NeuroImage 308 (2025) 121063



Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg





DeepCERES: A deep learning method for cerebellar lobule segmentation using ultra-high resolution multimodal MRI

Sergio Morell-Ortega [a,*], Marina Ruiz-Perez [a], Marien Gadea [b], Roberto Vivo-Hernando [c], Gregorio Rubio [d], Fernando Aparici [e], Maria de la Iglesia-Vaya [f], Gwenaelle Catheline [g], Boris Mansencal [h], Pierrick Coupé [h], José V. Manjón [a]

[a] Instituto de Aplicaciones de las Tecnologías de la Información y de las Comunicaciones Avanzadas (ITACA), Universitat Politècnica de València, Camino de Vera s/n, 46022, Valencia, Spain
[b] Department of Psychobiology, Faculty of Psychology, Universitat de Valencia, Valencia, Spain
[c] Instituto de Automática e Informática Industrial, Universitat Politècnica de València, Camino de Vera s/n, 46022, Valencia, Spain
[d] Departamento de matemática aplicada, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain
[e] Área de Imagen Médica. Hospital Universitario y Politécnico La Fe. Valencia, Spain
[f] Unidad Mixta de Imagen Biomédica FISABIO-CIPF. Fundación para el Fomento de la Investigación Sanitario y Biomédica de la Comunidad Valenciana - Valencia, Spain
[g] Univ. Bordeaux, CNRS, UMR 5287, Institut de Neurosciences Cognitives et Intégratives d'Aquitaine, Bordeaux, France
[h] CNRS, Univ. Bordeaux, Bordeaux INP, LABRI, UMR5800, in2brain, F-33400 Talence, France

ABSTRACT

This paper introduces a novel multimodal and high-resolution human brain cerebellum lobule segmentation method. Unlike current tools that operate at standard resolution (1 mm^3) or using mono-modal data, the proposed method improves cerebellum lobule segmentation through the use of a multimodal and ultra-high resolution (0.125 mm^3) training dataset. To develop the method, first, a database of semi-automatically labelled cerebellum lobules was created to train the proposed method with ultra-high resolution T1 and T2 MR images. Then, an ensemble of deep networks has been designed and developed, allowing the proposed method to excel in the complex cerebellum lobule segmentation task, improving precision while being memory efficient. Notably, our approach deviates from the traditional U-Net model by exploring alternative architectures. We have also integrated deep learning with classical machine learning methods incorporating a priori knowledge from multi-atlas segmentation which improved precision and robustness. Finally, a new online pipeline, named DeepCERES, has been developed to make available the proposed method to the scientific community requiring as input only a single T1 MR image at standard resolution.

1. Introduction

The modern age in the anatomical study of the cerebellum started 120 years ago when Santiago Ramón y Cajal published his first paper with Golgi-impregnated material. His pioneering work marked the inception of an extensive effort to unravel the intricate organization of the central nervous system (Sotelo, 2008). Despite representing a small portion of the total volume of the brain, the cerebellum itself contains approximately 50% of all the neurons in the brain (Laidi et al., 2015), allowing it to play a crucial role in cognitive, emotional, and behavioural functions, as well as its traditional role in movement coordination. Macroscopically, the cerebellum comprises two cerebellar hemispheres connected by a narrow median vermis. It is connected to the posterior aspect of the brainstem by three symmetrical bundles of nerve fibres called the superior, middle, and inferior cerebellar peduncles. The Schmahmann atlas (Schmahmann et al., 2000) presents a standardized nomenclature and unifying terminology. This atlas hierarchically differentiates the regions of the cerebellum into groups by its fissures and then into individual folds called lobules, which are identified by Roman numerals from I to X.

In the last decades, clinical observations and neuroimaging studies have demonstrated the importance of the cerebellum and its communication with the cerebral cortex in humans (Desmond and Fiez, 1998; King et al., 2019). In 1998 it was published a first description of the so-called cerebellar cognitive-affective syndrome (Schmahmann and Sherman, 1998), a constellation of cognitive and behavioural impairments such as problems with abstract reasoning and emotional control. This condition, which some now refer to as Schmahmann's syndrome, helped establish an appreciation of the cerebellum's role beyond coordinating movement. Nowadays we know its role has also been evidenced in cognitive operations such as learning (Guell et al., 2018), memory (Desmond and Fiez, 1998), language (Mariën et al., 2014), and emotional behaviour (Adamaszek et al., 2016). Hence, by quantifying the structure and function of the cerebellum, researchers

* Corresponding author.
E-mail address: sermoor1@teleco.upv.es (S. Morell-Ortega).

https://doi.org/10.1016/j.neuroimage.2025.121063
Received 22 May 2024; Received in revised form 27 January 2025; Accepted 27 January 2025
Available online 6 February 2025
1053-8119/© 2025 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

gain valuable insights into the role it plays in both motor control and cognitive processes.

The field of its study has been and remains very active. Recent research has highlighted that cerebellar lesions' have profound a impact on motor and cognitive functions. Lesions in the sensorimotor cerebellar lobules, particularly the anterior lobule, can lead to movement dysmetria, as seen in cerebellar motor syndrome. Conversely, damage to the cognitive-emotional cerebellar lobules in the posterior lobule can result in dysmetria of thought and emotion, giving rise to cerebellar cognitive affective syndrome. This linkage between the cerebellum's roles in motor control and cognition advances our understanding of cognitive mechanisms and holds promise for therapeutic interventions in behavioural neurology and neuropsychiatry, potentially benefiting conditions like schizophrenia or autism (Schmahmann et al., 2019).

In addition to applications in motor impairments such as cerebellar motor syndrome (Schmahmann et al., 2019), quantification of brain volumes has also been employed in other diseases such as spinocerebellar ataxia or amyotrophic lateral sclerosis (Chandrasekaran et al., 2022; Bede et al., 2021). Progress has been made in understanding the involvement of the cerebellum in particular mechanisms of cognition and behavioural and neuropsychiatric neurology, with studies on autism (Laidi et al., 2022), epilepsy (Warren et al., 2022), schizophrenia or bipolar disorder (Laidi et al., 2019). Its implication in dementia has also been studied by examining cerebellar cortical thickness (McKenna et al., 2021) or alterations in white and grey matter in frontotemporal dementia (Clouston et al., 2022).

Therefore, there is no doubt about the importance and impact of the precise quantification of cerebellum function and anatomy. Specifically, the accurate estimation of cerebellum volumetry through magnetic resonance imaging (MRI) has become an important research line in the last years. Although manual segmentation is considered the gold standard (Huo et al., 2019), it has several drawbacks, such as the inter-rater variability (Bogovic et al., 2013), the expertise required or the time needed to complete the task (Carass et al., 2018). These limitations have led to the need to design semi-automatic or automatic techniques for cerebellum segmentation.

One of the first fully automatic cerebellum segmentation methods was the Spatially Unbiased Infra-tentorial Template (SUIT) (Diedrichsen, 2006). It was based on the nonlinear registration of a probabilistic cerebellum atlas to the case to be segmented. Similarly, the Multiple Automatically Generated Templates (MAGeT) brain segmentation algorithm (Park et al., 2014) relied on creating a library of templates nonlinearly registered the target image. The final segmentation was achieved by merging multiple segmentations based on a majority voting scheme.

With a slightly different perspective, RASCAL (Weier et al., 2014) used an approach based on multi-atlas non-local patch-based label fusion method (Coupé et al., 2011) using a priori information via majority voting for label fusion and nonlinear registration. Similarly, another multi-atlas-based method called CERES (Romero et al., 2017a) was proposed using an ultra-fast patch-matching technique called Optimized PatchMatch ALgorithm (OPAL) (Giraud et al., 2016). An incremental version of this method was later proposed, CERES2 (Carass et al., 2018) which improved the results of CERES by adding a patch-based neural network method for systematic error correction named PEC (Patch-based Ensemble Corrector) (Romero et al., 2017b).

A review paper on cerebellum segmentation methods was published in 2018 (Carass et al., 2018) summarizing the results of the cerebellum segmentation challenge of the MICCAI2017 conference, where CERES2 method was found to be the best-performing method in all categories. It is worth noting that competing methods in this challenge also included Deep Learning methods such as LiviaNET (Dolz et al., 2018) and two others U-Net-based (Ronneberger et al., 2015) convolutional neural networks.

More recently, new methods for cerebellum segmentation based on deep learning have been proposed. One of them is named ACAPULCO (Han et al., 2020), an acronym for Anatomical Parcellation using a U-Net with Locally Constrained Optimization. This method uses a two-step deep learning strategy with two 3D convolutional neural networks (CNNs), one to localize the cerebellum and another to segment it. This method was capable of achieving state-of-the-art results yielding similar accuracy that the state-of-the-art method CERES2. Finally, the most recent method for cerebellum lobule segmentation is CerebNet (Faber et al., 2022), which combines the architecture of FastSurferCNN (Henschel et al., 2020), a 2.5D segmentation network based on U-Net, with an extensive data augmentation approach (i.e. realistic nonlinear deformations to increase anatomical variability) eliminating the need of preprocessing steps, such as spatial normalization or bias field correction.

All the methods described above have two significant limitations. Firstly, they exclusively operate with mono-modal data, typically T1, owing to its good contrast between cerebellum grey and white matter. However, this can occasionally result in false positives in peri-cerebellar veins and meninges. Secondly, they all operate at standard resolution, typically no greater than 1 mm^3, which appears inadequate due to the intricately convoluted nature of the cerebellum and the presence of partial volume artefacts at this resolution.

Therefore, in this paper, we propose a new cerebellum lobule segmentation method that uses multimodal data (T1 and T2-weighted images) to avoid false positives by adding extra information in the classification process, and that works at ultra-high resolution (0.125 mm^3, i.e. $0.5 \times 0.5 \times 0.5$ mm^3) to improve the segmentation accuracy.

Finally, a new online pipeline, named DeepCERES, has been developed to make available the proposed segmentation method to the scientific community requiring as input only a single T1 MR image at standard resolution.

2. Material and methods

For the development of the proposed method two MRI datasets were used. The first one was a subset of the Human Connectome Project (HCP) database (Van Essen et al., 2012), including multimodal T1 and T2 images. Specifically, we randomly selected 75 subjects (from the nearly 1200 dataset images). Those images consisted of T1 and T2 MRI volumes with dimensions of $260 \times 311 \times 260$ voxels and a resolution of 0.343 mm^3 (0.7 mm isotropic resolution). These subjects ranged from 22 to 36 years old (41 women and 34 men). This HCP dataset was used to train our proposed method after a semiautomatic labelling process described in the next section.

The second dataset consisted of 4857 T1 weighted MRI subjects from a recent lifespan paper (Coupé et al., 2017) from various publicly available datasets. This dataset was used for data augmentation to increase the training data variability by including individuals from early infancy to old age. These databases are the following:

- NDAR (N = 493): The Database for Autism Research (NDAR) is a national database funded by NIH dataset. This database included samples of different cohorts acquired on 1.5T MRI and 3T scanners.
- ABIDE (N = 905): The images from the Autism Brain Imaging Data Exchange (ABIDE) dataset were acquired at 20 different sites on 3T scanner.
- ICBM (N = 294): The images from the International Consortium for Brain Mapping (ICBM) dataset were obtained through the LONI website.
- OASIS (N = 393): The subject images come from the Open Access Series of Imaging Studies (OASIS) dataset.
- IXI (N = 549): The images from the Information eXtraction from Images (IXI) dataset consist of normal subjects from 1.5 and 3T scanners.
- ADNI (ADNI1 and ADNI2) (N = 1649): The images from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset consisted of subjects from the 1.5T and 3 T MR in collection 1 and 2. These images were acquired at different sites across the United States and Canada.

Table 1
Databases description and demographics. (CN: Cognitive normal, ASD: Autism spectrum disorder, AD: Alzheimer's disease, EMCI: early mild cognitive impairment, LMCI: Late mild cognitive impairment, MCI: mild cognitive impairment, SMC: subject memory complaints).

Data base	Gender	N	Age	Diagnostic						
				CN	ASD	AD	EMCI	LMCI	MCI	SMC
HCP	Male	34	26,59 ± 3,28	34						
	Female	41	27,93 ± 3,66	41						
ABIDE	Male	773	17,95 ± 8,39	408	365					
	Female	132	16,13 ± 7,67	84	48					
ADNI	Male	907	74 ± 7	201		181	144	91	246	44
	Female	742	72 ± 7	203		151	113	77	137	61
AIBL	Male	160	72,4 ± 6,96	112		18			30	
	Female	178	74,19 ± 7,74	120		29			29	
ICBM	Male	152	32,2 ± 11,91	152						
	Female	142	35,4 ± 16,4	142						
IRC0	Male	86	9,68 ± 4,4	86						
	Female	108	7,81 ± 4,63	108						
IXI	Male	242	46,97 ± 16,56	242						
	Female	307	50,18 ± 16,31	307						
NDAR	Male	303	13,29 ± 6,18	208	95					
	Female	190	12,21 ± 5,1	174	16					
OASIS	Male	150	63,91 ± 11,76	111		16			23	
	Female	243	67,56 ± 13,2	187		29			27	
UCLA	Male	21	7,75 ± 2,42	21						
	Female	21	7,4 ± 3,04	21						

- AIBL (N = 338) The Australian Imaging, Biomarkers and Lifestyle (AIBL) dataset. The imaging protocol was defined to follow ADNI's guideline on the 3T scanner and a MPRAGE sequence on the 1.5T scanner.
- C-MIND (N = 236): All the images of the dataset were acquired at the same site on a 3T scanner. The MRI are 3D T1-weighted MPRAGE high-resolution anatomical scan of the entire brain with a spatial resolution of 1 mm^3.

Table 1 summarizes the whole dataset in terms of the number of samples and the demographics of each database.

2.1. Ethics statement

This manuscript uses de-identified MRI scans from publicly available databases. All participants provided written informed consent according to the enrolment protocols of their respective open data sources. To maintain privacy and confidentiality, no personal identifying information was included in the analysis.

2.2. Ultra-high resolution cerebellum lobule labelling

To create our own library with the corresponding ground truth segmentations, we used the CERES method (Romero et al., 2017a), considering it one of the best automatic methods based on the comparison of Carass et al. (2018). It is worth noting that the CERES method's accuracy was reported to be close to the intra-ratter accuracy.

To generate the HCP data segmentations, the images underwent the specific preprocessing steps of CERES, which are part of the preprocessing in the proposed method, as we will describe later. This preprocessing aims to locate the images in a standardized geometric and intensity space. The preprocessing steps are as follows. First, noise reduction in the native space is performed using the Spatially Adaptive Non-Local Means Filter method (Manjón et al., 2010), followed by an inhomogeneity correction using the N4 method (Tustison et al., 2010). Then, an affine registration to the MNI152 space at a 0.125 mm^3 resolution is carried out using the ITKElastix tool (Ntatsis et al., 2023; Shamonin et al., 2014) to put the volumes into a common coordinate space. Next, a second inhomogeneity correction was applied only to the brain using the MNI152 intracranial mask using the N4 method. Lastly, a final step involves cropping the cerebellum area to reduce computational requirements. It is worth noting that this process was applied to both T1 and T2 images of the dataset (although only the T1 were used within the CERES method).

After applying all of these preprocessing steps, the resulting volume had dimensions of 252 × 156 × 162 voxels with a resolution of 0.125 mm^3. However, CERES method operates at a resolution of 1 mm^3. To be able to apply CERES method we employed a stride volume decomposition technique (Manjón et al., 2020). This approach is based on the decomposition into 8 volumes at a 1 mm^3 resolution of a single 0.125 mm^3 volume by sampling the 3D volume eight times with a step size of 2 voxels in each dimension with eight different offsets. Subsequently, these 8 volumes were segmented, and finally, the resulting segmentations were mapped back to the original 0.125 mm^3 space by reversing the stride operation.

The resulting segmentations had an overall good quality but suffered from three types of errors. First, CSF voxels within neighbour lobules were misclassified as GM since, at 1 mm^3 resolution, those voxels are affected by partial volume effects and typically not assigned to CSF tissue. Second, similarly, cerebellar white matter (also known as the "tree of life" or "*arbor vitae*" by anatomists due to its branching structure) was underestimated close to the cerebellar cortex again due to partial volume problems. Finally, some lobule parts were misclassified due to other reasons. The first two problems were partially solved using intensity information of both T1 and T2 images, adding voxels with intensities coherent with the corresponding tissue in each case (CSF or WM). The remaining labelling errors were manually corrected by an expert using ITK-SNAP software (Yushkevich et al., 2006). We also manually removed the cerebellar peduncles from the WM label as we do not consider them as part of the cerebellum. An example of the semiautomatic correction described is shown in Fig. 1.

As a result of the labelling process, a training dataset of 75 T1 and T2 images at 0.125 mm^3 MNI space with their corresponding labels was generated. To double the size of the training library, the left–right mirrored versions were also included, resulting in a final library of 150 cases. This library was divided into 130 cases for training, 10 cases for validation and 10 cases for test (the data split was done subject-wise so for each subset each case and its mirrored version was included).

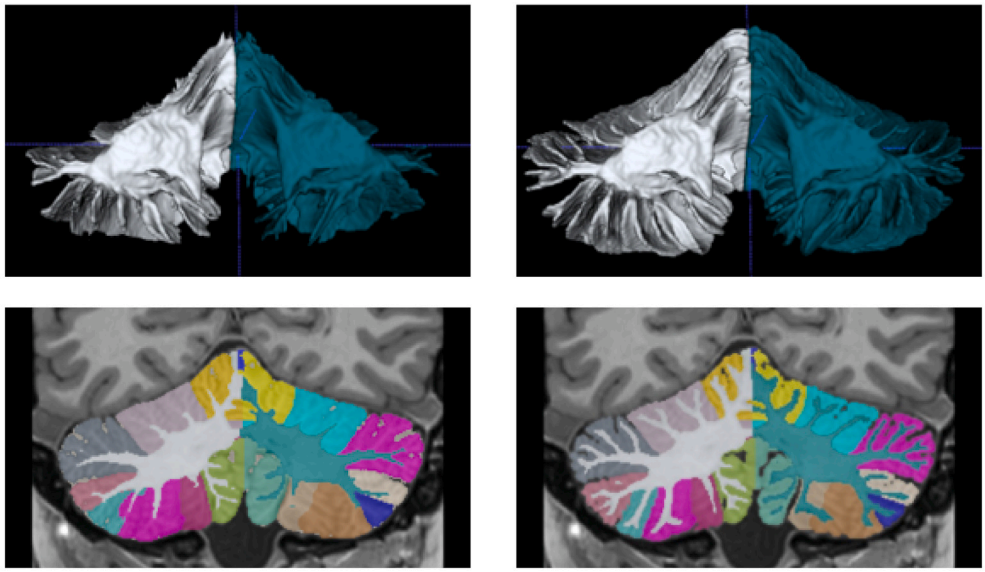


Fig. 1. Left: 3D reconstruction of WM label and a coronal view of segmentation before semiautomatic correction. Right: 3D reconstruction of WM label and a coronal segmentation view after semiautomatic correction. The WM "*arbor vitae*" is better defined in the corrected version.

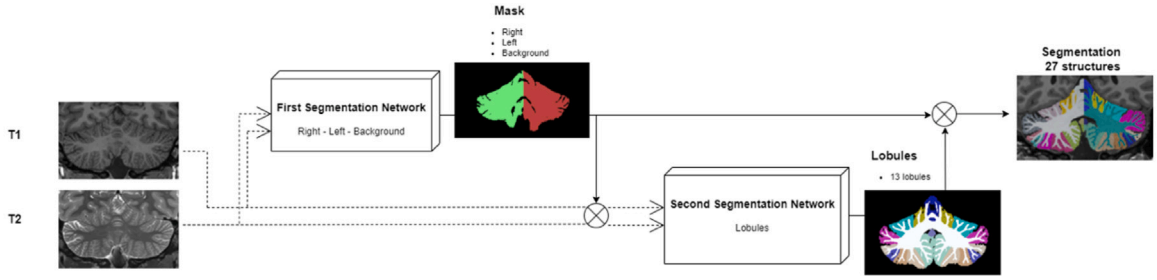


Fig. 2. Scheme of the proposed segmentation method. The first network segments left–right cerebellum and background and the second network segments the lobules using the T1 and T2 images and the output of the first network.

2.3. Segmentation method

The challenges of segmenting 27 cerebellar structures with ultra-high resolution multimodal volumes are the computational complexity and the memory consumption due to 3D convolutional neural networks. Thus, the right choice of network architecture plays a crucial role in the method's implementation. The U-Net-based (Ronneberger et al., 2015) architecture has been the most used topology in medical image analysis tasks. Previous methods for cerebellum segmentation have used 2.5D or 3D U-Net versions, being the last one the preferred option due to its better context analysis, which usually improves classification accuracy. Due to memory limitations of current GPUs, using a 3D U-Net at 0.125 mm^3 resolution is not feasible for segmenting all 27 structures simultaneously.

To deal with this limitation while maintaining the 3D nature of the proposed method, a two-stage cascade architecture is proposed, taking advantage of the pseudo-symmetry of the human cerebellum. Firstly, the left and right hemispheres are segmented using a first network, allowing the generation of a binary mask that classifies voxels into the left–right cerebellum and the background. Subsequently, a binary mask cerebellum/background from the output of this first network is multiplied to the T1 and T2 volumes to generate the input of a second network, thereby facilitating intra-cerebellar attention of the 14 classes (12 lobules, WM and background) segmentation network (regardless of whether they are left or right). Finally, to generate the 27 labels segmentation, outputs from both networks are combined to identify each hemisphere's left and right structures. The scheme of this process is summarized in Fig. 2.

We initially used for both networks a classical U-Net 3D architecture as reference model. However, even with the cascade approach, the maximum number of filters for the 3D U-Net was limited to 16 to fit into memory.

To reduce the network complexity, we explored the design of a more streamlined novel network characterized by fewer parameters. This approach aimed to facilitate effective learning of segmentations while mitigating the risk of succumbing to overfitting or underfitting. Therefore, we experimented with a self-designed architecture based on a coarse-to-fine approach that we called DPN (Deep Pyramidal Network), which enables a lighter and less memory demanding model. Unlike U-Net architecture, where an encoder–decoder strategy is used, the DPN network uses an 8 times subsampled version (using Average3DPooling) of the original volume as initial input (see Fig. 3). It concatenates the upsampled generated features obtained through three blocks (convolution + ReLU + BachNotmalization) with higher-resolution features obtained from the upper-resolution level image (1/4 resolution) after its pass through a feature extractor based on a (convolution + ReLU + BachNotmalization) block to refine frequency details. This process is repeated until the original resolution is reached, and a final softmax layer is used to generate output class probabilities. At the end of each resolution block, a dropout layer was added for the resolution 1/8, 1/4

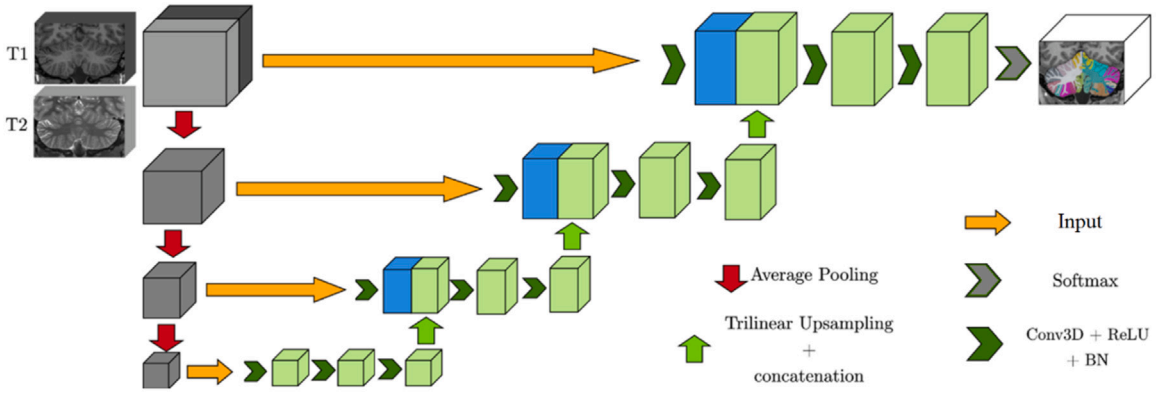


Fig. 3. Proposed DPN architecture. The original input is 3 times sub-sampled to generate a 1/8 resolution input using average pooling (red arrows). The network extracts features at each resolution level and concatenates them (blue boxes) to upsampled features (green boxes) from the previous resolution level (skip connections). All convolutional blocks have the same number of filters.

and 1/2 to minimize overfitting problems (rate 0.25). This architecture reduces the number of parameters to almost one-fifth compared to a U-Net model. A scheme of the proposed DPN architecture is shown in Fig. 3.

In the training process, a loss function composed of binary cross-entropy and Dice loss is minimized. To prevent gradient vanishing problems when the values of the loss function are close to 0, the logarithm operation is applied to the loss function to scale the gradients.

$$L = \log\left(\frac{1 - \sum_{i=1}^{C} \text{Dice}(y_c, \hat{y}_c) + \text{BCE}(y, \hat{y})}{2}\right) \quad (1)$$

Where C represents the classes (labels), Dice stands for the Dice index, and BCE is the Binary Cross-Entropy.

2.4. Atlas guided segmentation

On the other hand, we realized that in our lobule segmentation problem, different lobules have similar intensity patterns but different locations with sometimes no clear limits between neighbour lobules. Atlas-based methods (Carass et al., 2018) benefit from the location information by locally searching for similar patches (for example CERES method) after a registration process. Therefore, we hypothesized that including a priori information (i.e., a subject-specific atlas) could assist in the segmentation process, reducing labelling errors and making the method more robust. Using this a priori information could reduce the sensitivity of the models to noise, artefacts or variability of the input data, avoiding catastrophic errors and non-plausible topologies.

A multi-atlas strategy is used to create this atlas, as it has been proven more robust than the single-atlas. In this approach, a nonlinear registration transform is estimated for each anatomical volume (T1) in the training library to the volume to be segmented. Subsequently, these non-linear transforms are applied to the volumes and their corresponding segmentations, resulting in a subject-specific library of segmentations. Finally, these segmentations are merged voxel-wise according to a rule that assigns a value (label) to each voxel in the volume.

Nonetheless, the primary limitation of this methodology, as previously identified in the article by Park et al. (2014), is the time cost associated with non-linear registrations and label fusion. Fortunately, deep learning-based methods, such as VoxelMorph, have drastically reduced the time cost of non-linear registration from minutes to less than one second (Balakrishnan et al., 2018). In the proposed method, we used a similar deep non-linear registration network than Coupé et al. (2020). This replaces the costly registration process by evaluating a convolutional neural network previously trained on the training cerebellum dataset, substituting a computationally expensive minimization process with a memory-based one, with a time cost of 0.7 s per cerebellar volume.

To perform the label fusion, an intensity-based weighted majority voting is used where voxels of the library with intensities similar to the voxel to be classified have a higher weight than those being more dissimilar. Each weight is calculated as follows:

$$w_p = \frac{1}{1 + d \cdot ||I_p - L_{p,j}||} \quad (2)$$

Where d is a scaling coefficient, I_p is the intensity of a voxel at a position p, and L_p, j is the intensity of a voxel at a position p of an image j from library L. We set d as 0.5 experimentally.

To deal with the temporal limitations of multi-atlas approaches, the fusion algorithm was implemented in C language and parallelized using the OpenMP multiprocessing library (Anon). Another aspect that impacts the temporal cost of the atlas generation process is the number of templates used to generate it (the bigger the library, the more time will be spent on the atlas creation). Therefore, given the availability of 130 T1/T2 images with their corresponding segmentations in the library, we select only the N cases that are more similar (based on the L1 norm of image intensities) to perform the non-linear registration and the later label fusion.

To determine the optimal number of elements in the library, we used the 130 segmentations, from which we extracted a sample to calculate the specific atlas (leave-one-out). Empirically, it is observed that a plateau is reached in terms of the Dice between the atlas and the original segmentation (0.82) when the size of the library is N = 20 samples more similar to the sample under analysis.

As shown in Fig. 8 in Section Appendix, the Dice coefficient stabilizes while saving 50% of time, indicating that adding more samples beyond this point provides diminishing returns in accuracy. Finally, the resulting subject-specific atlas was normalized in the range [0,1] by dividing it by its maximum intensity before being fed to the networks as a third channel (concatenated with the T1 and T2 images). This provides a voxel-wise prior that gives a rough reference to guide the segmentation process and make it more robust.

Table 2
DPN and U-Net architecture comparison. Mean Values (Standard Deviation).

Model architecture	Parameters	Mean lobule Dice	Mean whole cerebellum Dice	Mean computational time (s)
2 DPN	696 177	0.9254 (0.036)	0.9890 (0.0022)	6.3029 (1.646)
2 U-Net	3 433 473	0.9286 (0.042)	0.9911 (0.0028)	7.5103 (1.611)

2.5. Extensive data augmentation

Since our models were trained in an age-limited training dataset, we decided to augment the training data to include more diverse anatomy patterns. To extent the training dataset in a realistic manner without having more manually labelled cases we used an approach based on non-linear registration. We used the lifespan dataset (N = 4857) cases to obtain a diverse shape population from samples from childhood to old age.

Specifically, for each case of the lifespan dataset we look for the most similar case (T1) in our semiautomatically labelled training dataset and this case is non-linear registered to the reference lifespan case using Greedy algorithm (Battle et al., 2016). Once the new 4857 new cases are generated (T1, T2 and Atlas) a fine tunning process is performed using as training data, both, the original training set (130 samples) and the extended dataset (4857 samples) with a 50% of probability each to balance both datasets.

3. Experiments and results

In this section, we present the results of several experiments to evaluate the performance of the proposed method on the test partition data. All the experiments have been performed using the TensorFlow 1.15 framework, Keras 2.2.4, and a batch size of 1 on a GPU V100 with 32 GB of memory. For training Adam optimizer was used during the first 2000 epochs and then Adamax (King et al., 2019) is used for an additional 1000 epochs. In the case of finetuning process, Adamax was directly used during 1000 epochs. The rationale for this choice also finds theoretical support in the Adam paper, which highlights the properties of Adamax. Specifically, Adamax employs the L^{∞} norm for gradient scaling, as opposed to the L^2 norm used by Adam. This characteristic enhances numerical stability, particularly in scenarios prone to vanishing or exploding gradients. In our case, this was especially relevant during extended training, where gradients are expected to be small due to the prior knowledge already present in the network. By combining these optimizers, we took advantage of Adam's adaptability and Adamax's robustness.

As image inputs, we used the T1 and T2 volumes normalized to have mean 0 and variance 1 (subtracting the mean and dividing by the standard deviation). To further increase input data variability during the first part of the training (HCP data only), we have used random data augmentation through the Torchio library (Pérez-García et al., 2021), enabling the online generation of typical transformations within the medical image domain. Intensity transformations such as random anisotropy, bias field, blur, ghosting and gamma have been applied alongside geometric transformations, including random affine and elastic deformation.

3.1. U-Net vs DPN

Both network architectures were trained independently to evaluate the choice of the optimal architecture (U-Net vs DPN). Two configurations were considered within a cascade framework (2 DPN and 2 U-Net) for right-left background and lobule networks. For both networks in the cascade, the number of filters of DPN was set to 32 for all resolutions, while for U-Net, the initial number of filters was set to 16, doubling this number at each down-sampling step.c As shown in Table 2, both models present a similar performance with no statistical difference. Although the U-Net-based option is slightly superior, it has proven the utility of DPN as an efficient alternative to classical U-Net, but with only a fifth of the parameters.

Table 3
Atlas vs non-atlas DPN and U-Net architecture comparison. Best results in bold. Mean Values (Standard Deviation).

Model architecture	Mean lobule Dice	Mean whole cerebellum Dice
2 DPN without atlas	0.9254 (0.0364)	0.9890 (0.0028)
2 DPN with atlas	0.9308** (0.0427)	0.9905 (0.0015)
2 U-Net without atlas	0.9286 (0.0420)	**0.9911 (0.0022)**
2 U-Net with atlas	**0.9322‡ (0.0392)**	0.9898 (0.0020)

(**,‡) Significant difference of models with atlas with respect to those not including it with $p < 0.01$, two-sided Wilcoxon test.

Table 4
Impact of modality and resolution on segmentation results. (HR) High Resolution and (SR) Standard Resolution. Mean Values (Standard Deviation). The best results are in bold.

Resolution/ Multimodality	Model	Structure Dice	Whole cerebellum Dice
SR, T1	DPN	0.9117 (0.0458)**	0.9889 (0.0036)
SR, T1	U-Net	0.9079 (0.0520)‡	0.9878 (0.0023)
SR, T1+T2	DPN	0.9138 (0.0467)**	0.9892 (0.0016)
SR, T1+T2	U-Net	0.9088 (0.0468)‡	0.9884 (0.0015)
HR, T1	DPN	0.9239 (0.0353)**	0.9896 (0.0025)
HR, T1	U-Net	0.9276 (0.0389)‡	0.9902 (0.0025)
HR, T1 + T2	DPN	0.9308 (0.0427)	**0.9905 (0.0022)**
HR, T1 + T2	U-Net	**0.9322 (0.0392)**	0.9898 (0.0020)

(**,‡) Significant difference of models with respect to its same architecture with high resolution and multimodal data with $p < 0.01$, two-sided Wilcoxon test.

3.2. Use of a subject-specific atlas as an additional channel

Once both architectures were validated, we decided to train both architectures with an additional input channel in the form of a subject-specific atlas to validate the suitability of introducing a priori information into the networks. Table 3 shows the results of the U-Net and DPN networks.

As expected, both architectures benefited from the inclusion of a priori information. The U-Net based model without atlas obtained the best results for whole cerebellum, while the U-Net based model with atlas got the best lobule segmentation Dice. Based on these results, we further investigated the atlas' impact in Section 3.5.

3.3. Multimodality and ultra-high resolution impact

Previous studies have demonstrated that using data from multiple modalities improves model performance in identifying prostate (Nai et al., 2020) or tumours (Zhou et al., 2019; Ren et al., 2021). However, the effectiveness of this approach with ultra-high resolution and multimodal data has not been confirmed to the best of our knowledge. To validate this, we retrained two models with the same approach (3 input channels and cascade approach), but changing the input data to the networks using only monomodal and/or standard resolution (SR) data. Results can be analysed in Table 4.

Table 5
Results of the different robustness settings. Loss of performance when having a bad T2. Mean Values (Standard Deviation).

Name	Mean Dice	Δ Mean Dice
DPN + ATLAS + bad T2	0.8255 (0.0722)	10.8368
DPN + ATLAS + original T2	0.9308 (0.0427)	
DPN + bad T2	0.7066 (0.1589)	21.8804
DPN + original T2	0.9254 (0.0361)	
U-Net + ATLAS + bad T2	0.8786 (0.0825)	5.5946
U-Net + ATLAS + original T2	0.9322 (0.0392)	
U-Net + bad T2	0.8346 (0.1355)	9.3977
U-Net + original T2	0.9286 (0.0428)	

On the one hand, the increased resolution has a significant impact on the Dice index at the lobule level (0. 9322 vs 0.9088 in U-Net and 0.9308 vs 0.9138 in DPN) and somewhat less as in the whole cerebellum (0.9904 vs 0.9884 in U-Net and 0.9907 vs 0.9892 in DPN). This can be seen in the trend of the Dice index with increasing resolution and the introduction of the channel in T2. It is worth remembering the intricate structure of the cerebellum, where there is a high degree of tissue packing and a highly textured surface, the folia, with submillimeter cortical thicknesses. This causes partial volume problems in standard resolution images (1 mm^3) where different tissues provide information on the same voxel, preventing their independent classification. The partial volume problems are reduced at ultra-high resolution, allowing the model to extract more information about the tissues and their textures as they are better differentiated. In addition, better performance is observed for the U-Net-based structure for HR (but not for SR), under the hypothesis that having a higher capacity (a greater number of parameters) than the DPN-based structure can extract more information from the input data.

On the other hand, the introduction of T2 benefits to a lesser extent than the increase in resolution (0.9321 vs. 0.9276 in U-Net and 0. 9308 vs. 0.9239 in DPN). Although the gain is smaller, it reduces some false positives and negatives by providing information from another modality for the same anatomical structure.

3.4. Robustness study

Traditional methods often rely on data augmentation to ensure robustness against anomalous or low-quality data. We have followed this technique by introducing (in-domain transformation using TorchIO Pérez-García et al., 2021). However, we believe that incorporating an atlas offers additional advantages when ensuring the robustness and stability of the models. This inclusion provides the segmentation network with valuable anatomical prior information and differentiates regions with similar brightness that belong to different structures.

To further understand the impact of the atlas on the models' robustness, we investigated the drop in the Dice coefficient when models were fed with low-quality inputs. A common scenario in the clinic involves acquiring only the T1 image. Sometimes, synthesis techniques are used to obtain the T2 image. However, the quality of the synthetic image does not match the actual acquisition of a T2 image. To simulate these conditions, we substituted the T2 images with T1 images in our test set. This method allowed us to assess the performance difference when the model had to work with less ideal input data.

Results in Table 5 clearly show that networks with an atlas have approximately half the Dice loss compared to models without the atlas as an additional channel. Besides, these results also reveal that U-Net is more robust than DPN, with a smaller loss in terms of Dice, which can be explained by the higher number of parameters. Understanding that the network's performance highly relies on the quality of input data further justifies the integration of methods for super-resolution and synthesis of T2 images into the complete pipeline. In fact, tests were conducted to showcase the capability of these synthesis and super-resolution networks to enhance image quality and thereby mitigate the impact on Dice loss (see Table 8).

3.5. Ensemble of architectures

As shown in Table 3, the U-Net model had better mean lobule Dice while the DPN model had better whole cerebellum Dice. Deeply analysing the differences among models, we found some complementarity behaviour for different lobules, where sometimes one of the networks is better than the other (see Table 6). Therefore, instead of choosing one model, we combined both by averaging their predictions using a classical bagging approach. This approach is part of ensemble learning, a widely recognized concept in machine learning employed to improve the overall performance of a technique. This improvement is achieved by training multiple models before combining them (Kamnitsas et al., 2018).

Table 6 shows the results of the ensemble of U-Net and DPN models. As can be noticed, the ensemble results for each individual lobule are better than those of the original networks.

3.6. Fine tuning with realistic data augmentation

As previously mentioned, to enhance the robustness of the proposed technique against anatomical variations, we fine-tuned the trained networks using a realistically augmented dataset created by non-linearly registering samples from the lifespan dataset. This fine-tuning was applied to both tasks: right-left-background hemisphere segmentation and lobule segmentation, using Adamax during 1000 epochs. To mitigate the risk of forgetting (Lopez-Paz and Ranzato, 2017; Li and Hoiem, 2018), cases from the original library were intermittently presented throughout the training process with a 50% probability, alongside an equal 50% probability of cases from the extended library. Results of the fine-tuning are shown in Table 7. The table shows that there are no significantly different values for any of the labels, demonstrating that the fine-tuning process has not altered the performance of the original method.

As observed, the test results after fine-tuning are slightly lower than before fine-tuning on the age-limited HCP test set. This decrease can be attributed to the trade-off between bias and variance introduced during fine-tuning. By incorporating diverse populations and acquisition conditions, the fine-tuning process reduces the model's bias towards the HCP dataset, improving its ability to generalize to out-of-domain data. However, this also makes the model less tailored to the specific characteristics of the HCP dataset, leading to a slight reduction in performance on this test set. Importantly, this trade-off aligns with the goal of enhancing robustness and generalization across diverse datasets.

Additionally to the test set evaluation, we wanted to measure the generalization gain of the fine-tuning process. Since we have no ground truth labels for the lifespan dataset, we decided to use a population-based validation approach. Specifically, we selected the 2962 healthy cases from the lifespan dataset and segmented them with the "before" and "after" fine-tuning models to generate age-volume curves for each structure. These data were fitted to their corresponding lifespan models to compute volume trajectory over the entire lifespan as done in Coupé et al. (2017). Our hypothesis was that the improved generalization would result in more accurate segmentations for each age range and structure, thus reducing the age-related variability of the derived volumes models through the obtention of more coherent results. Results of this analysis showed that 16 out of the 26 GM structures had a reduced variability, with statistical significance observed in 15 ($p < 0.05$) Wilcoxon two-sided test. Details are shown in Fig. 7 in Section Appendix.

Table 6
Dice results for each lobule for the U-Net, DPN and Ensemble models.

Label	2 U-Nets with atlas	2 DPN with atlas	Ensemble
Whole cerebellum	0.9898 (0.0025)	0.9905 (0.0024)	**0.9906 (0.0024)**
Right Lobule I-II	0.8333 (0.0614)	0.8303 (0.0588)	**0.8426 (0.0484)**
Right Lobule III	0.9104** (0.0237)	0.9107** (0.0238)	**0.9157 (0.0217)**
Right Lobule IV	0.9213** (0.0108)	0.9258** (0.0102)	**0.9309 (0.0083)**
Right Lobule V	0.9257** (0.0131)	0.9295** (0.0167)	**0.9352 (0.0142)**
Right Lobule VI	0.9536** (0.0083)	0.9552** (0.0091)	**0.9579 (0.0078)**
Right Crus I	0.9597** (0.0072)	0.9608** (0.0059)	**0.9629 (0.0063)**
Right Crus II	0.9502** (0.0122)	0.9502** (0.0104)	**0.9541 (0.0105)**
Right Lobule VIIB	0.9277** (0.0240)	0.9288** (0.0215)	**0.9362 (0.0206)**
Right Lobule VIIIA	0.9421** (0.0183)	0.9428** (0.0185)	**0.9480 (0.0178)**
Right Lobule VIIIB	0.9459** (0.0085)	0.9418** (0.0102)	**0.9479 (0.0080)**
Right Lobule IX	0.9459** (0.0217)	0.9420** (0.0267)	**0.9481 (0.0212)**
Right Lobule X	0.9470** (0.0101)	0.9440** (0.0114)	**0.9488 (0.0100)**
Right Grey Matter	0.9597** (0.0031)	0.9601** (0.0037)	**0.9622 (0.0031)**
Left Lobules I-II	0.8245 (0.0561)	0.8166 (0.0869)	**0.8259 (0.0717)**
Left Lobule III	0.9159** (0.0183)	0.9137** (0.0185)	**0.9197 (0.0188)**
Left Lobule IV	0.9201** (0.0127)	0.9184** (0.0157)	**0.9236 (0.0140)**
Left Lobule V	0.9212** (0.0149)	0.9160** (0.0201)	**0.9267 (0.0159)**
Left Lobule VI	0.9533** (0.0079)	0.9534** (0.0082)	**0.9577 (0.0065)**
Left Crus I	0.9586** (0.0062)	0.9592** (0.0079)	**0.9614 (0.0070)**
Left Crus II	0.9544** (0.0051)	0.9536** (0.0060)	**0.9573 (0.0050)**
Left Lobule VIIB	0.9348** (0.0151)	0.9284** (0.0227)	**0.9365 (0.0191)**
Left Lobule VIIIA	0.9389** (0.0171)	0.9338** (0.0243)	**0.9403 (0.0200)**
Left Lobule VIIIB	0.9399** (0.0107)	0.9377** (0.0119)	**0.9432 (0.0103)**
Left Lobule IX	0.9474** (0.0127)	0.9437** (0.0145)	**0.9493 (0.0126)**
Left Lobule X	0.9442** (0.0129)	0.9432** (0.0122)	**0.9465 (0.0121)**
Left White Matter	0.9612** (0.0026)	0.9614** (0.0027)	**0.9633 (0.0027)**
Average	0.9322** (0.0392)	0.9308** (0.0427)	**0.9362 (0.0388)**

** Significant difference of the individual models compared to the ensemble model with $p < 0.01$, two-sided Wilcoxon test.

Table 7
Test Dice before and after fine-tuning with the extended dataset.

Label	Ensemble	*Ensemble after fine-tuning*
Whole cerebellum	0.9906 (0.0024)	0.9900 (0.0025)
Right Lobule I-II	0.8426 (0.0484)	0.8319 (0.0551)
Right Lobule III	0.9157 (0.0217)	0.9132 (0.0218)
Right Lobule IV	0.9309 (0.0083)	0.9283 (0.0052)
Right Lobule V	0.9352 (0.0142)	0.9317** (0.0130)
Right Lobule VI	0.9579 (0.0078)	0.9558** (0.0079)
Right Crus I	0.9629 (0.0063)	0.9606** (0.0072)
Right Crus II	0.9541 (0.0105)	0.9516** (0.0122)
Right Lobule VIIB	0.9362 (0.0206)	0.9319** (0.0232)
Right Lobule VIIIA	0.9480 (0.0178)	0.9441** (0.0185)
Right Lobule VIIIB	0.9479 (0.0080)	0.9453** (0.0085)
Right Lobule IX	0.9481 (0.0212)	0.9443** (0.0263)
Right Lobule X	0.9488 (0.0100)	0.9452** (0.0083)
Right Grey Matter	0.9622 (0.0031)	0.9605** (0.0029)
Left Lobules I-II	0.8259 (0.0717)	0.8234 (0.0442)
Left Lobule III	0.9197 (0.0188)	0.9159** (0.0165)
Left Lobule IV	0.9236 (0.0140)	0.9206** (0.0140)
Left Lobule V	0.9267 (0.0159)	0.9240 (0.0136)
Left Lobule VI	0.9577 (0.0065)	0.9556** (0.0065)
Left Crus I	0.9614 (0.0070)	0.9600** (0.0061)
Left Crus II	0.9573 (0.0050)	0.9544** (0.0053)
Left Lobule VIIB	0.9365 (0.0191)	0.9319** (0.0207)
Left Lobule VIIIA	0.9403 (0.0200)	0.9370** (0.0214)
Left Lobule VIIIB	0.9432 (0.0103)	0.9392** (0.0107)
Left Lobule IX	0.9493 (0.0126)	0.9459** (0.0136)
Left Lobule X	0.9465 (0.0121)	0.9436** (0.0121)
Left White Matter	0.9633 (0.0027)	0.9619** (0.0026)
Average	0.9362 (0.0388)	0.9330** (0.0386)

** Significant difference of the individual models compared to the ensemble model with $p < 0.01$, two-sided Wilcoxon test.

3.7. Comparison with state-of-the-art methods

Regarding the comparison with other methods, we cannot make a direct quantitative comparison because either labelling protocols or the resolution of the input data are different (biasing the Dice coefficient). However, by making a pseudo-quantitative comparison of the mean structure Dice reported at each publication, it is possible to observe a gain in terms of the accuracy of the proposed method. For instance, CERES got an average lobule Dice of 0.7729, Acapulco 0.77 and more recently, CerebNet got a Dice of 0.87. All these results are quite far from the 0.9330 obtained by the proposed method.

Nevertheless, we acknowledge the limitations of this pseudo-comparison. Therefore, we conducted a practical comparison between the state-of-the-art method, CerebNet, and our approach, using the complete pipeline described in the next section. We evaluated their abilities to distinguish populations using the open neuroimage ADNI dataset, illustrating the impact of both high-resolution imaging and detailed segmentation protocols on a real clinical problem. Previous studies have identified changes in cerebellar volume over the course of Alzheimer's disease progression (Jacobs et al., 2018; Lin et al., 2020; Colloby et al., 2014) highlighting the clinical relevance of cerebellum volumetry analysis.

For the experiment, we used 250 curated cases from the ADNI dataset, which included samples from individuals with Alzheimer's disease (AD) and cognitively normal (CN) controls. The subjects were 115 female and 135 male, with 104 diagnosed with AD at an average age of 74.56 ± 7.9 and 139 CN individuals at an average age of 73.51 ± 6.47.

In order to make the comparison, we calculated the effect size using Cohen's d (Cohen, 2013), a measure based on the differences between two populations. We have chosen this metric because while statistical significance assesses the probability of obtaining an observed difference by random chance, effect size quantifies the magnitude of that difference, providing valuable information about the practical significance

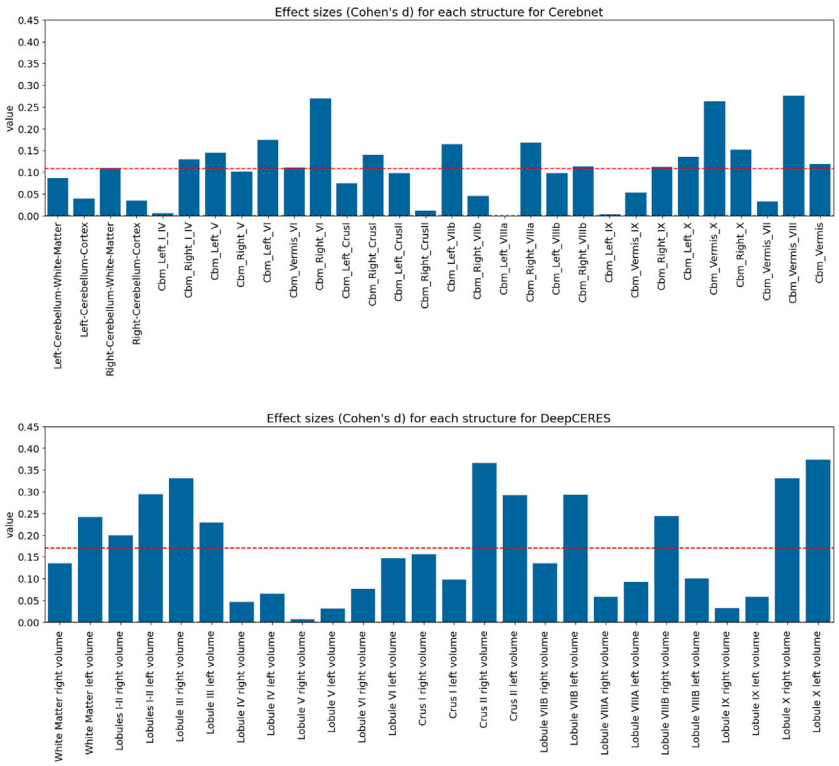


Fig. 4. Cohen's d for CN/AD for volumes measured with DeepCERES and CerebNet. The dotted orange line represents the mean Cohen's d for both CerebNet and DeepCERES methods.

of the findings. As can be seen in Fig. 4, DeepCERES is capable of measuring bigger differences than CerebNet between populations both in max (0.3731 vs 0.2761) and mean values (0.1703 vs 0.1090).

Furthermore, comparing the statistical significance (p) of the mean differences for each structure reveals that DeepCERES identifies more significant structures than CerebNet (11 vs 3). Additionally, the segmentation quality of each method for an AD case is visually compared in the appendix (see Fig. 9 in Section Appendix), where it can be observed how DeepCERES outperforms CerebNet in delineating lobules and sharply segmenting white matter, a structure typically underestimated in the rest of the methods.

3.8. DeepCERES pipeline

Despite the proposed method's good performance with high-resolution multimodal data, we are aware of the limitations of the availability of such data in research and clinical settings. Therefore, we decided to develop a full pipeline able to work on standard resolution T1 MR images. We named the proposed pipeline DeepCERES (summarized in Fig. 5).

The DeepCERES pipeline consists on the following steps :

- Denoise: The Spatially Adaptive Non-local means (SANLM) filter (Manjón et al., 2010) was used to reduce random noise naturally present in the images.
- Registration: Affine registration to MNI152 space (1 mm^3 resolution). ITKElastix software (Ntatsis et al., 2023; Shamonin et al., 2014) was employed.
- Bias correction: The N4 bias correction method was used to correct the inhomogeneity of the images (Tustison et al., 2010).
- (Intensity) Normalization: We normalized the T1 images by applying a piecewise linear tissue mapping based on the TMS method (Manjón et al., 2008) as described in the study by Manjón and Coupé (2016).
- Brain Extraction: To compute normalized volumes, we segmented the ICV using the Deep ICE method (Manjón et al., 2020).
- Second bias correction and (intensity) normalization: This was performed using the ICV extracted volume instead of the original image to further improve the image quality.
- Super-resolution T1: The T1 image was super-resolved to 0.125 mm^3 resolution (factor 2) using an in-house ResNet (He et al., 2015) like super-resolution network trained using the HCP dataset. This step generated a volume of 362 × 434 × 362 voxels with synthetic 0.125 mm^3 resolution.
- Crop: A crop of the cerebellar region of the super-resolved T1 image is made (crop size of 162 × 182 × 168 voxels). This region was estimated from the limits of the library's manual segmentations in the MNI152 space (including 10 voxel margin in each dimension to accommodate cerebellum location and size variability).
- Atlas generation: The case-specific atlas is calculated as previously described from the training library, using a network to estimate the deformation fields of the 20 most similar cases and a parallel implementation for the weighted label fusion.
- T2 Synthesis: We used a cropped version of a recently proposed full volume convolutional neural network for MR image synthesis (Manjón et al., 2021).
- Segmentation: Once the cropped T1, the synthetized T2 and the subject specific atlas are generated we run the proposed ensemble-based method to generate the segmentation.

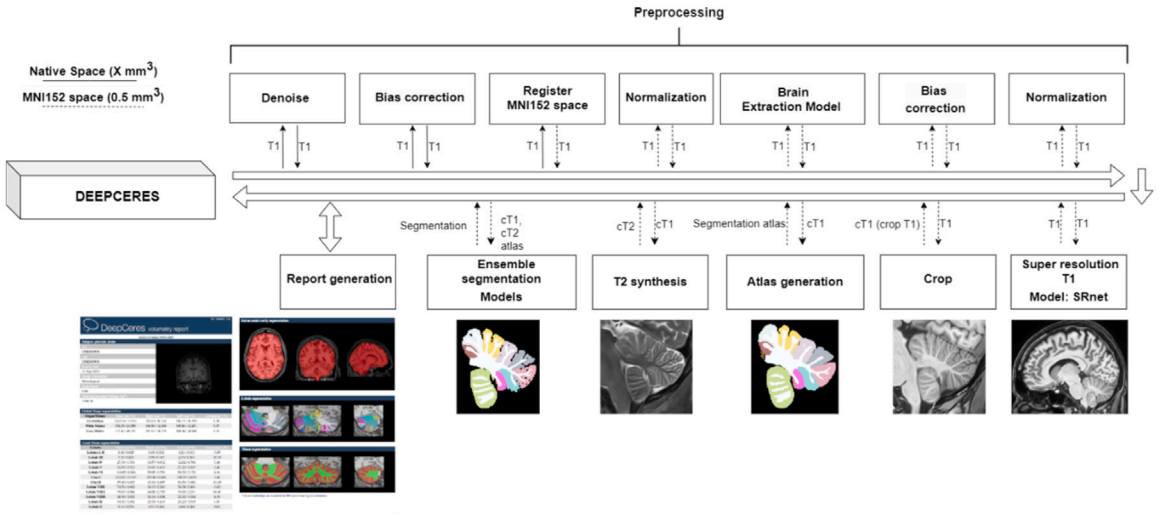


Fig. 5. Scheme of the DeepCERES pipeline.

- Report generation: Finally, to make the analysis of the results more user friendly we generate a pdf report (and a CSV file) summarizing the volumetric results derived from the obtained segmentation. This report includes the volumes of the different lobes in absolute value and normalized to the ICV (asymmetric indexes are also supplied). If the user supplies the age and sex of the subject a comparison with the lifespan model is performed, highlighting if each specific structure is inside the population bounds for a given age and sex.

We are aware that the outcomes for super-resolved T1 and synthetic T2 images may not match the quality of native multimodal high-resolution data. To estimate this performance drop, we applied the proposed pipeline to monomodal and down-sampled T1 data from the test set. The results are listed in Table 8. The results indicate a performance drop (0.9330 vs 0.9290), which reflects the impact of super-resolution and synthetic T2 on segmentation accuracy. Although this result is based on the in-domain test data and likely represents a lower performance loss bound, we believe the results remain competitive.

Finally, we assessed the efficiency of the proposed pipeline and its usability. We verified that the proposed pipeline can run for inference on a Titan Xp with 12 GB of memory. The complete pipeline has a total temporal cost (from denoising to report generation) of approximately 2 and a half minutes (details can be seen in Fig. 6). We have made DeepCERES available to the scientific community through our online system volBrain.net[1].

Table 8
Impact of not having high resolution (HR) and multimodal data: super resolving T1 and synthesizing T2.

Labels	Ensemble of models with original multimodal HR data	Ensemble of models with synthetic multimodal HR data
Whole cerebellum	0.9900 (0.0025)	0.9877 (0.0022)
Right Lobule I-II	0.8319 (0.0551)	0.8308 (0.0574)
Right Lobule III	0.9132 (0.0218)	0.9077 (0.0221)
Right Lobule IV	0.9283 (0.0052)	0.9217 (0.0086)
Right Lobule V	0.9317 (0.0130)	0.9278 (0.0141)
Right Lobule VI	0.9558 (0.0079)	0.9516 (0.0078)
Right Crus I	0.9606 (0.0072)	0.9574 (0.0054)
Right Crus II	0.9516 (0.0122)	0.9479 (0.0101)
Right Lobule VIIB	0.9319 (0.0232)	0.9282 (0.0235)
Right Lobule VIIIA	0.9441 (0.0185)	0.9404 (0.0190)
Right Lobule VIIIB	0.9453 (0.0085)	0.9425 (0.0085)
Right Lobule IX	0.9443 (0.0263)	0.9404 (0.0224)
Right Lobule X	0.9452 (0.0083)	0.9404 (0.0118)
Right Grey Matter	0.9605 (0.0029)	0.9567 (0.0030)
Left Lobules I-II	0.8234 (0.0442)	0.8253 (0.0685)
Left Lobule III	0.9159 (0.0165)	0.9121 (0.0168)
Left Lobule IV	0.9206 (0.0140)	0.9161 (0.0124)
Left Lobule V	0.9240 (0.0136)	0.9164 (0.0161)
Left Lobule VI	0.9556 (0.0065)	0.9505 (0.0075)
Left Crus I	0.9600 (0.0061)	0.9547 (0.0086)
Left Crus II	0.9544 (0.0053)	0.9502 (0.0058)
Left Lobule VIIB	0.9319 (0.0207)	0.9309 (0.0192)
Left Lobule VIIIA	0.9370 (0.0214)	0.9328 (0.0210)
Left Lobule VIIIB	0.9392 (0.0107)	0.9356 (0.0108)
Left Lobule IX	0.9459 (0.0136)	0.9420 (0.0133)
Left Lobule X	0.9436 (0.0121)	0.9409 (0.0134)
Left White Matter	0.9619 (0.0026)	0.9577 (0.0026)
Average	0.9330 (0.0386)	0.9290(0.0392)

4. Discussion

In this paper we have proposed an ultra-high resolution multimodal cerebellum lobule segmentation method based on deep learning. The key contributions of the proposed method include the development of an ultra-high resolution labelling protocol and dataset, the use of efficient neural architectures, an efficient cascade based segmentation approach, the use of apriori information in the form of a multiatlas as extra input to improve accuracy and robustness and the use of realistic data augmentation. Besides, we have also proposed a standard resolution T1-based segmentation pipeline (DeepCERES) to make available the proposed technology to users with clinical quality data.

To train the proposed method we have created a semi-automatically segmented ultra-high resolution multimodal library. Thanks to the enhanced resolution of this training set a detailed delineation of the

[1] https://volbrain.net/services/DeepCERES

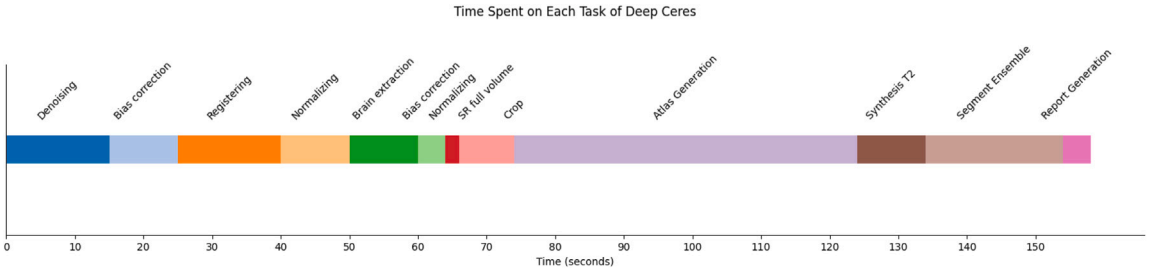


Fig. 6. Temporal profile of DeepCERES method.

"arbor vitae" has been made possible allowing a more precise definition of the white matter and lobular grey matter.

To process the full cerebellar volume, the proposed method splits the problem into two sub-problems using a cascade approach. This choice was motivated by both practical and experimental considerations. Practically, a single network capable of segmenting all structures simultaneously could not fit into memory. Experimentally, we observed that the Dice score for the whole cerebellum was higher when using a dedicated network for cerebellum segmentation compared to using the same network designed for individual structures. This improvement aligns with expectations, as cerebellum segmentation is a simpler problem than distinguishing 27 fine structures. To minimize false negatives in the second network, we carefully cropped the cerebellum in the MNI space with sufficient offsets in all dimensions, ensuring the entire structure is included. Additionally, we applied geometric and intensity data augmentation strategies to make the model more robust. While cascade networks may introduce errors that are challenging to recover in the second network, such as false negatives, they also allow the recovery of false positives, further enhancing the results of the first network. These findings validate the cascade approach as effective for tackling the complex task of high-resolution multimodal segmentation into 27 structures.

We have also demonstrated that combining deep learning and classical machine learning approaches (multi-atlas) can be a successful strategy for enhancing performance and robustness in both architectures (using a priori knowledge through the atlas). Combining these two philosophies has made it possible to create more accurate models with a higher Dice index value, more efficient when compared to other more powerful models (with more parameters), and more robust with this a priori information.

The use of alternative models to the classic U-Net network, such as the DPN model, has also been explored, achieving outstanding results despite having 5 times fewer parameters. Since it is commonly known that simpler models are less prone to overfit and, hence, generalize better, this information is crucial. However, the best outcomes in this experiment came from combining different models (U-Net and DPN), demonstrating that the use of an ensemble-based approach can help to further reduce the classification error. This finding is consistent with current literature showing that ensemble methods can improve robustness and accuracy by leveraging diverse model architectures, DeepCERES uses this approach to maximize segmentation performance.

Given the limited number and age range of the training dataset, a strong realistic data augmentation approach using lifespan data was used (including pathological cases). The proposed approach has been subjected to several analyses in terms of robustness and generality, showing excellent results. Notably, the fine-tuning strategy employed in this study effectively addresses biases inherent in single-dataset training (e.g., HCP-specific acquisition protocols), enabling the model to generalize better across diverse datasets. While this resulted in a slight performance reduction on in-domain test data, it was accompanied by substantial improvements in out-of-domain performance, highlighting the trade-off between bias and variance in the model's design.

We have also proposed a full pipeline named DeepCERES that is able to segment standard resolution T1 MRI using superresolution and image synthesis deep learning-based techniques. Since a direct comparison with other state-of-the-art methods was not possible due to differences in protocols and resolutions, we performed a practical comparison to measure the sensitivity of DeepCERES and CerebNet in distinguishing between populations. Through the analysis of the ADNI dataset, we observed that DeepCERES has a demonstrated superior capability in separating individuals with Alzheimer's disease (AD) from cognitively normal (CN) controls, as evidenced by more significant structures identified and stronger effect sizes measured using Cohen's d (mean and max values). Notably, DeepCERES identified more structures with small effect sizes, providing valuable guidance for future research in Alzheimer's disease progression. These findings underscore the practical significance and potential clinical utility of our proposed method. Despite the high quality of the proposed method, we are aware that the proposed method has several limitations that warrant further exploration. Although we have demonstrated that our realistic data augmentation improves the model generalization we are aware that we are only simulating different brain shapes using the original HCP data intensity distribution. A richer augmentation would also include intensity patterns and textures not present on HCP data. Regarding the DeepCERES pipeline the quality of the resulting segmentation heavily depends on the quality of the synthesis and superresolution methods. Better synthesis and superresolution methods would definitively improve the results.

Future research lines are related to the commented limitations. We plan to improve superresolution and synthesis methods to reduce the gap between synthetic and real data. We will work also on transferring intensity patterns from real data to simulated data to enhance data augmentation. We also plan to use semi-supervised methods to improve the method generalization further. Additionally, we aim to include data from clinical populations with cerebellar pathologies in our training dataset, as this would enhance the applicability of the pipeline in diagnostic and research contexts. However, despite incorporating a more diverse range of training data, the method may still be susceptible to failure in cases involving large lesions, such as tumours (as seen in Fig. 10 in the Appendix section). In such instances, alternative lesion inpainting strategies will be explored to remove these regions early in the pipeline. This approach would allow the methodology to proceed as designed, minimizing the impact of out-of-domain data, and then reintroducing the lesion later in the process.

These additions would also allow the tool to support advanced clinical decision-making by identifying and analysing pathological structures.

Finally, we have made the proposed method openly accessible to the scientific community through our online system volBrain.net. This

system is very easy to use and requires no prior expertise or significant computational resources, unlike other methods currently available. Besides, its high efficiency (lower than 3 min) makes it ideal for large datasets analysis. We believe that DeepCERES pipeline bridges the gap between research-grade tools and clinical applications by enabling segmentation on standard-resolution images, making advanced cerebellar analyses (even legacy data) more accessible".

5. Conclusion

In this study, we introduce a novel method for cerebellum lobule segmentation utilizing deep learning on ultra-high resolution multimodal MR images. Our experiments confirm the hypothesis that leveraging ultra-high resolution and multimodal data enhances the precision of cerebellum lobule segmentation. Furthermore, we demonstrate the efficacy of combining classical methods with deep learning, incorporating a specific atlas as a priori information.

We introduce a new pipeline, DeepCERES, already publicly accessible through the online service volBrain.net. This pipeline is designed to seamlessly process standard-resolution T1 images, making it particularly valuable for analysing numerous existing legacy datasets.

The upcoming availability of DeepCERES opens avenues for in-depth analysis of normal and pathological cerebellar patterns, shedding new light on this structure's involvement in various neurological diseases.

CRediT authorship contribution statement

Sergio Morell-Ortega: Writing – original draft, Visualization, Validation, Methodology, Investigation, Data curation. **Marina Ruiz-Perez:** Software, Methodology, Investigation. **Marien Gadea:** Writing – review & editing, Data curation. **Roberto Vivo-Hernando:** Writing – review & editing, Formal analysis. **Gregorio Rubio:** Writing – review & editing, Formal analysis. **Fernando Aparici:** Writing – review & editing, Data curation. **Maria de la Iglesia-Vaya:** Writing – review & editing, Data curation. **Gwenaelle Catheline:** Writing – review & editing, Supervision, Data curation. **Boris Mansencal:** Software, Methodology. **Pierrick Coupé:** Writing – review & editing, Supervision, Project administration, Investigation, Formal analysis, Conceptualization. **José V. Manjón:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work has been developed thanks to the project PID2020-118608RB-I00 and to the project PID2023-152127OB-I00 of the Ministerio de Ciencia, Innovacion y Universidades de España. This work also benefited from the support of the project DeepvolBrain and HoliBrain of the French National Research Agency (ANR-18-CE45-0013 and ANR-23-CE45-0020-01). This study was achieved within the Laboratory of Excellence TRAIL ANR-10-LABX-57 for the BigDataBrain project. Moreover, we thank the Investments for the future Program IdEx Bordeaux (ANR-10- IDEX- 03- 02, HL-MRI Project), Cluster of excellence CPU and the CNRS.

Moreover, this work is based on multiple samples. We wish to thank all investigators of these projects who collected these datasets and made them freely accessible. The C-MIND data used in the preparation of this article were obtained from the C-MIND Data Repository (accessed in February 2015) created by the C-MIND study of Normal Brain Development. This is a multisite, longitudinal study of typically developing children from ages newborn through young adulthood conducted by Cincinnati Children's Hospital Medical Center and UCLA A listing of the participating sites and a complete listing of the study investigators can be found at link.

The NDAR data used in the preparation of this manuscript were obtained from the NIH-supported National Database for Autism Research (NDAR). NDAR is a collaborative informatics system created by the National Institutes of Health to provide a national resource to support and accelerate research in autism. The NDAR dataset includes data from the NIH Pediatric MRI Data Repository created by the NIH MRI Study of Normal Brain Development. This is a multisite, longitudinal study of typically developing children from ages newborn through young adulthood conducted by the Brain Development Cooperative Group A listing of the participating sites and a complete listing of the study investigators can be found at link.

The ADNI data used in the preparation of this manuscript were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI). The ADNI is funded by the National Institute on Ageing and the National Institute of Biomedical Imaging and Bioengineering and through generous contributions from the following: Abbott, AstraZeneca AB, Bayer Schering Pharma AG, Bristol-Myers Squibb, Eisai Global Clinical Development, Elan Corporation, Genentech, GE Healthcare, GlaxoSmithKline, Innogenetics NV, Johnson & Johnson, Eli Lilly and Co., Medpace, Inc., Merck and Co., Inc., Novartis AG, Pfizer Inc., F. Hoffmann-La Roche, Schering-Plough, Synarc Inc., as well as nonprofit partners, the Alzheimer's Association and Alzheimer's Drug Discovery Foundation, with participation from the U.S. Food and Drug Administration. Private sector contributions to the ADNI are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study was coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for NeuroImaging at the University of California, Los Angeles.

The OASIS data used in the preparation of this manuscript were obtained from the OASIS project. See link for more details. The AIBL data used in the preparation of this manuscript were obtained from the AIBL study of ageing. See www.aibl.csiro.au link for further details. The ICBM data used in the preparation of this manuscript. The IXI data used in the preparation of this manuscript were supported by the Brain Development.

The ABIDE data used in the preparation of this manuscript were supported by ABIDE funding resources listed at link. ABIDE primary support for the work by Adriana Di Martino. Primary support for the work by Michael P. Milham and the INDI team was provided by gifts from Joseph P. Healy and the Stavros Niarchos Foundation to the Child Mind Institute, link

Appendix

See Figs. 7–10.

Data availability

Data will be made available on request.

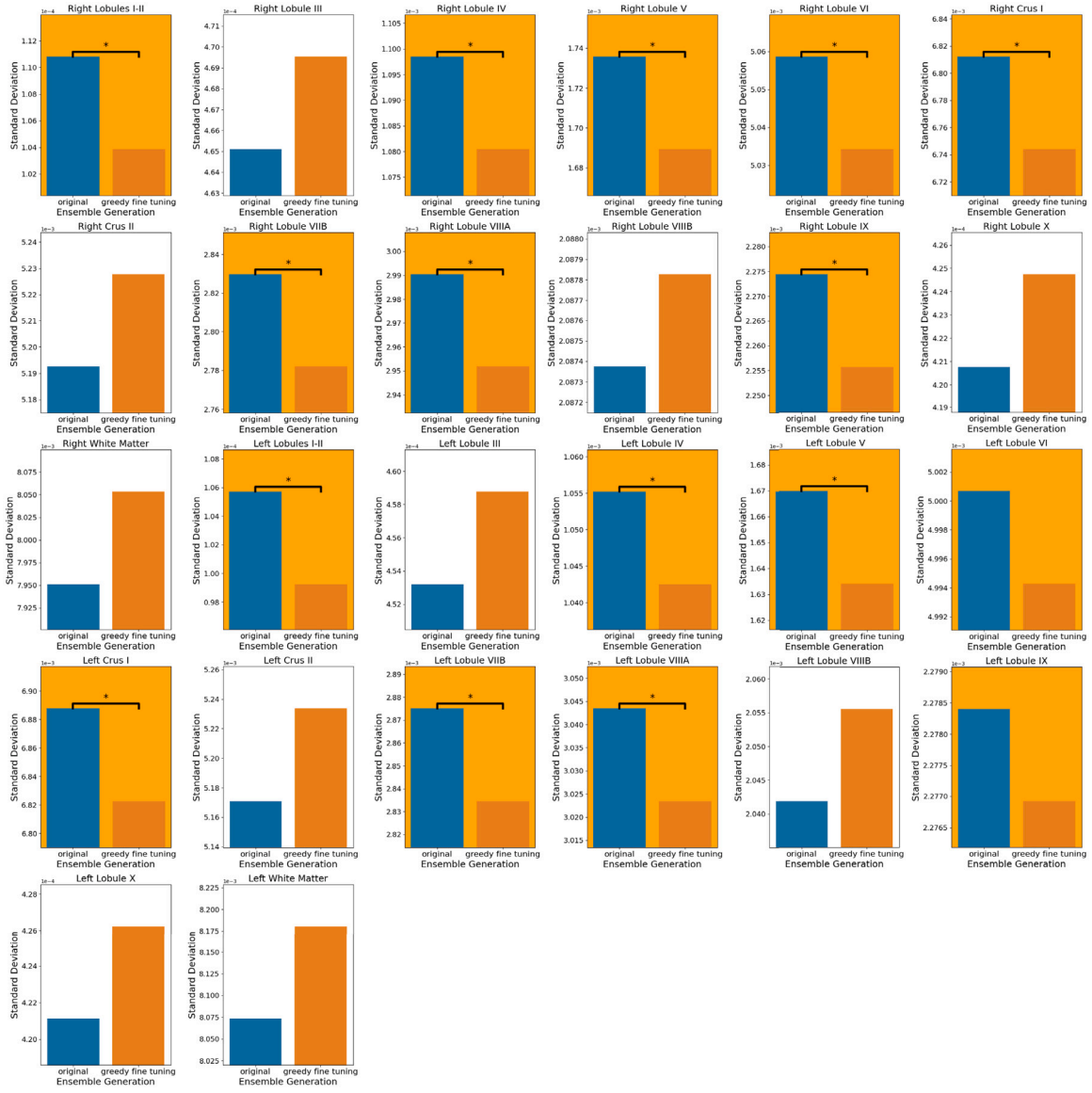


Fig. 7. Standard deviation of the volumes measured for each structure. * Significant differences in the mean values of the standard deviation of both methods when using the Wilcoxon two-sided test with $p < 0.05$, also marked with orange background.

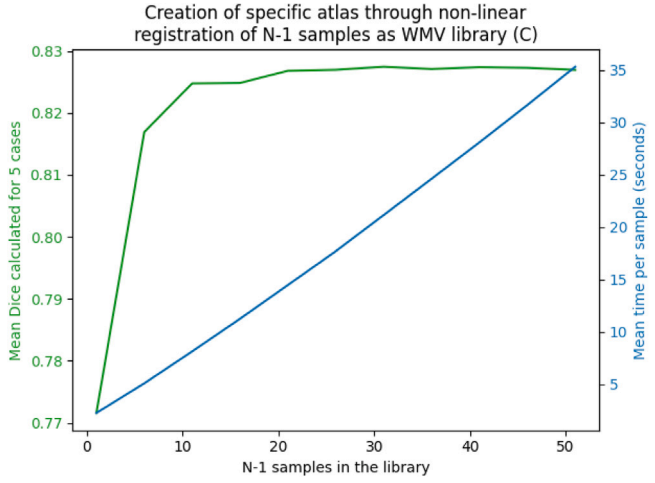


Fig. 8. Relationship between the number of samples in the library (N-1) and performance metrics: the mean Dice coefficient (green line) and the mean time per sample in seconds (blue line).

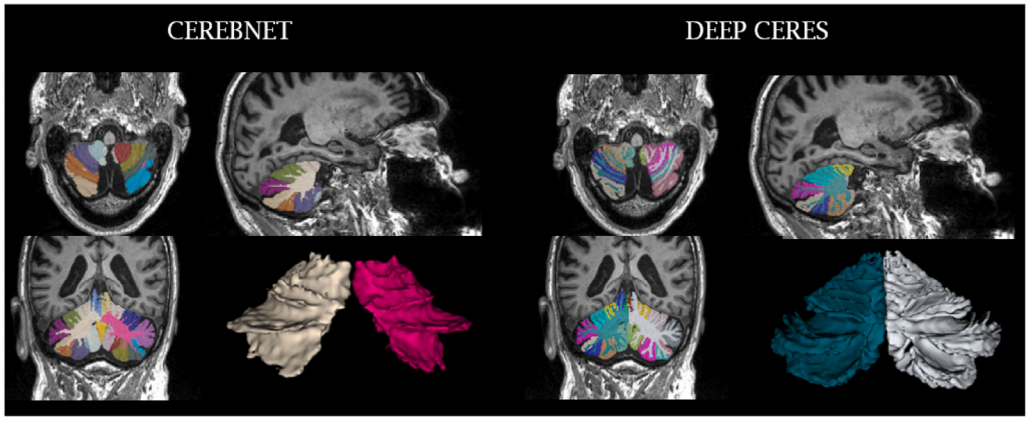


Fig. 9. Qualitative comparison of the cerebellar segmentation and 3D render of cerebellar white matter (*arbor vitae*) done by Cerebnet and DeepCERES methods of an Alzheimer-diagnosticated ADNI sample.

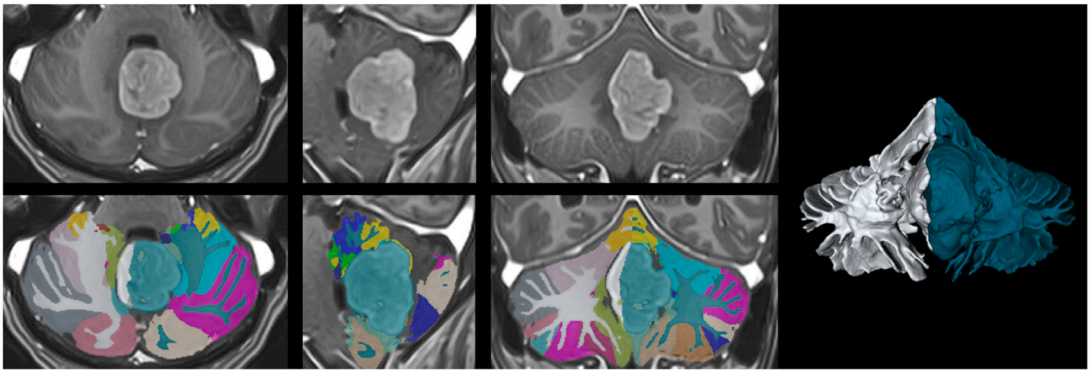


Fig. 10. Example of an out-of-domain case (a big tumour) in the cerebellum and its impact in the segmentation.

References

Adamaszek, M., D'Agata, F., Ferrucci, R., Habas, C., Keulen, S., Kirkby, K.C., Leggio, M., Mariën, P., Molinari, M., Moulton, E., Orsi, L., Van Overwalle, F., Papadelis, C., Priori, A., Sacchetti, B., Schutter, D.J., Styliadis, C., Verhoeven, J., 2016. Consensus Paper: Cerebellum and Emotion. Cerebellum 2016 16: 2 16 (2), 552–576.

Anon, Using OpenMP with C — Research Computing University of Colorado Boulder documentation.

Balakrishnan, Guha, Zhao, Amy, Sabuncu, Mert R., Guttag, John, Dalca, Adrian V., 2018. VoxelMorph: A Learning Framework for Deformable Medical Image Registration. IEEE Trans. Med. Imaging 38 (8), 1788–1800.

Battle, Mark R., Buckley, Chris J., Smith, Adrian, Van Laere, Koen, Vandenberghe, Rik, Lowe, Val J., Yushkevich, Paul A., Pluta, John, Wang, Hongzhi, Wisse, Laura E.M., Das, Sandhitsu, Wolk, David, 2016. Fast Automatic Segmentation of Hippocampal Subfields and Medial Temporal Lobe Subregions In 3 Tesla and 7 Tesla T2-Weighted MRI. Alzheimer' s & Dement. 12, P126–P127, 7S_Part_2.

Bede, Peter, Chipika, Rangariroyashe H., Christidi, Foteini, Hengeveld, Jennifer C., Karavasilis, Efstratios, Argyropoulos, Georgios D., Lope, Jasmin, Li Hi Shing, Stacey, Velonakis, Georgios, Dupuis, Léonie, Doherty, Mark A., Vajda, Alice, McLaughlin, Russell L., Hardiman, Orla, 2021. Genotype-associated cerebellar profiles in ALS: focal cerebellar pathology and cerebro-cerebellar connectivity alterations. J. Neurol. Neurosurg. Psychiatry 92 (11), 1197–1205.

Bogovic, John A., Jedynak, Bruno, Rigg, Rachel, Du, Annie, Landman, Bennett A., Prince, Jerry L., Ying, Sarah H., 2013. Approaching expert results using a hierarchical cerebellum parcellation protocol for multiple inexpert human raters. NeuroImage 64 (1), 616–629.

Carass, Aaron, Cuzzocreo, Jennifer L., Han, Shuo, Hernandez-Castillo, Carlos R., Rasser, Paul E., Ganz, Melanie, Beliveau, Vincent, Dolz, Jose, Ben Ayed, Ismail, Desrosiers, Christian, Thyreau, Benjamin, Romero, José E., Coupé, Pierrick, Manjón, José V., Fonov, Vladimir S., Collins, D. Louis, Ying, Sarah H., Onyike, Chiadi U., Crocetti, Deana, Landman, Bennett A., Mostofsky, Stewart H., Thompson, Paul M., Prince, Jerry L., 2018. Comparing fully automated state-of-the-art cerebellum parcellation from magnetic resonance images. NeuroImage 183, 150–172.

Chandrasekaran, Jayashree, Petit, Emilien, Park, Young Woo, du Montcel, Sophie Tezenas, Joers, James M., Deelchand, Dinesh K., Považan, Michal, Banan, Guita, Valabregue, Romain, Ehses, Philipp, Faber, Jennifer, Coupé, Pierrick, Onyike, Chiadi U., Barker, Peter B., Schmahmann, Jeremy D., Ratai, Eva Maria, Subramony, S.H., Mareci, Thomas H., Bushara, Khalaf O., Paulson, Henry, Durr, Alexandra, Klockgether, Thomas, Ashizawa, Tetsuo, Lenglet, Christophe, Öz, Gülin, Rosenthal, Liana, Burns, Matthew, Grobe-Einsler, Marcus, Oender, Demet, Kimmich, Okka, Roy, Nina, Ewenczyk, Claire, Heinzmann, Anna, Lallemant, Pauline, Heide, Solveig, Charles, Perrine, Coarelli, Giulia, Cunha, Paulina, Sayah, Sabrina, Wilmot, George, Scorr, Laura, Opal, Puneet, Sha, Sharon, Santini, Veronica, Sampson, Jacinda, Perlman, Susan, Geschwind, Michael, Nelson, Alexandra, Dietiker, Cameron, Gomez, Christopher, Hawkins, Trevor, Morrison, Peter, 2022. Clinically Meaningful Magnetic Resonance Endpoints Sensitive to Preataxic Spinocerebellar Ataxia Types 1 and 3. Ann. Neurol. 93 (4), 686–701.

Clouston, Sean A.P., Kritikos, Minos, Huang, Chuan, Kuan, Pei Fen, Vaska, Paul, Pellecchia, Alison C., Santiago-Michels, Stephanie, Carr, Melissa A., Gandy, Sam, Sano, Mary, Bromet, Evelyn J., Lucchini, Roberto G., Luft, Benjamin J., 2022. Reduced cerebellar cortical thickness in World Trade Center responders with cognitive impairment. Transl. Psychiatry 12 (1).

Cohen, Jacob, 2013. Statistical power analysis for the behavioral sciences. Stat. Power Anal. Behav. Sci..

Colloby, Sean J., O'Brien, John T., Taylor, John Paul, 2014. Patterns of cerebellar volume loss in dementia with Lewy bodies and Alzheimer's disease: A VBM-DARTEL study. Psychiatry Research: Neuroimaging 223, 187–191.

Coupé, Pierrick, Catheline, Gwenaelle, Lanuza, Enrique, Manjón, José Vicente, 2017. Towards a unified analysis of brain maturation and aging across the entire lifespan: A MRI analysis. Hum. Brain Mapp. 38 (11), 5501–5518.

Coupé, Pierrick, Manjón, José V., Fonov, Vladimir, Pruessner, Jens, Robles, Montserrat, Collins, D. Louis, 2011. Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation. NeuroImage 54 (2), 940–954.

Coupé, Pierrick, Mansencal, Boris, Clément, Michaël, Giraud, Rémi, Denis de Senneville, Baudouin, Ta, Vinh Thong, Lepetit, Vincent, Manjon, José V., 2020. AssemblyNet: A large ensemble of CNNs for 3D whole brain MRI segmentation. NeuroImage 219, 117026.

Desmond, John E., Fiez, Julie A., 1998. Neuroimaging studies of the cerebellum: language, learning and memory. Trends Cogn. Sci. 2 (9), 355–362.

Diedrichsen, Jörn, 2006. A spatially unbiased atlas template of the human cerebellum. NeuroImage 33 (1), 127–138.

Dolz, Jose, Desrosiers, Christian, Ben Ayed, Ismail, 2018. 3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study. NeuroImage 170, 456–470.

Faber, Jennifer, Kügler, David, Bahrami, Emad, Heinz, Lea Sophie, Timmann, Dagmar, Ernst, Thomas M., Deike-Hofmann, Katerina, Klockgether, Thomas, van de Warrenburg, Bart, van Gaalen, Judith, Reetz, Kathrin, Romanzetti, Sandro, Oz, Gulin, Joers, James M., Diedrichsen, Jorn, Giunti, Paola, Garcia-Moreno, Hector, Jacobi, Heike, Jende, Johann, de Vries, Jeroen, Povazan, Michal, Barker, Peter B., Steiner, Katherina Marie, Krahe, Janna, Reuter, Martin, 2022. CerebNet: A fast and reliable deep-learning pipeline for detailed cerebellum sub-segmentation. NeuroImage 264, 119703.

Giraud, Rémi, Ta, Vinh Thong, Papadakis, Nicolas, Manjón, José V., Collins, D. Louis, Coupé, Pierrick, 2016. An Optimized PatchMatch for multi-scale and multi-feature label fusion. NeuroImage 124, 770–782.

Guell, Xavier, Gabrieli, John D.E., Schmahmann, Jeremy D., 2018. Triple representation of language, working memory, social and emotion processing in the cerebellum: convergent evidence from task and seed-based resting-state fMRI analyses in a single large cohort. NeuroImage 172, 437–449.

Han, Shuo, Carass, Aaron, He, Yufan, Prince, Jerry L., 2020. Automatic cerebellum anatomical parcellation using U-Net with locally constrained optimization. NeuroImage 218, 116819.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, Sun, Jian, 2015. Deep Residual Learning for Image Recognition. Proc. the IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 770–778, 2016-December,

Henschel, Leonie, Conjeti, Sailesh, Estrada, Santiago, Diers, Kersten, Fischl, Bruce, Reuter, Martin, 2020. FastSurfer - A fast and accurate deep learning based neuroimaging pipeline. NeuroImage 219.

Huo, Yuankai, Xu, Zhoubing, Xiong, Yunxi, Aboud, Katherine, Parvathaneni, Prasanna, Bao, Shunxing, Bermudez, Camilo, Resnick, Susan M., Cutting, Laurie E., Landman, Bennett A., 2019. 3D whole brain segmentation using spatially localized atlas network tiles. NeuroImage 194, 105–119.

Jacobs, Heidi I.L., Hopkins, David A., Mayrhofer, Helen C., Bruner, Emiliano, Leeuwen, Fred W. Van, Raaijmakers, Wijnand, Schmahmann, Jeremy D., 2018. The cerebellum in Alzheimer's disease: evaluating its role in cognitive decline. Brain 141, 37–47.

Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., Rajchl, M., Lee, M., Kainz, B., Rueckert, D., Glocker, B., 2018. Ensembles of multiple models and architectures for robust brain tumour segmentation. Lect. Notes Comput. Sci. (Incl. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 10670 LNCS, 450–462,

King, Maedbh, Hernandez-Castillo, Carlos R., Poldrack, Russell A., Ivry, Richard B., Diedrichsen, Jörn, 2019. Functional boundaries in the human cerebellum revealed by a multi-domain task battery. Nat. Neurosci. 2019 22: 8 22 (8), 1371–1378.

Laidi, C., D'Albis, M.A., Wessa, M., Linke, J., Phillips, M.L., Delavest, M., Bellivier, F., Versace, A., Almeida, J., Sarrazin, S., Poupon, C., Le Dudal, K., Daban, C., Hamdani, N., Leboyer, M., Houenou, J., 2015. Cerebellar Volume in Schizophrenia and Bipolar I Disorder with and without Psychotic Features. Acta Psychiatr. Scand. 131 (3), 223.

Laidi, Charles, Floris, Dorothea L., Tillmann, Julian, Elandaloussi, Yannis, Zabihi, Mariam, Charman, Tony, Wolfers, Thomas, Durston, Sarah, Moessnang, Carolin, Dell'Acqua, Flavio, Ecker, Christine, Loth, Eva, Murphy, Declan, Baron-Cohen, Simon, Buitelaar, Jan K., Marquand, Andre F., Beckmann, Christian F., Frouin, Vincent, Leboyer, Marion, Duchesnay, Edouard, Coupé, Pierrick, Houenou, Josselin, 2022. Cerebellar Atypicalities in Autism? Biol. Psychiatry 92 (8), 674–682.

Laidi, C., Hajek, T., Spaniel, F., Kolenic, M., D'Albis, M.A., Sarrazin, S., Mangin, J.F., Duchesnay, E., Brambilla, P., Wessa, M., Linke, J., Polosan, M., Favre, P., Versace, A.L., Phillips, M.L., Manjon, J.V., Romero, J.E., Hozer, F., Leboyer, M., Coupe, P., Houenou, J., 2019. Cerebellar parcellation in schizophrenia and bipolar disorder. Acta Psychiatr. Scand. 140 (5), 468–476.

Li, Zhizhong, Hoiem, Derek, 2018. Learning without forgetting. IEEE Trans. Pattern Anal. Mach. Intell. 40, 2935–2947.

Lin, Chi Ying, Chen, Chi Hua, Tom, Sarah E., Kuo, Sheng Han, 2020. Cerebellar volume is associated with cognitive decline in mild cognitive impairment: results from ADNI. Cerebellum (Lond. England) 19, 217.

Lopez-Paz, David, Ranzato, Marc' Aurelio, 2017. Gradient episodic memory for continual learning. Adv. Neural Inf. Process. Syst. 30.

Manjón, José V., Coupé, Pierrick, 2016. Volbrain: An online MRI brain volumetry system. Front. Neuroinformatics 10 (JUL), 197669.

Manjón, José V., Coupé, Pierrick, Martí-Bonmatí, Luis, Collins, D. Louis, Robles, Montserrat, 2010. Adaptive non-local means denoising of MR images with spatially varying noise levels. J. Magn. Reson. Imaging 31 (1), 192–203.

Manjón, José V., Romero, José E., Coupe, Pierrick, 2021. Deep learning based MRI contrast synthesis using full volume prediction using full volume prediction. Biomed. Phys. Eng. Express 8 (1), 015013.

Manjón, José V., Romero, Jose E., Vivo-Hernando, Roberto, Rubio-Navarro, Gregorio, De la Iglesia-Vaya, María, Aparici-Robles, Fernando, Coupé, Pierrick, 2020. Deep ICE: A Deep learning approach for MRI Intracranial Cavity Extraction.

Manjón, José V., Tohka, Jussi, García-Martí, Gracian, Carbonell-Caballero, José, Lull, Juan J., Martí-Bonmatí, Luís, Robles, Montserrat, 2008. Robust MRI brain tissue parameter estimation by multistage outlier rejection. Magn. Reson. Med. 59 (4), 866–873.

Mariën, Peter, Ackermann, Herman, Adamaszek, Michael, Barwood, Caroline H.S., Beaton, Alan, Desmond, John, De Witte, Elke, Fawcett, Angela J., Hertrich, Ingo, Küper, Michael, Leggio, Maria, Marvel, Cherie, Molinari, Marco, Murdoch, Bruce E., Nicolson, Roderick I., Schmahmann, Jeremy D., Stoodley, Catherine J., Thürling, Markus, Timmann, Dagmar, Wouters, Ellen, Ziegler, Wolfram, 2014. Consensus paper: Language and the cerebellum: An ongoing enigma. Cerebellum 13 (3), 386–410.

McKenna, Mary Clare, Chipika, Rangariroyashe H., Li Hi Shing, Stacey, Christidi, Foteini, Lope, Jasmin, Doherty, Mark A., Hengeveld, Jennifer C., Vajda, Alice, McLaughlin, Russell L., Hardiman, Orla, Hutchinson, Siobhan, Bede, Peter, 2021. Infratentorial pathology in frontotemporal dementia: cerebellar grey and white matter alterations in FTD phenotypes. J. Neurol. 268 (12), 4687–4697.

Nai, Ying Hwey, Teo, Bernice W., Tan, Nadya L., Chua, Koby Yi Wei, Wong, Chun Kit, O'Doherty, Sophie, Stephenson, Mary C., Schaefferkoetter, Josh, Thian, Yee Liang, Chiong, Edmund, Reilhac, Anthonin, 2020. Evaluation of multimodal algorithms for the segmentation of multiparametric MRI prostate images. Comput. Math. Methods Med. 2020.

Ntatsis, Konstantinos, Dekker, Niels, van der Valk, Viktor, Birdsong, Tom, Zukić, Dženan, Klein, Stefan, Staring, Marius, McCormick, Matthew, 2023. itk-elastix: Medical image registration in Python. Proc. the 22nd Python Sci. Conf. 101–105.

Park, Min Tae M., Pipitone, Jon, Baer, Lawrence H., Winterburn, Julie L., Shah, Yashvi, Chavez, Sofia, Schira, Mark M., Lobaugh, Nancy J., Lerch, Jason P., Voineskos, Aristotle N., Chakravarty, M. Mallar, 2014. Derivation of high-resolution MRI atlases of the human cerebellum at 3 T and segmentation using multiple automatically generated templates. NeuroImage 95, 217–231.

Pérez-García, Fernando, Sparks, Rachel, Ourselin, Sébastien, 2021. TorchIO: A Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. Comput. Methods Programs Biomed. 208.

Ren, Jintao, Eriksen, Jesper Grau, Nijkamp, Jasper, Korreman, Stine Sofia, 2021. Comparing different CT, PET and MRI multi-modality image combinations for deep learning-based head and neck tumor segmentation. Acta Oncol. (Stock. Sweden) 60, 1399–1406.

Romero, Jose E., Coupé, Pierrick, Giraud, Rémi, Ta, Vinh Thong, Fonov, Vladimir, Park, Min Tae M., Chakravarty, M. Mallar, Voineskos, Aristotle N., Manjón, Jose V., 2017a. CERES: A new cerebellum lobule segmentation method. NeuroImage 147, 916–924.

Romero, José E., Coupé, Pierrick, Manjón, José V., 2017b. HIPS: A new hippocampus subfield segmentation method. NeuroImage 163, 286–295.

Ronneberger, Olaf, Fischer, Philipp, Brox, Thomas, 2015. U-net: Convolutional networks for biomedical image segmentation. Lect. Notes Comput. Sci. (Incl. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 9351, 234–241,

Schmahmann, J.D., Doyon, J., Petrides, M., Evans, A.C., 2000. MRI atlas of the human cerebellum.

Schmahmann, Jeremy D., Guell, Xavier, Stoodley, Catherine J., Halko, Mark A., 2019. The Theory and Neuroscience of Cerebellar Cognition. 42, 337–364. http://dx.doi.org/10.1146/ANNUREV-NEURO-070918-050258.

Schmahmann, Jeremy D., Sherman, Janet C., 1998. The cerebellar cognitive affective syndrome. Brain : A J. Neurol. 121 (Pt 4) (4), 561–579.

Shamonin, Denis P., Bron, Esther E., Lelieveldt, Boudewijn P.F., Smits, Marion, Klein, Stefan, Staring, Marius, 2014. Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer's disease. Front. Neuroinformatics 7 (JAN).

Sotelo, Constantino, 2008. Viewing the cerebellum through the eyes of Ramón Y Cajal. Cerebellum (Lond. England) 7 (4), 517–522.

Tustison, Nicholas J., Avants, Brian B., Cook, Philip A., Zheng, Yuanjie, Egan, Alexander, Yushkevich, Paul A., Gee, James C., 2010. N4ITK: improved N3 bias correction. IEEE Trans. Med. Imaging 29 (6), 1310–1320.

Van Essen, D.C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T.E.J., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S.W., Della Penna, S., Feinberg, D., Glasser, M.F., Harel, N., Heath, A.C., Larson-Prior, L., Marcus, D., Michalareas, G., Moeller, S., Oostenveld, R., Petersen, S.E., Prior, F., Schlaggar, B.L., Smith, S.M., Snyder, A.Z., Xu, J., Yacoub, E., 2012. The Human Connectome Project: a data acquisition perspective. NeuroImage 62 (4), 2222–2231.

Warren, Aaron E.L., Dalic, Linda J., Bulluss, Kristian J., BAppSci, Annie Roten, Thevathasan, Wesley, Archer, John S., 2022. The Optimal Target and Connectivity for Deep Brain Stimulation in Lennox-Gastaut Syndrome. Ann. Neurol. 92 (1), 61–74.

Weier, Katrin, Fonov, Vladimir, Lavoie, Karyne, Doyon, Julien, Louis Collins, D., 2014. Rapid automatic segmentation of the human cerebellum and its lobules (RASCAL)–implementation and application of the patch-based label-fusion technique with a template library to segment the human cerebellum. Hum. Brain Mapp. 35 (10), 5026–5039.

Yushkevich, Paul A., Piven, Joseph, Hazlett, Heather Cody, Smith, Rachel Gimpel, Ho, Sean, Gee, James C., Gerig, Guido, 2006. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. NeuroImage 31 (3), 1116–1128.

Zhou, Tongxue, Ruan, Su, Canu, Stéphane, 2019. A review: Deep learning for medical image segmentation using multi-modality fusion. Array 3–4, 100004.