphic transformation, obtained as follows: first, a small vector field of dimensions $10 \times 10 \times 10 \times 3$ is sampled from a zero-mean Gaussian distribution with a standard deviation $\sigma_{SVF}$, where $\sigma_{SVF}$ is drawn from the uniform distribution $\mathcal{U}(0, b_{\text{nonlin}})$. This vector field is then upscaled to the full image size using trilinear interpolation to produce a stationary velocity field (SVF). The SVF is subsequently integrated using a scale-and-square method (Arsigny et al., 2006) [citation] to generate a diffeomorphic deformation field, ensuring that the resulting transformation is free of holes and folds:

$$\begin{aligned} \sigma_{\text{SVF}} &\sim \mathcal{U}(0, b_{\text{nonlin}}), \\ \text{SVF}' &\sim \mathcal{N}_{10\times10\times10\times3}(0, \sigma_{\text{SVF}}), \\ \text{SVF} &= \text{Resample}(\text{SVF}; r_H R), \\ \phi_{\text{nonlin}} &= \text{Integrate}(\text{SVF}), \end{aligned}$$

Finally, the augmented map $L$ is generated by applying the spatial transform $\phi$ to $S_i$ through nearest-neighbour interpolation:

$$L = S_i \circ \phi = S_i \circ (\phi_{\text{aff}} \circ \phi_{\text{nonlin}})$$

**Initial high resolution synthetic image** Following the deformation of the input segmentation, an initial synthetic scan $G$ at high resolution (HR) is generated by sampling from a Gaussian Mixture Model (GMM) conditioned on the augmented map $L$. For convenience, the means and standard deviations of the GMM are grouped into $\mathcal{M}_G = \{\mu_k\}_{1\leq k\leq K}$ and $\Sigma_G = \{\sigma_k\}_{1\leq k\leq K}$, respectively. To introduce variability in the contrast of $G$, the parameters in $\mathcal{M}_G$ and $\Sigma_G$ are sampled for each mini-batch from uniform distributions within the ranges $\{a_\mu, b_\mu\}$ and $\{a_\sigma, b_\sigma\}$, respectively. It is important to note that $\Sigma_G$ simultaneously accounts for tissue heterogeneities and scanner thermal noise. $G$ is then formed by independently sampling at each location $(x, y, z)$ the distribution indexed by $L(x, y, z)$:

$$\begin{aligned} \mu_k &\sim \mathcal{U}(a_\mu, b_\mu), \\ \sigma_k &\sim \mathcal{U}(a_\sigma, b_\sigma), \\ G(x, y, z) &\sim \mathcal{N}(\mu, \sigma^2), \end{aligned}$$

**Bias field and intensity augmentation** To enhance the robustness of SynthSeg against bias field artifacts, a simulated bias field is introduced into the synthetic scan. First, a small volume of size 43 is sampled from a zero-mean Gaussian distribution with a randomly varying standard deviation $\sigma_B$. This volume is then upsampled to the full image size, and the voxel-wise exponential of this upscaled volume is computed to generate a smooth and non-negative bias field $B$. The biased image $G_B$ is obtained by multiplying the synthetic scan $G$ by the bias field $B$. The exponential transformation ensures that division and multiplication by the same factor are equally likely.

$$\begin{aligned} \sigma_B &\sim \mathcal{U}(0, b_B), \\ B' &\sim \mathcal{N}_{4\times4\times4}(0, \sigma_B^2), \\ B &= \text{Upsample}(B'), \\ G_B(x, y, z) &= G(x, y, z) \times \exp[B(x, y, z)], \end{aligned}$$

Subsequently, a final high-resolution image $I_{HR}$ is generated by rescaling the biased image $G_B$ to the range $[0, 1]$ and applying a random Gamma transformation (voxel-wise exponentiation) to further enhance the intensity distribution of the synthetic scans. This transformation allows for the skewing of the intensity distribution while ensuring that the pixel values remain within the $[0, 1]$ interval. In practice, the exponent for this transformation is sampled from a

zero-mean Gaussian distribution with a standard deviation $\sigma_\gamma$ in the logarithmic domain. As a result, the final high-resolution image $I_{HR}$ is given by:

$$\gamma \sim \mathcal{N}(0, \sigma_\gamma^2)$$

$$I_{HR}(x, y, z) = \frac{(G(x, y, z) - \min_{x,y,z} G)^{\exp(\gamma)}}{\max_{x,y,z} G - \min_{x,y,z} G}$$

**Simulation of resolution variability** To enhance the network robustness against variations in image resolution, the authors have modeled differences in acquisition direction (i.e., axial, coronal, sagittal), slice spacing, and slice thickness. Initially, a random acquisition direction is selected. Subsequently, the slice spacing $r_{spacing}$ and slice thickness $r_{thick}$ are sampled from uniform distributions: $r_{spacing} \sim \mathcal{U}(r_{HR}, b_{res})$ and $r_{thick} \sim \mathcal{U}(r_{HR}, r_{spacing})$. It is important to note that $r_{thick}$ is constrained by $r_{spacing}$ to reflect the practical scenario where slices rarely overlap.

After sampling the resolution parameters, the effect of slice thickness is first simulated by applying a Gaussian blur to $I_{HR}$, resulting in $I_\sigma$. The Gaussian kernel used for this blurring is designed to approximate the real slice excitation profile, with a standard deviation $\sigma_{thick}$ set to reduce the power of the high-resolution signal by a factor of 10 at the cut-off frequency, as described by Billot et al. (2020b) [citation]. To introduce slight deviations from the nominal slice thickness and to address deviations from the Gaussian assumption, $\sigma_{thick}$ is multiplied by a random coefficient $\alpha$.

Next, slice spacing is modeled by downsampling $I_\sigma$ to $I_{LR}$ at the designated low resolution $r_{spacing}$ using trilinear interpolation (Van Leemput et al., 2003) [citation]. Finally, $I_{LR}$ is upsampled back to the high resolution $r_{HR}$ (typically 1 mm). This approach ensures that the convolutional neural network is trained to produce accurate high-resolution segmentations regardless of the simulated resolution variations.

$$\begin{aligned}
r_{spac} &\sim \mathcal{U}(r_{HR}, b_{res}), \\
r_{thick} &\sim \mathcal{U}(r_{HR}, r_{spac}), \\
\alpha &\sim \mathcal{U}(a_\alpha, b_\alpha), \\
\sigma_{thick} &= 2\alpha \log(10) \, (2\pi)^{-1} \, \tfrac{r_{thick}}{r_{HR}}, \\
I_\sigma &= I_{HR} * \mathcal{N}(0, \sigma_{thick}), \\
I &= \text{Resample}(I_\sigma; r_{spac}), \\
I &= \text{Resample}(I_{LR}; r_{HR}),
\end{aligned}$$

**Model output and segmentation target** In the SynthSeg method, two volumes are generated at each training step: an image $I$ created by the generative model, and its corresponding segmentation target $T$. The segmentation target $T$ is produced from the deformed map $L$. To focus on specific structures, label values that do not need to be segmented are reset to the background. As a result, $T$ contains $K' \leq K$ labels.

The main contribution of the SynthSeg method is the domain randomisation strategy.

### 5.3.2 SynthSeg - Segmentation network for brain structures segmentation

The segmentation network of SynthSeg is a 3D U-Net model architecture. It consists of five levels, each separated by a batch normalisation layer along with a max-pooling, or upsampling operation. All levels comprise two convolution layers with $3 \times 3 \times 3$ kernels. Every convolutional layer is associated with an Exponential Linear Unit activation, except for the last one, which uses a softmax.

For the training, the loss function used is the soft Dice (Milletari et al., 2016) [30]:

$$Loss(Y,T) = 1 - \sum_{k=1}^{K'} \frac{2\times\sum_{x,y,z} Y_k(x,y,z)T_k(x,y,z)}{\sum_{x,y,z} Y_k(x,y,z)^2+T_k(x,y,z)^2}$$

where $Y_k$ is the soft prediction for the label $k \in [1, ..., K']$ and $T_k$ is its associated ground truth in one-hot encoding, and $(x, y, z)$ are the spatial coordinates.

### 5.3.3 Training characteristics

In the work of Billot and al. [5], the authors precised that they have trained their model over 300,000 steps with a learning rate of $10^{-4}$ and a batch size of 1.
In the code, the optimiser implemented is the Adam optimiser.
Two distinct loss functions are utilized to optimize performance: the weighted L2 loss function is employed during the pre-training phase (optional), while the soft Dice loss function is applied during the subsequent training phase. The weighted L2 loss is designed to minimize the squared differences between predicted and true voxel values, which helps in fine-tuning the model parameter estimates early in the training process. The soft Dice, as previously mentioned, is used to enhance the model ability to achieve accurate segmentation by focusing on the overlap between predicted and true segmentation masks.

The patch size used is $128 \times 128 \times 128$, which is the recommended size according to the authors.

The loss functions in the SynthSeg code are implemented in the following way: depending on which function is used, at the end of the model architecture is added a new layer that calculates the associated loss function.

To evaluate the models and monitor training progress, the Dice score was utilized. Since this metric was not included in the original SynthSeg code, it was subsequently incorporated into the evaluation framework.

# 6 Results and discussion

In order to compare the performances of the different models in segmenting the ventricles, several scores were used: structure Dice score and mean Dice score over all the structures, Recall and Precision scores and calculating ventricular volume, multiple implementations of each model were tested. The means and standard deviation were considered. The best-performing implementation from each model type was selected for the final comparison.

The tests are conducted on the seven-label segmentation task, but given that manual segmentation is time-consuming for the doctor, this study also evaluates the models on the single-label task of segmenting the ventricles. This approach is particularly relevant when obtaining labels for multiple structures proves too challenging.

## 6.1 Results of original SynthSeg on fetal and preborn MRI

The pre-trained SynthSeg model was tested on several Necker MRI scans to evaluate its performance, as originally trained on adult and young subjects by Billot et al. [5]. Figure 11 displays the segmentations containing the most labels achieved on two different subjects from the Necker cohort.

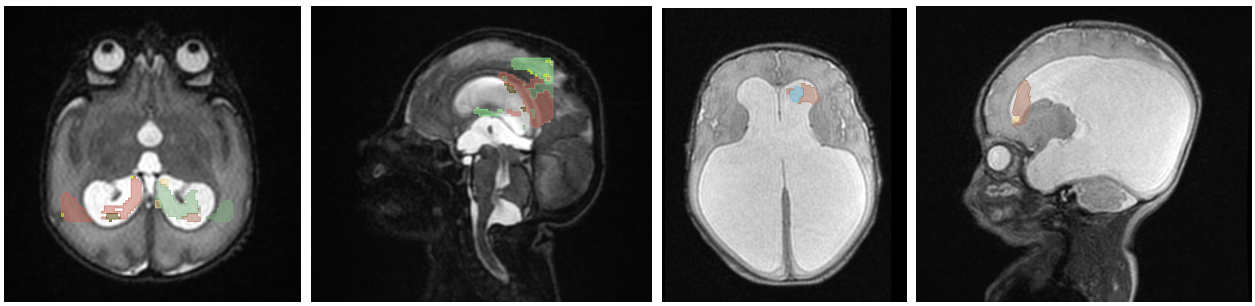

Figure 11: Examples of the predicted segmentation of the pre-trained SynthSeg model on two Necker subjects - axial and saggital plans.

The SynthSeg model was initially trained to segment 37 structures, including the ventricles. However, as illustrated in Figure 11, the pre-trained model completely fails to segment any structures on the Necker subjects. Consequently, it was not feasible to use this pre-trained model, necessitating the retraining of the SynthSeg model.

## 6.2 Results with FeTA 2022: best 2D U-Net model searching

To explore potential improvements to the 2D U-Net model, experiments were conducted to evaluate two factors: the inclusion of slices without the structure(s) of interest in the training set and the impact of having at least half of the slices of the training batch containing the structure(s) of interest.

To thoroughly evaluate the 2D models, two types of inferences were performed. The first type is the '2D inference', where the model is tested on individual 2D slices, and the patient scores are calculated as the mean scores from all the patient's slices. This method provides insights into the model performance on each slice independently, reflecting its ability to segment structures in isolated 2D images. The second type is the '3D reconstruction inference', where predicted segmentations from individual slices are combined to reconstruct a 3D volume. This reconstructed volume is then compared to the manually segmented 3D volume. This approach allows for evaluating the model performance in the context of the entire 3D volume, offering a more integrated assessment of how well the model captures and preserves spatial relationships across slices.

### 6.2.1 2D U-Net model with 1 dataloader on random MRI vs non-empty MRI

To evaluate the impact of including data without any structure or data without ventricles segmentation, two tests were conducted. These experiments involved two versions of the model: one trained to segment seven anatomical labels (External Cerebrospinal Fluid, Grey Matter, White Matter, Ventricles, Cerebellum, Deep Grey Matter, Brainstem), and another trained to segment a single label (ventricles).
For both model variations, the single dataloader approach involved using a unified training dataset. During training, batches were randomly selected from this single dataloader.
The MRI slices in the training and validation datasets are from different subjects compared to those in the test dataset, ensuring no overlap of subjects between the training and test data.

**Seven-Label Segmentation Model:** In the test on the seven-label segmentation model, two training approaches were compared:

- First Approach: The model was trained on a dataset comprising random MRI slices, where the segmentation files either lacked any structures or contained some structures.

- Second Approach: The model was trained on a dataset consisting exclusively of MRI slices with associated segmentation files that contain at least one structure label. It will be referred as 'non-empty' segmentations.

For the seven-label segmentation model, the first implementation was trained over 3,687 2D MRI slices and their associated segmentation (the all train dataset extracted from the University Children's Hospital Zurich dataset). The second implementation was trained over 3,310 2D MRI slices and their associated segmentation (only the slices with associated segmentation files that contain at least one structure label in the train dataset extracted from the University Children's Hospital Zurich dataset).

**One-Label Segmentation Model:** In the test on the one-label segmentation model, two training approaches were compared:

- First Approach: The model was trained on a dataset comprising random MRI slices, where the segmentation files either lacked ventricles structures or contained them.
- Second Approach: The model was trained on a dataset consisting exclusively of MRI slices with associated segmentation files that contain ventricles structures label. It will be referred as 'non-empty' segmentations.

For the one-label segmentation model, the first implementation was trained over 3,687 2D MRI slices and their associated segmentation (the all train dataset extracted from the University Children's Hospital Zurich dataset). The second implementation was trained over 1,716 2D MRI slices and their associated segmentation (only the slices that contain ventricles in the train dataset extracted from the University Children's Hospital Zurich dataset).

**Results for Seven-Label Segmentation Model** The results were obtained by testing the models on 1,554 2D MRI slices and their associated segmentations from the University Children's Hospital Zurich (Kispi) dataset. Over 1,554 2D MRI slices, 1,334 ($\sim$ 86%) contain at least one structure label. The models have been trained over 500 epochs with a batch size of 8.
The results of the inference of the 2D U-Net model over 2D slices are displayed in Figures 12 and 13. Figures 14 and 15 show the results of the 2D U-Net model inference on 3D segmentation reconstructions

| | Dice_score | Precision_score | Recall_score | V_Dice_score |
|---|---|---|---|---|
| Total Mean | 0.781 | 0.842 | 0.810 | 0.832 |
| Total Standard deviation | 0.080 | 0.051 | 0.074 | 0.120 |

Figure 12: Scores of 7 labels segmentation model with 1 dataloader on random files

| | Dice_score | Precision_score | Recall_score | V_Dice_score |
|---|---|---|---|---|
| Total Mean | 0.655 | 0.695 | 0.822 | 0.543 |
| Total Standard Deviation | 0.103 | 0.087 | 0.057 | 0.130 |

Figure 13: Scores of 7 labels segmentation model with 1 dataloader on only files with non-empty segmentation

| | Dice_score | Precision_score | Recall_score | V_Dice_score |
|---|---|---|---|---|
| Total Mean | 0.481 | 0.533 | 0.474 | 0.562 |
| Total Standard Deviation | 0.219 | 0.218 | 0.216 | 0.263 |

Figure 14: Scores on 3D segmentation reconstruction of 7 labels segmentation model with 1 dataloader on random files

| | Dice_score | Precision_score | Recall_score | V_Dice_score |
|---|---|---|---|---|
| Total Mean | 0.472 | 0.498 | 0.488 | 0.595 |
| Total Standard Deviation | 0.212 | 0.210 | 0.215 | 0.233 |

Figure 15: Scores on 3D segmentation reconstruction of 7 labels segmentation model with 1 dataloader on only files with non-empty segmentations

The total mean Dice score and the total mean precision score are significantly higher for the model trained on random files (with both empty and non-empty segmentations) for the 2D and 3D reconstruction inferences. This finding confirms that a model not trained on files lacking structures will fail to learn to refrain from segmenting structures when they are absent. Consequently, the model that was not trained on MRI images without structures tends to over-segment those structures.

**Results for One-Label Segmentation Model** The results were obtained by testing the models on 1,554 2D MRI slices and their associated segmentations from the University Children's Hospital Zurich (Kispi) dataset. Over 1,554 2D MRI slices, 724 ($\sim 47\%$) contain ventricles. The models have been trained over 500 epochs with a batch size of 8.
The results of the inference of the 2D U-Net model over 2D slices are displayed in Figures 16 and 17. Figures 18 and 19 show the results of the 2D U-Net model inference on 3D segmentation reconstructions

| | Dice_score | Precision_score | Recall_score | V_Dice_score |
|---|---|---|---|---|
| Total Mean | 0.840 | 0.943 | 0.837 | 0.840 |
| Total Standard deviation | 0.151 | 0.031 | 0.158 | 0.151 |

Figure 16: Scores of 1 label (ventricles) segmentation model with 1 dataloader on random files

| | Dice_score | Precision_score | Recall_score | V_Dice_score |
|---|---|---|---|---|
| Total Mean | 0.725 | 0.777 | 0.865 | 0.725 |
| Total Standard Deviation | 0.112 | 0.109 | 0.112 | 0.112 |

Figure 17: Scores of 1 label (ventricles) segmentation model with 1 dataloader on only files with non-empty segmentation

| | Dice_score | Precision_score | Recall_score | V_Dice_score |
|---|---|---|---|---|
| Total Mean | 0.562 | 0.668 | 0.544 | 0.562 |
| Total Standard Deviation | 0.290 | 0.249 | 0.298 | 0.290 |

Figure 18: Scores on 3D segmentation reconstruction of 1 label (ventricles) segmentation model with 1 dataloader on random files

| | Dice_score | Precision_score | Recall_score | V_Dice_score |
|---|---|---|---|---|
| Total Mean | 0.592 | 0.652 | 0.564 | 0.592 |
| Total Standard Deviation | 0.228 | 0.254 | 0.240 | 0.228 |

Figure 19: Scores on 3D segmentation reconstruction of 1 label (ventricles) segmentation model with 1 dataloader on only files with non-empty segmentations

Similarly to the results observed with the seven-label segmentation model, the ventricles segmentation model achieved the highest mean Dice and mean precision scores when trained on random images. Training exclusively on slices containing the structure to be segmented (ventricles) without including slices where the structure is absent leads to a significant loss in precision. The model which did not encounter 2D slices without ventricles structures during training, exhibited a tendency to over-segment.

To conclude, the experiments conducted to assess the impact of including empty data or data without ventricles segmentation reveal notable insights for both the seven-label and single-label

(ventricles) segmentation models. For both model types, training with a dataset containing random MRI slices—comprising both empty and non-empty segmentations—resulted in superior Dice and precision scores compared to training on datasets with exclusively non-empty segmentations.
The observed performance improvements indicate that models trained on a diverse set of MRI slices, including those with empty segmentations, are better at generalizing and avoiding over-segmentation. By contrast, models trained only on images with non-empty segmentations tend to over-segment when encountering slices lacking the targeted structures.

### 6.2.2 2D U-Net model with 2 dataloaders vs 1 dataloader

To explore potential enhancements to the 2D U-Net model, experiments were conducted utilizing either one or two dataloaders. These experiments were performed on two versions of the model: one trained to segment seven anatomical labels (External Cerebrospinal Fluid, Grey Matter, White Matter, Ventricles, Cerebellum, Deep Grey Matter, Brainstem) and another trained to segment a single label (ventricles).
For both model variations, the single dataloader approach involved using a unified training dataset. During training, batches were randomly selected from this single dataloader. The two-dataloaders implementation implies that at least half of the slices in the training batch contains the structures of interest.
The tests were conducted on a total of 3,687 2D MRI slices and their corresponding segmentations. Out of these, 1,716 slices (approximately 47%) had the ventricles label present in their segmentations. Additionally, 3,310 slices (approximately 90%) included at least one structure label in their segmentations.
Given that only about 10% of the 2D MRI slices had segmentations devoid of any structure labels, it was deemed impractical to use separate dataloaders exclusively for "empty" and "non-empty" segmentations. Instead, a dataloader containing exclusively non-empty segmentations (respectively exclusively with ventricles label) was compared against a random dataloader that included both empty and non-empty segmentations (respectively both with and without ventricles label). This approach ensured that at least half of the files in each batch would have non-empty (respectively with ventricles label) segmentations with structure labels.
The MRI slices in the training and validation datasets are from different subjects compared to those in the test dataset, ensuring no overlap of subjects between the training and test data.

**Seven-Label Segmentation Model:** In the two-dataloaders setup for the seven-label segmentation model, the following configurations were implemented:

- Dataloader 1: Contained only data where the segmentation file included at least one of the seven anatomical structure labels.
- Dataloader 2: Contained data where the segmentation file could be either empty or include any of the anatomical structure labels.

The datasets in dataloader 1 were exclusive and did not overlap with those in dataloader 2. During training, each batch was composed of an equal number of samples from both dataloader 1 and dataloader 2.
Specifically, the training dataset 1 comprised 1,658 2D MRI slices along with their associated segmentations, while the validation dataset 1 included 185 2D MRI slices and their corresponding segmentations. Similarly, the training dataset 2 contained 1,659 2D MRI slices with their

segmentations, and the validation dataset 2 comprised 185 2D MRI slices and their corresponding segmentations.

**One-Label Segmentation Model:** In the two-dataloaders setup for the one-label segmentation model, the following configurations were implemented:

- Dataloader 1: Contained only data where the segmentation file included the Ventricles label.
- Dataloader 2: Contained data where the segmentation file could either include or exclude the Ventricles label.

Similarly to the seven-label model, the datasets in dataloader 1 were exclusive and did not overlap with those in dataloader 2. During training, each batch was equally split between samples from dataloader 1 and dataloader 2.
Since less than half of the dataset contains the ventricles structure—specifically, 1,716 out of 3,687 ($\sim$ 47%) 2D MRI slices with ventricles labeled in their associated segmentation—the first dataloader includes all images from the dataset that contain ventricles, while the second dataloader consists solely of images without ventricles.
Specifically, the training dataset 1 comprised 1,544 2D MRI slices along with their associated segmentations, while the validation dataset 1 included 172 2D MRI slices and their corresponding segmentations. Similarly, the training dataset 2 contained 1,773 2D MRI slices with their segmentations, and the validation dataset 2 comprised 198 2D MRI slices and their corresponding segmentations.

**Results for Seven-Label Segmentation Model** The results were obtained by testing the models on 1,554 2D MRI slices and their associated segmentations from the University Children's Hospital Zurich (Kispi) dataset. The models have been trained over 500 epochs with a batch size of 8.
The results of the inference of the 2D U-Net model over 2D slices are displayed in Figures 20 and 21. Figures 22 and 23 show the results of the 2D U-Net model inference on 3D segmentation reconstructions.
The detailed table of the results of the 3D reconstruction inference of the 2D U-Net with two dataloaders on seven labels can be found in the appendix (8) in the Figures 44.

| | Dice_score | Precision_score | Recall_score | V_Dice_score |
|---|---|---|---|---|
| Total Mean | 0.781 | 0.842 | 0.810 | 0.832 |
| Total Standard deviation | 0.080 | 0.051 | 0.074 | 0.120 |

Figure 20: Scores of 7 labels segmentation model with 1 dataloader

| | Dice_score | Precision_score | Recall_score | V_Dice_score |
|---|---|---|---|---|
| Total Mean | 0.792 | 0.852 | 0.815 | 0.885 |
| Total Standard Deviation | 0.077 | 0.049 | 0.071 | 0.074 |

Figure 21: Scores of 7 labels segmentation model with 2 dataloaders

| | Dice_score | Precision_score | Recall_score | V_Dice_score |
|---|---|---|---|---|
| Total Mean | 0.481 | 0.533 | 0.474 | 0.562 |
| Total Standard Deviation | 0.219 | 0.218 | 0.216 | 0.263 |

Figure 22: Scores on 3D segmentation reconstruction of 7 labels segmentation model with 1 dataloader

| | Dice_score | Precision_score | Recall_score | V_Dice_score |
|---|---|---|---|---|
| Total Mean | 0.494 | 0.547 | 0.487 | 0.543 |
| Total Standard Deviation | 0.220 | 0.1995 | 0.220 | 0.233 |

Figure 23: Scores on 3D segmentation reconstruction of 7 labels segmentation model with 2 dataloaders

The results for the seven-label segmentation model are similar for the one-dataloader and two-dataloaders implementations (see Figures 22 and 23). The results do not provide sufficient evidence to conclude that one approach is superior to the other.

**Results for One-Label Segmentation Model** The results were obtained by testing the models on 1,554 2D MRI slices and their associated segmentations from the University Children's Hospital Zurich dataset. The models have been trained over 500 epochs with a batch size of 8.
The results of the inference of the 2D U-Net model over 2D slices are displayed in Figures 24 and 25. Figures 26 and 27 show the results of the 2D U-Net model inference on 3D segmentation reconstructions.
The detailed table of the results of the 3D reconstruction inference of the 2D U-Net with two dataloaders on one label (ventricles) can be found in the appendix (8) in the Figures 45.

| | Dice_score | Precision_score | Recall_score | V_Dice_score |
|---|---|---|---|---|
| Total Mean | 0.840 | 0.943 | 0.837 | 0.840 |
| Total Standard deviation | 0.151 | 0.031 | 0.158 | 0.151 |

Figure 24: Scores of 1 label (ventricles) segmentation model with 1 dataloader

| | Dice_score | Precision_score | Recall_score | V_Dice_score |
|---|---|---|---|---|
| Total Mean | 0.838 | 0.941 | 0.832 | 0.838 |
| Total Standard Deviation | 0.141 | 0.032 | 0.153 | 0.141 |

Figure 25: Scores of 1 label (ventricles) segmentation model with 2 dataloaders

| | Dice_score | Precision_score | Recall_score | V_Dice_score |
|---|---|---|---|---|
| Total Mean | 0.562 | 0.668 | 0.544 | 0.562 |
| Total Standard Deviation | 0.290 | 0.249 | 0.298 | 0.290 |

Figure 26: Scores on 3D segmentation reconstruction of 1 label (ventricles) segmentation model with 1 dataloader

| | Dice_score | Precision_score | Recall_score | V_Dice_score |
|---|---|---|---|---|
| Total Mean | 0.545 | 0.642 | 0.521 | 0.545 |
| Total Standard Deviation | 0.292 | 0.273 | 0.303 | 0.292 |

Figure 27: Scores on 3D segmentation reconstruction of 1 label (ventricles) segmentation model with 2 dataloaders

The results for the one-label (ventricles) segmentation task do not provide clear evidence that one implementation of the model is superior to the other. Specifically, there is no indication that the two-dataloaders implementation performs better for the 2D model. This outcome may be attributed to the fact that the two-dataloaders model processes 200 fewer images per epoch due to the smaller dataloader containing only data with ventricles. Over 500 epochs, this results in the two-dataloaders model seeing 100,000 fewer images compared to the one-dataloader model.

To conclude, the experiments conducted on the 2D model do not provide a definitive conclusion regarding which implementation is superior. However, the two-dataloaders implementation offers better control over the training process by ensuring that at least half of the images in each batch contain the structure of interest. Therefore, for the remainder of this work, for the 2D U-Net model, the two-dataloaders implementation will be utilized.

## 6.3 Results with FeTA 2024: best 3D U-Net model searching

To explore potential improvements to the 3D U-Net model, experiments were conducted to assess the impact of ensuring that at least half of the patches in each training batch contained a certain percentage of the structure(s) of interest. These experiments were performed on two versions of the model: one trained to segment seven anatomical labels (External Cerebrospinal

Fluid, Grey Matter, White Matter, Ventricles, Cerebellum, Deep Grey Matter, Brainstem) and another trained to segment a single label (ventricles).

For two models were trained over 100 MRI and their associated manual segmentation from the University Children's Hospital Zurich dataset (70 train subjects) and on the Hospital Vienna dataset (30 train subjects). The images were divided into patches of size $128 \times 128 \times 128$ to align with the patch size used in the SynthSeg model (see section 5.3). During training, a stride of 32 was used, resulting in 125 patches per image, resulting in $125 \times 100 = 12,500$ patches in the train dataset. For testing, a larger stride of 64 was applied, yielding 27 patches per image, with 20 test images (10 from the University Children's Hospital Zurich dataset and 10 from the Hospital Vienna dataset), resulting in $27 \times 20 = 540$ patches in the test dataset. The smaller stride during training produced more overlapping patches, allowing the model to be exposed to slightly shifted views of regions, thereby enhancing feature learning and improving robustness. The model was trained over 500 epochs, with each epoch consisting of 500 steps and a batch size of 2, leading to the model seeing 500 patches per epoch.
For the one-dataloader implementation, the model sees random patches among all the epochs.

For the two-dataloaders model, the implementation of the dataloaders depends on a threshold of the amount of ventricles in the patches, with a threshold defined at 5% in this work.

**2-dataloaders models:** In the two-dataloaders setup for the seven-label and one-label segmentation models, the following configurations were implemented:

- Dataloader 1: Contained only patches where the segmentation file included at least 5% of the ventricles label.
- Dataloader 2: Contained data where the segmentation file could have any pourcentage of ventricles label.

The datasets in dataloader 1 were exclusive and did not overlap with those in dataloader 2. During training, each batch was composed of an equal number of samples from both dataloader 1 and dataloader 2.
Specifically, in the training dataset, 1,406 patches containing at least 5% of ventricles structures were extracted (against 11,094 patches containing less than 5% of ventricles structures). The dataset 1 comprised 1,124 patches along with their associated segmentations, while the validation dataset 1 included 282 patches and their corresponding segmentations. Similarly, the training dataset 2 contained 8,875 patches with their segmentations, and the validation dataset 2 comprised 2,219 patches and their corresponding segmentations.

The patches in the training and validation datasets are from different subjects compared to those in the test dataset, ensuring no overlap of subjects between the training and test data.

**Results for Seven-Label Segmentation Model** The results were obtained by testing the 3D U-Net models on 20 MRI and their associated manual segmentations from the University Children's Hospital Zurich dataset (10 test subjects) and on the Hospital Vienna dataset (10 test subjects). The models have been trained over 500 epochs with a batch size of 2.

| | Dice_score | Precision_score | Recall_score | V_Dice_score |
|---|---|---|---|---|
| **Mean of FeTA Zurich dataset** | 0.648 | 0.698 | 0.634 | 0.788 |
| **Std Dev of FeTA Zurich dataset** | 0.244 | 0.247 | 0.245 | 0.266 |
| **Mean of FeTA Vienna dataset** | 0.689 | 0.759 | 0.677 | 0.745 |
| **Std Dev of FeTA Vienna dataset** | 0.169 | 0.13 | 0.171 | 0.14 |
| **Total Mean** | **0.694** | **0.759** | **0.682** | **0.789** |
| **Total Standard deviation** | **0.15** | **0.114** | **0.153** | **0.144** |

Figure 28: Scores of 3D U-Net model on 7 labels with 1 dataloader

| | Dice_score | Precision_score | Recall_score | V_Dice_score |
|---|---|---|---|---|
| **Mean of FeTA Zurich dataset** | 0.705 | 0.75 | 0.701 | 0.844 |
| **Std Dev of FeTA Zurich dataset** | 0.129 | 0.104 | 0.131 | 0.105 |
| **Mean of FeTA Vienna dataset** | 0.594 | 0.717 | 0.632 | 0.67 |
| **Std Dev of FeTA Vienna dataset** | 0.233 | 0.211 | 0.183 | 0.255 |
| **Total Mean** | **0.65** | **0.734** | **0.667** | **0.757** |
| **Total Standard deviation** | **0.181** | **0.158** | **0.157** | **0.18** |

Figure 29: Scores of 3D U-Net model on 7 labels with 2 dataloaders

In the seven-label segmentation task, the 3D U-Net model did not show a significant difference between the one-dataloader (see Figure 28) and two-dataloaders implementations (see Figure 29). While the precision score was slightly higher for the two-dataloaders model, the recall was slightly better for the one-dataloader model.

**Results for One-Label Segmentation Model** The results were obtained by testing the 3D U-Net models on 20 MRI and their associated manual segmentations from the University Children's Hospital Zurich dataset (10 test subjects) and on the Hospital Vienna dataset (10 test subjects). The models have been trained over 500 epochs with a batch size of 2.

| | Dice_score | Precision_score | Recall_score | V_Dice_score |
|---|---|---|---|---|
| **Mean of FeTA Zurich dataset** | 0.767 | 0.82 | 0.734 | 0.767 |
| **Std Dev FeTA Zurich dataset** | 0.27 | 0.281 | 0.275 | 0.27 |
| **Mean of FeTA Vienna dataset** | 0.593 | 0.906 | 0.512 | 0.593 |
| **Std Dev of FeTA Vienna dataset** | 0.279 | 0.063 | 0.291 | 0.279 |
| **Total Mean** | **0.689** | **0.911** | **0.628** | **0.689** |
| **Total Standard deviation** | **0.269** | **0.066** | **0.295** | **0.269** |

Figure 30: Scores of 3D U-Net model on 1 label (ventricles) with 1 dataloader

| | Dice_score | Precision_score | Recall_score | V_Dice_score |
|---|---|---|---|---|
| **Mean of FeTA Zurich dataset** | 0.786 | 0.827 | 0.754 | 0.786 |
| **Std Dev FeTA Zurich dataset** | 0.267 | 0.28 | 0.263 | 0.267 |
| **Mean of FeTA Vienna dataset** | 0.647 | 0.913 | 0.545 | 0.647 |
| **Std Dev of FeTA Vienna dataset** | 0.214 | 0.069 | 0.234 | 0.214 |
| **Total Mean** | **0.733** | **0.919** | **0.659** | **0.733** |
| **Total Standard deviation** | **0.212** | **0.061** | **0.249** | **0.212** |

Figure 31: Scores of 3D U-Net model on 1 label (ventricles) with 2 dataloaders

In the one-label (ventricles) segmentation task, the 3D U-Net model demonstrated better performance with the two-dataloaders implementation (see Figure 31) compared to the one-dataloader implementation (see Figure 30), particularly in terms of the Dice score. Consequently, the two-dataloaders 3D U-Net model will be used in the remainder of this report.

## 6.4 Results with FeTA 2024: best SynthSeg model searching

To determine the optimal SynthSeg model, three tests were conducted. These tests aimed to evaluate the utility of pre-training (one epoch of 5000 steps) with a pixel-wise loss (L2 weighted loss function) and to assess whether training the model with multiple structures segmentation, as opposed to only the ventricles segmentation, yielded better results. Training the SynthSeg model with only the ventricles segmentation implies that the synthetic images generated for training would contain only the ventricles structure, requiring the model to learn to segment

these ventricles against the rest of the image (considered as background).
The three models were trained over 100 manual segmentation from the University Children's Hospital Zurich dataset (70 train subjects) and from the Hospital Vienna dataset (30 train subjects). All the models have been trained over 100 epochs of 5000 steps with a batch size of 1.

The results in Figure 32 were obtained by testing the SynthSeg models on 20 MRI and their associated manual segmentations from the University Children's Hospital Zurich dataset (10 test subjects) and on the Hospital Vienna dataset (10 test subjects).

| | | Dice_score | Precision_score | Recall_score | V_Dice_score |
|---|---|---|---|---|---|
| **SynthSeg 0 pre-training** | **Total mean** | 0.168 | 0.297 | 0.175 | 0.171 |
| | **Total standard deviation** | 0.162 | 0.161 | 0.187 | 0.21 |
| **SynthSeg 1 pre-training** | **Total mean** | 0.514 | 0.711 | 0.504 | 0.801 |
| | **Total standard deviation** | 0.693 | 0.699 | 0.754 | 0.835 |
| **SynthSeg only Ventricles - 1 pre-training** | **Total mean** | 0.257 | 0.175 | 0.71 | 0.257 |
| | **Total standard deviation** | 0.174 | 0.134 | 0.269 | 0.174 |

Figure 32: Scores of SynthSeg models

The model that was trained with only the ventricles segmentation has the best total mean Dice score but the worst precision score. This model over-segments with the label ventricles, likely because it has never learn to segment the ventricles when the ventricles are in a brain image (with other structures in the image than just the ventricles), as Figure 33 and 34 shows. It appears to segment cerebrospinal fluid primarily based on the image contrast, without effectively distinguishing between cerebrospinal fluid within the ventricles and the cerebrospinal fluid external to them. This suggest that this model is not effectively learning to recognise the anatomical position of the cerebrospinal fluid. Thus, even if it has the best ventricles Dice score, it will not be used in the remainder of this report. It could be used if a post-processing was made in order to cropped the image around the ventricles area, leading in a priori good scores on the ventricles structures.

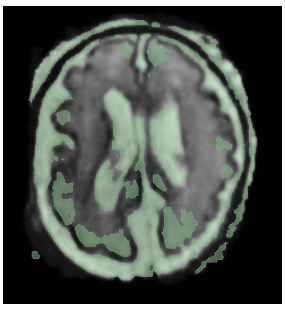

Figure 33: Example of predicted segmentation of the SynthSeg model trained with only the ventricles segmentation

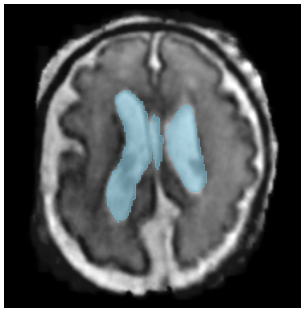

Figure 34: Manual segmentation of the example from the University Children's Hospital Zurich dataset

The ventricles Dice score and the recall score are similar for the model with pre-training and the model without. As the model with pre-training has a better precision score, it is the one that will be used in the remainder of this report.

## 6.5 Results with FeTA 2024 - 7 labels

After identifying the best version of each model, a comparison was conducted to evaluate its performance. Given the superior performance of the 3D U-Net model compared to the 2D U-Net model, the 3D U-Net model has been selected as the representative U-Net architecture for the comparisons in this study.
The ventricles volume and the ventricles error, defined as the difference between the predicted ventricles volume and the manually segmented ventricles volume, are included in the comparison. If the ventricles error is negative, it indicates that the predicted ventricles volume is smaller than the manually segmented volume. Therefore, if the mean ventricles error in a model scores is negative, it suggests that the model tends to undersegment the ventricles volume across the tested dataset.
Initially, the comparison focused on models trained to segment seven labels. These models were trained using a dataset that included 100 manually segmented cases: 70 subjects from the University Children's Hospital Zurich and 30 subjects from the Hospital Vienna dataset. The models were then tested on 44 subjects: 10 from the University Children's Hospital Zurich (only T2 images), 10 from the Hospital Vienna (only T2 images), and 24 from the Necker dataset (12 T2 and 12 T1 images). This diverse testing set was crucial for assessing the models generalizability and robustness across different datasets.
Figure 35 presents the overall scores obtained by each model, while Figure 36 provides a detailed breakdown of the results by test groups.
The detailed tables of results can be found in the appendix (8) in the Figures 46 and 49.

| | | Dice_score | Precision_score | Recall_score | V_Dice_score | Ventricles_volume | Ventricles_error |
|---|---|---|---|---|---|---|---|
| SynthSeg 1 pre-training | Total mean | 0.336 | 0.477 | 0.545 | 0.394 | 103620.514 | -182.715 |
| | Total standard deviation | 0.17 | 0.165 | 0.151 | 0.199 | 70250.632 | 243.824 |
| 3D U-Net 2 dataloaders | Total mean | 0.556 | 0.54 | 0.717 | 0.628 | 189131.689 | -97.2 |
| | Total standard deviation | 0.246 | 0.176 | 0.138 | 0.245 | 161175.4 | 161.563 |

Figure 35: Comparison of the best version of the 3D models trained over the FeTA datasets on 7 labels. The ventricles volume is displayed in $mm^3$ and the ventricles error in $cm^3$.

| | | Zurich dataset | Vienna dataset | Necker dataset | Necker T1 | Necker T2 | Pre-op T2 Necker | Post-op T2 Necker |
|---|---|---|---|---|---|---|---|---|
| SynthSeg 1 pre-training | Mean V_Dice_score | 0.683 | 0.204 | 0.293 | 0.231 | 0.356 | 0.304 | 0.409 |
| | Std V_Dice_score | 0.257 | 0.124 | 0.215 | 0.214 | 0.206 | 0.258 | 0.143 |
| 3D U-Net 2 dataloaders | Mean V_Dice_score | 0.844 | 0.67 | 0.37 | 0.092 | 0.648 | 0.605 | 0.691 |
| | Std V_Dice_score | 0.105 | 0.255 | 0.376 | 0.25 | 0.246 | 0.305 | 0.188 |

Figure 36: Comparison of the best version of the 3D models trained over the FeTA datasets on 7 labels on the different test groups.

However, the 3D U-Net model outperformed the SynthSeg model overall, despite the fact that the test set including T1 images. While SynthSeg performed relatively consistently on both T1 and T2 images, the 3D U-Net model achieved significantly higher scores on T2 images but struggled with T1 images, as it was only trained on T2 sequences. This discrepancy is highlighted in Figure 36.

Both models achieved better performances on the Zurich dataset.

Both models achieve similar performances over the pre-surgery operation and post-surgery operation on the T2 images of Necker, but are more robust on the post-surgery operation images.

Ultimately, the 3D U-Net model, trained on T2 FeTA images, emerged as the best model for segmenting seven labels T2 images. However, it is not reliable for T1 image segmentation.

The three models tend to undersegment the ventricles, as indicated by the negative mean ventricles error.

Examples of segmentations can be found in the Figures 52, 53, 54 and 55 in the appendix (8).

## 6.6 Results with FeTA 2024 - 1 label

Given the superior performance of the 3D U-Net model compared to the 2D U-Net model, the 3D U-Net model has been selected as the representative U-Net architecture for the comparisons in this study.

Figure 37 presents the overall scores obtained by each model, while Figure 38 provides a detailed breakdown of the results by test groups.
The detailed tables of results can be found in the appendix (8) in the Figures 47 and 50.

| | | Dice_score | Precision_score | Recall_score | V_Dice_score | Ventricles_volume | Ventricles_error |
|---|---|---|---|---|---|---|---|
| SynthSeg 1 pre-training | Total mean | 0.213 | 0.143 | 0.615 | 0.213 | 1089345.875 | 803.01 |
| | Total standard deviation | 0.175 | 0.129 | 0.322 | 0.175 | 267876.206 | 386.573 |
| 3D U-Net 2 dataloaders | Total mean | 0.572 | 0.773 | 0.52 | 0.572 | 154176.772 | -132.159 |
| | Total standard deviation | 0.267 | 0.236 | 0.276 | 0.267 | 132134.1 | 42.362 |

Figure 37: Comparison of the best version of the 3D models trained over the FeTA datasets on 1 label (ventricles). The ventricles volume is displayed in $mm^3$ and the ventricles error in $cm^3$.

| | | Zurich dataset | Vienna dataset | Necker dataset | Necker T1 | Necker T2 | Pre-op T2 Necker | Post-op T2 Necker |
|---|---|---|---|---|---|---|---|---|
| SynthSeg 1 pre-training | Mean V_Dice_score | 0.209 | 0.305 | 0.125 | 0.019 | 0.232 | 0.357 | 0.107 |
| | Std V_Dice_score | 0.114 | 0.233 | 0.178 | 0.037 | 0.199 | 0.161 | 0.156 |
| 3D U-Net 2 dataloaders | Mean V_Dice_score | 0.819 | 0.564 | 0.333 | 0.094 | 0.572 | 0.529 | 0.616 |
| | Std V_Dice_score | 0.18 | 0.276 | 0.345 | 0.251 | 0.247 | 0.314 | 0.177 |

Figure 38: Comparison of the best version of the 3D models trained over the FeTA datasets on 1 label (ventricles) on the different test groups.

The 3D U-Net model trained on the seven-label segmentation task achieves a better Dice score on the ventricles segmentation than the model trained on the single-label (ventricles) segmentation task (see Figures 35 and 37). As the network learns to recognize structures other than the ventricles, it becomes better at distinguishing the ventricles from other structures, reducing the likelihood of confusion and improving the accuracy of ventricles segmentation. This finding is consistent with results reported in the literature [19].

Globally, as Figure 37 highlighted, the performance of the 3D models trained on a single label is not significantly different from that of the models trained on seven labels.

Once again, the 3D U-Net model outperformed the SynthSeg model on the test set. As observed in the single-label segmentation task, Figure 38 highlights that the 3D U-Net model performed well on T2 images but poorly on T1 images. Surprisingly, the SynthSeg model did not demonstrate the more consistent performance across both T1 and T2 images that was observed when it was trained on seven labels.

While the SynthSeg model achieve better performances on the pre-surgery operation T2 images of the Necker dataset, the 3D U-Net model achieve better performances on the T2 post-surgery operation, and is more robust.

The SynthSeg model shows a tendency to oversegment the ventricles when trained on a single label, by contrast to its performance on the seven-label task, where it tended to undersegment. On the other hand, the 3D U-Net model consistently undersegments the ventricles, as evidenced by the negative mean ventricles error.

Examples of segmentations can be found in the Figures 56, 57, 58 and 59 in the appendix (8).

## 6.7 Results with FeTA 2024 and Necker - 1 label

To assess the added value of including Necker data in the training process, three tests were conducted. The Necker dataset used contained 48 T2 images for the train set and 24 images (T1 and T2) for the test set. In the 24 images in the test set, 12 are T1-weighted images and 12 are T1-weighted images. The test set is composed of subject that had a couple of T1 and T2 images in the exam before the operation and a couple of T1 and T2 in the exam after the operation.

Given the superior performance of the 3D U-Net model compared to the 2D U-Net model, the 3D U-Net model has been selected as the representative U-Net architecture for the comparisons in this study. The best-performing models from each family —U-Net and SynthSeg—were trained on 148 manually segmented images, from the University Children's Hospital Zurich dataset (70 subjects), the Hospital Vienna dataset (30 subjects), and the Necker Hospital dataset (48 subjects).
For training the 3D U-Net model, a total of 18,500 patches were extracted (125 patches per image), with 2,345 patches containing at least 5% ventricles. These were divided into 1,876 patches for the first training dataloader and 469 patches for the first validation dataloader.
The 3D U-Net model was trained over 500 epochs with a batch size of 2. The SynthSeg model was trained over 100 epochs of 5,000 steps each, with a batch size of 1.

Figure 39 presents the overall scores obtained by each model, while Figure 40 provides a detailed breakdown of the results by test groups.
The detailed tables of results can be found in the appendix (8) in the Figures 48 and 51.

| | | Dice_score | Precision_score | Recall_score | V_Dice_score | Ventricles_volume | Ventricles_error |
|---|---|---|---|---|---|---|---|
| SynthSeg 1 pre-training | Total mean | 0.221 | 0.144 | 0.789 | 0.221 | 1663398.253 | 1377.063 |
| | Total standard deviation | 0.161 | 0.119 | 0.192 | 0.161 | 524553.539 | 513.79 |
| 3D U-Net 2 dataloaders | Total mean | 0.605 | 0.82 | 0.556 | 0.605 | 158513.3 | -127.822 |
| | Total standard deviation | 0.294 | 0.23 | 0.305 | 0.294 | 138596.267 | 199.884 |

Figure 39: Comparison of the best version of the 3D models trained over all the datasets on 1 label (ventricles). The ventricles volume is displayed in $mm^3$ and the ventricles error in $cm^3$.

| | | Zurich dataset | Vienna dataset | Necker dataset | Necker T1 | Necker T2 | Pre-op T2 Necker | Post-op T2 Necker |
|---|---|---|---|---|---|---|---|---|
| SynthSeg 1 pre-training | Mean V_Dice_score | 0.209 | 0.305 | 0.125 | 0.019 | 0.232 | 0.357 | 0.107 |
| | Std V_Dice_score | 0.114 | 0.233 | 0.178 | 0.037 | 0.199 | 0.161 | 0.156 |
| 3D U-Net 2 dataloaders | Mean V_Dice_score | 0.819 | 0.564 | 0.333 | 0.094 | 0.572 | 0.529 | 0.616 |
| | Std V_Dice_score | 0.18 | 0.276 | 0.345 | 0.251 | 0.247 | 0.314 | 0.177 |

Figure 40: Comparison on the different test groups of the best version of the 3D models trained over all datasets on 1 label (ventricles).

When comparing the models trained only on the FeTA datasets to those trained on all three datasets, the SynthSeg model showed similar performances. However, adding the Necker dataset to the training did not improve results and even led to an increase in the oversegmentation of the ventricles. For the 3D U-Net model, performance slightly improved when trained on the three datasets, but the difference is not significant.

Consistent with the results from the models trained on one label from the FeTA datasets, the 3D U-Net model trained on all three datasets with one label outperformed the SynthSeg model. Figure 40 shows that the 3D U-Net model continued to perform well on T2 images but poorly on T1 images, just like the SynthSeg model.

While the SynthSeg model achieve better performances on the pre-surgery operation T2 images of the Necker dataset, the 3D U-Net model achieve better performances on the T2 post-surgery operation, and is more robust.

Examples of segmentations can be found in the Figures 60, 61, 62 and 63 in the appendix (8).

# 7 Conclusion and perspectives

After testing several implementations of the U-Net models, the best version was the one trained on more heterogeneous data patches, particularly including patches without the structure(s) of interest. This diversity allowed the model to generalize better. Additionally, the U-Net models performed better when at least half of the training batch contained the structure(s) of interest, ensuring that the model consistently learned relevant features. These findings underscore the importance of including a varied dataset, including images with absent structures, to enhance model robustness and accuracy in segmentation tasks.
The 3D U-Net model outperformed the 2D U-Net, which can be attributed to its ability to capture spatial information in three dimensions. This capability enables the 3D U-Net to better analyze volumetric data and maintain contextual relationships across different layers, leading to more accurate segmentation results.

After testing several implementations of the SynthSeg model, the best version was achieved through pre-training using a pixel-wise loss function (L2 weighted loss). Additionally, the SynthSeg model performed better when multiple structures were used to create the synthetic images, and when the segmentation task involved multiple labels. This result aligns with expectations: since the SynthSeg model is trained solely on synthetic images, having images that contain only ventricles and background leads to a significant drop in performance on real data. By including multiple structures in the training process, the model learns more comprehensive spatial information, which enhances its ability to accurately segment ventricles in real-world scenarios. The inclusion of various labels in the training data likely helps the model better understand the overall anatomy, resulting in improved segmentation performance for the ventricles.

The performance of the 3D U-Net model on the Necker data varies significantly depending on the MRI sequence used as input: it performs well on T2 sequence images, achieving a mean Dice score of up to 0.648 for the T2 Necker group MRI, but performs poorly on T1 sequence images, with a maximum mean Dice score of only 0.094 for the T1 Necker group MRI. This outcome is expected since the 3D U-Net model was trained exclusively on T2 images. Consequently, while the 3D U-Net model is suitable for segmenting or determining the volume of ventricles on T2 MRI, it is not directly applicable for T1 MRI. On the other hand, the SynthSeg model, designed to be contrast-agnostic, confirms this property when trained on multiple structures but fails to maintain it when trained on just one structure (ventricles). This

indicates that while SynthSeg can handle variations in image contrast effectively when exposed to diverse anatomical structures during training, its contrast-agnostic capability diminishes when its training is limited to a single structure.

In general, the 3D U-Net outperforms the SynthSeg model, except when it comes to segmenting T1 images in scenarios where the models are trained on multiple structures. Valabregue et al. (2023) [44] also observed that the SynthSeg model performance on neonatal brain segmentation was not as strong as expected. They particularly noted a clear influence of the infant's age on the model predictions, which could further reinforce the conclusions of this work. Given that the subjects of this study are very young infants with hydrocephalus, this age-related variability in SynthSeg performance may explain some of the challenges encountered in segmenting these complex cases. This observation suggests that the model robustness may be compromised when applied to very young or diverse patient populations, emphasizing the need for further refinement or alternative approaches when dealing with such specific clinical conditions.
The SynthSeg model relatively poor performance can be attributed to its training and evaluation being conducted exclusively on synthetic data, without exposure to real-world data. This limitation suggests that while the SynthSeg model is robust in synthetic environments, it struggles to generalize effectively to real-world images. A potential direction for future work could involve modifying the SynthSeg model to incorporate training and evaluation on real data. This adjustment could enhance the model ability to perform better over real-world images.

Finally, the 3D U-Net model trained on the Zurich and Vienna datasets over seven labels, using two dataloaders, achieved the best performances in this study. It obtained a Dice score of 0.844 for ventricles segmentation with a standard deviation of 0.105 on the Zurich dataset, and a Dice score of 0.705 for all structures with a standard deviation of 0.129 on the Zurich dataset. In comparison, the winning team of the FeTA 2021 challenge, which trained its model solely on the Zurich dataset and tested it on the challenge test set (which is not publicly accessible), achieved a Dice score of 0.9 for ventricles segmentation and a Dice score of 0.786 for overall structures segmentation [34].

A key consideration is that the number of MRI scans of preborn infants with hydrocephalus and potential intraventricular stroke used in this study is quite limited. This scarcity of data, a common challenge in medical research, inevitably impacts the development of a highly effective model for this specific task. To enhance the 3D U-Net model, incorporating data augmentation techniques could be beneficial, particularly to increase the number of images with hydrocephalus. This would likely improve the model ability to generalize across various pathological conditions. Additionally, this study did not account for strokes due to the insufficient number of subjects with hemorrhage (only 25 out of 48 training subjects), which was too small a sample for a statistically significant analysis. Future studies could benefit from including a larger number of subjects with hydrocephalus and potentially intraventricular hemorrhage, even with poorer quality or with more artefacts. This would likely lead to improved performance in the segmentation task and allow for the inclusion of intraventricular hemorrhage in the segmentation process.

Finally, it is important to note that the calculation of the Dice score warrants special consideration. If an MRI slice or patch in the test set lacks the structure(s) to be segmented and the model does not segment the missing structure(s), the Dice score will be 1. This outcome can artificially inflate the average Dice score. However, it is essential to encourage the model during training to correctly refrain from segmenting when the structure is absent, thereby reinforcing its ability to avoid over-segmentation. Considering an alternative approach to calculating the Dice score in such situations could be a valuable direction for future work.

# 8 Appendix

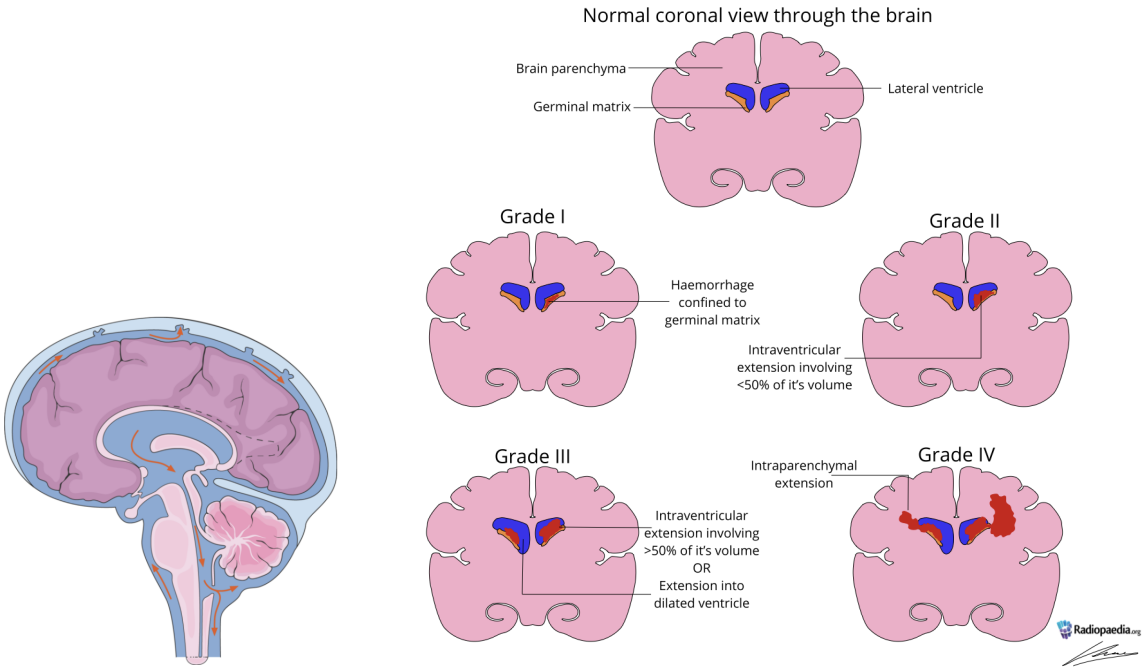


Figure 41: On the left: Schema of external spaces of cerebrospinal fluid and cerebral ventricles - Servier Medical Art. *https : //smart.servier.com/smart_image/cerebrospinalfluid/*
On the right: Classification of germinal matrix bleeding according to Papile - Gendy D, Germinal matrix haemorrhage grading. Case study, Radiopaedia.org (Accessed on 25 Jun 2024), https://doi.org/10.53347/rID-79252

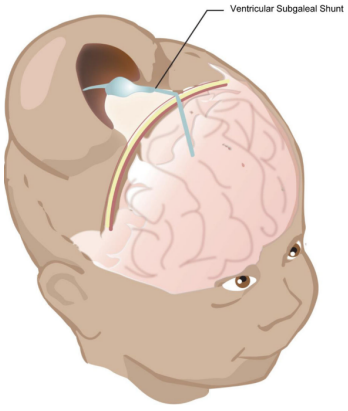


Figure 42: Illustration of the ventriculo-subgaleal shunt [13]


Figure 43: Illustration of the subcutaneous reservoir insertion [13]

| Patient | Dice_score | Precision_score | Recall_score | ECF_Dice_score | GM_Dice_score | WM_Dice_score | V_Dice_score | C_Dice_score | DGM_Dice_score | B_Dice_score |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.815 | 0.831 | 0.8 | 0.815 | 0.616 | 0.896 | 0.84 | 0.853 | 0.849 | 0.834 |
| 2 | 0.638 | 0.648 | 0.637 | 0.693 | 0.374 | 0.767 | 0.67 | 0.657 | 0.777 | 0.53 |
| 3 | 0.106 | 0.108 | 0.105 | 0.157 | 0.1 | 0.28 | 0.148 | 0.0 | 0.057 | 0.0 |
| 4 | 0.578 | 0.658 | 0.557 | 0.125 | 0.426 | 0.846 | 0.862 | 0.462 | 0.828 | 0.494 |
| 5 | 0.5 | 0.521 | 0.537 | 0.165 | 0.379 | 0.751 | 0.617 | 0.518 | 0.662 | 0.404 |
| 6 | 0.589 | 0.615 | 0.569 | 0.501 | 0.352 | 0.652 | 0.718 | 0.603 | 0.665 | 0.632 |
| 7 | 0.15 | 0.34 | 0.128 | 0.023 | 0.109 | 0.439 | 0.242 | 0.01 | 0.23 | 0.0 |
| 8 | 0.473 | 0.51 | 0.451 | 0.131 | 0.261 | 0.662 | 0.606 | 0.682 | 0.653 | 0.313 |
| 9 | 0.445 | 0.59 | 0.449 | 0.358 | 0.275 | 0.477 | 0.599 | 0.479 | 0.37 | 0.56 |
| 10 | 0.644 | 0.651 | 0.639 | 0.675 | 0.349 | 0.708 | 0.711 | 0.667 | 0.704 | 0.694 |
| **Total Mean** | **0.494** | **0.547** | **0.487** | **0.362** | **0.294** | **0.552** | **0.543** | **0.495** | **0.584** | **0.446** |
| **Total Standard Deviation** | **0.220** | **0.1995** | **0.220** | **0.286** | **0.151** | **0.194** | **0.233** | **0.281** | **0.268** | **0.276** |

Figure 44: Detailed scores of the 3D reconstruction inference of the 2D U-Net model trained and tested over the Zurich dataset on seven labels, with two dataloaders

| Patient | Dice_score | Precision_score | Recall_score | V_Dice_score |
|---|---|---|---|---|
| 1 | 0.839 | 0.871 | 0.809 | 0.839 |
| 2 | 0.633 | 0.735 | 0.555 | 0.633 |
| 3 | 0.196 | 0.208 | 0.186 | 0.196 |
| 4 | 0.883 | 0.864 | 0.902 | 0.883 |
| 5 | 0.589 | 0.571 | 0.608 | 0.589 |
| 6 | 0.721 | 0.793 | 0.66 | 0.721 |
| 7 | 0.063 | 0.122 | 0.042 | 0.063 |
| 8 | 0.64 | 0.638 | 0.643 | 0.64 |
| 9 | 0.177 | 0.907 | 0.098 | 0.177 |
| 10 | 0.712 | 0.712 | 0.712 | 0.712 |
| **Total Mean** | **0.545** | **0.642** | **0.521** | **0.545** |
| **Total Standard Deviation** | **0.292** | **0.273** | **0.303** | **0.292** |

Figure 45: Detailed scores of the 3D reconstruction inference of the 2D U-Net trained and tested over the Zurich dataset on one label (ventricles), with two dataloaders

| Patient | Dice_score | Precision_score | Recall_score | ECF_Dice_score | GM_Dice_score | WM_Dice_score | V_Dice_score | C_Dice_score | DGM_Dice_score | B_Dice_score | Ventricles_volume | Ventricles_error |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.829 | 0.825 | 0.838 | 0.881 | 0.703 | 0.883 | 0.842 | 0.837 | 0.902 | 0.754 | 172488.0 | -41.779 |
| 2 | 0.799 | 0.844 | 0.777 | 0.815 | 0.704 | 0.873 | 0.905 | 0.819 | 0.859 | 0.616 | 406363.0 | -38.89 |
| 3 | 0.657 | 0.74 | 0.655 | 0.189 | 0.538 | 0.874 | 0.893 | 0.638 | 0.839 | 0.628 | 126529.0 | 1.757 |
| 4 | 0.654 | 0.661 | 0.706 | 0.375 | 0.6 | 0.858 | 0.785 | 0.618 | 0.723 | 0.621 | 69265.0 | 6.269 |
| 5 | 0.856 | 0.869 | 0.845 | 0.841 | 0.75 | 0.894 | 0.909 | 0.892 | 0.844 | 0.861 | 361044.0 | -57.927 |
| 6 | 0.558 | 0.592 | 0.568 | 0.114 | 0.496 | 0.834 | 0.855 | 0.637 | 0.745 | 0.223 | 103419.0 | 3.582 |
| 7 | 0.675 | 0.705 | 0.655 | 0.327 | 0.523 | 0.86 | 0.909 | 0.814 | 0.776 | 0.514 | 252554.0 | -7.641 |
| 8 | 0.633 | 0.715 | 0.613 | 0.511 | 0.404 | 0.63 | 0.837 | 0.652 | 0.741 | 0.653 | 1076337.0 | -262.574 |
| 9 | 0.883 | 0.897 | 0.873 | 0.926 | 0.785 | 0.922 | 0.933 | 0.897 | 0.889 | 0.832 | 338473.0 | 17.522 |
| 10 | 0.506 | 0.648 | 0.48 | 0.73 | 0.589 | 0.487 | 0.573 | 0.475 | 0.179 | 0.508 | 276865.0 | -362.385 |
| **Mean of FeTA Zurich dataset** | **0.705** | **0.75** | **0.701** | **0.571** | **0.609** | **0.812** | **0.844** | **0.728** | **0.75** | **0.621** | **318333.7** | **-74.201** |
| **Std Dev of FeTA Zurich dataset** | **0.129** | **0.104** | **0.131** | **0.305** | **0.123** | **0.139** | **0.105** | **0.142** | **0.21** | **0.184** | **289996** | **130.058** |
| 101 | 0.774 | 0.807 | 0.763 | 0.784 | 0.694 | 0.842 | 0.864 | 0.774 | 0.682 | 0.777 | 190031.0 | -27.997 |
| 102 | 0.502 | 0.785 | 0.473 | 0.734 | 0.661 | 0.719 | 0.809 | 0.0 | 0.21 | 0.383 | 340895.0 | -131.304 |
| 103 | 0.764 | 0.835 | 0.745 | 0.874 | 0.711 | 0.835 | 0.89 | 0.81 | 0.429 | 0.8 | 288827.0 | -44.103 |
| 104 | 0.469 | 0.742 | 0.433 | 0.587 | 0.574 | 0.576 | 0.525 | 0.0 | 0.43 | 0.593 | 269876.0 | -429.762 |
| 105 | 0.797 | 0.833 | 0.783 | 0.884 | 0.707 | 0.832 | 0.89 | 0.835 | 0.637 | 0.796 | 353607.0 | -38.374 |
| 106 | 0.366 | 0.604 | 0.325 | 0.485 | 0.341 | 0.303 | 0.646 | 0.081 | 0.06 | 0.644 | 315193.0 | -324.067 |
| 107 | 0.487 | 0.713 | 0.475 | 0.78 | 0.58 | 0.749 | 0.612 | 0.0 | 0.673 | 0.011 | 39326.0 | -31.101 |
| 108 | 0.812 | 0.863 | 0.778 | 0.921 | 0.767 | 0.884 | 0.726 | 0.869 | 0.704 | 0.812 | 57051.0 | -23.404 |
| 109 | 0.825 | 0.834 | 0.82 | 0.836 | 0.715 | 0.902 | 0.706 | 0.909 | 0.864 | 0.842 | 32456.0 | -5.064 |
| 110 | 0.148 | 0.157 | 0.722 | 0.0 | 0.004 | 0.0 | 0.035 | 1.0 | 0.0 | 0.0 | 57870.0 | -201.193 |
| **Mean of FeTA Vienna dataset** | **0.594** | **0.717** | **0.632** | **0.689** | **0.575** | **0.664** | **0.67** | **0.528** | **0.469** | **0.566** | **194513.2** | **-125.63** |
| **Std Dev of FeTA Vienna dataset** | **0.233** | **0.211** | **0.183** | **0.278** | **0.234** | **0.295** | **0.255** | **0.442** | **0.295** | **0.326** | **134911** | **147.321** |
| 8010761051_2017_11_06_T2 | 0.768 | 0.123 | 0.822 | | | | 0.768 | | | | 122891.0 | -28.262 |
| 8010761051_2017_11_06_T1 | 0.039 | 0.16 | 0.718 | | | | 0.039 | | | | 12548.0 | -52.927 |
| 8010761051_2017_11_30_T2 | 0.726 | 0.109 | 0.813 | | | | 0.726 | | | | 17375.0 | -1.879 |
| 8010761051_2017_11_30_T1 | 0.049 | 0.015 | 0.719 | | | | 0.049 | | | | 22831.0 | -47.674 |
| 8011759516_2018_10_15_T2 | 0.633 | 0.141 | 0.835 | | | | 0.633 | | | | 53280.0 | -39.038 |
| 8011759516_2018_10_15_T1 | 0.069 | 0.007 | 0.731 | | | | 0.069 | | | | 53818.0 | 31.857 |
| 8011759516_2018_10_22_T2 | 0.33 | 0.03 | 0.821 | | | | 0.33 | | | | 40040.0 | 28.713 |
| 8011759516_2018_10_22_T1 | 0.01 | 0.001 | 0.716 | | | | 0.01 | | | | 99756.0 | 21.61 |
| 8013327626_2019_07_29_T2 | 0.635 | 0.093 | 0.946 | | | | 0.635 | | | | 49188.0 | -2.182 |
| 8013327626_2019_07_29_T1 | 0.047 | 0.014 | 0.719 | | | | 0.047 | | | | 6067.0 | -13.114 |
| 8013327626_2019_08_11_T2 | 0.749 | 0.093 | 0.983 | | | | 0.749 | | | | 109704.0 | 28.932 |
| 8013327626_2019_08_11_T1 | 0.0 | 0.0 | 0.714 | | | | 0.0 | | | | 1304.0 | -155.964 |
| 8017618189_2023_11_02_T2 | 0.818 | 0.108 | 0.9 | | | | 0.818 | | | | 114500.0 | 19.258 |
| 8017618189_2023_11_02_T1 | 0.0 | 0.429 | 0.714 | | | | 0.0 | | | | 285.0 | -81.507 |
| 8017618189_2023_11_19_T2 | 0.828 | 0.103 | 0.996 | | | | 0.828 | | | | 79260.0 | 20.543 |
| 8017618189_2023_11_19_T1 | 0.0 | 0.0 | 0.857 | | | | 0.0 | | | | 445.0 | -704.152 |
| 8017871237_2024_02_20_T2 | 0.775 | 0.135 | 0.951 | | | | 0.775 | | | | 146113.0 | -63.664 |
| 8017871237_2024_02_20_T1 | 0.0 | 0.0 | 0.857 | | | | 0.0 | | | | 238.0 | -10.609 |
| 8017871237_2024_02_26_T2 | 0.672 | 0.1 | 0.95 | | | | 0.672 | | | | 24759.0 | -1.893 |
| 8017871237_2024_02_26_T1 | 0.0 | 0.429 | 0.714 | | | | 0.0 | | | | 9.0 | -743.305 |
| 8017896414_2024_03_25_T2 | 0.002 | 0.568 | 0.714 | | | | 0.002 | | | | 260.0 | -252.739 |
| 8017896414_2024_03_25_T1 | 0.007 | 0.011 | 0.717 | | | | 0.007 | | | | 10292.0 | -177.392 |
| 8017896414_2024_03_29_T2 | 0.843 | 0.117 | 0.84 | | | | 0.843 | | | | 202793.0 | 13.912 |
| 8017896414_2024_03_29_T1 | 0.882 | 0.869 | 0.864 | | | | 0.882 | | | | 141400.0 | 9.039 |
| **Mean of Necker dataset** | **0.37** | **0.152** | **0.817** | | | | **0.37** | | | | **54548.167** | **-91.768** |
| **Std Dev of Necker dataset** | **0.376** | **0.213** | **0.099** | | | | **0.376** | | | | **58619.2** | **207.311** |

| | Dice_score | Precision_score | Recall_score | ECF_Dice_score | GM_Dice_score | WM_Dice_score | V_Dice_score | C_Dice_score | DGM_Dice_score | B_Dice_score | Ventricles_volume | Ventricles_error |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Total Mean** | **0.556** | **0.54** | **0.717** | **0.63** | **0.592** | **0.738** | **0.628** | **0.628** | **0.61** | **0.594** | **189131.689** | **-97.2** |
| **Total Standard deviation** | **0.246** | **0.176** | **0.138** | **0.292** | **0.179** | **0.217** | **0.245** | **0.92** | **0.253** | **0.255** | **161175.4** | **161.563** |

Figure 46: Detailed scores of the 3D U-Net model trained over the FeTA datasets and tested over the three datasets on seven labels, with two dataloaders

| Patient | Dice_score | Precision_score | Recall_score | V_Dice_score | Ventricles_volume | Ventricles_error |
|---|---|---|---|---|---|---|
| **1** | 0.88 | 0.951 | 0.818 | 0.88 | 184329.0 | -29.938 |
| **2** | 0.916 | 0.959 | 0.877 | 0.916 | 407331.0 | -37.922 |
| **3** | 0.888 | 0.895 | 0.882 | 0.888 | 122982.0 | -1.79 |
| **4** | 0.831 | 0.824 | 0.838 | 0.831 | 64039.0 | 1.043 |
| **5** | 0.908 | 0.99 | 0.839 | 0.908 | 354998.0 | -63.973 |
| **6** | 0.822 | 0.851 | 0.794 | 0.822 | 93178.0 | -6.659 |
| **7** | 0.908 | 0.912 | 0.904 | 0.908 | 257981.0 | -2.214 |
| **8** | 0.768 | 0.965 | 0.638 | 0.768 | 884954.0 | -453.957 |
| **9** | 0.938 | 0.927 | 0.949 | 0.938 | 328589.0 | 7.638 |
| **10** | 0.327 | 0.961 | 0.197 | 0.327 | 131090.0 | -508.16 |
| **Mean of FeTA Zurich dataset** | **0.819** | **0.924** | **0.774** | **0.819** | **282947.1** | **-109.593** |
| **Std Dev FeTA Zurich dataset** | **0.18** | **0.053** | **0.219** | **0.18** | **242233** | **197.41** |
| **101** | 0.687 | 0.965 | 0.534 | 0.687 | 120602.0 | -97.426 |
| **102** | 0.597 | 0.979 | 0.429 | 0.597 | 206987.0 | -265.212 |
| **103** | 0.854 | 0.978 | 0.759 | 0.854 | 258343.0 | -74.587 |
| **104** | 0.516 | 0.967 | 0.352 | 0.516 | 254466.0 | -445.172 |
| **105** | 0.866 | 0.96 | 0.788 | 0.866 | 321838.0 | -70.143 |
| **106** | 0.139 | 0.945 | 0.075 | 0.139 | 50919.0 | -588.341 |
| **107** | 0.503 | 0.813 | 0.364 | 0.503 | 31481.0 | -38.946 |
| **108** | 0.72 | 0.872 | 0.613 | 0.72 | 56535.0 | -23.92 |
| **109** | 0.705 | 0.779 | 0.644 | 0.705 | 31011.0 | -6.509 |
| **110** | 0.052 | 0.141 | 0.032 | 0.052 | 59061.0 | -200.002 |
| **Mean of FeTA Vienna dataset** | **0.564** | **0.84** | **0.459** | **0.564** | **139124.3** | **-181.026** |
| **Std Dev of FeTA Vienna dataset** | **0.276** | **0.256** | **0.261** | **0.276** | **110623** | **196.939** |
| **8010761051_2017_11_06_T2** | 0.816 | 0.9 | 0.746 | 0.816 | 125146.0 | -26.007 |
| **8010761051_2017_11_06_T1** | 0.061 | 0.185 | 0.036 | 0.061 | 12875.0 | -52.6 |
| **8010761051_2017_11_30_T2** | 0.676 | 0.604 | 0.769 | 0.676 | 24507.0 | 5.253 |
| **8010761051_2017_11_30_T1** | 0.071 | 0.12 | 0.05 | 0.071 | 29319.0 | -41.186 |
| **8011759516_2018_10_15_T2** | 0.558 | 0.889 | 0.407 | 0.558 | 42247.0 | -50.071 |
| **8011759516_2018_10_15_T1** | 0.047 | 0.033 | 0.079 | 0.047 | 52204.0 | 30.243 |
| **8011759516_2018_10_22_T2** | 0.273 | 0.18 | 0.566 | 0.273 | 35701.0 | 24.374 |
| **8011759516_2018_10_22_T1** | 0.001 | 0.002 | 0.0 | 0.001 | 13253.0 | -64.893 |
| **8013327626_2019_07_29_T2** | 0.702 | 0.81 | 0.619 | 0.702 | 39228.0 | -12.142 |
| **8013327626_2019_07_29_T1** | 0.054 | 0.099 | 0.037 | 0.054 | 7131.0 | -12.05 |
| **8013327626_2019_08_11_T2** | 0.652 | 0.644 | 0.66 | 0.652 | 82809.0 | 2.037 |
| **8013327626_2019_08_11_T1** | 0.0 | 0.0 | 0.0 | 0.0 | 1789.0 | -155.479 |
| **8017618189_2023_11_02_T2** | 0.772 | 0.812 | 0.736 | 0.772 | 86321.0 | -8.921 |
| **8017618189_2023_11_02_T1** | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | -81.792 |
| **8017618189_2023_11_19_T2** | 0.784 | 0.697 | 0.895 | 0.784 | 75462.0 | 16.745 |
| **8017618189_2023_11_19_T1** | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | -704.597 |
| **8017871237_2024_02_20_T2** | 0.328 | 0.963 | 0.198 | 0.328 | 43065.0 | -166.712 |
| **8017871237_2024_02_20_T1** | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | -10.847 |
| **8017871237_2024_02_26_T2** | 0.619 | 0.635 | 0.603 | 0.619 | 25325.0 | -1.327 |
| **8017871237_2024_02_26_T1** | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | -743.313 |
| **8017896414_2024_03_25_T2** | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | -252.999 |
| **8017896414_2024_03_25_T1** | 0.003 | 0.036 | 0.002 | 0.003 | 8224.0 | -179.46 |
| **8017896414_2024_03_29_T2** | 0.689 | 0.84 | 0.584 | 0.689 | 131354.0 | -57.527 |
| **8017896414_2024_03_29_T1** | 0.886 | 0.877 | 0.895 | 0.886 | 135053.0 | 2.692 |

Figure 47: Detailed scores of the 3D U-Net model trained over the FeTA datasets and tested over the three datasets on one label (ventricles), with two dataloaders

| Patient | Dice_score | Precision_score | Recall_score | V_Dice_score | Ventricles_volume | Ventricles_error |
|---|---|---|---|---|---|---|
| **1** | 0.88 | 0.951 | 0.818 | 0.88 | 184329.0 | -29.938 |
| **2** | 0.916 | 0.959 | 0.877 | 0.916 | 407331.0 | -37.922 |
| **3** | 0.888 | 0.895 | 0.882 | 0.888 | 122982.0 | -1.79 |
| **4** | 0.831 | 0.824 | 0.838 | 0.831 | 64039.0 | 1.043 |
| **5** | 0.908 | 0.99 | 0.839 | 0.908 | 354998.0 | -63.973 |
| **6** | 0.822 | 0.851 | 0.794 | 0.822 | 93178.0 | -6.659 |
| **7** | 0.908 | 0.912 | 0.904 | 0.908 | 257981.0 | -2.214 |
| **8** | 0.768 | 0.965 | 0.638 | 0.768 | 884954.0 | -453.957 |
| **9** | 0.938 | 0.927 | 0.949 | 0.938 | 328589.0 | 7.638 |
| **10** | 0.327 | 0.961 | 0.197 | 0.327 | 131090.0 | -508.16 |
| **Mean of FeTA Zurich dataset** | **0.819** | **0.924** | **0.774** | **0.819** | **282947.1** | **-109.593** |
| **Std Dev FeTA Zurich dataset** | **0.18** | **0.053** | **0.219** | **0.18** | **242233** | **197.41** |
| **101** | 0.687 | 0.965 | 0.534 | 0.687 | 120602.0 | -97.426 |
| **102** | 0.597 | 0.979 | 0.429 | 0.597 | 206987.0 | -265.212 |
| **103** | 0.854 | 0.978 | 0.759 | 0.854 | 258343.0 | -74.587 |
| **104** | 0.516 | 0.967 | 0.352 | 0.516 | 254466.0 | -445.172 |
| **105** | 0.866 | 0.96 | 0.788 | 0.866 | 321838.0 | -70.143 |
| **106** | 0.139 | 0.945 | 0.075 | 0.139 | 50919.0 | -588.341 |
| **107** | 0.503 | 0.813 | 0.364 | 0.503 | 31481.0 | -38.946 |
| **108** | 0.72 | 0.872 | 0.613 | 0.72 | 56535.0 | -23.92 |
| **109** | 0.705 | 0.779 | 0.644 | 0.705 | 31011.0 | -6.509 |
| **110** | 0.052 | 0.141 | 0.032 | 0.052 | 59061.0 | -200.002 |
| **Mean of FeTA Vienna dataset** | **0.564** | **0.84** | **0.459** | **0.564** | **139124.3** | **-181.026** |
| **Std Dev of FeTA Vienna dataset** | **0.276** | **0.256** | **0.261** | **0.276** | **110623** | **196.939** |
| **8010761051_2017_11_06_T2** | 0.936 | 0.944 | 0.928 | 0.936 | 148495.0 | -2.658 |
| **8010761051_2017_11_06_T1** | 0.072 | 0.334 | 0.04 | 0.072 | 7902.0 | -57.573 |
| **8010761051_2017_11_30_T2** | 0.806 | 0.912 | 0.722 | 0.806 | 15236.0 | -4.018 |
| **8010761051_2017_11_30_T1** | 0.031 | 0.092 | 0.019 | 0.031 | 14535.0 | -55.97 |
| **8011759516_2018_10_15_T2** | 0.927 | 0.953 | 0.903 | 0.927 | 87417.0 | -4.901 |
| **8011759516_2018_10_15_T1** | 0.017 | 0.016 | 0.019 | 0.017 | 25442.0 | 3.481 |
| **8011759516_2018_10_22_T2** | 0.411 | 0.294 | 0.681 | 0.411 | 26224.0 | 14.897 |
| **8011759516_2018_10_22_T1** | 0.0 | 0.0 | 0.0 | 0.0 | 72670.0 | -5.476 |
| **8013327626_2019_07_29_T2** | 0.883 | 0.855 | 0.912 | 0.883 | 54745.0 | 3.375 |
| **8013327626_2019_07_29_T1** | 0.061 | 0.113 | 0.041 | 0.061 | 7044.0 | -12.137 |
| **8013327626_2019_08_11_T2** | 0.859 | 0.838 | 0.881 | 0.859 | 84915.0 | 4.143 |
| **8013327626_2019_08_11_T1** | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | -157.268 |
| **8017618189_2023_11_02_T2** | 0.89 | 0.911 | 0.87 | 0.89 | 90935.0 | -4.307 |
| **8017618189_2023_11_02_T1** | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | -81.792 |
| **8017618189_2023_11_19_T2** | 0.887 | 0.818 | 0.97 | 0.887 | 69597.0 | 10.88 |
| **8017618189_2023_11_19_T1** | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | -704.597 |
| **8017871237_2024_02_20_T2** | 0.954 | 0.973 | 0.937 | 0.954 | 202128.0 | -7.649 |
| **8017871237_2024_02_20_T1** | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | -10.847 |
| **8017871237_2024_02_26_T2** | 0.714 | 0.808 | 0.639 | 0.714 | 21080.0 | -5.572 |
| **8017871237_2024_02_26_T1** | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | -743.314 |
| **8017896414_2024_03_25_T2** | 0.139 | 1.0 | 0.075 | 0.139 | 18914.0 | -234.085 |
| **8017896414_2024_03_25_T1** | 0.003 | 0.071 | 0.002 | 0.003 | 4280.0 | -183.404 |
| **8017896414_2024_03_29_T2** | 0.904 | 0.871 | 0.939 | 0.904 | 203509.0 | 14.628 |
| **8017896414_2024_03_29_T1** | 0.873 | 0.888 | 0.86 | 0.873 | 128176.0 | -4.185 |

Figure 48: Detailed scores of the 3D U-Net model trained and tested over the three datasets on one label (ventricles), with two dataloaders

| Patient | Dice_score | Precision_score | Recall_score | ECF_Dice_score | GM_Dice_score | WM_Dice_score | V_Dice_score | C_Dice_score | DGM_Dice_score | B_Dice_score | Ventricles_volume | Ventricles_error |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 0.514 | 0.711 | 0.504 | 0.491 | 0.43 | 0.826 | 0.801 | 0.128 | 0.509 | 0.416 | 175074 | -39.193 |
| **2** | 0.693 | 0.699 | 0.754 | 0.501 | 0.556 | 0.81 | 0.835 | 0.86 | 0.604 | 0.686 | 363455 | -81.798 |
| **3** | 0.492 | 0.604 | 0.572 | 0.242 | 0.515 | 0.812 | 0.818 | 0.391 | 0.27 | 0.396 | 134017 | 9.245 |
| **4** | 0.477 | 0.475 | 0.573 | 0.04 | 0.29 | 0.748 | 0.696 | 0.72 | 0.556 | 0.287 | 93691 | 30.695 |
| **5** | 0.667 | 0.705 | 0.727 | 0.543 | 0.63 | 0.796 | 0.772 | 0.854 | 0.399 | 0.677 | 273584 | -145.387 |
| **6** | 0.355 | 0.542 | 0.396 | 0.072 | 0.388 | 0.768 | 0.778 | 0 | 0.476 | 0 | 106377 | 6.54 |
| **7** | 0.36 | 0.546 | 0.39 | 0.077 | 0.376 | 0.756 | 0.766 | 0.313 | 0.168 | 0.062 | 201328 | -58.867 |
| **8** | 0.259 | 0.364 | 0.292 | 0.396 | 0.329 | 0.4 | 0.475 | 0.031 | 0 | 0.182 | 429382 | -909.529 |
| **9** | 0.786 | 0.804 | 0.813 | 0.811 | 0.707 | 0.88 | 0.87 | 0.869 | 0.657 | 0.711 | 276233 | -44.718 |
| **10** | 0.142 | 0.688 | 0.168 | 0.584 | 0.272 | 0.114 | 0.021 | 0 | 0 | 0 | 7002 | -632.248 |
| **Mean of FeTA Zurich dataset** | **0.475** | **0.614** | **0.519** | **0.376** | **0.449** | **0.691** | **0.683** | **0.417** | **0.364** | **0.342** | **206014.3** | **-186.526** |
| **Std Dev of FeTA Zurich dataset** | **0.202** | **0.133** | **0.21** | **0.259** | **0.147** | **0.241** | **0.257** | **0.376** | **0.242** | **0.282** | **130166.525** | **318.845** |
| **101** | 0.297 | 0.665 | 0.277 | 0.624 | 0.368 | 0.361 | 0.372 | 0 | 0.239 | 0.116 | 52084 | -165.944 |
| **102** | 0.207 | 0.593 | 0.229 | 0.655 | 0.37 | 0.271 | 0.151 | 0 | 0.004 | 0 | 38823 | -433.376 |
| **103** | 0.313 | 0.731 | 0.325 | 0.796 | 0.505 | 0.597 | 0.219 | 0 | 0.076 | 0 | 41894 | -291.036 |
| **104** | 0.2 | 0.695 | 0.232 | 0.584 | 0.336 | 0.321 | 0.131 | 0 | 0.031 | 0 | 50611 | -649.027 |
| **105** | 0.306 | 0.667 | 0.305 | 0.779 | 0.448 | 0.532 | 0.332 | 0 | 0 | 0.05 | 82144 | -309.837 |
| **106** | 0.121 | 0.617 | 0.161 | 0.519 | 0.209 | 0.054 | 0.054 | 0 | 0.013 | 0 | 17801 | -621.459 |
| **107** | 0.266 | 0.641 | 0.273 | 0.732 | 0.401 | 0.545 | 0.186 | 0 | 0 | 0 | 8950 | -61.477 |
| **108** | 0.31 | 0.75 | 0.287 | 0.813 | 0.534 | 0.638 | 0.179 | 0 | 0.004 | 0 | 10177 | -70.278 |
| **109** | 0.323 | 0.716 | 0.291 | 0.683 | 0.529 | 0.681 | 0.031 | 0 | 0.003 | 0.334 | 1453 | -36.067 |
| **110** | 0.055 | 0.088 | 0.758 | 0 | 0.002 | 0 | 0.385 | 0 | 0 | 0 | 117315 | -141.748 |
| **Mean of FeTA Vienna dataset** | **0.24** | **0.616** | **0.314** | **0.619** | **0.37** | **0.4** | **0.204** | **0,000** | **0.037** | **0.05** | **42125.2** | **-278.025** |
| **Std Dev of FeTA Vienna dataset** | **0.092** | **0.192** | **0.163** | **0.237** | **0.163** | **0.24** | **0.124** | **0,000** | **0.075** | **0.107** | **36215.912** | **226.233** |
| **8010761051_2017_11_06_T2** | 0.344 | 0.1 | 0.802 | | | | 0.344 | | | | 50083 | -101.07 |
| **8010761051_2017_11_06_T1** | 0.275 | 0.052 | 0.746 | | | | 0.275 | | | | 40083 | -25.392 |
| **8010761051_2017_11_30_T2** | 0.345 | 0.042 | 0.774 | | | | 0.345 | | | | 27744 | 8.49 |
| **8010761051_2017_11_30_T1** | 0.15 | 0.182 | 0.731 | | | | 0.15 | | | | 26548 | -43.957 |
| **8011759516_2018_10_15_T2** | 0.202 | 0.181 | 0.749 | | | | 0.202 | | | | 58393 | -33.925 |
| **8011759516_2018_10_15_T1** | 0.141 | 0.016 | 0.743 | | | | 0.141 | | | | 39790 | 17.829 |
| **8011759516_2018_10_22_T2** | 0.211 | 0.019 | 0.787 | | | | 0.211 | | | | 43494 | 32.167 |
| **8011759516_2018_10_22_T1** | 0.096 | 0.013 | 0.731 | | | | 0.096 | | | | 104420 | 26.274 |
| **8013327626_2019_07_29_T2** | 0.152 | 0.014 | 0.909 | | | | 0.152 | | | | 196324 | 144.954 |
| **8013327626_2019_07_29_T1** | 0.019 | 0.445 | 0.716 | | | | 0.019 | | | | 1740 | -17.441 |
| **8013327626_2019_08_11_T2** | 0.518 | 0.072 | 0.933 | | | | 0.518 | | | | 85627 | 4.855 |
| **8013327626_2019_08_11_T1** | 0.396 | 0.221 | 0.791 | | | | 0.396 | | | | 90688 | -66.58 |
| **8017618189_2023_11_02_T2** | 0.786 | 0.114 | 0.866 | | | | 0.786 | | | | 93788 | -1.454 |
| **8017618189_2023_11_02_T1** | 0.1 | 0.314 | 0.724 | | | | 0.1 | | | | 27207 | -54.585 |
| **8017618189_2023_11_19_T2** | 0.595 | 0.214 | 0.963 | | | | 0.595 | | | | 87583 | 28.866 |
| **8017618189_2023_11_19_T1** | 0.242 | 0.406 | 0.877 | | | | 0.242 | | | | 119064 | -585.533 |
| **8017871237_2024_02_20_T2** | 0.29 | 0.404 | 0.882 | | | | 0.29 | | | | 44675 | -165.102 |
| **8017871237_2024_02_20_T1** | 0.266 | 0.33 | 0.89 | | | | 0.266 | | | | 8096 | -2.751 |
| **8017871237_2024_02_26_T2** | 0.471 | 0.353 | 0.925 | | | | 0.471 | | | | 26962 | 0.31 |
| **8017871237_2024_02_26_T1** | 0.071 | 0.068 | 0.72 | | | | 0.071 | | | | 60066 | -683.248 |
| **8017896414_2024_03_25_T2** | 0.048 | 0.175 | 0.718 | | | | 0.048 | | | | 30903 | -222.096 |
| **8017896414_2024_03_25_T1** | 0.187 | 0.224 | 0.742 | | | | 0.187 | | | | 37842 | -149.842 |
| **8017896414_2024_03_29_T2** | 0.315 | 0.22 | 0.749 | | | | 0.315 | | | | 77623 | -111.258 |
| **8017896414_2024_03_29_T1** | 0.823 | 0.672 | 0.778 | | | | 0.823 | | | | 126586 | -5.775 |
| **Mean of Necker dataset** | **0.293** | **0.202** | **0.802** | | | | **0.293** | | | | **62722.042** | **-83.594** |
| **Std Dev of Necker dataset** | **0.215** | **0.169** | **0.08** | | | | **0.215** | | | | **44369.457** | **186.392** |

| | Dice_score | Precision_score | Recall_score | ECF_Dice_score | GM_Dice_score | WM_Dice_score | V_Dice_score | C_Dice_score | DGM_Dice_score | B_Dice_score | Ventricles_volume | Ventricles_error |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Total Mean** | **0.336** | **0.477** | **0.545** | **0.497** | **0.41** | **0.546** | **0.394** | **0.208** | **0.2** | **0.196** | **103620.514** | **-182.715** |
| **Total Standard deviation** | **0.17** | **0.165** | **0.151** | **0.248** | **0.155** | **0.241** | **0.199** | **0.188** | **0.158** | **0.194** | **70250.632** | **243.824** |

Figure 49: Detailed scores of the SynthSeg model with pre-training, trained over the FeTA datasets and tested over the three datasets on seven labels

| Patient | Dice_score | Precision_score | Recall_score | V_Dice_score | Ventricles_volume | Ventricles_error |
|---|---|---|---|---|---|---|
| **1** | 0.88 | 0.951 | 0.818 | 0.88 | 184329.0 | -29.938 |
| **2** | 0.916 | 0.959 | 0.877 | 0.916 | 407331.0 | -37.922 |
| **3** | 0.888 | 0.895 | 0.882 | 0.888 | 122982.0 | -1.79 |
| **4** | 0.831 | 0.824 | 0.838 | 0.831 | 64039.0 | 1.043 |
| **5** | 0.908 | 0.99 | 0.839 | 0.908 | 354998.0 | -63.973 |
| **6** | 0.822 | 0.851 | 0.794 | 0.822 | 93178.0 | -6.659 |
| **7** | 0.908 | 0.912 | 0.904 | 0.908 | 257981.0 | -2.214 |
| **8** | 0.768 | 0.965 | 0.638 | 0.768 | 884954.0 | -453.957 |
| **9** | 0.938 | 0.927 | 0.949 | 0.938 | 328589.0 | 7.638 |
| **10** | 0.327 | 0.961 | 0.197 | 0.327 | 131090.0 | -508.16 |
| **Mean of FeTA Zurich dataset** | **0.819** | **0.924** | **0.774** | **0.819** | **282947.1** | **-109.593** |
| **Std Dev FeTA Zurich dataset** | **0.18** | **0.053** | **0.219** | **0.18** | **242233** | **197.41** |
| **101** | 0.687 | 0.965 | 0.534 | 0.687 | 120602.0 | -97.426 |
| **102** | 0.597 | 0.979 | 0.429 | 0.597 | 206987.0 | -265.212 |
| **103** | 0.854 | 0.978 | 0.759 | 0.854 | 258343.0 | -74.587 |
| **104** | 0.516 | 0.967 | 0.352 | 0.516 | 254466.0 | -445.172 |
| **105** | 0.866 | 0.96 | 0.788 | 0.866 | 321838.0 | -70.143 |
| **106** | 0.139 | 0.945 | 0.075 | 0.139 | 50919.0 | -588.341 |
| **107** | 0.503 | 0.813 | 0.364 | 0.503 | 31481.0 | -38.946 |
| **108** | 0.72 | 0.872 | 0.613 | 0.72 | 56535.0 | -23.92 |
| **109** | 0.705 | 0.779 | 0.644 | 0.705 | 31011.0 | -6.509 |
| **110** | 0.052 | 0.141 | 0.032 | 0.052 | 59061.0 | -200.002 |
| **Mean of FeTA Vienna dataset** | **0.564** | **0.84** | **0.459** | **0.564** | **139124.3** | **-181.026** |
| **Std Dev of FeTA Vienna dataset** | **0.276** | **0.256** | **0.261** | **0.276** | **110623** | **196.939** |
| **8010761051_2017_11_06_T2** | 0.816 | 0.9 | 0.746 | 0.816 | 125146.0 | -26.007 |
| **8010761051_2017_11_06_T1** | 0.061 | 0.185 | 0.036 | 0.061 | 12875.0 | -52.6 |
| **8010761051_2017_11_30_T2** | 0.676 | 0.604 | 0.769 | 0.676 | 24507.0 | 5.253 |
| **8010761051_2017_11_30_T1** | 0.071 | 0.12 | 0.05 | 0.071 | 29319.0 | -41.186 |
| **8011759516_2018_10_15_T2** | 0.558 | 0.889 | 0.407 | 0.558 | 42247.0 | -50.071 |
| **8011759516_2018_10_15_T1** | 0.047 | 0.033 | 0.079 | 0.047 | 52204.0 | 30.243 |
| **8011759516_2018_10_22_T2** | 0.273 | 0.18 | 0.566 | 0.273 | 35701.0 | 24.374 |
| **8011759516_2018_10_22_T1** | 0.001 | 0.002 | 0.0 | 0.001 | 13253.0 | -64.893 |
| **8013327626_2019_07_29_T2** | 0.702 | 0.81 | 0.619 | 0.702 | 39228.0 | -12.142 |
| **8013327626_2019_07_29_T1** | 0.054 | 0.099 | 0.037 | 0.054 | 7131.0 | -12.05 |
| **8013327626_2019_08_11_T2** | 0.652 | 0.644 | 0.66 | 0.652 | 82809.0 | 2.037 |
| **8013327626_2019_08_11_T1** | 0.0 | 0.0 | 0.0 | 0.0 | 1789.0 | -155.479 |
| **8017618189_2023_11_02_T2** | 0.772 | 0.812 | 0.736 | 0.772 | 86321.0 | -8.921 |
| **8017618189_2023_11_02_T1** | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | -81.792 |
| **8017618189_2023_11_19_T2** | 0.784 | 0.697 | 0.895 | 0.784 | 75462.0 | 16.745 |
| **8017618189_2023_11_19_T1** | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | -704.597 |
| **8017871237_2024_02_20_T2** | 0.328 | 0.963 | 0.198 | 0.328 | 43065.0 | -166.712 |
| **8017871237_2024_02_20_T1** | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | -10.847 |
| **8017871237_2024_02_26_T2** | 0.619 | 0.635 | 0.603 | 0.619 | 25325.0 | -1.327 |
| **8017871237_2024_02_26_T1** | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | -743.313 |
| **8017896414_2024_03_25_T2** | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | -252.999 |
| **8017896414_2024_03_25_T1** | 0.003 | 0.036 | 0.002 | 0.003 | 8224.0 | -179.46 |
| **8017896414_2024_03_29_T2** | 0.689 | 0.84 | 0.584 | 0.689 | 131354.0 | -57.527 |
| **8017896414_2024_03_29_T1** | 0.886 | 0.877 | 0.895 | 0.886 | 135053.0 | 2.692 |

Figure 50: Detailed scores of the SynthSeg model with pre-training, trained over the FeTA datasets and tested over the three datasets on one label (ventricles)

| Patient | Dice_score | Precision_score | Recall_score | V_Dice_score | Ventricles_volume | Ventricles_error |
|---|---|---|---|---|---|---|
| **1** | 0.145 | 0.079 | 0.928 | 0.145 | 2531590 | 2317.323 |
| **2** | 0.282 | 0.166 | 0.949 | 0.282 | 2552616 | 2107.363 |
| **3** | 0.116 | 0.062 | 0.943 | 0.116 | 1906036 | 1781.264 |
| **4** | 0.05 | 0.026 | 0.953 | 0.05 | 2349046 | 2286.05 |
| **5** | 0.316 | 0.193 | 0.875 | 0.316 | 1901571 | 1482.6 |
| **6** | 0.115 | 0.061 | 0.975 | 0.115 | 1593944 | 1494.107 |
| **7** | 0.18 | 0.099 | 0.945 | 0.18 | 2474213 | 2214.018 |
| **8** | 0.53 | 0.369 | 0.941 | 0.53 | 3414262 | 2075.351 |
| **9** | 0.23 | 0.131 | 0.938 | 0.23 | 2292211 | 1971.26 |
| **10** | 0.518 | 0.397 | 0.747 | 0.518 | 1203731 | 564.481 |
| **Mean of FeTA Zurich dataset** | **0.248** | **0.158** | **0.919** | **0.248** | **2221922** | **1829.382** |
| **Std Dev FeTA Zurich dataset** | **0.166** | **0.129** | **0.066** | **0.166** | **609659.038** | **536.184** |
| **101** | 0.368 | 0.229 | 0.937 | 0.368 | 891648 | 673.62 |
| **102** | 0.517 | 0.368 | 0.867 | 0.517 | 1111545 | 639.346 |
| **103** | 0.377 | 0.238 | 0.918 | 0.377 | 1286692 | 953.762 |
| **104** | 0.609 | 0.446 | 0.959 | 0.609 | 1504914 | 805.276 |
| **105** | 0.415 | 0.269 | 0.909 | 0.415 | 1326351 | 934.37 |
| **106** | 0.578 | 0.455 | 0.792 | 0.578 | 1113119 | 473.859 |
| **107** | 0.104 | 0.056 | 0.744 | 0.104 | 938290 | 867.863 |
| **108** | 0.086 | 0.045 | 0.766 | 0.086 | 1358406 | 1277.951 |
| **109** | 0.032 | 0.017 | 0.673 | 0.032 | 1522629 | 1485.109 |
| **110** | 0.103 | 0.058 | 0.446 | 0.103 | 1994209 | 1735.146 |
| **Mean of FeTA Vienna dataset** | **0.319** | **0.218** | **0.801** | **0.319** | **1304780.3** | **984.63** |
| **Std Dev of FeTA Vienna dataset** | **0.22** | **0.169** | **0.156** | **0.22** | **324132.87** | **398.038** |
| **8010761051_2017_11_06_T2** | 0.215 | 0.121 | 0.968 | 0.215 | 1210305 | 1059.152 |
| **8010761051_2017_11_06_T1** | 0.026 | 0.013 | 0.449 | 0.026 | 2191211 | 2125.736 |
| **8010761051_2017_11_30_T2** | 0.032 | 0.016 | 0.873 | 0.032 | 1026491 | 1007.237 |
| **8010761051_2017_11_30_T1** | 0.047 | 0.026 | 0.255 | 0.047 | 696540 | 626.035 |
| **8011759516_2018_10_15_T2** | 0.144 | 0.078 | 0.985 | 0.144 | 1169149 | 1076.831 |
| **8011759516_2018_10_15_T1** | 0.015 | 0.008 | 0.413 | 0.015 | 1162114 | 1140.153 |
| **8011759516_2018_10_22_T2** | 0.017 | 0.009 | 0.752 | 0.017 | 998395 | 987.068 |
| **8011759516_2018_10_22_T1** | 0.01 | 0.005 | 0.144 | 0.01 | 2216827 | 2138.681 |
| **8013327626_2019_07_29_T2** | 0.071 | 0.037 | 0.998 | 0.071 | 1399122 | 1347.752 |
| **8013327626_2019_07_29_T1** | 0.024 | 0.012 | 0.394 | 0.024 | 613246 | 594.065 |
| **8013327626_2019_08_11_T2** | 0.088 | 0.046 | 0.983 | 0.088 | 1729378 | 1648.606 |
| **8013327626_2019_08_11_T1** | 0.015 | 0.008 | 0.114 | 0.015 | 2180998 | 2023.73 |
| **8017618189_2023_11_02_T2** | 0.13 | 0.07 | 0.947 | 0.13 | 1291398 | 1196.156 |
| **8017618189_2023_11_02_T1** | 0.011 | 0.006 | 0.101 | 0.011 | 1404630 | 1322.838 |
| **8017618189_2023_11_19_T2** | 0.07 | 0.036 | 0.999 | 0.07 | 1628054 | 1569.337 |
| **8017618189_2023_11_19_T1** | 0.145 | 0.089 | 0.383 | 0.145 | 3016505 | 2311.908 |
| **8017871237_2024_02_20_T2** | 0.301 | 0.178 | 0.966 | 0.301 | 1137867 | 928.09 |
| **8017871237_2024_02_20_T1** | 0.018 | 0.009 | 0.653 | 0.018 | 756053 | 745.206 |
| **8017871237_2024_02_26_T2** | 0.018 | 0.009 | 0.97 | 0.018 | 2787641 | 2760.989 |
| **8017871237_2024_02_26_T1** | 0.154 | 0.118 | 0.22 | 0.154 | 1387721 | 644.407 |
| **8017896414_2024_03_25_T2** | 0.333 | 0.201 | 0.981 | 0.333 | 1235162 | 982.163 |
| **8017896414_2024_03_25_T1** | 0.053 | 0.033 | 0.128 | 0.053 | 723064 | 535.38 |
| **8017896414_2024_03_29_T2** | 0.269 | 0.156 | 0.984 | 0.269 | 1190713 | 1001.832 |
| **8017896414_2024_03_29_T1** | 0.109 | 0.058 | 0.864 | 0.109 | 1971235 | 1838.874 |

Figure 51: Detailed scores of the SynthSeg model with pre-training, trained and tested over the three datasets on one label (ventricles)

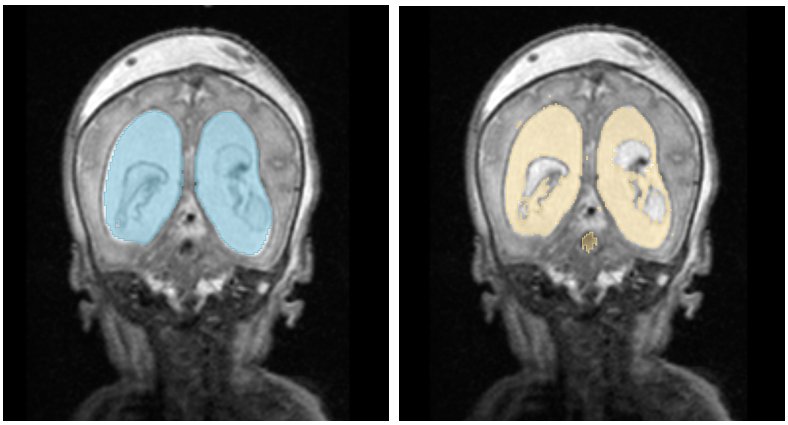

Figure 52: 8017896414_2024_03_29_T2 subject predicted segmentation (on the left) with its associated manual segmentation (on the right) for the 3D U-Net model trained over the FeTA datasets with two dataloaders and seven labels

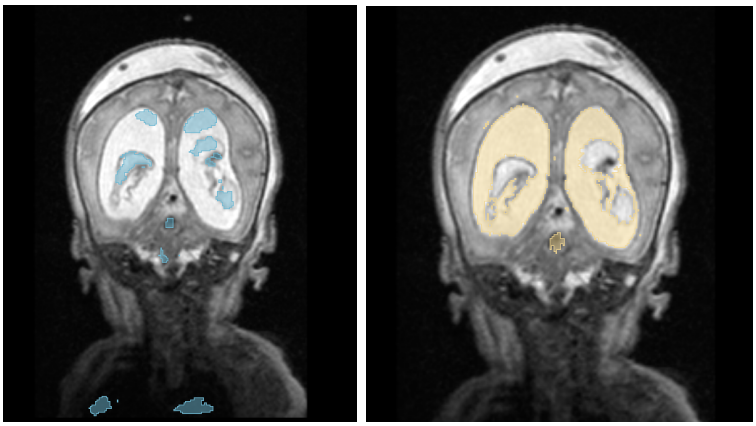

Figure 53: 8017896414_2024_03_29_T2 subject predicted segmentation (on the left) with its associated manual segmentation (on the right) for the SynthSeg model trained over the FeTA datasets with two dataloaders and seven labels

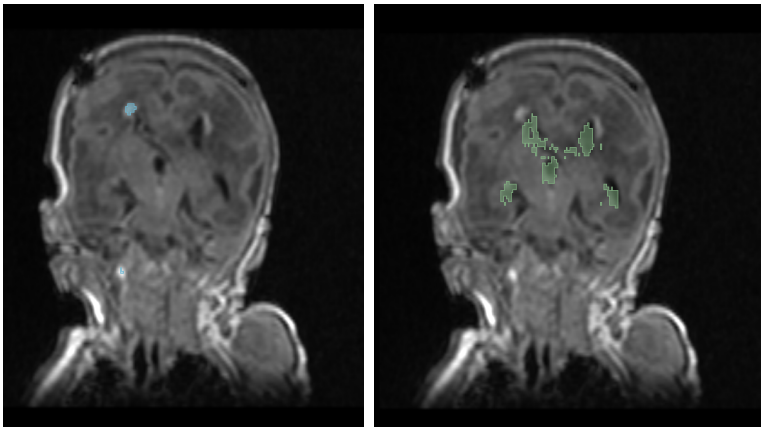

Figure 54: 8017871237_2024_02_26_T1 subject predicted segmentation (on the left) with its associated manual segmentation (on the right) for the 3D U-Net model trained over the FeTA datasets with two dataloaders and seven labels

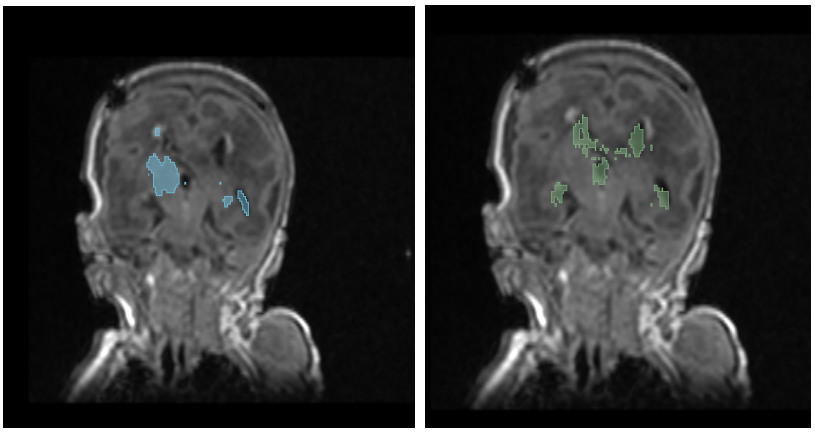

Figure 55: 8017871237_2024_02_26_T1 predicted segmentation (on the left) with its associated manual segmentation (on the right) for the SynthSeg model trained over the FeTA datasets with two dataloaders and seven labels

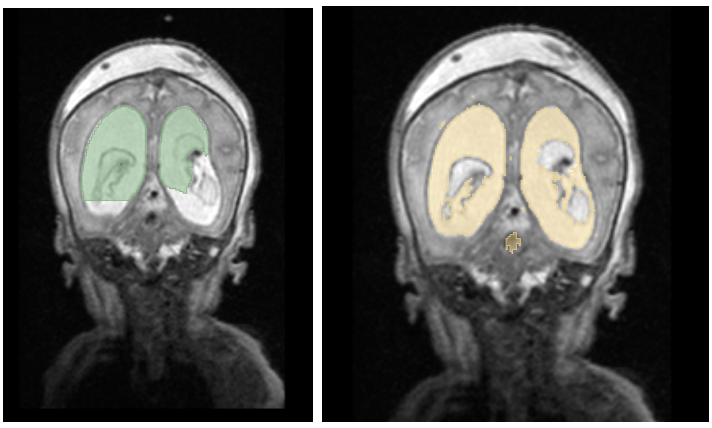

Figure 56: 8017896414_2024_03_29_T2 subject predicted segmentation (on the left) with its associated manual segmentation (on the right) for the 3D U-Net model trained over the FeTA datasets with two dataloaders and one label (ventricles)

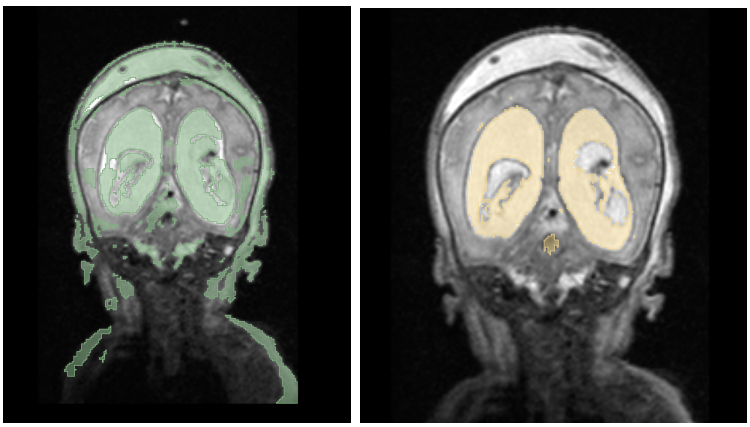

Figure 57: 8017896414_2024_03_29_T2 subject predicted segmentation (on the left) with its associated manual segmentation (on the right) for the SynthSeg model trained over the FeTA datasets with two dataloaders and one label (ventricles)

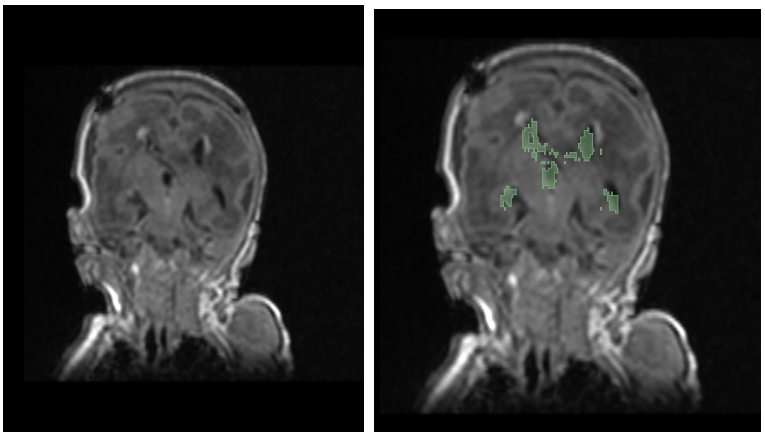

Figure 58: 8017871237_2024_02_26_T1 subject predicted segmentation (on the left) with its associated manual segmentation (on the right) for the 3D U-Net model trained over the FeTA datasets with two dataloaders and one label (ventricles)

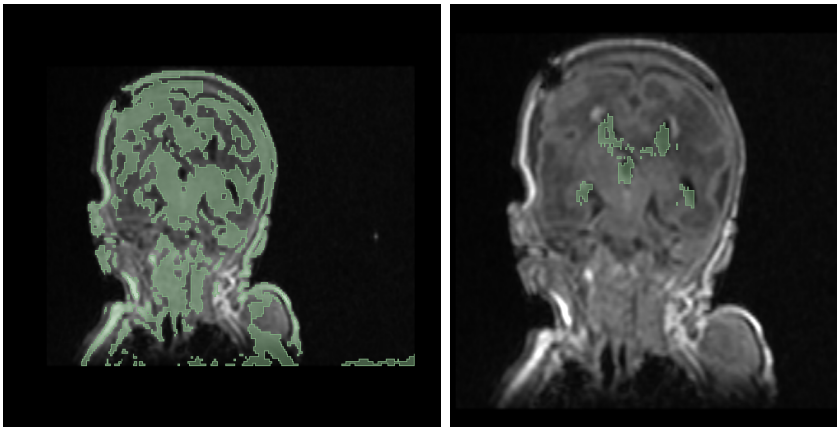

Figure 59: 8017871237_2024_02_26_T1 predicted segmentation (on the left) with its associated manual segmentation (on the right) for the SynthSeg model trained over the FeTA datasets with two dataloaders and one label (ventricles)

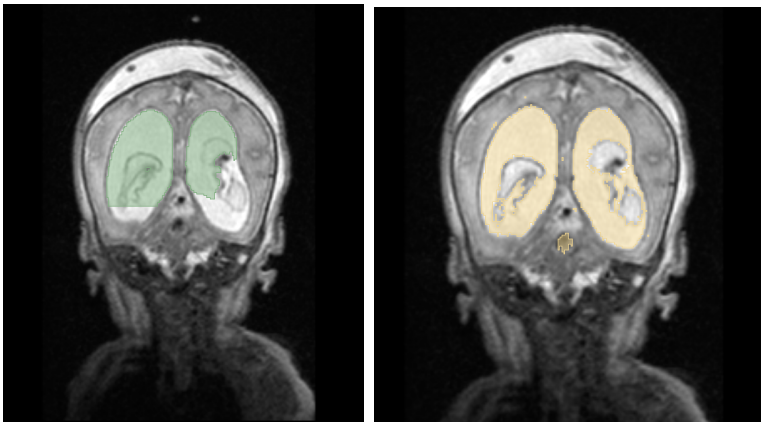

Figure 60: 8017896414_2024_03_29_T2 subject predicted segmentation (on the left) with its associated manual segmentation (on the right) for the 3D U-Net model trained over the three datasets with two dataloaders and one label (ventricles)

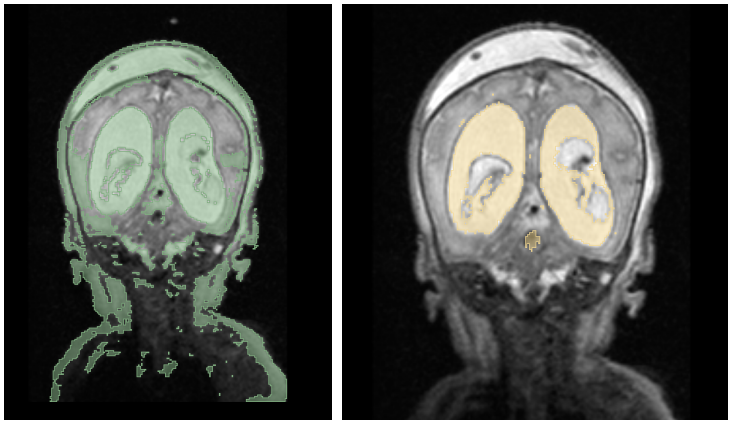

Figure 61: 8017896414_2024_03_29_T2 subject predicted segmentation (on the left) with its associated manual segmentation (on the right) for the SynthSeg model trained over the three datasets with two dataloaders and one label (ventricles)

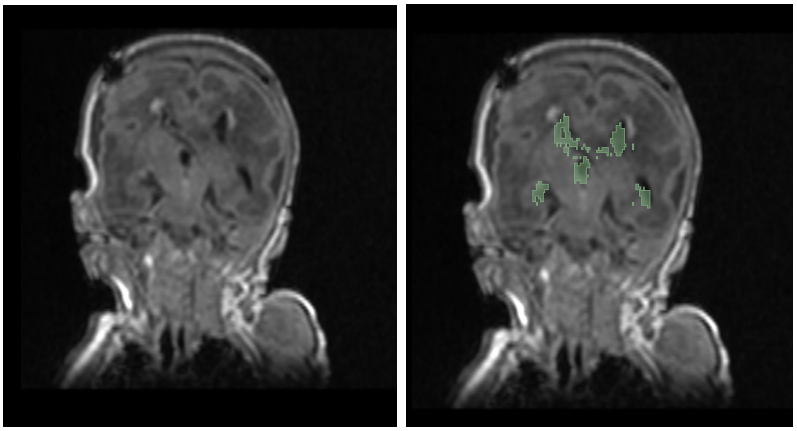

Figure 62: 8017871237_2024_02_26_T1 subject predicted segmentation (on the left) with its associated manual segmentation (on the right) for the 3D U-Net model trained over the three datasets with two dataloaders and one label (ventricles)

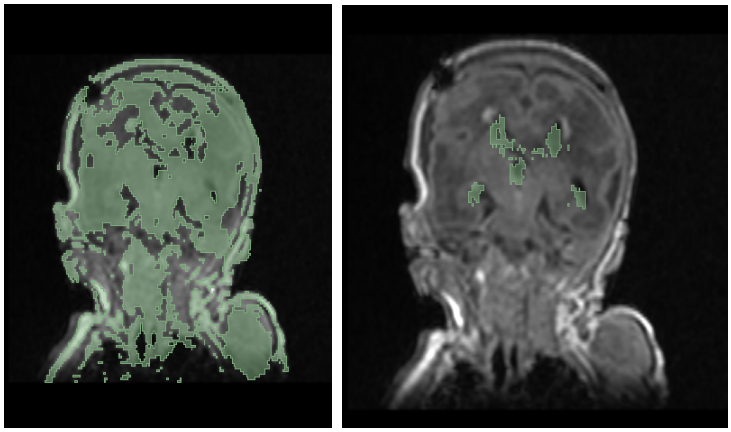

Figure 63: 8017871237_2024_02_26_T1 predicted segmentation (on the left) with its associated manual segmentation (on the right) for the SynthSeg model trained over the three datasets with two dataloaders and one label (ventricles)

# Acknowledgements

I thank Pr Isabelle Bloch for her excellent supervision and for enabling this project. Her impressive knowledge and her kindness were extremely precious during this project.

I also thank my work partener in this project Dr Sarah Stricker. Her expertise, determination and her sympathy helped a lot having instructive exchanges.

I am very much appreciate the collaboration with all the Imagine Institute's laboratory IMAG2: Giammarco La Barbera, Enzo Bonnot and Thomas Isla for their support, time and relevant advices.

I also want to thanks Joy-Rose Dunoyer de Segonzac and all the members of the Forme et Croissance du Crâne laboratory: Margaux, Maxime, Thomas, Ezra and all the others for the great moments shared.

A special thanks to the members of the LFI team of the LIP6 laboratory, especially Guillaume Gervois, Garance Martin, Yann Munro, Leandro Nascimento, Yuxuan Xi, Duane Fernandes Duarte, Ege Sendogan, Arun Nadaradjane, Yona Mellul, Christophe Marsala, Marie-Jeanne Lesot, Jean-Noël Vittaut and also other members of the 26-00 corridor such as Mélina Verger, Gauvain Bourgne and Camilo Sarmiento. Thank you for welcoming me in your incredible team, for your support and making me laugh.

I would like to extend my thanks to Andrea Pinna and Bertrand Granado for their trust and for giving me the opportunity to start a new chapter in the LIP6 team.

Last but not least I am very grateful to all my friends for their very much appreciated comments and support through my internship and more.

# References


[1] Pierre-Yves Ancel, François Goffinet, Pierre Kuhn, Bruno Langer, Jacqueline Matis, Xavier Hernandorena, Pierre Chabanier, Laurence Joly-Pedespan, Bénédicte Lecomte, Françoise Vendittelli, Michel Dreyfus, Bernard Guillois, Antoine Burguet, Pierre Sagot, Jacques Sizun, Alain Beuchée, Florence Rouget, Amélie Favreau, Elie Saliba, Nathalie Bednarek, Patrice Morville, Gérard Thiriez, Loïc Marpeau, Stéphane Marret, Gilles Kayem, Xavier Durrmeyer, Michèle Granier, Olivier Baud, Pierre-Henri Jarreau, Delphine Mitanchez, Pascal Boileau, Pierre Boulot, Gilles Cambonie, Hubert Daudé, Antoine Bédu, Fabienne Mons, Jeanne Fresson, Rachel Vieux, Catherine Alberge, Catherine Arnaud, Christophe Vayssière, Patrick Truffert, Véronique Pierrat, Damien Subtil, Claude D'Ercole, Catherine Gire, Umberto Simeoni, André Bongain, Loïc Sentilhes, Jean-Christophe Rozé, Jean Gondry, André Leke, Michel Deiber, Olivier Claris, Jean-Charles Picaud, Anne Ego, Thierry Debillon, Anne Poulichet, Eliane Coliné, Anne Favre, Olivier Fléchelles, Sylvain Samperiz, Duksha Ramful, Bernard Branger, Valérie Benhammou, Laurence Foix-L'Hélias, Laetitia Marchand-Martin, and Monique Kaminski. Survival and morbidity of preterm children born at 22 through 34 weeks' gestation in france in 2011: Results of the EPIPAGE-2 cohort study. 169(3):230.

[2] David F Bauer, Lissa C Baird, Paul Klimo, Catherine A Mazzola, Dimitrios C Nikas, Mandeep S Tamber, and Ann Marie Flannery. Congress of neurological surgeons systematic review and evidence-based guidelines on the treatment of pediatric hydrocephalus: Update of the 2014 guidelines. 87(6):1071–1075.

[3] Oualid M. Benkarim, Gerard Sanroma, Veronika A. Zimmer, Emma Muñoz-Moreno, Nadine Hahner, Elisenda Eixarch, Oscar Camara, Miguel Angel González Ballester, and Gemma Piella. Toward the automatic quantification of in utero brain development in 3d structural MRI: A review. 38(5):2772–2787.

[4] Benjamin Billot, Douglas Greve, Koen Van Leemput, Bruce Fischl, Juan Eugenio Iglesias, and Adrian V. Dalca. A learning strategy for contrast-agnostic MRI segmentation.

[5] Benjamin Billot, Douglas N. Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V. Dalca, and Juan Eugenio Iglesias. SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining. 86:102789.

[6] Aj Brouwer, Mj Brouwer, F Groenendaal, Mjnl Benders, A Whitelaw, and Ls De Vries. European perspective on the diagnosis and treatment of posthaemorrhagic ventricular dilatation. 97(1):F50–F55.

[7] Agisilaos Chartsias, Thomas Joyce, Mario Valerio Giuffrida, and Sotirios A. Tsaftaris. Multimodal MR synthesis via modality-invariant latent representation. 37(3):803–814.

[8] Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng-Ann Heng. Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. 33(1):865–872.

[9] Tommaso Ciceri, Luca Casartelli, Florian Montano, Stefania Conte, Letizia Squarcina, Alessandra Bertoldo, Nivedita Agarwal, Paolo Brambilla, and Denis Peruzzo. Fetal brain MRI atlases and datasets: A review. 292:120603.

[10] Michael C. Dewan, Abbas Rattani, Rania Mekary, Laurence J. Glancz, Ismaeel Yunusa, Ronnie E. Baticulon, Graham Fieggen, John C. Wellons, Kee B. Park, and Benjamin C.

Warf. Global hydrocephalus epidemiology and incidence: systematic review and meta-analysis. 130(4):1065–1079.

[11] DOLPHIN-UK Collaborators and Saniya Mediratta. A standardised protocol for neuroendoscopic lavage for post-haemorrhagic ventricular dilatation: A delphi consensus approach.

[12] A. David Edwards, Daniel Rueckert, Stephen M. Smith, Samy Abo Seada, Amir Alansary, Jennifer Almalbis, Joanna Allsop, Jesper Andersson, Tomoki Arichi, Sophie Arulkumaran, Matteo Bastiani, Dafnis Batalle, Luke Baxter, Jelena Bozek, Eleanor Braithwaite, Jacqueline Brandon, Olivia Carney, Andrew Chew, Daan Christiaens, Raymond Chung, Kathleen Colford, Lucilio Cordero-Grande, Serena J. Counsell, Harriet Cullen, John Cupitt, Charles Curtis, Alice Davidson, Maria Deprez, Louise Dillon, Konstantina Dimitrakopoulou, Ralica Dimitrova, Eugene Duff, Shona Falconer, Seyedeh-Rezvan Farahibozorg, Sean P. Fitzgibbon, Jianliang Gao, Andreia Gaspar, Nicholas Harper, Sam J. Harrison, Emer J. Hughes, Jana Hutter, Mark Jenkinson, Saad Jbabdi, Emily Jones, Vyacheslav Karolis, Vanessa Kyriakopoulou, Gregor Lenz, Antonios Makropoulos, Shaihan Malik, Luke Mason, Filippo Mortari, Chiara Nosarti, Rita G. Nunes, Camilla O'Keeffe, Jonathan O'Muircheartaigh, Hamel Patel, Jonathan Passerat-Palmbach, Maximillian Pietsch, Anthony N. Price, Emma C. Robinson, Mary A. Rutherford, Andreas Schuh, Stamatios Sotiropoulos, Johannes Steinweg, Rui Pedro Azeredo Gomes Teixeira, Tencho Tenev, Jacques-Donald Tournier, Nora Tusor, Alena Uus, Katy Vecchiato, Logan Z. J. Williams, Robert Wright, Julia Wurie, and Joseph V. Hajnal. The developing human connectome project neonatal data release. 16:886772.

[13] Mohamed El-Dib, David D. Limbrick, Terrie Inder, Andrew Whitelaw, Abhaya V. Kulkarni, Benjamin Warf, Joseph J. Volpe, and Linda S. De Vries. Management of post-hemorrhagic ventricular dilatation in the infant born preterm. 226:16–27.e3.

[14] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks.

[15] Ali Gholipour, Alireza Akhondi-Asl, Judy A. Estroff, and Simon K. Warfield. Multi-atlas multi-shape segmentation of fetal brain MRI for volumetric and morphometric analysis of ventriculomegaly. 60(3):1819–1831.

[16] Ali Gholipour, Caitlin K. Rollins, Clemente Velasco-Annis, Abdelhakim Ouaalam, Alireza Akhondi-Asl, Onur Afacan, Cynthia M. Ortinau, Sean Clancy, Catherine Limperopoulos, Edward Yang, Judy A. Estroff, and Simon K. Warfield. A normative spatiotemporal MRI atlas of the fetal brain for automatic segmentation and analysis of early brain growth. 7(1):476.

[17] Maximilian Ilse, Jakub M. Tomczak, and Patrick Forré. Selecting data augmentation for simulating interventions. In *Proceedings of the 38th International Conference on Machine Learning*, pages 4555–4562. PMLR.

[18] Terrie E. Inder, Linda S. De Vries, Donna M. Ferriero, P. Ellen Grant, Laura R. Ment, Steven P. Miller, and Joseph J. Volpe. Neuroimaging of the preterm brain: Review and recommendations. 237:276–287.e4.

[19] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-net: a self-configuring method for deep learning-based biomedical image segmentation. 18(2):203–211.

[20] Konstantinos Kamnitsas, Christian Baumgartner, Christian Ledig, Virginia F. J. Newcombe, Joanna P. Simpson, Andrew D. Kane, David K. Menon, Aditya Nori, Antonio Criminisi, Daniel Rueckert, and Ben Glocker. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks.

[21] Neerav Karani, Krishna Chaitanya, Christian Baumgartner, and Ender Konukoglu. A lifelong learning approach to brain MR segmentation across scanners and protocols.

[22] I. Kostović, G. Sedmak, and M. Judaš. Neural histology and neurogenesis of the human fetal and infant brain. 188:743–773.

[23] Grace Y. Lai, William Chu-Kwan, Annie B. Westcott, Abhaya V. Kulkarni, James M. Drake, and Sandi K. Lam. Timing of temporizing neurosurgical treatment in relation to shunting and neurodevelopmental outcomes in posthemorrhagic ventricular dilatation of prematurity: A meta-analysis. 234:54–64.e20.

[24] Gang Li, Li Wang, Pew-Thian Yap, Fan Wang, Zhengwang Wu, Yu Meng, Pei Dong, Jaeil Kim, Feng Shi, Islem Rekik, Weili Lin, and Dinggang Shen. Computational neuroanatomy of baby brains: A review. 185:906–925.

[25] Karen Luyt, Sally L Jary, Charlotte L Lea, Grace J. Young, David E Odd, Helen E Miller, Grazyna Kmita, Cathy Williams, Peter S Blair, William Hollingworth, Michelle Morgan, Adam P Smith-Collins, Steven Walker-Cox, Kristian Aquilina, Ian Pople, and Andrew G Whitelaw. Drainage, irrigation and fibrinolytic therapy (DRIFT) for posthaemorrhagic ventricular dilatation: 10-year follow-up of a randomised controlled trial. 105(5):466–473.

[26] Antonios Makropoulos, Serena J. Counsell, and Daniel Rueckert. A review on automatic fetal and neonatal brain MRI segmentation. 170:231–248.

[27] Antonios Makropoulos, Ioannis S. Gousias, Christian Ledig, Paul Aljabar, Ahmed Serag, Joseph V. Hajnal, A. David Edwards, Serena J. Counsell, and Daniel Rueckert. Automatic whole brain MRI segmentation of the developing neonatal brain. 33(9):1818–1831.

[28] Antonios Makropoulos, Emma C. Robinson, Andreas Schuh, Robert Wright, Sean Fitzgibbon, Jelena Bozek, Serena J. Counsell, Johannes Steinweg, Katy Vecchiato, Jonathan Passerat-Palmbach, Gregor Lenz, Filippo Mortari, Tencho Tenev, Eugene P. Duff, Matteo Bastiani, Lucilio Cordero-Grande, Emer Hughes, Nora Tusor, Jacques-Donald Tournier, Jana Hutter, Anthony N. Price, Rui Pedro A.G. Teixeira, Maria Murgasova, Suresh Victor, Christopher Kelly, Mary A. Rutherford, Stephen M. Smith, A. David Edwards, Joseph V. Hajnal, Mark Jenkinson, and Daniel Rueckert. The developing human connectome project: A minimal processing pipeline for neonatal cortical surface reconstruction. 173:88–112.

[29] Samuel G. McClugage, Nicholas M. B. Laskay, Brian N. Donahue, Anastasia Arynchyna, Kathrin Zimmerman, Inmaculada B. Aban, Elizabeth N. Alford, Myriam Peralta-Carcelen, Jeffrey P. Blount, Curtis J. Rozzelle, James M. Johnston, and Brandon G. Rocque. Functional outcomes at 2 years of age following treatment for posthemorrhagic hydrocephalus of prematurity: what do we know at the time of consult? 25(5):453–461.

[30] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571. IEEE.

[31] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. 22(10):1345–1359.

[32] Lu-Ann Papile, Jerome Burstein, Rochelle Burstein, and Herbert Koffler. Incidence and evolution of subependymal and intraventricular hemorrhage: A study of infants with birth weights less than 1,500 gm. 92(4):529–534.

[33] Kelly Payette, Priscille De Dumast, Hamza Kebiri, Ivan Ezhov, Johannes C. Paetzold, Suprosanna Shit, Asim Iqbal, Romesa Khan, Raimund Kottke, Patrice Grehten, Hui Ji, Levente Lanczi, Marianna Nagy, Monika Beresova, Thi Dao Nguyen, Giancarlo Natalucci, Theofanis Karayannis, Bjoern Menze, Meritxell Bach Cuadra, and Andras Jakab. An automatic multi-tissue human fetal brain segmentation benchmark using the fetal tissue annotation dataset. 8(1):167.

[34] Kelly Payette, Hongwei Bran Li, Priscille De Dumast, Roxane Licandro, Hui Ji, Md Mahfuzur Rahman Siddiquee, Daguang Xu, Andriy Myronenko, Hao Liu, Yuchen Pei, Lisheng Wang, Ying Peng, Juanying Xie, Huiquan Zhang, Guiming Dong, Hao Fu, Guotai Wang, ZunHyan Rieu, Donghyeon Kim, Hyun Gi Kim, Davood Karimi, Ali Gholipour, Helena R. Torres, Bruno Oliveira, João L. Vilaça, Yang Lin, Netanell Avisdris, Ori Ben-Zvi, Dafna Ben Bashat, Lucas Fidon, Michael Aertsen, Tom Vercauteren, Daniel Sobotka, Georg Langs, Mireia Alenyà, Maria Inmaculada Villanueva, Oscar Camara, Bella Specktor Fadida, Leo Joskowicz, Liao Weibin, Lv Yi, Li Xuesong, Moona Mazher, Abdul Qayyum, Domenec Puig, Hamza Kebiri, Zelin Zhang, Xinyi Xu, Dan Wu, Kuanlun Liao, Yixuan Wu, Jintai Chen, Yunzhi Xu, Li Zhao, Lana Vasung, Bjoern Menze, Meritxell Bach Cuadra, and Andras Jakab. Fetal brain tissue annotation and segmentation challenge results. 88:102833.

[35] Philippa Rees, Caitriona Callan, Karan R. Chadda, Meriel Vaal, James Diviney, Shahad Sabti, Fergus Harnden, Julian Gardiner, Cheryl Battersby, Chris Gale, and Alastair Sutcliffe. Preterm brain injury and neurodevelopmental outcomes: A meta-analysis. 150(6):e2022057442.

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation.

[37] Sarah Stricker, Raphael Guzman, Thomas Blauwblomme, and Moise Danielpour. Is the choroid plexus needed? 57(5):301–305.

[38] Ulrich-Wilhelm Thomale, Giuseppe Cinalli, Abhaya V. Kulkarni, Sara Al-Hakim, Jonathan Roth, Andreas Schaumann, Christoph Bührer, Sergio Cavalheiro, Spyros Sgouros, Shlomi Constantini, and Hans Christoph Bock. TROPHY registry study design: a prospective, international multicenter study for the surgical treatment of posthemorrhagic hydrocephalus in neonates. 35(4):613–619.

[39] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world.

[40] Devavrat Tomar, Behzad Bozorgtabar, Manana Lortkipanidze, Guillaume Vray, Mohammad Saeed Rad, and Jean-Philippe Thiran. Self-supervised generative style transfer for one-shot medical image segmentation.

[41] Sébastien Tourbier, Xavier Bresson, Patric Hagmann, Jean-Philippe Thiran, Reto Meuli, and Meritxell Bach Cuadra. An efficient total variation algorithm for super-resolution in fetal brain MRI with adaptive regularization. 118:584–597.

[42] Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Training deep networks with synthetic data: Bridging the reality gap by domain randomization.

[43] Andrea Urru, Ayako Nakaki, Oualid Benkarim, Francesca Crovetto, Laura Segalés, Valentin Comte, Nadine Hahner, Elisenda Eixarch, Eduard Gratacos, Fàtima Crispi, Gemma Piella, and Miguel A. González Ballester. An automatic pipeline for atlas-based fetal and neonatal brain segmentation and analysis. 230:107334.

[44] R Valabregue, F Girka, A Pron, F Rousseau, and G Auzias. Comprehensive analysis of synthetic learning applied to neonatal brain MRI segmentation.

[45] Andrew Whitelaw, David Evans, Michael Carter, Marianne Thoresen, Jolanta Wroblewska, Marek Mandera, Janusz Swietlinski, Judith Simpson, Constantinos Hajivassiliou, Linda P. Hunt, and Ian Pople. Randomized clinical trial of prevention of hydrocephalus after intraventricular hemorrhage in preterm infants: Brain-washing versus tapping fluid. 119(5):e1071–e1078.

[46] Andrew Whitelaw, Sally Jary, Grazyna Kmita, Jolanta Wroblewska, Ewa Musialik-Swietlinska, Marek Mandera, Linda Hunt, Michael Carter, and Ian Pople. Randomized trial of drainage, irrigation and fibrinolytic therapy for premature infants with posthemorrhagic ventricular dilatation: Developmental outcome at 2 years. 125(4):e852–e858.

[47] Zizhao Zhang, Lin Yang, and Yefeng Zheng. Translating and segmenting multimodal medical volumes with cycle- and shape-consistency generative adversarial network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9242–9251. IEEE.

[48] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. pages 1–20.