Computational and Structural Biotechnology Journal 28 (2025) 128–140

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

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

Research article

Deep learning-based automatic segmentation of brain structures on MRI: A test-retest reproducibility analysis

Tomasz Puzio [a], Katarzyna Matera [a], Jan Karwowski [b], Joanna Piwnik [c], Sebastian Białkowski [b], Marek Podyma [b], Kosma Dunikowski [b], Małgorzata Siger [d], Mariusz Stasiołek [d], Piotr Grzelak [a], Ernest J. Bobeff [e,f,*,1]

[a] Department of Diagnostic Imaging, Polish Mothers' Memorial Hospital - Research Institute, Lodz, Poland
[b] Pixel Technology LLC, Lodz, Poland
[c] Department of Biostatistics and Translational Medicine, Medical University of Lodz, Lodz, Poland
[d] Department of Neurology, Medical University of Lodz, Lodz, Poland
[e] Department of Neurosurgery and Neuro-Oncology, Barlicki University Hospital, Medical University of Lodz, Lodz, Poland
[f] Department of Sleep Medicine and Metabolic Disorders, Medical University of Lodz, Lodz, Poland

ARTICLE INFO

Keywords:
Automated segmentation
Deep-learning neural network
Brain MRI
Test-retest
Brain atrophy

ABSTRACT

Objective: The aim of our study was to assess the reproducibility of deep learning-based automatic segmentation of brain structures in MRI scans across different scanner types and magnetic field strengths, particularly focusing on the comparison between 1.5 T and 3 T MRI scanners.

Methods: Our analysis encompassed a comprehensive examination of MRI images, focusing on the consistency of volumetric segmentation. We utilized advanced deep learning techniques with human-in-the-loop as a part of the workflow for segmenting brain structures and compared results across subsequent scans using the same and different scanner types.

Results: Our findings revealed high consistency in volumetric segmentation when comparing scans conducted on the same type of scanner (1.5 T to 1.5 T or 3 T to 3 T). The study revealed slightly better segmentation results for 1.5 T scanners compared to 3 T scanners when each was used independently. However, cross-comparisons between different scanner types (1.5 T vs. 3 T) demonstrated slightly less consistency, highlighting the influence of magnetic field strength on segmentation accuracy.

Conclusion: This study emphasizes the necessity of using the same scanner type and protocol for reliable MRI studies, particularly for brain atrophy monitoring. The high repeatability of deep learning-based segmentation under these conditions confirms its efficacy for clinical and research applications.

1. Introduction

Magnetic resonance imaging (MRI) is a highly precise diagnostic method and is widely used in clinical practice for visualizing brain anatomical structures and pathologies. The age-related decrease in brain tissue volume, including both white matter (WM) and grey matter (GM), and the corresponding expansion of the ventricular system and subarachnoid cerebrospinal fluid (CSF) reserve, is a normal process associated with aging [1]. However, in many diseases — particularly neurodegenerative ones such as Alzheimer's disease (AD), psychiatric disorders such as schizophrenia or inflammatory diseases like multiple sclerosis (MS) — the atrophy progresses more rapidly than in the general population [2–4].

Cerebral atrophy is visually assessed on T1-weighted MRI scans. This involves comparing consecutive studies of an individual taken over an extended time period, often spanning years. In clinical practice, qualitative descriptions such as 'mild,' 'moderate,' or 'severe' cortical/subcortical atrophy are commonly used. Additionally, semi-quantitative scales like the Scheltens scale, which is specifically designed for assessing medial temporal lobe atrophy (MTA) in AD, can be also employed [5].

With the advances made in imaging methods, it has become possible

* Corresponding author at: Department of SleepMedicine and Metabolic Disorders, Medical University of Lodz, Kopcinskiego St. 22, Lodz 90-153, Poland.
E-mail address: ernest.bobeff@umed.lodz.pl (E.J. Bobeff).
[1] ORCID: 0000-0003-1891-3791

https://doi.org/10.1016/j.csbj.2025.04.007
Received 30 December 2024; Received in revised form 1 April 2025; Accepted 3 April 2025
Available online 6 April 2025

to routinely image the brain using MR protocols that include isometric imaging series allowing the resulting image data to be analysed and the volume of brain structures to be calculated with high accuracy. Many atlas-based automated methods of segmentation [6–8] as well as deep neural networks (DNN) have been successfully applied for this task [6–12]. Cerebral atrophy is considered as one of the factors important for evaluating the course of aforementioned diseases and making therapeutic decisions [13–19].

The most critical issue is the accuracy of such segmentations and, specifically, its reproducibility. Given that the rate of atrophy in a normally aging brain is estimated to be around –0.05 % at 20–30 years of age, increasing to –0.3 % at 60–70 years of age [1], the average error in brain structure segmentation in longitudinal studies of an individual should be at most of the same order of magnitude. This precision is essential to ensure that an increased rate of atrophy can be reliably detected.

To assess the reproducibility of automated brain segmentation, a test-retest approach is often employed [20,21]. This involves a subject undergoing multiple MRI scans within a brief time period, allowing for a comparison of the volumetric measurements. The settings for these tests can vary significantly. They may include scans performed on a single MRI scanner or across various scanners, each potentially introducing different variables into the evaluation. This variability in the test-retest setup plays a crucial role in determining the robustness and generalizability of the segmentation process. This is particularly relevant for results produced by a DNN, which might yield more reliable outcomes on images from a specific MRI scanner if that scanner was predominantly used to prepare the training dataset. The optimal scenario is to train the DNN model using ground truth (GT) studies from different MRI scanners and then evaluate its reproducibility on studies obtained from other MRI scanners.

In this paper, we present an innovative approach for the segmentation of brain structures, employing a combination of DNN with manual corrections performed by radiologists in a human-in-the-loop (HITL) strategy. We then assessed the reproducibility of the resulting DNN in a test-retest setting on an independent sample. This evaluation included examinations conducted on the same MRI scanner, as well as on two scanners with a different field strength.

2. Methods

Our work comprised two stages. In the first stage, we used a sample of online-available MRI studies to create a DNN model for brain segmentation. In the second stage, we applied the developed DNN model to automatically segment MRI studies performed multiple times in the same patients, using both the same and different MRI scanners. We then evaluated the reproducibility of the segmentation. The study complied with declarations and regulations regarding human rights and was approved by the Institutional Review Board of Polish Mothers' Memorial Hospital - Research Institute (No. KB-58/2023). Patient data was anonymized prior to analysis. The manuscript was prepared in accordance with the CLAIM guidelines.

2.1. SAMSEG ground truth

We used MRI studies in healthy adults from the publicly available external IXI dataset [27]. FreeSurfer 7.2 tool, which utilizes the functionality of Sequence Adaptive Multimodal Segmentation (SAMSEG), was employed [7]. Automated SAMSEG segmentation of the brain structures was performed on T1- and T2-weighted isometric series of 522 MRI studies and served as the ground truth for the first DNN model training.

2.2. First DNN model

The GT (ground truth) images were split into the training and validation subsets, the latter containing 10 % of the sample. We utilized a basic U-Net neural network architecture [22]. We employed some popular modifications, including 3D convolutions (as MR images are three-dimensional), five resolution levels with 32, 32, 64, 128 and 256 features respectively, leaky ReLU activation function and layer normalization [23]. The input MRI scans are single-channel images, while target segmentations contain 40 classes, which enforces the model to have 1 channel at the input and 40 channels at the output.

Prior to being processed by the model, all the input images and segmentations were resampled to the common space, with isometric voxel size of 1 mm x 1 mm x 1 mm and RAS (Right, Anterior, Superior) orientation. The image voxels intensity was z-normalized. For each GT segmentation and each label, only the greatest connected component was preserved. In this paper, we refer to a group of pixels that are connected based on their intensity or colour values as a Connected Component Region (CCR). All the holes in the segmentations were filled with the greatest surrounding labels. After the pre-processing step, we employed some common, random augmentation techniques, including Gaussian noise and Gaussian blur, adding Gibb's artifacts, contrast modification, rotation and scaling. Since the model needs a constant input size and due to the limited GPU VRAM size, the 180 mm x 180 mm x 180 mm random crop is cut out from the image and segmentation pair at every training step. These crops are used as a direct input to the model during the training. This particular size meets the GPU constraints, while it gives possibly large and profitable context to the model.

During the 300 epochs of the training (~ 700 000 steps), the model's parameters (5.7 million) were adapted with the AdamW optimizer to minimize the sum of the Dice Loss and the pixel-wise Cross Entropy functions. The training was performed using the PyTorch framework in a distributed environment, which allowed an effective batch size of 8 samples.

The trained model was evaluated using Dice Coefficient on the test subset. During the evaluation, the whole images were used as a direct input to the model, i.e., all the augmentations including random cropping were skipped.

2.3. Amendments to the first DNN model

The automated segmentation by the first DNN model did not include a dedicated label for the corpus callosum (CC). As a result, voxels corresponding to this structure were consistently assigned to the right and left cerebral white matter, symmetrically divided around the mid-sagittal plane. To introduce a specific CC label, we applied a morphological operation (binary dilation) based on the existing grey and white matter labels of the left and right hemispheres. After dilation applied to both hemispheres, a shared area was created. The mid-sagittal plane was defined as the plane that minimizes the sum of squared distances of voxels within the shared area using Singular Value Decomposition. This approach allowed for the CC segmentation to be independent of the angle of brain inclination in the image. The largest connected component of the shared area and pertaining to the white matter within 3 mm from the mid-sagittal plane was labelled as CC Fig. 1.

The automated segmentation by the first DNN model also misrepresented the intracranial space boundary, which was observed most often on studies with hyperintense parts of the cranium on T1-weighted images due to increased amounts of fat in the bone marrow of the skull. Additionally, there were repeating errors regarding the segmentation of prepontine cistern/premedullary cistern which were generally undersegmented.

To address these issues, we selected 75 studies for further review by the radiologist. The 75 studies used in the second stage of training were selected based on their cerebrospinal fluid (CSF) segmentation results from the initial model. Specifically, the entire dataset was sorted by the total CSF volume predicted, and samples representing outliers on both the high and low ends of the distribution were identified. These outliers

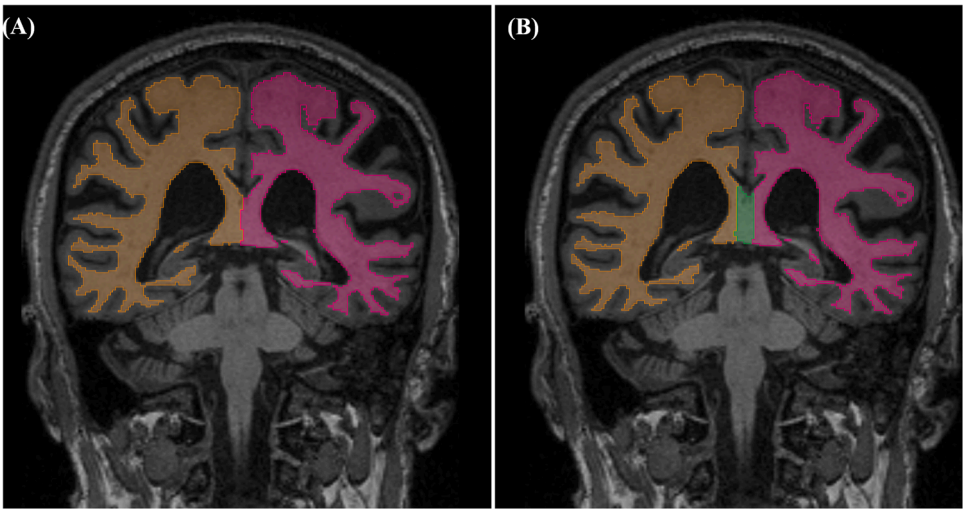

Fig. 1. DNN segmentation with the segmentation of only white matter of hemispheres visible, superimposed on T1-weighted sample scan from the training dataset. (A) Before applying the algorithm incorporating the corpus callosum label. (B) After applying the algorithm. The segmentation of the corpus callosum is visible in the mid-sagittal plane.

were then visually inspected to assess whether the segmentation errors appeared to be systematic. Final selection included 38 MRIs with inaccurately segmented intracranial space boundaries. Due to insufficient contrast for robust manual segmentation in the skull/CSF subarachnoid space interface on T1-weighted scans, we co-registered corresponding T2-weighted images with the T1-weighted images. The segmentations requiring correction were then transferred onto these images. The T2-weighted images, chosen for their clear contrast—marked by high signal intensity of the CSF and medium to low intensity of the skull—were exclusively utilized for skull/CSF boundary correction. Fig. 2. These corrections were performed using designed in-house tools available in Exhibeon3 DICOM Viewer (Pixel Technology LLC., Lodz, Poland) by two independent radiologists, T.P. and P.G. An example of a systematic error made by the first iteration of the DNN, along with the manually corrected segmentation, is shown in Fig. 3.

2.4. Final DNN model

The 75 manually corrected samples were utilized to perform the second training – 7 samples were randomly chosen as a test set (for model evaluation), while the remaining 68 were used for training. The trained model and the training procedure are very similar to the first training. The main differences are: the smaller but more accurate dataset (better annotations quality and added a missing label for CC), a slightly shorter training (6000 epochs long, around 400 000 steps), a simple learning rate scheduling with short warm up and slowly descending value, stronger regularization implemented as stronger augmentations, especially rotation, and scaling. Finally, after the training phase, the model was evaluated using a sliding window inference approach with 180 mm x 180 mm x 180 mm window size. Both the sliding window inference and stronger regularization turned out to be crucial to achieve satisfying results, when the new, improved dataset was utilized. The final model evaluated using the Dice coefficient gained an accuracy level equivalent to the average for all brain structures: 0.87. Example renderings of the segmentations are shown in Fig. 4. To assess whether the manual corrections effectively reduced systematic segmentation errors, we visually inspected all studies that had previously exhibited such errors—specifically those involving misrepresentation of the

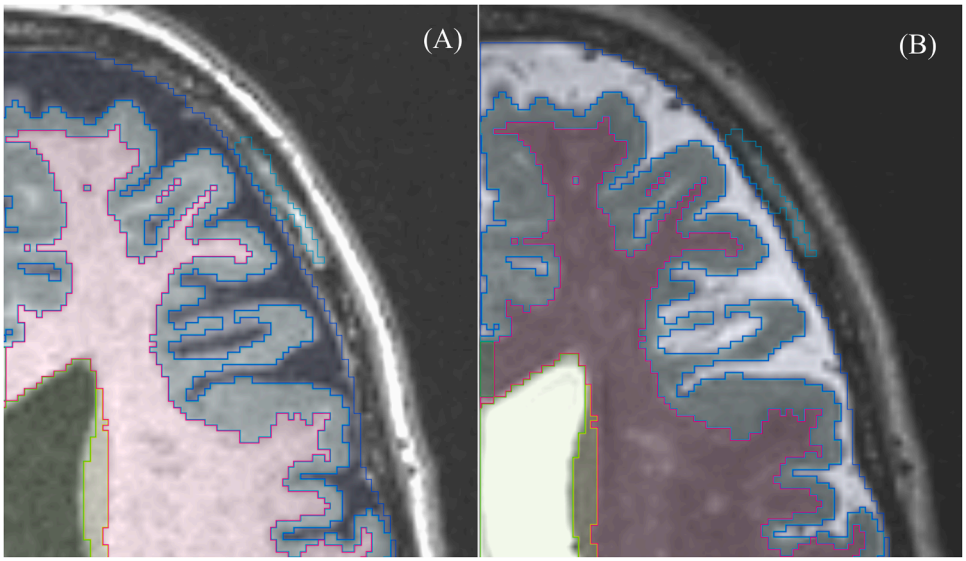

Fig. 2. DNN segmentation superimposed on (A) T1W scan (B) co-registered T2W scan. Incorrectly segmented fatty bone marrow within skull bone is seen. The boundary between CSF and skull is much better visible on T2W scan (B) which was subsequently used for manual correction.

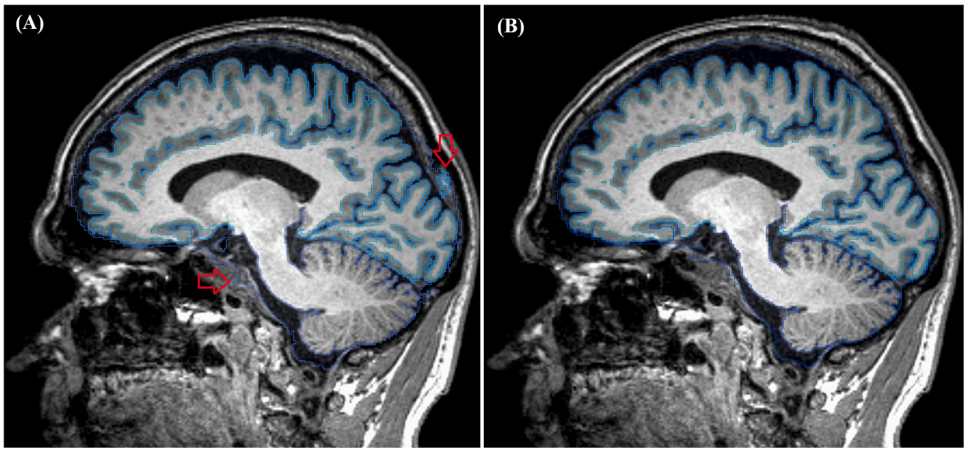

Fig. 3. DNN segmentation superimposed on T1W scan (A) with CSF-skull boundary errors indicated by red arrows, (B) after the manual correction.

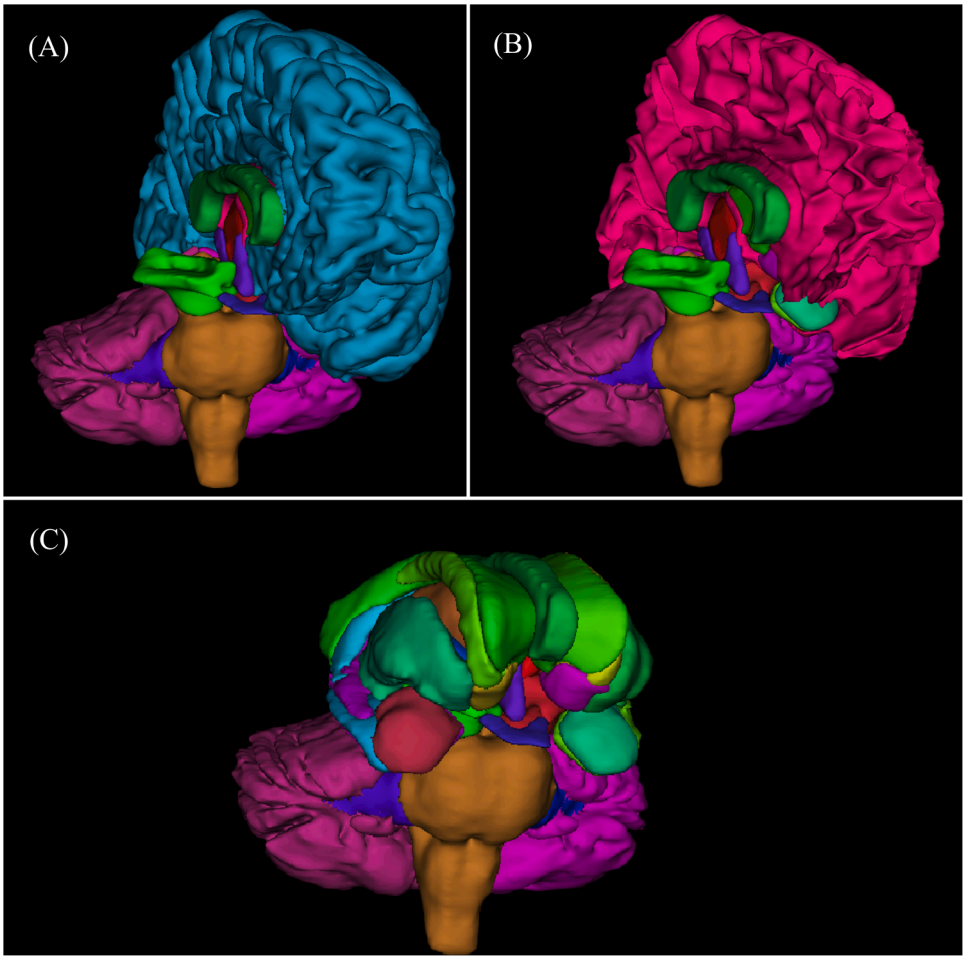

Fig. 4. 3D rendering of sample segmentation with (A) right hemisphere hidden, (B) right hemisphere (cortical grey matter, subcortical white matter, subcortical grey matter) and left cortical grey matter hidden, (C) both hemispheres hidden.

intracranial boundary and under-segmentation of the prepontine/premedullary cisterns. The updated model no longer demonstrated consistent errors in these regions, supporting the effectiveness of the human-in-the-loop refinement process.

2.5. Test-retest of the final DNN model

To evaluate the repeatability and subsequent clinical utility of the resulting DNN model, we decided to assess it using a test-retest approach. The study group consisted of 22 volunteers (10 females, 12 males) aged 25–73 (median 38 years) with no previous history of central nervous system disease. The patients gave their informed consent before the study.

Each participant underwent a total of four examinations, two on each MR scanner: 1.5 T (Ingenia, Philips, Netherlands) and 3 T (Achieva Smart Path to dStream, Philips, Netherlands). A T1-weighted 3-

dimensional fast field echo (FFE) scan was performed on both scanners. On 1.5 T scanner parameters were as follows: echo time (TE), 26 ms; repetition time (TR), 600 ms; flip angle (FA), 90°; and field of view (FOV), 230 mm with 1–1 contiguous sagittal 1-mm slices. On the 3 T scanner, these parameters had the following values: [TE], 4 ms; [TR], 8 ms; FA, 8°; and FOV, 256 mm with 1–1 contiguous sagittal 1-mm slices. Additionally, a 3D FLAIR sequence was performed during each study, in order to exclude any major brain pathologies. Resulting FLAIR images were not used for the volumetric analysis at any point. Repositioning occurred between the examinations, lasting from 5 to 10 min. During the break, participants walked to the other scanner (about 500 m), were not allowed to eat and remained in an upright position.

The statistical analysis was performed using Statistica 13.3 software, R programming, and the Python's Pandas, Pingouin, Statsmodels and Matplotlib packages. The comparison of intraclass correlation coefficient (ICC) obtained for 1.5 T and 3 T scanners was conducted using the Wilcoxon signed-rank test for paired samples.

We decided to use a two-way random model for average measurements (ICC (2, k)), as we are not concerned with the compatibility between two specific devices, utilized in our study, but rather with the repeatability of the results of the DNN model operating on randomly selected results of imaging studies.

Our research is based on calculations of the volume of 38 brain structures, which constitutes its added value and may be a valuable source of statistical data for other researchers. For this reason, we decided to include all obtained volumes together with some anatomically relevant aggregated volumes (Table 1).

Table 1
Components of aggregated structures from Table 2.

Lateral Ventricle with Choroid Plexus – Left	Lateral Ventricle, Temporal Horn of the Lateral Ventricle, Choroid plexus (for all only left)
Lateral Ventricle with Choroid Plexus – Right	Lateral Ventricle, Temporal Horn of the Lateral Ventricle, Choroid plexus (for all only right)
Brain Structures Supratentorial - Left	Cerebral White Matter, Thalamus, Caudate, Pallidum, Amygdala, Nucleus accumbens, Putamen, Cerebral Cortex, Hippocampus (for all only left)
Brain Structures Supratentorial - Right	Cerebral White Matter, Thalamus, Caudate, Pallidum, Amygdala, Nucleus accumbens, Putamen, Cerebral Cortex, Hippocampus (for all only right)
Cerebrospinal Fluid with Vessels – CSF	Cerebrospinal fluid – CSF, Vessel - Left, Vessel – Right
Cerebral Cortex	Cerebral Cortex – Left, Cerebral Cortex – Right
Cerebral and Cerebellar Cortex	Cerebral Cortex – Left, Cerebral Cortex – Right, Cerebellum Gray Matter – Left, Cerebellum Gray Matter – Right
Subcortical Gray Matter	Thalamus, Caudate, Pallidum, Amygdala, Nucleus accumbens, Putamen, Hippocampus (all both left and right)
Cerebral White Matter	Cerebral White Matter (left and right)
White Matter	Cerebral White Matter, Cerebellum White Matter (all both left and right)
Whole Brain (no CSF)	Cerebral White Matter, Thalamus, Caudate, Pallidum, Amygdala, Nucleus accumbens, Putamen, Cerebral Cortex, Hippocampus, Cerebellum White Matter, Cerebellum Gray Matter, Ventral Diencephalon (all both left and right) and Corpus Callosum, Stem, Cerebrospinal fluid – CSF, Optic Chiasm
Ventricular System	Lateral Ventricle, Temporal Horn of the Lateral Ventricle, Choroid plexus (all both left and right) and 3rd Ventricle, 4th Ventricle
Ventricular System + Subarachnoid Space	Lateral Ventricle, Temporal Horn of the Lateral Ventricle, Choroid plexus, Vessel (all both left and right) and 3rd Ventricle, 4th Ventricle, Cerebrospinal fluid – CSF,
Intracranial Volume	Cerebral White Matter, Thalamus, Caudate, Pallidum, Amygdala, Nucleus accumbens, Putamen, Cerebral Cortex, Hippocampus, Lateral Ventricle, Temporal Horn of the Lateral Ventricle, Choroid plexus, Cerebellum White Matter, Cerebellum Gray Matter, Ventral Diencephalon, Vessel (all both left and right) and Corpus Callosum, Stem, Cerebrospinal fluid – CSF, 3rd Ventricle, 4th Ventricle, Optic Chiasm
Brain Structures - Infratentorial	Cerebellum White Matter, Cerebellum Gray Matter, Ventral Diencephalon (all both left and right) and Stem

3. Results

All results are shown in Tables 2–5. A relevant commentary on the results is provided in the discussion section.

Tables 3–5 show the ICC results for comparing segmentations performed on two studies from the same 1.5 T scanner (Table 3), two studies from the same 3 T scanner (Table 4) and for the above four studies combined (Table 5).

For all structures and all scanner combinations, we obtained statistically significant ICCs, with the lowest ICC value being 0.8717. The mean value of ICC, calculated for all segmented structures (without aggregated structures) equalled 0.9921 (SD=0.0099) for 1.5 T scanner, 0.9855 (SD=0.0141) for 3 T scanner and 0.9785 (SD=0.0231) for both scanners together. The ICC value was, on average, higher for two studies conducted on a 1.5 T scanner than for two studies from a 3 T scanner, $p < .0001$.

Figs. 5–7 show Bland-Altman plots illustrating comparisons between different MRI machines (1.5 T vs. 3 T, left columns in each figure) and repeated measures on the same MRI scanner (Test vs. Retest, right columns in each figure). The Y-axes depict the differences between the two measurements for each intracranial structure, where one dot represents the difference in one patient. The two parallel dashed red lines represent the limits of agreement for each intracranial structure. A reduced spread in the limits of agreement in the right columns suggests greater agreement between measurements on the same MRI scanner for the majority of intracranial structures, including intracranial volume, cortex, subcortical grey matter, white matter, CSF, subarachnoid space, lateral ventricles, third ventricle, fourth ventricle, hippocampus, putamen, and pallidum. The somewhat equal limits of agreement in both the right and left columns were observed for the corpus callosum, thalamus, and caudate, indicating no significant difference, regardless of whether these structures were assessed on the same MRI scanner or on two scanners with different T value.

4. Discussion

Our study's findings underscore the reliability of volumetric segmentation in brain MRI examinations when conducted with uniformity in scanner type and protocols. Notably, our analyses consistently indicated that deviations in volumetric measurements of brain structures, when comparing subsequent scans performed on identical scanner types (1.5 T or 3 T), were significantly lower than the divergence in results when scans were cross-compared between different scanner types, specifically from 1.5 T to 3 T. This observation raises important considerations about the comparability and consistency of volumetric data across different MRI scanner types, suggesting a nuanced approach when interpreting such data across scanners of various field strengths, including those from different vendors or within generally heterogeneous scanning environments. The clinical implications of these differences, particularly in longitudinal studies using heterogeneous scanning setups, are further discussed in the Clinical application section.

Recent advancements in MRI technology have facilitated the use of 3 T machines, which offer higher magnetic field strengths. This enhancement leads to improved signal-to-noise ratios, allowing for higher resolution imaging and better clarity [24]. In contrast, 1.5 T

Table 2
Average volumes (calculated as a mean value from all examinations of all patients) with SD of all segmented structures plus calculated additional volumes of potential clinical significance.

Mean volumes obtained from DNN segmentation [ml]				
Symmetrical structures	**Left**		**Right**	
	Mean	**SD**	**Mean**	**SD**
Cerebral White Matter	219.139	22.397	218.303	22.225
Thalamus	7.042	0.694	6.646	0.551
Caudate	3.422	0.390	3.755	0.480
Pallidum	1.820	0.176	1.712	0.209
Amygdala	1.754	0.197	1.928	0.217
Nucleus accumbens	0.645	0.067	0.556	0.060
Putamen	5.572	0.585	5.157	0.597
Cerebral Cortex	233.481	18.518	231.543	18.564
Hippocampus	4.519	0.433	4.502	0.472
Lateral Ventricle	10.587	3.616	10.126	3.369
Temporal Horn of the Lateral Ventricle	0.334	0.204	0.535	0.272
Choroid plexus	0.472	0.145	0.479	0.142
Cerebellum White Matter	13.903	1.535	13.300	1.660
Cerebellum Gray Matter	53.305	4.445	52.918	4.609
Ventral Diencephalon	4.154	0.422	4.056	0.438
Vessel	0.063	0.029	0.045	0.022
Single structures	**Mean**		**SD**	
Corpus Callosum	3.948		0.528	
Stem	22.421		2.719	
Cerebrospinal fluid – CSF	335.393		69.151	
3rd Ventricle	1.500		0.613	
4th Ventricle	1.613		0.474	
Optic Chiasm	0.268		0.153	
Some additional aggregated volumes [ml]				
Symmetrical structures	**Left**		**Right**	
	Mean	**SD**	**Mean**	**SD**
Lateral Ventricle with Choroid Plexus	11.389	3.717	11.1419	3.522
Brain Structures Supratentorial	477.466	40.099	474.137	39.635
Single structures	**Mean**		**SD**	
Cerebrospinal Fluid with Vessels – CSF	335.538		69.095	
Cerebral Cortex	465.157		37.064	
Cerebral and Cerebellar Cortex	571.353		42.602	
Subcortical Gray Matter	49.024		4.043	
Cerebral White Matter	437.464		44.627	
White Matter	464.651		46.627	
Whole Brain (no CSF)	1145.528		91.890	
Ventricular System	25.645		7.507	
Ventricular System + Subarachnoid Space	361.091		73.839	
Intracranial Volume	1480.999		143.174	
Brain Structures - Infratentorial	164.069		13.985	

Table 3
ICC and percentage change for all segmented structures based on two 1.5 T scanner examinations. The percentage change was calculated for each patient as an absolute difference between two measurements of the volume of the same structure divided by the smaller of the two volumes, and then averaged across all patients.

Concordance of 1.5 T scanner-based segmentations (22 pairs)				
Symmetrical structures	**Left**		**Right**	
	ICC	**Percentage change**	**ICC**	**Percentage change**
Cerebral White Matter	0.9994	0.42 %	0.9996	0.29 %
Thalamus	0.9968	0.74 %	0.9958	0.86 %
Caudate	0.9966	1.05 %	0.9954	1.31 %
Pallidum	0.9916	1.25 %	0.9923	1.87 %
Amygdala	0.9890	1.88 %	0.9919	1.52 %
Nucleus accumbens	0.9753	2.96 %	0.9509	3.46 %
Putamen	0.9981	0.77 %	0.9961	1.22 %
Cerebral Cortex	0.9987	0.43 %	0.9982	0.60 %
Hippocampus	0.9948	1.17 %	0.9903	1.74 %
Lateral Ventricle	0.9997	0.92 %	0.9997	1.02 %
Choroid plexus	0.9949	3.40 %	0.9925	3.88 %
Cerebellum White Matter	0.9945	1.19 %	0.9968	1.04 %
Cerebellum Gray Matter	0.9985	0.50 %	0.9975	0.71 %
Ventral Diencephalon	0.9944	1.29 %	0.9971	0.91 %
Vessel	0.9830	11.54 %	0.9708	12.11 %
Single structures	**ICC**		**Percentage change**	
Corpus Callosum	0.9836		2.73 %	
Stem	0.999		0.56 %	
Cerebrospinal fluid – CSF	0.9989		1.10 %	
3rd Ventricle	0.9992		2.24 %	
4th Ventricle	0.9927		3.76 %	
Optic Chiasm	0.972		13.76 %	
Some additional aggregated volumes [ml]				
Symmetrical structures	**Left**		**Right**	
	ICC	**Percentage change**	**ICC**	**Percentage change**
Lateral Ventricle with Choroid Plexus	0.9997	0.90 %	0.9997	0.99 %
Brain Structures Supratentorial	0.9993	0.32 %	0.9993	0.35 %
Single structures	**ICC**		**Percentage change**	
Cerebrospinal Fluid with Vessels – CSF	0.9989		1.04 %	
Cerebral Cortex	0.9982		0.53 %	
Cerebral and Cerebellar Cortex	0.9988		0.40 %	
Subcortical Gray Matter	0.9988		0.45 %	
Cerebral White Matter	0.9995		0.34 %	
White Matter	0.9996		0.31 %	
Whole Brain (no CSF)	0.9996		0.24 %	
Ventricular System	0.9998		0.79 %	
Ventricular System + Subarachnoid Space	0.9990		1.00 %	
Intracranial Volume	0.9999		0.13 %	
Brain Structures - Infratentorial	0.9995		0.27 %	

machines, while more commonplace and accessible, exhibit lower sensitivity to artifacts. This attribute proves beneficial in certain scenarios, such as imaging patients with metal implants [25,26]. Interestingly, our study observed that results were more consistent when comparing segmentations in 1.5 T to 1.5 T scans than in 3 T to 3 T scans. This discrepancy may be attributed to a potential bias in the original ground-truth dataset, where more 1.5 T scans were present compared to 3 T scans [27]. Such a bias could have influenced the segmentation algorithms, favouring the 1.5 T results. Although exact proportions of 1.5 T and 3 T scans in the training dataset are not formally specified, metadata from the contributing datasets (such as IXI) indicate a likely overrepresentation of 1.5 T studies. This imbalance may have inadvertently shaped the model's performance, leading to greater consistency in 1.5 T segmentations. While the dataset was constructed to promote generalization, this observation underscores the need for more quantitatively balanced data across scanner types. Addressing this issue will be a key objective in future iterations of model development.

The integration of studies from both 1.5 T and 3 T scanners aligns with a fundamental principle of dataset creation for neural network training: the inclusion of highly diversified data. Numerous scientific publications indicate that increased dataset diversification mitigates overfitting during training and enhances model generalization. In our study, while we compare model performance on 1.5 T and 3 T studies, separate training was not considered a viable solution to the observed inferior performance on 3 T studies. Instead, in future research, we plan to augment the dataset with additional studies from other 3 T scanners to further diversify the data. We are confident that separate training on 1.5 T and 3 T studies would not yield superior results compared to our current approach. Moreover, our strategy aims to improve the model's robustness and applicability across various scanning conditions, ensuring better performance in real-world scenarios where both 1.5 T and 3 T scans are encountered.

In our investigation, we observed some degree of variability in the repeatability of segmentations for various brain structures, which warrants a nuanced examination. Although the vast majority of structures are segmented with an Intraclass Correlation Coefficient (ICC) of over 0.99, indicating high consistency, there are notable exceptions.

Table 4
ICC and percentage change for all segmented structures based on two 3 T scanner examinations. The percentage change was calculated for each patient as an absolute difference between two measurements of the volume of the same structure divided by the smaller of the two volumes, and then averaged across all patients.

Concordance of 3 T scanner-based segmentations (22 pairs)				
Symmetrical structures	**Left**		**Right**	
	ICC	**Percentage change**	**ICC**	**Percentage change**
Cerebral White Matter	0.9993	0.46 %	0.9991	0.41 %
Thalamus	0.9917	1.30 %	0.9868	1.61 %
Caudate	0.9978	0.89 %	0.9957	1.33 %
Pallidum	0.9685	3.10 %	0.9682	3.16 %
Amygdala	0.9763	3.04 %	0.9799	2.98 %
Nucleus accumbens	0.9546	3.55 %	0.9584	3.95 %
Putamen	0.9924	1.60 %	0.9944	1.56 %
Cerebral Cortex	0.9978	0.57 %	0.9981	0.57 %
Hippocampus	0.9844	1.84 %	0.9812	2.32 %
Lateral Ventricle	0.9997	1.11 %	0.9995	1.36 %
Choroid plexus	0.9921	4.50 %	0.9884	3.98 %
Cerebellum White Matter	0.9950	1.18 %	0.9878	2.53 %
Cerebellum Gray Matter	0.9965	0.83 %	0.9890	1.35 %
Ventral Diencephalon	0.9817	2.18 %	0.9936	1.37 %
Vessel	0.9811	15.96 %	0.9427	17.88 %
Single structures	**ICC**		**Percentage change**	
Corpus Callosum	0.9797		3.17 %	
Stem	0.9980		0.78 %	
Cerebrospinal fluid – CSF	0.9984		1.45 %	
3rd Ventricle	0.9990		2.41 %	
4th Ventricle	0.9889		4.76 %	
Optic Chiasm	0.9637		13.87 %	
Some additional aggregated volumes [ml]				
Symmetrical structures	**Left**		**Right**	
	ICC	**Percentage change**	**ICC**	**Percentage change**
Lateral Ventricle with Choroid Plexus	0.9996	1.17 %	0.9992	1.70 %
Brain Structures Supratentorial	0.9992	0.39 %	0.9993	0.36 %
Single structures	**ICC**		**Percentage change**	
Cerebrospinal Fluid with Vessels – CSF	0.9985		1.44 %	
Cerebral Cortex	0.9983		0.50 %	
Cerebral and Cerebellar Cortex	0.9984		0.47 %	
Subcortical Gray Matter	0.9980		0.54 %	
Cerebral White Matter	0.9995		0.40 %	
White Matter	0.9996		0.37 %	
Whole Brain (no CSF)	0.9995		0.30 %	
Ventricular System	0.9997		0.96 %	
Ventricular System + Subarachnoid Space	0.9986		1.35 %	
Intracranial Volume	0.9999		0.17 %	
Brain Structures - Infratentorial	0.9983		0.57 %	

Table 5
ICC and percentage change for all segmented structures based on two 1.5 T scanner and two 3 T scanner examinations. The percentage change was calculated for each patient as an absolute difference between maximum and minimum volume of the same structure divided by the minimum volume, and then averaged across all patients.

Concordance of 3 T and 1.5 T scanners-based segmentations (22 patients, 4 measurements each)				
Symmetrical structures	**Left**		**Right**	
	ICC	**Percentage change**	**ICC**	**Percentage change**
Cerebral White Matter	0.9969	1.81 %	0.9963	2.07 %
Thalamus	0.9820	4.98 %	0.9713	5.39 %
Caudate	0.9977	2.22 %	0.9976	2.54 %
Pallidum	0.9707	6.28 %	0.9525	9.93 %
Amygdala	0.9613	8.73 %	0.9746	6.93 %
Nucleus accumbens	0.9680	7.17 %	0.9476	9.80 %
Putamen	0.9927	3.50 %	0.9888	4.76 %
Cerebral Cortex	0.9972	1.57 %	0.9972	1.57 %
Hippocampus	0.9781	5.14 %	0.9844	4.89 %
Lateral Ventricle	0.9989	4.13 %	0.9991	3.88 %
Choroid plexus	0.9703	21.37 %	0.9710	18.09 %
Cerebellum White Matter	0.9757	6.42 %	0.9360	10.53 %
Cerebellum Gray Matter	0.9925	2.47 %	0.9830	3.27 %
Ventral Diencephalon	0.9936	3.27 %	0.9921	3.70 %
Vessel	0.9885	34.01 %	0.9717	35.47 %
Single structures	**ICC**		**Percentage change**	
Corpus Callosum	0.9726		8.76 %	
Stem	0.9820		5.89 %	
Cerebrospinal fluid – CSF	0.9946		5.67 %	
3rd Ventricle	0.9983		7.31 %	
4th Ventricle	0.9758		18.46 %	
Optic Chiasm	0.8717		69.62 %	
Some additional aggregated volumes [ml]				
Symmetrical structures	**Left**		**Right**	
	ICC	**Percentage change**	**ICC**	**Percentage change**
Lateral Ventricle with Choroid Plexus	0.9987	4.39 %	0.999	4.14 %
Brain Structures Supratentorial	0.9982	1.10 %	0.9985	1.09 %
Single structures	**ICC**		**Percentage change**	
Cerebrospinal Fluid with Vessels – CSF	0.9946		5.64 %	
Cerebral Cortex	0.9975		1.53 %	
Cerebral and Cerebellar Cortex	0.998		1.29 %	
Subcortical Gray Matter	0.9968		1.63 %	
Cerebral White Matter	0.9967		1.92 %	
White Matter	0.9958		2.14 %	
Whole Brain (no CSF)	0.9986		0.97 %	
Ventricular System	0.999		3.51 %	
Ventricular System + Subarachnoid Space	0.9954		5.25 %	
Intracranial Volume	0.9984		1.30 %	
Brain Structures - Infratentorial	0.9955		1.88 %	

Specifically, structures such as the ‘choroid plexus’, ‘vessels’, and ‘optic chiasm’ exhibit lower repeatability, with percentage changes ranging from 3.4 % (left choroid plexus at 1.5 T) to as high as 35.47 % (right vessels, across different field strengths). The observed inaccuracy can be attributed to several factors. For instance, the structure labelled as ‘vessels’, which mainly contained arteries of the circle of Willis and middle cerebral arteries within the Sylvian fissure, shows variable signal intensity on T1W sequences. This variation arises due to blood flow effects, as T1W sequences without gadolinium contrast are not optimized for arterial evaluation. In the case of the choroid plexus, its subtle motility within the lateral ventricles leads to variability in imaging. It’s important to note that volumetric measurements of these structures hold limited clinical significance in the context of brain atrophy assessment. Consequently, in our aggregated list of structures, they have been merged with the larger structures that encompass them - vessels with subarachnoid space, and choroid plexus with lateral ventricles.

The ’optic chiasm’ presents a different challenge, being the second smallest structure on our list with an average volume of only 0.268 ml. Generally, we found that the smaller the structure being segmented, the greater the changes observed between two scans. This is likely because, for smaller structures, the volumes of voxels on the edge constitute a significant proportion relative to the entire structure. Variations in voxel signal, which may shift slightly between studies, can significantly impact the segmentation accuracy, especially for voxels at the interface of different tissues like the grey-white matter boundary. These variations are partly due to the averaging algorithms within MRI scanners and the relative placement of the voxel grid to the tissue interfaces. This phenomenon could pose issues in scenarios where monitoring or using a single, small structure, such as the nucleus accumbens (averaging 0.6 ml

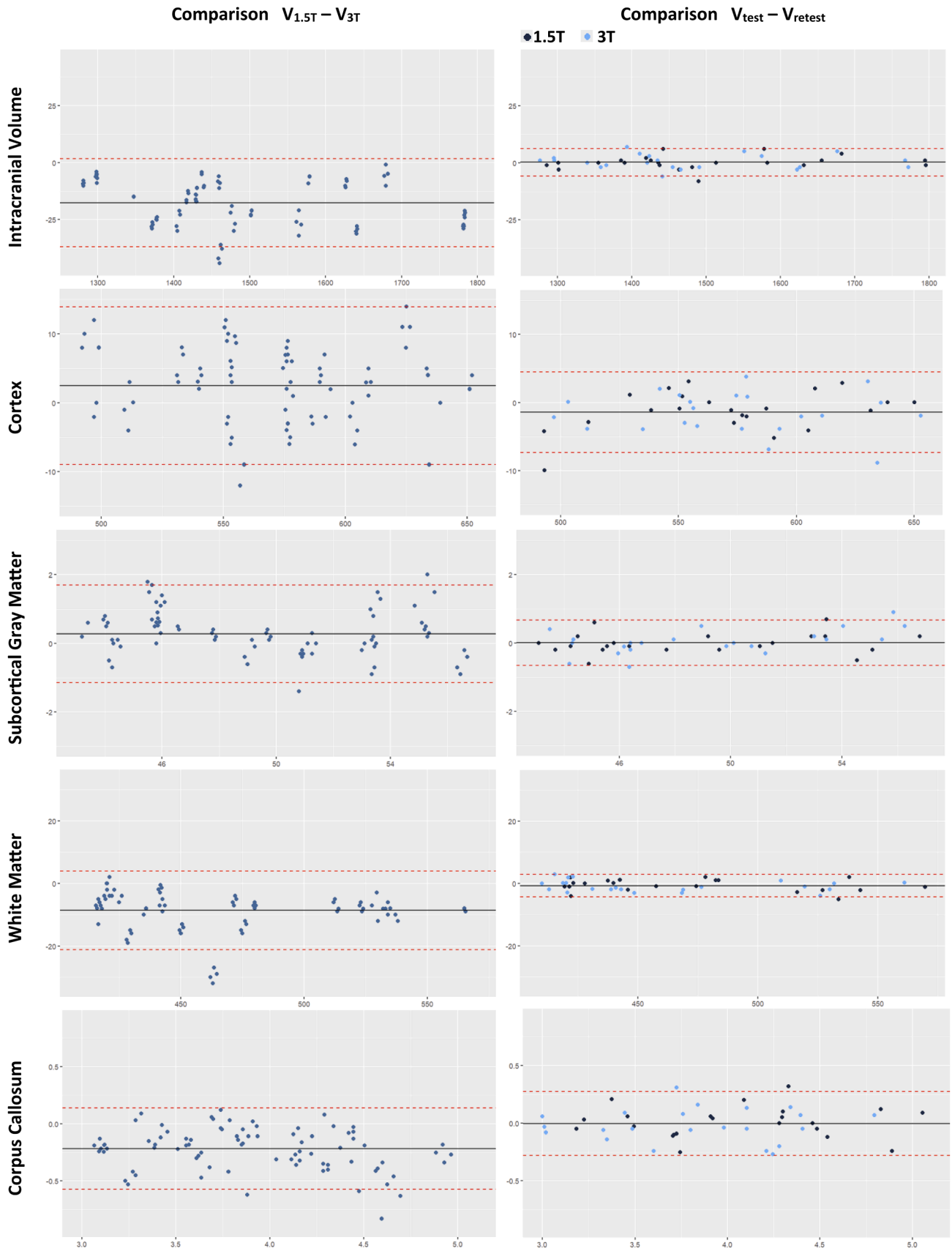

Fig. 5. Bland-Altman plots illustrating the comparisons between different MRI machines (1.5 T vs. 3 T) and repeated measures on the same scanner (Test vs. Retest) for various intracranial regions, including Intracranial Volume, Cortex, Subcortical Gray Matter, White Matter, and Corpus Callosum. The differences between the two measurements are depicted on the Y-axis, while the averages of the two measurements are represented on the X-axis, all values being provided in millilitres.

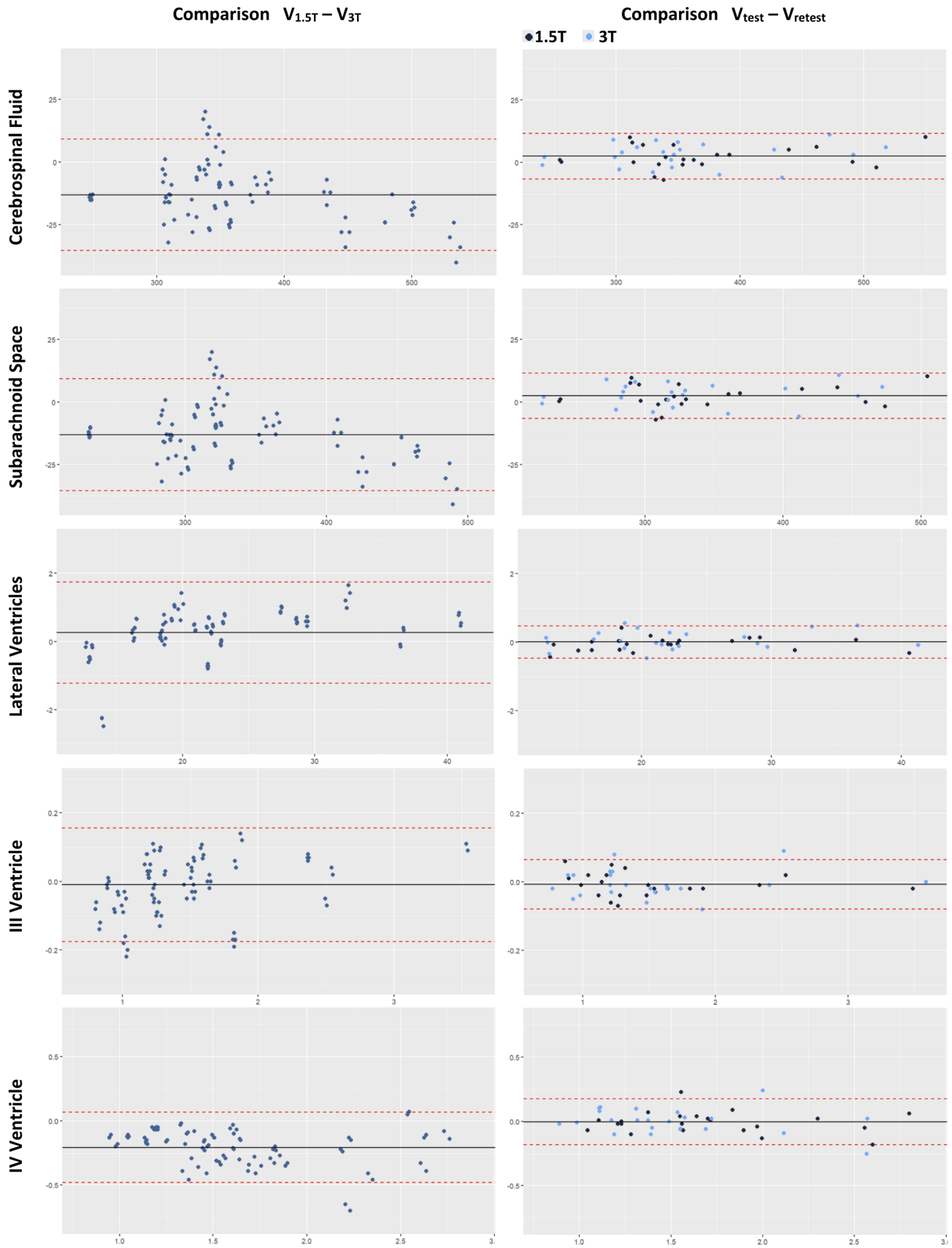

Fig. 6. Bland-Altman plots illustrating the comparisons between different MRI machines (1.5 T vs. 3 T) and repeated measures (Test vs. Retest) for various intracranial regions, including Cerebrospinal Fluid, Subarachnoid Space, Lateral ventricles, III Ventricle, and IV Ventricle. The differences between the two measurements are depicted on the Y-axis, while the averages of the two measurements are represented on the X-axis, all values being provided in millilitres.

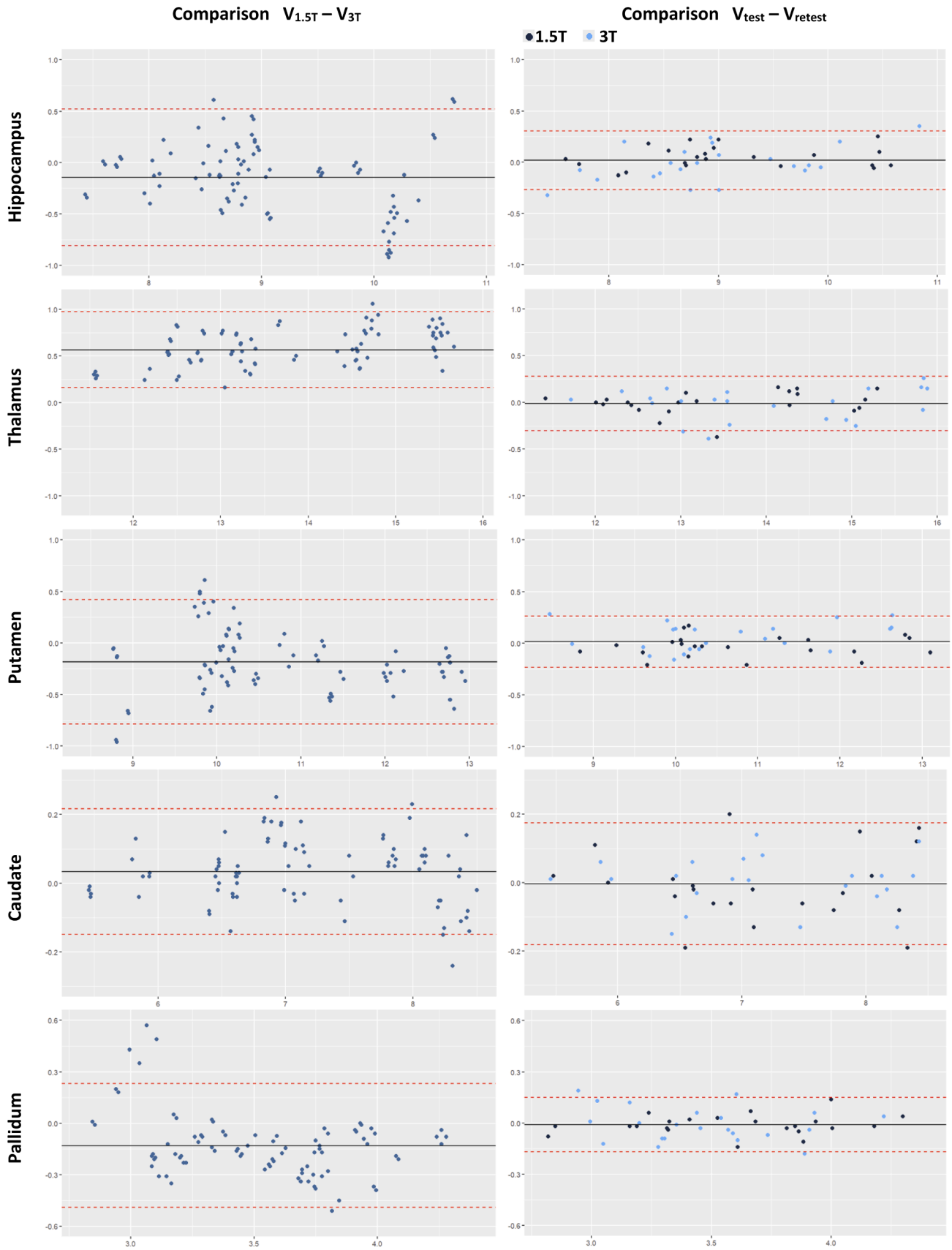

Fig. 7. Bland-Altman plots illustrating the comparisons between different MRI machines (1.5 T vs. 3 T) and repeated measures (Test vs. Retest) for various intracranial regions, including Hippocampus, Thalamus, Putamen, Caudate, and Pallidum. The differences between the two measurements are depicted on the Y-axis, while the averages of the two measurements are represented on the X-axis, all values being provided in millilitres.

across all studies) or aforementioned optic chiasm, is critical as a marker for disease.

Although the Dice coefficient is a widely used metric for evaluating segmentation accuracy, we do not consider it an appropriate measure of reproducibility—particularly in the context of brain MRI and longitudinal studies. Dice is sensitive to minor variations along anatomical boundaries, which are often inherently ambiguous in structures such as grey and white matter, and especially problematic in small-volume regions like the optic chiasm or vessels. Even small differences in voxel classification at structure edges can cause substantial changes in Dice scores, despite having little to no clinical significance. For this reason, while we report overall Dice values for completeness, our focus is on test-retest reproducibility as a more robust and clinically meaningful indicator of model reliability.

4.1. Clinical application

According to the 'MAGNIMS consensus recommendations on the use of brain and spinal cord atrophy measures in clinical practice,' measuring global brain volume loss is currently recommended as the most reliable predictor of disability progression in MS [28]. In our study, we quantified the mean difference in global brain volume (excluding cerebrospinal fluid, CSF) between two scans. The differences were expressed as a percentage of the total brain volume. We observed a mean difference of 0.24 % for 1.5 T MRI scans compared with other 1.5 T scans, 0.30 % for 3 T scans compared with other 3 T scans, and 0.97 % when comparing 1.5 T scans to 3 T scans. The repeatability of our method for scans conducted on the same type of scanner (either 1.5 T or 3 T) is comparable to the yearly brain atrophy rate observed in normal aging. However, it is lower than the accelerated atrophy annual rate seen in MS for which the cut-off rate is estimated at −0.4 % [1,4]. This suggests that our approach could be considered for monitoring disease progression in MS patients. While segmentation performance remains high across different scanners, the slight reduction in ICC and increased variability for different scanners further highlights a broader challenge in quantitative MRI, and underscores the importance of consistent scanning protocols in longitudinal studies. Our findings are in line with existing literature and reaffirm the need for harmonized acquisition settings when volumetric measurements are used for tracking changes over time.

Furthermore, there is growing evidence that region specific atrophy can be also used for monitoring MS, i.e., deep grey matter (GM), temporal cortical GM, bilateral thalamus, pre/postcentral regions and cingulate gyrus [29,30]. Our solution has an excellent repeatability for subcortical grey matter as a whole: mean change 0,45 % (1,5 T) and 0, 54 % (3 T) and slightly worse for thalamus: mean change 0,74 % and 0, 86 % (left and right, respectively, for 1.5 T) and 1.30 % and 1,61 % (left and right, respectively, for 3 T). Therefore, it could be a viable tool used for research further evaluating those region-specific biomarkers of MS progression.

One of the consequences of Alzheimer's disease (AD) is an increased rate of atrophy in both the cerebral cortex and hippocampus [2]. In our approach, the cortex is segmented with high repeatability for scanners of the same field strength; the mean change is 0.40 % for 1.5 T and 0.47 % for 3 T scanners. The hippocampus, being a much smaller structure, presents slightly lower repeatability in its segmentation. Nonetheless, the mean change for the hippocampus was 1.17 % and 1.74 % (left and right, respectively, for 1.5 T) and 1.84 % and 2.32 % (left and right, respectively, for 3 T), which remains lower than the estimated annualized rate of hippocampal atrophy reported in AD patients—3.0 % and 3.6 % for slow and fast progressors, respectively, as reported by Jack et al. [15]. Additionally, the same study found that the annualized volume increase of the lateral ventricles is among the best structural MRI biomarkers for distinguishing AD, mild cognitive impairment (MCI), and normal aging, with rates ranging from +1.7 % (normal cognition) to +6.4 % (AD, fast progression). These values are also well within the test-retest repeatability range of our model, where the mean change in ventricular system volume was 0.79 % for 1.5 T and 0.96 % for 3 T.

Recent studies have also demonstrated that the quality of structural MRI images can be significantly enhanced using super-resolution techniques, which can further improve the detection of MCI by deep learning models trained on perceptually optimized inputs [31]. Such approaches are complementary to our focus on segmentation repeatability, and together highlight the importance of both high-quality input data and robust model consistency for clinical applications in early neurodegenerative disease detection. We acknowledge, however, that our model currently does not segment hippocampal subfields, which are recognized as critical for early diagnosis and staging of AD. Recent research has shown that specific subregions—such as the hippocampal fissure, dentate gyrus, and CA4—undergo atrophy at different stages of disease progression, and that inter-hemispheric asymmetries in these patterns are closely linked with memory decline [32]. Incorporating subfield-level analysis remains an important area for future development to enhance the sensitivity of our method in detecting early neurodegenerative changes.

5. Limitations

There are several limitations of our study concerning both the test-retest procedure and the broader applicability of our findings in clinical practice.

First, although we evaluated reproducibility across both 1.5 T and 3 T field strengths, the paired scans were always acquired on scanners from the same vendor (Philips). This design limits our ability to assess scanner-specific effects that might arise from different manufacturers. A more robust evaluation would include test-retest scans from multiple vendors to fully capture the variability introduced by differing hardware and reconstruction pipelines.

Second, while the validation dataset used in our study was independent from the training data and originated from a different clinical population and imaging protocol, it is not publicly available. This restricts the ability of other researchers to replicate our findings or benchmark against our results directly. We recognize that publicly available datasets play a critical role in promoting transparency, reproducibility, and progress in the field.

To advance the field of quantitative neuroimaging, we believe there is a strong need for the creation of a large-scale, well-annotated, publicly available test-retest dataset that includes scans from multiple vendors, institutions, and patient populations. Such a dataset would allow for standardized assessment of segmentation reproducibility and generalizability, facilitating more direct comparisons between different models and fostering the development of clinically robust solutions. We hope future initiatives, potentially involving multicentre collaboration, will address this important gap.

To the best of our knowledge, no publicly available dataset currently provides structural T1-weighted MRI with a true test-retest design across multiple scanner vendors. Existing resources such as Kirby 21, HCP Test-Retest, and MASiVar either rely on a single scanner or are limited to diffusion imaging [33–35]. This highlights the need for more comprehensive, publicly accessible datasets that incorporate scanner and site variability in test-retest settings—something we aim to contribute to in future work.

Another important limitation is that both the training/validation dataset and the scans in the test-retest dataset were obtained from healthy subjects. Further evaluation of our DNN's performance on scans from individuals with conditions associated with accelerated atrophy rates and/or other lesions, such as white matter hypointensities on T1-weighted images ('black holes') typically seen in MS, is necessary.

6. Conclusions

A pivotal conclusion from our study is the recommendation to

conduct follow-up MRI studies on the same scanner for longitudinal atrophy measurements. In situations where using the same scanner is not feasible, such as due to scanner replacement or upgrade, it is advisable to conduct a new baseline study. This approach ensures the reliability and comparability of the data across different time points.

Furthermore, our study demonstrates that segmentation achieved through a DNN, which has been refined via a human-in-the-loop process, exhibits repeatability that is sufficiently robust for assessing the rate of atrophy. The inclusion of a human-in-the-loop process significantly improved the reliability of deep learning-based segmentation, bridging the gap between automated analysis and expert-level precision. This applies not only to the brain as a whole but also to its separate tissues, particularly in clinical scenarios where a condition increases the rate of atrophy.

CRediT authorship contribution statement

Matera Katarzyna: Writing – original draft, Project administration, Methodology, Investigation, Data curation, Conceptualization. **Bobeff Ernest J.:** Writing – review & editing, Visualization, Validation, Supervision, Investigation, Formal analysis, Data curation. **Puzio Tomasz:** Writing – original draft, Visualization, Resources, Methodology, Investigation, Data curation, Conceptualization. **Dunikowski Kosma:** Software, Resources. **Siger Małgorzata:** Writing – review & editing, Resources. **Stasiołek Mariusz:** Writing – review & editing. **Grzelak Piotr:** Writing – review & editing, Data curation. **Karwowski Jan:** Writing – original draft, Software, Methodology, Data curation. **Piwnik Joanna:** Writing – original draft, Visualization, Formal analysis. **Białkowski Sebastian:** Software, Resources. **Podyma Marek:** Writing – review & editing, Conceptualization.

Funding

The study detailed in this article received financial backing from the "RADi –Asystent Radiologa" project. This project was jointly funded by the European Union and the European Regional Development Fund, as part of the Smart Growth Operational Programme for the years 2014–2020. It falls under the Priority Axis focusing on "Enterprise R&D Support", specifically within Action 1.1, titled "Enterprise R&D Projects", and Sub-measure 1.1.1, which deals with "Industrial Research and Development Activities Conducted by Enterprises". The funding contributed to both the development of the deep learning neural network and the technical assistance essential for this research.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used OpenAI's ChatGPT to improve language and readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of Competing Interest

The authors declare a potential conflict of interest and state it below

Author Sebastian Bialkowski, Kosma Dunikowski, Marek Podyma and Jan Karwowski are employed by Pixel Technology. The remaining authors affirm that the research was conducted without any existing commercial or financial affiliations that could be construed as potential conflicts of interest. The article's presented results are exclusively derived from the data collected and analysed in alignment with the study's objectives and methodologies. The study findings were unaffected by the funding source, and the authors affirm the absence of any conflicts of interest associated with the research.

References

[1] Battaglini M, Gentile G, Luchetti L, Giorgio A, Vrenken H, Barkhof F, et al. Lifespan normative data on rates of brain volume changes. Neurobiol Aging 2019;81:30–7.
[2] Bakkour A, Morris JC, Wolk DA, Dickerson BC. The effects of aging and Alzheimer's disease on cerebral cortical anatomy: specificity and differential relationships with cognition. NeuroImage 2013;76:332–44.
[3] Bethlehem RAI, Seidlitz J, White SR, Vogel JW, Anderson KM, Adamson C, et al. Brain charts for the human lifespan. Nature 2022;604(7906):525–33.
[4] De Stefano N, Stromillo ML, Giorgio A, Bartolozzi ML, Battaglini M, Baldini M, et al. Establishing pathological cut-offs of brain atrophy rates in multiple sclerosis. J Neurol Neurosurg Psychiatry 2016;87(1):93–9.
[5] Scheltens P, Launer LJ, Barkhof F, Weinstein HC, Gool WA. Visual assessment of medial temporal lobe atrophy on magnetic resonance imaging: Interobserver reliability. J Neurol 1995;242(9):557–60.
[6] Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, et al. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 2002;33(3):341–55.
[7] Puonti O, Iglesias JE, Van Leemput K. Fast and sequence-adaptive whole-brain segmentation using parametric bayesian modeling. NeuroImage 2016;143:235–49.
[8] Zhang Y, Brady M, Smith S. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans Med Imaging 2001;20(1):45–57.
[9] Guha Roy A, Conjeti S, Navab N, Wachinger C. QuickNAT: a fully convolutional network for quick and accurate segmentation of neuroanatomy. NeuroImage 2019;186:713–27.
[10] Henschel L, Conjeti S, Estrada S, Diers K, Fischl B, Reuter M. FastSurfer - a fast and accurate deep learning based neuroimaging pipeline. NeuroImage 2020;219:117012.
[11] Mehta R, Majumdar A, Sivaswamy J. BrainSegNet: a convolutional neural network architecture for automated segmentation of human brain structures. J Med Imaging 2017;4(2):024003.
[12] Wachinger C, Reuter M, Klein T. DeepNAT: deep convolutional neural network for segmenting neuroanatomy. NeuroImage 2018;170:434–45.
[13] Contador J, Pérez-Millán A, Tort-Merino A, Balasa M, Falgàs N, Olives J, et al. Longitudinal brain atrophy and CSF biomarkers in early-onset Alzheimer's disease. NeuroImage Clin 2021;32:102804.
[14] Di Filippo M, Anderson VM, Altmann DR, Swanton JK, Plant GT, Thompson AJ, et al. Brain atrophy and lesion load measures over 1 year relate to clinical status after 6 years in patients with clinically isolated syndromes. J Neurol Neurosurg Psychiatry 2010;81(2):204–8.
[15] Jack CR, Shiung MM, Gunter JL, O'Brien PC, Weigand SD, Knopman DS, et al. Comparison of different MRI brain atrophy rate measures with clinical disease progression in AD. Neurology 2004;62(4):591–600.
[16] Mak E, Su L, Williams GB, Firbank MJ, Lawson RA, Yarnall AJ, et al. Longitudinal whole-brain atrophy and ventricular enlargement in nondemented Parkinson's disease. Neurobiol Aging 2017;55:78–90.
[17] ten Kate M, Dicks E, Visser PJ, van der Flier WM, Teunissen CE, Barkhof F, et al. Atrophy subtypes in prodromal Alzheimer's disease are associated with cognitive decline. Brain 2018;141(12):3443–56.
[18] Vemuri P, Wiste HJ, Weigand SD, Knopman DS, Trojanowski JQ, Shaw LM, et al. Serial MRI and CSF biomarkers in normal aging, MCI, and AD. Neurology 2010;75(2):143–51.
[19] Whitwell JL, Josephs KA, Murray ME, Kantarci K, Przybelski SA, Weigand SD, et al. MRI correlates of neurofibrillary tangle pathology at autopsy. Neurology 2008;71(10):743–9.
[20] de Boer R, Vrooman HA, Ikram MA, Vernooij MW, Breteler MMB, van der Lugt A, et al. Accuracy and reproducibility study of automatic MRI brain tissue segmentation methods. NeuroImage 2010;51(3):1047–56.
[21] Liu S, Hou B, Zhang Y, Lin T, Fan X, You H, et al. Inter-scanner reproducibility of brain volumetry: influence of automated brain segmentation software. BMC Neurosci 2020;21(1):35.
[22] Ronneberger O., Fischer P., Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation [Internet]. arXiv; 2015 [cited 2024 Jan 3]. Available from: http://arxiv.org/abs/1505.04597.
[23] Isensee F, Petersen J, Klein A, Zimmerer D, Jaeger PF, Kohl S, et al. nnU-Net: Self-adapting Framework for U-Net-Based. Med Image Segm [Internet] 2018 (arXiv).
[24] Tanenbaum LN. Clinical 3T MR imaging: mastering the challenges. Magn Reson Imaging Clin N Am 2006;14(1):1–15.
[25] Olsrud J, Lätt J, Brockstedt S, Romner B, Björkman-Burtscher IM. Magnetic resonance imaging artifacts caused by aneurysm clips and shunt valves: dependence on field strength (1.5 and 3 T) and imaging parameters. J Magn Reson Imaging 2005;22(3):433–7.
[26] Sundseth J, Jacobsen EA, Kolstad F, Nygaard OP, Zwart JA, Hol PK. Magnetic resonance imaging evaluation after implantation of a titanium cervical disc prosthesis: a comparison of 1.5 and 3 Tesla magnet strength. Eur Spine J 2013;22(10):2296–302.
[27] IXI Dataset – Brain Development [Internet]. [cited 2024 Jan 7]. Available from: http://brain-development.org/ixi-dataset/.
[28] on behalf of the MAGNIMS study group, Sastre-Garriga J, Pareto D, Battaglini M, Rocca MA, Ciccarelli O, et al. MAGNIMS consensus recommendations on the use of brain and spinal cord atrophy measures in clinical practice. Nat Rev Neurol 2020;16(3):171–82.
[29] Eshaghi A, Prados F, Brownlee WJ, Altmann DR, Tur C, Cardoso MJ, et al. Deep gray matter volume loss drives disability worsening in multiple sclerosis. Ann Neurol 2018;83(2):210–22.

[30] Lansley J, Mataix-Cols D, Grau M, Radua J, Sastre-Garriga J. Localized grey matter atrophy in multiple sclerosis: a meta-analysis of voxel-based morphometry studies and associations with functional disability. Neurosci Biobehav Rev 2013;37(5):819–30.
[31] Grigas O, Damaševičius R, Maskeliūnas R. Positive effect of super-resolved structural magnetic resonance imaging for mild cognitive impairment detection. Brain Sci 2024;14(4):381. https://doi.org/10.3390/brainsci14040381.
[32] Xu J, Tan S, Wen J, Zhang M. Alzheimer's Disease Neuroimaging Initiative, Xu X. Progression of hippocampal subfield atrophy and asymmetry in Alzheimer's disease. Eur J Neurosci 2023. https://doi.org/10.1111/ejn.16543.
[33] Landman BA, Huang AJ, Gifford HC, et al. Multi-parametric neuroimaging reproducibility: A 3-T resource study. Neuroimage 2011;54(4):2854–66. https://doi.org/10.1016/j.neuroimage.2010.11.047.
[34] Van Essen DC, Smith SM, Barch DM, et al. The WU-Minn human connectome project: an overview. Neuroimage 2013;80:62–79. https://doi.org/10.1016/j.neuroimage.2013.05.041.
[35] Cai LY, Yang Q, Kanakaraj P, et al. MASiVar: multisite, multiscanner, and multisubject acquisitions for studying variability in diffusion weighted magnetic resonance imaging. Neuroimage 2021;224:117007. https://doi.org/10.1016/j.neuroimage.2020.117007.