




Assessing the Accuracy and Reproducibility of PARIETAL: A Deep Learning Brain Extraction Algorithm

Sergi Valverde, PhD,^{1*} Lluïa Coll, MSc,¹ Liliana Valencia, MSc,¹  Albert Clèrigues, MSc,¹
 Arnau Oliver, PhD,^{1,2*}  Joan C. Vilanova, PhD,³  Lluís Ramió-Torrentà, PhD,^{2,4,5}
 Àlex Rovira, MD,⁶ and Xavier Lladó, PhD^{1,2}

Background: Manual brain extraction from magnetic resonance (MR) images is time-consuming and prone to intra- and inter-rater variability. Several automated approaches have been developed to alleviate these constraints, including deep learning pipelines. However, these methods tend to reduce their performance in unseen magnetic resonance imaging (MRI) scanner vendors and different imaging protocols.

Purpose: To present and evaluate for clinical use PARIETAL, a pre-trained deep learning brain extraction method. We compare its reproducibility in a scan/rescan analysis and its robustness among scanners of different manufacturers.

Study Type: Retrospective.

Population: Twenty-one subjects (12 women) with age range 22–48 years acquired using three different MRI scanner machines including scan/rescan in each of them.

Field Strength/Sequence: T1-weighted images acquired in a 3-T Siemens with magnetization prepared rapid gradient-echo sequence and two 1.5 T scanners, Philips and GE, with spin-echo and spoiled gradient-recalled (SPGR) sequences, respectively.

Assessment: Analysis of the intracranial cavity volumes obtained for each subject on the three different scanners and the scan/rescan acquisitions.

Statistical Tests: Parametric permutation tests of the differences in volumes to rank and statistically evaluate the performance of PARIETAL compared to state-of-the-art methods.

Results: The mean absolute intracranial volume differences obtained by PARIETAL in the scan/rescan analysis were 1.88 mL, 3.91 mL, and 4.71 mL for Siemens, GE, and Philips scanners, respectively. PARIETAL was the best-ranked method on Siemens and GE scanners, while decreasing to Rank 2 on the Philips images. Intracranial differences for the same subject between scanners were 5.46 mL, 27.16 mL, and 30.44 mL for GE/Philips, Siemens/Philips, and Siemens/GE comparison, respectively. The permutation tests revealed that PARIETAL was always in Rank 1, obtaining the most similar volumetric results between scanners.

Data Conclusion: PARIETAL accurately segments the brain and it generalizes to images acquired at different sites without the need of training or fine-tuning it again. PARIETAL is publicly available.

Level of Evidence: 2

Technical Efficacy Stage: 2

J. MAGN. RESON. IMAGING 2021.

Magnetic resonance imaging (MRI) is one of the most widely used imaging techniques in neuroimaging pipelines.^{1–7} In recent years, several automated methods have been proposed to support the diagnosis, treatment, or evaluation of diverse brain diseases such as Alzheimer's,¹ multiple sclerosis,² or brain tumors.³ However, in most of the studies, brain scans have to be pre-processed beforehand with different tasks such as brain extraction,⁴ noise reduction,⁵

View this article online at wileyonlinelibrary.com. DOI: 10.1002/jmri.27776

Received Nov 10, 2020, Accepted for publication Jun 1, 2021.

*Address reprint requests to: A.O. or S.V., Ed. P-IV, Campus Montilivi, 17003 Girona, Spain. E-mail: arnau.oliver@udg.edu; sergivalv@gmail.com
 Sergi Valverde, Lluïa Coll, and Liliana Valencia contributed equally to this study.

From the ¹Research Institute of Computer Vision and Robotics, University of Girona, Girona, Spain; ²REEM, Red Española de Esclerosis Múltiple; ³Girona Magnetic Resonance Center, Girona, Spain; ⁴Multiple Sclerosis and Neuroimmunology Unit, Neurology Department, Dr. Josep Trueta University Hospital, Institut d'Investigació Biomèdica, Girona, Spain; ⁵Medical Sciences Department, University of Girona, Girona, Spain; and ⁶Magnetic Resonance Unit, Department of Radiology, Vall d'Hebron University Hospital, Barcelona, Spain

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

inhomogeneity correction,⁶ and registration.⁷ Hence, it is important to have a high accuracy and reliability of each of these tasks, since pre-processing errors will be propagated to subsequent methods, affecting the overall obtained accuracy.⁸

From the set of pre-processing tasks performed in computational neuroimaging, brain extraction, also termed skull stripping, consists of removing the skull, dura, and scalp from the brain. Manual annotation of skull stripping is time-consuming, increases processing time per volume, and is prone to intra and inter-rater variability.⁹ Consequently, several automated approaches have been developed to alleviate the manual variability and to reduce time constraints.¹⁰ Automated skull stripping techniques have to overcome several challenges such as image domain shifts caused by the use of different scanner vendors and image protocols,¹¹ and the variation of the brain structures that changes with individuals (i.e., vessels and muscles),¹² motion artifacts,¹³ or even pathological conditions and treatment-related changes (i.e., cavities resection, radiation effects, etc.) that may alter brain structure.^{14,15}

Traditional state-of-the-art methods for automated brain extraction can be classified into several categories¹⁶: intensity-based techniques such as Brain Surface Extractor (BSE)¹⁷; intensity and morphological methods such as Marker Based Watershed Scalper (MBWSS)¹⁸ and Multispectral Adaptive Region Growing Algorithm (MARGA)¹⁹; atlas-based methods such as Brain extraction based on nonlocal Segmentation Technique (BeaST),²⁰ where an atlas template is fitted to the MRI brain image in order to separate the brain from the skull; deformable model-based methods such as Brain Extraction Tool (BET),²¹ which uses a deformable model that evolves to fit the brain's surface by the application of a set of locally adaptive model forces; and hybrid methods such as ROBust, learning-based Brain EXtraction system (ROBEX),²² where several of the previous techniques are combined.

In recent years, however, the introduction of supervised deep learning strategies has provided an increase in the methods' performance compared to classical approaches for many medical imaging applications.^{2,23,24} In particular, convolutional neural networks (CNN)²⁵ and U-Net based architectures²³ are considered the most efficient models for biomedical image segmentation. In the case of skull stripping, several approaches have previously been proposed. These include voxelwise CNN models as those proposed in Kleesiek et al,¹⁴ U-Net architectures such as Isensee et al²⁶ and CONSNNet,¹⁶ or cascaded methods composed of both CNN and U-Net architectures as the one proposed by Salehi et al.²⁷ In all these studies, deep learning methods showed a superior performance not only in accuracy but also in computing (testing) time in comparison with state-of-the-art methods before the deep learning era. However, despite the astonishing performance of deep learning methods, the accuracy of those models tends to decrease significantly when

evaluated on different image protocols than those included in the training dataset,²⁸ which may limit their usability and applicability in clinical practice.

In order to validate the accuracy of these automated methods, several datasets have been used such as the Open Access Series of Imaging Studies (OASIS) dataset,²⁹ composed of non-demented and demented patient images; the LONI Probabilistic Brain Atlas (LPBA40),³⁰ composed of healthy subject images; and the Calgary-Campinas (CC-359) dataset,¹¹ which has been proposed more recently and incorporates a set of multi-vendor and multi-field-strength brain images along with silver masks (i.e., consensus brain extraction masks generated by the Simultaneous Truth And Performance Level Estimation (STAPLE) algorithm³¹ using different automated methods¹⁰).

The main aims of the present study are as follows. First, to propose a novel deep learning brain EXtraction tool (PARIETAL) trained on the publicly available CC-359 dataset¹¹ that is suitable for automated skull-stripping of MRI images. Second, to evaluate PARIETAL's performance against manual expert annotations, comparing it also to other 11 state-of-the-art methods including two recently proposed deep learning architectures. Third, to study PARIETAL's robustness to other MRI sites of those used during training, by analyzing the reproducibility across scan/rescan images and also images from the same subject scanned on three different unseen MRI sites. In order to increase the reproducibility of our results, PARIETAL software and its source code are currently available for downloading at our research website as an open-source toolbox (available at <https://github.com/NIC-VICOROB/PARIETAL>).

Materials and Methods

Datasets

Two datasets were used to develop the proposed method and its corresponding analysis, the Calgary-Campinas dataset and an in-house clinical dataset. Each of them consists of a collection of three-dimensional (3D) MR volumes or images, where each image corresponds to a different subject. The dataset characteristics are described as follows.

CALGARY-CAMPINAS DATASET. The Calgary-Campinas CC-359 dataset¹¹ contains 359 public available MR images of healthy adults (29–80 years old, 183 female, and 176 male subjects). Images were acquired on scanners from three different vendors (GE, Philips, and Siemens) and at two magnetic field strengths (1.5 T and 3 T). Detailed information about the brand, model, and magnetic field can be found in Table 1. Data were obtained using T1-weighted 3D imaging sequences: a 3D Magnetization Prepared RAPid Gradient Echo sequence (MP-RAGE; Philips, Siemens) and a comparable T1-weighted spoiled gradient-echo sequence (General Electrics). In all machines, image volumes had a spatial resolution of 1.0 mm × 1.0 mm × 1.0 mm.³²

TABLE 1. Brand, Model and Magnetic Field Strength for Each of the Scanners Used in the Calgary-Campinas CC-359 Dataset

Calgary-Campinas CC-359 Dataset		
MR Brand	MR Model	Magnetic Field Strength (T)
GE	Signa HDxT	1.5
GE	Discovery MR750	3
Philips	Achieva ^a	3
Siemens	Avanto	1.5
Siemens	Skyra	3

^aTwo sites contributed with images from a Philips Achieva scanner.

TABLE 2. Brand, Model and Magnetic Field Strength for Each of the Scanners Used in the Clinical In-House Dataset from Three Hospitals of the Catalan Public Health System

In-House Dataset		
MR Brand	MR Model	Magnetic Field Strength (T)
Siemens	Trio A Tim System	3
Philips	Intera	1.5
GE	Signa HDxT	1.5

From the set of released images, the dataset contains 12 ground-truth cases (two for each site) that have been manually segmented by three trained radiologists. For the rest of 347 scans, the dataset includes a *silver-mask* annotated brain mask computed using a supervised classification consensus³¹ from skull stripping methods ANTS,³³ BeAST,²⁰ BET,²¹ BSE,¹⁷ HWA,⁴ MBWSS,¹⁸ OPTIBET,¹⁴ and ROBEX.²²

IN-HOUSE CLINICAL DATASET. The protocol used to build the in-house dataset was approved by the hospital research and ethics committee, and informed consent was obtained from each participant before enrolment in the study. The dataset contains images of 21 subjects (12 women and 9 men, age range 22–48 years), all of them scanned in three different hospital sites of the Catalan health system: Hospital Vall d’Hebron (Barcelona), Hospital Santa Caterina—IAS (Girona) and Hospital Dr. Josep Trueta (Girona). Detailed information about the brand, model, and magnetic field can be found in Table 2. Furthermore, every subject was scanned twice in each scanner resulting in a dataset with 126 images.

Participants were asked to lie in a supine position, the center of the head-coil was aligned with the line of the eyes, they were instructed not to move, and the head was gently stabilized with cushions. Sequences were repeated after patient repositioning for the scan/rescan acquisition. The subjects moved among the three centers in a period of 1 month. For each subject, T1-weighted images were acquired at each hospital scanner along with scan-rescan acquisitions. The image protocols were the ones used in the clinical practice in each hospital. Scanner details for each hospital were:

- Hospital Vall d’Hebron: subjects were scanned on a 3-T Siemens Trio A Tim System with a 12-channel phased-array head coil, with acquired sagittal 3D T1 magnetization prepared rapid gradient-echo (MPRAGE) (repetition time [TR] = 2300 msec, echo time [TE] = 2 msec, inversion time [TI] = 900 msec; voxel size = 1 mm × 1 mm × 1.2 mm).
- Hospital Santa Caterina—IAS: subjects were scanned on 1.5 T Philips Intera (R12) head coil, with a 2D T1-weighted spin-echo sequence (TR = 653 msec, TE = 14 msec; voxel size = 1.0 mm × 1.0 mm × 1.0 mm).
- Hospital Dr. Josep Trueta: subjects were scanned on 1.5 T General Electrics (GE) HDxt, with acquired 3D fast T1-weighted spoiled gradient-recalled (SPGR) echo sequence (TR = 30 msec, TE = 9 msec; voxel size = 1.0 mm × 1.094 mm × 1.094 mm).

Proposed Method

CONVOLUTIONAL NEURAL NETWORK ARCHITECTURE.

The proposed method is based on a modified U-Net convolutional network architecture.²³ First, two-dimensional layers were replaced by 3D layers as shown in Çiçek et al.,²⁴ given the three-dimensional nature of MRI images that permits to extract more context information from voxels. Second, convolutional layers were replaced by residual layers as described in He et al.,³⁴ in order to improve the gradient back-propagation, improve optimization convergence speed, and allow for deeper network training.³⁵ Third, concatenation layers were replaced by summation layers, as these have been previously shown to reduce the model complexity without a significant decrease in the performance.³⁵ Finally, instead of using the entire 3D images as input, images were split into 3D patches of 32 × 32 × 32 voxels, in order to overcome the possible limitations in memory when trying to fit the entire MRI scan and to mitigate the class imbalance with respect to brain voxels.³⁵

The final architecture was composed of three encoding and decoding core layers, as shown in Fig. 1, with approximately 10.2 million parameters. Each core layer was based on residual convolutional 3D layers that produce 3 × 3 × 3 and 1 × 1 × 1 kernel convolution layers, normalized using batch normalization³² and afterwards added and passed through a rectified linear unit (RELU) activation function.³⁶ The encoder elements of the architecture were composed of k kernels of 32, 64, and 128, respectively, each one followed by a 2 × 2 × 2 downscale max pooling operation. After the encoder layers, a single residual core module with k = 256 was applied. For the decoder part, three successive 3D deconvolutions

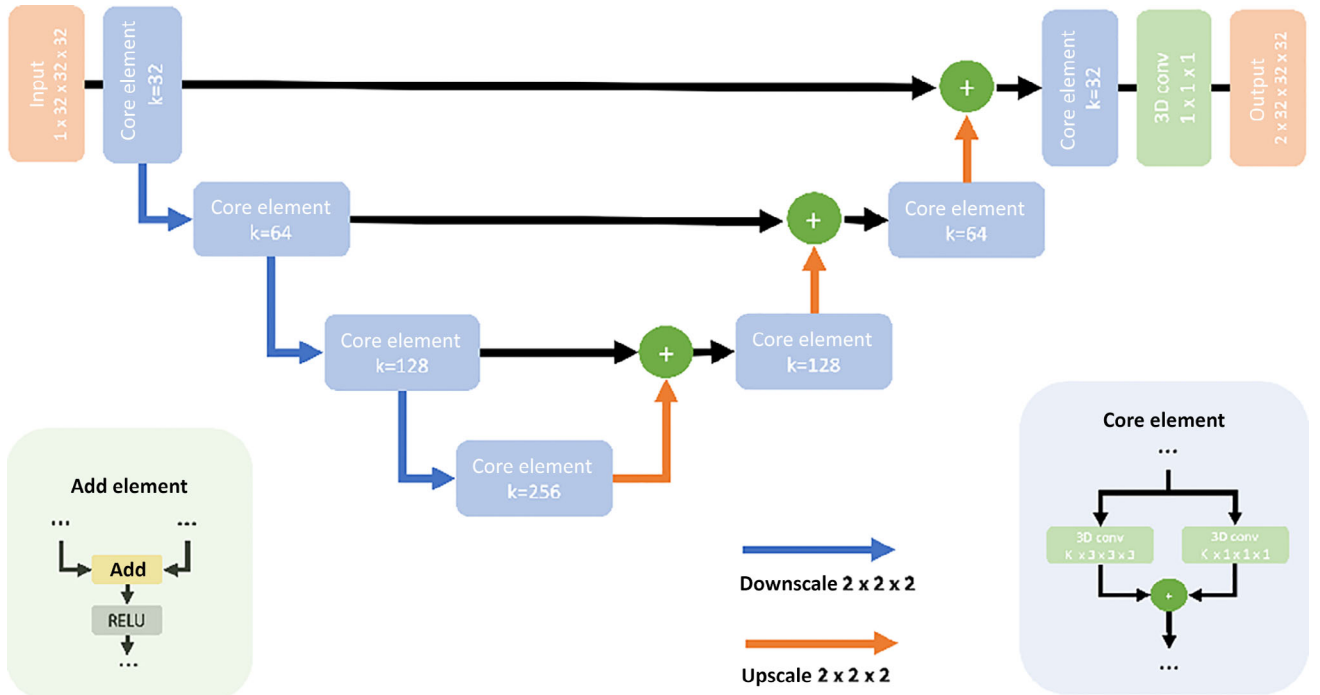


FIGURE 1: Proposed patch-wise 3D encoder/decoder architecture for brain skull stripping. Core elements modules are composed by residual layers as described in He et al.³⁴ Merged layers are implemented using summation instead of concatenation, using add elements modules.

(upscale $2 \times 2 \times 2$) were added to the output of the correspondent core layer (skip connections), followed by a core element with k kernels of 128 and 64 and 32, respectively. Finally, the output was passed through a 3D convolution layer followed by a softmax layer with two outputs that provided the probability of each voxel to pertain to the brain or non-brain class.

TRAINING PROCEDURE. To mitigate class imbalance with respect to brain voxels, we extracted the same number of patches for each class as proposed in previous studies.³⁵ We first created a head mask using a histogram filtering technique, where we eliminate all the voxels that were inside the first bin of the histogram (corresponding to background air), remaining only the head voxels depicting either brain or non-brain parts (bone, skin, fat, muscle, neck, and eyeballs). The ground truth mask was then used to sample all brain voxels (positive class) from the head mask with 3D patches of size $32 \times 32 \times 32$ voxels with a $16 \times 16 \times 16$ voxels overlap and centered around the brain voxel evaluated. All resulting head voxels not considered as positive by the ground-truth mask (negative class), were also sampled using 3D patches of size $32 \times 32 \times 32$ voxels centered around the brain voxel evaluated. The final training set was composed of all the sampled positive patches and the same number of negative patches randomly selected from the whole set of negative set of patches.

For all the experiments, we trained the model only once, using the 347 T1-weighted images and their correspondent silver-masks provided by the Calgary-Campinas-359.¹¹ This resulted in a total number of 150,000 training patches after sampling the images. We trained the model for 200 epochs, with a fixed batch size of 32, categorical cross-entropy as loss cost, and the adaptive learning rate method (ADADELTA)³⁷ as the optimizer. Furthermore, we

followed an early stopping strategy to prevent over-fitting by stopping training after an $N = 50$ epochs without a decrease in the validation error.

INFERENCE. Similar to training, new T1-weighted images used for inference were also split into 3D patches of size $32 \times 32 \times 32$ voxels with a $16 \times 16 \times 16$ overlap. Each of these overlapping patches was then passed through the trained model and the resulting output probabilities were averaged across patches to reconstruct the final probability map for the brain. Finally, binary brain masks were obtained by thresholding the probability map with $P > 0.5$.

IMPLEMENTATION. The proposed method was implemented in Python,¹ using the Pytorch library. We ran all the experiments on a GNU/Linux machine box running Ubuntu 18.04, with 32 GB RAM. For training the model, we used a single TITAN-X GPU (NVIDIA Corp, USA) with 12 GB VRAM memory.

Analysis

For performing the analysis, we pre-trained PARIETAL on the 347 public available silver masks of the CC-359 dataset.

We compared the performance of PARIETAL in three different experimental scenarios: 1) evaluating the accuracy of inferred brain cavities against manual brain annotations; 2) analyzing the robustness of the inferred intracranial cavity by minimizing differences between scan/rescan images of the same subject; and 3) evaluating the reproducibility of the intracranial brain volume measurements for the same subject scanned on different MRI scanner vendors.

¹<https://www.python.org/>

QUANTITATIVE EVALUATION OF THE BASELINE MODEL

The first quantitative analysis to evaluate the baseline performance of the model was done using the 12 images of the dataset that contained ground-truth annotations. We compared the intracranial volume mask of our proposed method against manual annotations in the CC359 dataset in terms of the sensitivity, specificity, and the Dice coefficient. The Dice coefficient was calculated by:

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

where X and Y represent the areas segmented manually and automatically. Furthermore, we compared the obtained Dice coefficient of PARIETAL with respect to several publicly available state-of-the-art tools: ANTS,³³ BEAST,²⁰ BET,²¹ BSE,¹⁷ HWA,⁴ MBWSS,¹⁸ OPTIBET,¹⁴ ROBEX,²² and two other recent deep learning methods such as CONSNNet¹⁶ and HD-BET.²⁶

REPRODUCIBILITY ANALYSIS. We computed the absolute differences of the intracranial cavity volume between scan/rescan images. Furthermore, we compared the performance of our method with four other state-of-the-art methods, two well-known and widely used conventional approaches: BET²¹ and ROBEX,²² and two recent deep learning models such as CONSNNet¹⁶ and HD-BET.²⁶ BET was run with the default configuration proposed in Ref. 22 with the additional flag “-B” to reduce the residual neck voxels and image bias. ROBEX was run under version 1.2 which did not contain any tunable parameters. CONSNNet was run using the pre-

trained model with a patch size of 128 voxels described in Lucena et al.¹⁶ HD-BET was run in accurate mode with test time data augmentation and enabled post-processing as proposed in Isensee et al.²⁶

VARIABILITY BETWEEN SCANNERS AND MRI DOMAINS.

We evaluated the robustness of the proposed method when analyzing the same subject acquired in three different MRI domains in less than 1-month period. The pre-trained version of PARIETAL on the CC-359 dataset was used here to infer the intracranial cavity volume for each of the 21 subjects acquired at the three MRI scanners of the in-house clinical dataset. Since scan/rescan images were available, we processed also each patient twice, therefore resulting in a testing set of 42 subjects.

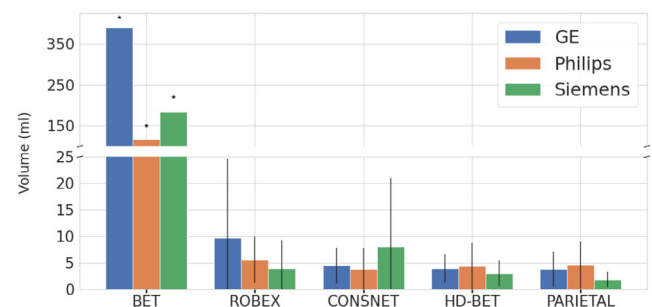


FIGURE 2: Differences in intracranial cavity volume (mL) for the scan/rescan images of the in-house clinical dataset. Results are shown for each of the skull stripping methods and magnetic resonance imaging scanners evaluated.

TABLE 3. Quantitative Evaluation of Skull Stripping Methods Using the 12 Testing Subjects With Manual Ground Truth Masks from the CC-359 Dataset.

Method	Dice (%)	Sensitivity	Specificity
ANTs	95.93 (0.009)	94.51 (0.016)	99.70 (0.001)
BEAST	95.77 (0.012)	93.84 (0.026)	99.76 (0.001)
BET	95.22 (0.009)	98.26 (0.016)	99.13 (0.002)
BSE	90.48 (0.070)	91.44 (0.050)	98.64 (0.020)
HWA	91.66 (0.011)	99.93 (0.001)	97.83 (0.008)
MBWSS	95.57 (0.015)	92.78 (0.027)	99.85 (0.004)
OPTIBET	95.43 (0.007)	96.13 (0.010)	99.37 (0.003)
ROBEX	95.61 (0.007)	98.42 (0.007)	99.13 (0.003)
STAPLE (previous)	96.80 (0.007)	98.98 (0.006)	99.38 (0.002)
Silver-masks	97.13	96.82	99.70
CONSNNet	97.18 (0.005)	98.91 (0.005)	99.46 (0.002)
HD-BET	96.66 (0.005)	99.21 (0.003)	94.21 (0.002)
PARIETAL	97.20 (0.004)	96.80 (0.009)	97.80 (0.008)

Obtained values for all shown methods but PARIETAL and HD-BET were extracted from the Lucena et al¹⁶ study. Highest values in each measure are marked in bold.

For each of the subjects, we computed its own differences in the intracranial cavity volume between the Siemens/Phillips, Siemens/GE, and Phillips/GE acquired images. As in the previous section, we compared the performance of our method with the same state-of-the-art methods BET,²¹ ROBEX,²² CONSNNet,¹⁶ and HD-BET.²⁶

Statistical Analysis

Unpaired *t*-test using mean and SD were used to analyze statistically the performance of the methods. We considered *P*-value less than 0.005 indicates a high statistical significance of a certain result when compared to another. In addition, pairwise permutation tests were used to statistically rank the methods.^{7,38} For all the tests, we set the number of comparisons between each pair of methods to $S = 1000$.

Execution Analysis

We compared the execution times of each of the methods evaluated in the article. All deep learning methods were run on GPU and CPU, while the classical methods were run in CPU only.

Results

Quantitative Evaluation Using CC-359

Table 3 summarizes the quantitative evaluation in terms of Dice, sensitivity, and specificity when using the 12 testing subjects with manual annotations of the CC-359 dataset. PARIETAL is compared with different classical state-of-the-art methods and deep learning methods CONSNNet and HD-BET. Obtained values for all reported methods except for PARIETAL and HD-BET were extracted from the Lucena et al¹⁶ study.

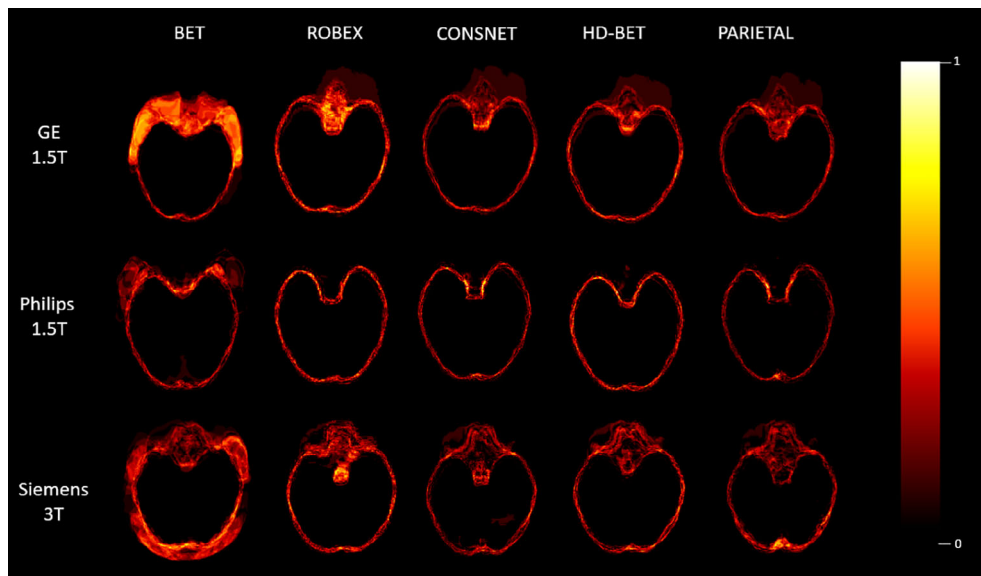


FIGURE 3: Graphical description of the average scan/rescan absolute error maps between all the subjects that composed the clinical dataset. Results are shown for each of the skull stripping methods and magnetic resonance imaging scanners evaluated.

TABLE 4. Permutation Test Results for Each of the Evaluated Methods in the Reproducibility Analysis Performed on the 21 Subjects of the In-House Clinical Dataset

	GE	μ (SD)	Philips	μ (SD)	Siemens	μ (SD)
Rank 1	PARIETAL	0.40 (0.55)	CONSNNet	0.60 (0.55)	PARIETAL	0.39 (0.53)
					ROBEX	0.20 (0.45)
					HD-BET	0.20 (0.45)
					CONSNNet	0.01 (0.68)
Rank 2	ROBEX	-0.40 (0.89)	PARIETAL	0.00 (0.71)		
			ROBEX	-0.20 (0.83)		
Rank 3	BET	-0.80 (0.45)	BET	-0.80 (0.45)	BET	-0.80 (0.45)

Final ranks are based on the intracranial volume differences of the scan/rescan images for each MRI scanner.

The performance of the deep learning methods was superior to other state-of-the-art methods and very similar to the consensus segmentation used as silver masks for training PARIETAL. In terms of the Dice coefficient, PARIETAL obtained a 97.20% agreement against manual annotations, while the performance of the deep learning methods such as CONSNNet and HD-BET methods was 97.18% and 96.66%, respectively. In comparison, the performance of the consensus mask was 97.18%, while the rest of the methods were below 95.93%. The small SD obtained makes the performance of PARIETAL significantly better than the other methods.

Reproducibility Analysis

Figure 2 shows the absolute differences in the intracranial cavity volume for PARIETAL, BET, ROBEX, CONSNNet, and HD-BET on the 21 scan/rescan T1-weighted subject images acquired at each of the MRI scanners of the in-house clinical dataset. PARIETAL provided the highest reproducibility for the images of the Siemens and GE scanners, showing a mean absolute intracranial volume difference of

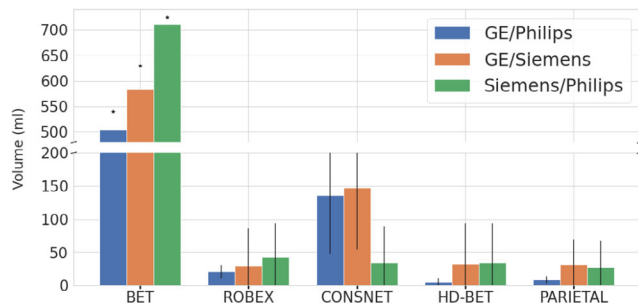


FIGURE 4: Differences in intracranial cavity volume (mL) for each of the 42 subjects of the in-house clinical dataset. Results are shown for each of the skull stripping methods and magnetic resonance imaging scanner pairs available: GE/Philips, GE/Siemens, and Siemens/Philips. Scan and re-scan images for each patient and scanner pair are considered as individual items (42 images).

1.88 ± 1.48 mL and 3.91 ± 3.33 mL, respectively. Differences with other methods in the Siemens images were significant while in the GE ones differences were significant for non-deep learning methods but not for CONSNNet and HD-BET). In contrast, the reproducibility of CONSNNet was better on the Philips scanner with a mean error of 3.92 ± 3.93 mL in contrast to the 4.71 ± 4.39 mL obtained by PARIETAL (differences not significant, $P = 0.54$). In the case of non-deep learning approaches, ROBEX provided the best results, almost similar to the ones obtained with deep learning approaches, while BET was not robust, providing a much bigger intracranial volume difference.

Figure 3 shows the absolute mean error difference map between each of the subjects and MR scanners after registering all of the T1-weighted scan/rescan images to a T1-weighted image template and resample the scan/rescan brain cavities returned by BET, ROBEX, and the deep learning models CONSNNet, HD-BET, and PARIETAL. All methods showed a similar trend with respect to regional differences between subjects and virtually all differences were found in the external borders of the brain cavity, where methods were more uncertain in scan/rescan performance. As shown in the figure, all deep learning methods reduced the differences in comparison to classical methods, especially BET.

In terms of the permutation tests (see Table 4), PARIETAL was the best-ranked method on Siemens and GE scanners, while decreasing to Rank 2 on the Phillips images with CONSNNet and HD-BET both ranked first. Deep learning methods were always ranked first, while both ROBEX and BET methods were mostly ranked second and third.

Robustness Across Scanners

Figure 4 shows the variation of intracranial volume measured between pairs of T1-weighted images of the same subject acquired with the different MRI scanners available

TABLE 5. Permutation Test Results for the Evaluated Methods

	GE/Philips	μ (SD)	GE/Siemens	μ (SD)	Siemens/Philips	μ (SD)
Rank 1	HD-BET	0.80 (0.45)	PARIETAL	0.40 (0.55)	PARIETAL	0.80 (0.45)
	PARIETAL	0.40 (0.89)	HD-BET	0.40 (0.55)		
			ROBEX	0.40 (0.55)		
Rank 2	ROBEX	0.00 (1.00)	CONSNNet	-0.40 (0.89)	CONSNNet	0.20 (0.84)
					HD-BET	0.20 (0.84)
Rank 3	CONSNNet	-0.40 (0.89)	BET	-0.80 (0.45)	ROBEX	-0.40 (0.89)
	BET	-0.80 (0.45)			BET	-0.80 (0.45)

Final ranks based on the intracranial volume differences between images of different MRI scanners. The analysis was done comparing pairs of images from different scanners.

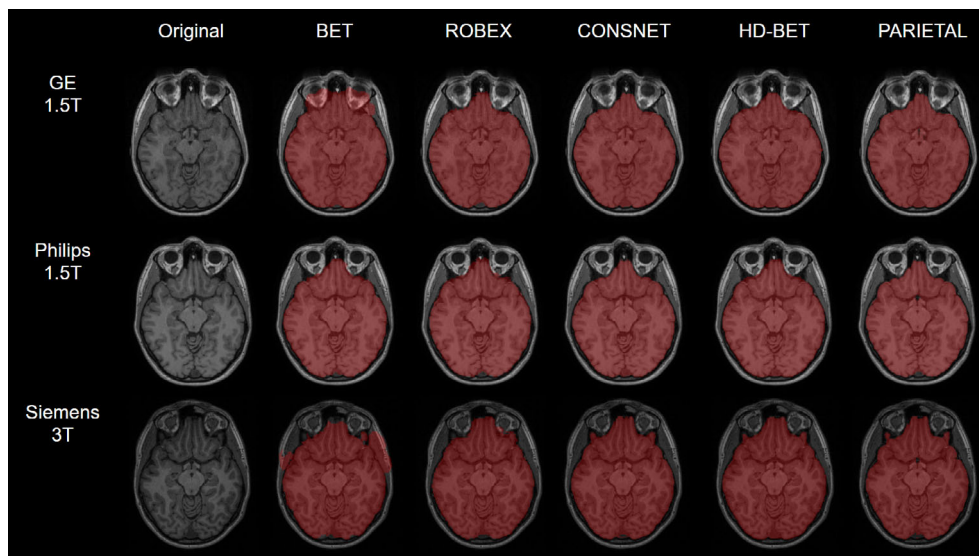


FIGURE 5: Qualitative examples of the in-house dataset from the same patient scanned at three different magnetic resonance imaging machines during a month period. For each scanner, intracranial cavities for BET, ROBEX, CONSNET, HD-BET and PARIETAL are shown.

TABLE 6. Average Running Times in Seconds for Each of the Methods Evaluated in the Article

Method	CC-359	In-House Dataset
ANTs	1378	-
BEAST	1128	-
BET	9	8
BSE	2	-
HWA	846	-
MBWSS	145	-
OPTIBET	773	-
ROBEX	60	59
CONSNET (GPU)	25	35
CONSNET (CPU)	351	301
HD-BET (GPU)	19	18
HD-BET (CPU)	72	77
PARIETAL (GPU)	17	13
PARIETAL (CPU)	129	138

In the CC-359 dataset, obtained values for all methods except for PARIETAL and HD-BET were extracted from Lucena et al.¹⁶ The rest of evaluated methods in the CC-359 and the in-house clinical dataset were computed with the same hardware described in subsection “Implementation.” Methods with GPU support were run using CPU and GPU.

in the in-house dataset. Obtained values are shown for PARIETAL, BET, ROBEX, CONSNET, and HD-BET. PARIETAL reported the lowest differences in volume

between Siemens/Philips (27.16 ± 40.56 mL), while ROBEX obtained the lowest errors between GE/Siemens scanners (30.44 ± 55.52 mL) and HD-BET between GE/Philips scanners (5.46 ± 5.47 mL). In all cases, differences were not significant due to the large SD obtained among cases. However, permutation tests shown in Table 5 revealed that PARIETAL was the best-ranked method on GE/Philips and Siemens/Philips pair of scanners, while HD-BET was the best method on GE/Philips scanner pair. Among the classical skull stripping methods, they were always ranked second and third except for ROBEX that was ranked first for the GE/Siemens scanner pair.

Figure 5 presents a qualitative example of the same subject scanned in the three MRI machines with the intracranial cavities by each of the evaluated methods.

Execution Analysis

Table 6 shows the average running times in seconds for each of the methods in the CC-359 and the clinical in-house dataset. Overall, the fastest method was BET, while ROBEX took around 1 minute in all the experiments performed. From the deep learning methods, PARIETAL reported the fastest running times in both datasets (17 and 13 seconds, respectively) when run using the GPU. In contrast, running times were lower on HD-BET when forced to use CPU in comparison with PARIETAL and CONSNET.

Discussion

The present study highlighted in general the superior performance of the deep learning methods, and in particular, the proposed PARIETAL method, with respect to classical state-

of-the-art techniques in all the experiments performed. Furthermore, our results showed that the performance of PARIETAL was consistent even on different MRI sites without the need of fine-tuning the network architecture. PARIETAL yielded a superior reproducibility of the inferred brain cavities in terms of both subject repositioning and multi-site subjects analysis, outperforming other state-of-the-art methods.

Among the evaluated deep learning methods, there were several remarkable differences in the training data and model. In terms of the data used for training, HD-BET counted with 1568 MRI exams for training, coming from a private dataset and containing a wide variety of MRI scanners and magnetic fields. In contrast, PARIETAL (and CONSNNet) stood on the valuable work of the authors of the Campinas CC-359 dataset,¹¹ which contained images from three different MRI vendors and five scanner models, permitting us to define a heterogeneous training dataset with manual annotations from where our model learned to extract the brain cavity. The results of this study suggest that at least with the available data, the performance of PARIETAL was not affected by the number of training subjects available for training, showing a superior performance to HD-BET in the reproducibility experiments.

Although all deep learning methods were patch-based, the size and dimensions of patches also differed between them. In the case of HD-BET, the authors justified the use of 3D patches of size 128 voxels in order to enable the network to correctly reconstruct the brain mask even when large portions of the brain were missing due to a disease or a traumatic brain injury. In the case of CONSNNet, the authors stated the use of 2.5D patches (sagittal, coronal, and axial planes centered on the voxel of interest) of size 128 voxels, arguing that this configuration solved the problem of the differences of matrix dimensions between each vendor. In contrast, given the limitations in the training data, PARIETAL used remarkably smaller 3D patches of size 32 voxels with the aim to solve the class imbalance problem and to provide enough spatial contexts to the network. In all the experiments performed, the use of smaller patches did not reflect a diminished performance of PARIETAL, and the proposed method showed a higher overlap against manual expert annotations and a superior or similar reproducibility in scan/rescan and multi-site subject experiments.

In terms of method architectures, several differences were also found across methods. While HD-BET and PARIETAL used variants of 3D U-Nets, CONSNNet used a different approach where 2.5D planes were fed to three parallel 2D CNNs. In all the experiments done, the results show a superior performance of the 3D U-net networks against the 2.5D approach. Existing differences in the performance of HD-BET and PARIETAL may be explained by the design decision strategies such as data imbalance mitigation or how training was optimized by the use of residual and summation layers.

In comparison to state-of-the-art methods such as BET or ROBEX, deep learning methods and especially HD-BET and PARIETAL show a higher capability to learn a better general representation of the brain tissues and scalp that could help to reduce the intrinsic reproducibility errors that most of past the state-of-the-art methods suffered. This can be crucial for some post-processing steps such as the automated quantification of brain atrophy measurements or the automatic lesion quantification that could be biased due to errors of the skull stripping process.

Limitations

Despite the fact that the obtained reproducibility results suggest that deep learning methods could be an interesting contribution to reduce the errors of brain tissue volume measurements, our study was limited only to understand the role of skull-stripping, leaving its effect on tissue volume for future work. Furthermore, our study shows that the performance of deep learning methods is consistent on other MR sites and image protocols than those used for training, although the analysis was limited to the same scanner vendors Siemens, Philips, and GE that were already present in the training datasets. Although in all the cases MRI protocols and scanner machines differed across hospitals between the pre-training phase and the experiments performed in this article, the lack of public data does not permit to evaluate the performance of deep learning methods not only in different MRI protocols but also in other MRI scanners.

Conclusion

This study proposed PARIETAL, an open-source available toolbox for fast and efficient MRI brain extraction. Our extensive analysis shows that PARIETAL provides good accuracy for segmenting the skull. The proposed deep learning architecture generalizes well to unseen MR image protocols, permitting to reduce the differences in intracavity volume between hospitals without fine-tuning at each new MRI site. Furthermore, the obtained reproducibility results suggest that PARIETAL may be considered to improve brain volume measurements or atrophy quantification in longitudinal studies, where intracranial volume variations can have a big influence in this quantification.

Acknowledgments

The authors thank Dr. Roberto Souza for sharing the information about the Calgary-Campinas dataset. This work has been partially supported by DPI2017-86696-R from the Ministerio de Ciencia, Innovación y Universidades. Albert Clèrigues also holds a FPI grant PRE2018-083507. The authors gratefully acknowledge the support of the NVIDIA Corporation with their donation of the TITAN X GPU used in this research.

References

1. Payan A, Montana G. Predicting Alzheimer's disease: A neuroimaging study with 3D convolutional neural networks. *arXiv preprint arXiv:150202506*, 2015.
2. Brosch T, Tang LY, Yoo Y, Li DK, Traboulsee A, Tam R. Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. *IEEE Trans Med Imaging* 2016;35(5):1229-1239.
3. Pereira S, Pinto A, Alves V, Silva CA. Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Trans Med Imaging* 2016;35(5):1240-1251.
4. Ségonne F, Dale AM, Busa E, et al. A hybrid approach to the skull stripping problem in MRI. *Neuroimage* 2004;22(3):1060-1075.
5. Vaishali S, Rao KK, Rao GVS. A review on noise reduction methods for brain MRI images. In: *2015 International Conference on Signal Processing and Communication Engineering Systems*; 2015, p 363-365.
6. Akkus Z, Galimzianova A, Hoogi A, Rubin DL, Erickson BJ. Deep learning for brain MRI segmentation: State of the art and future directions. *J Digit Imaging* 2017;30(4):449-459.
7. Klein A, Andersson J, Ardekani BA, et al. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *Neuroimage* 2009;46(3):786-802. <https://doi.org/10.1016/j.neuroimage.2008.12.037>.
8. Speier W, Iglesias JE, El-Kara L, Tu Z, Arnold C. Robust skull stripping of clinical glioblastoma multiforme data. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer; 2011, p 659-666.
9. Roy S, Butman JA, Pham DL, Alzheimer's Disease Neuroimaging Initiative. Robust skull stripping using multiple MR image contrasts insensitive to pathology. *Neuroimage* 2017;146:132-147.
10. Lucena O, Souza R, Rittner L, Frayne R, Lotufo R. Silver standard masks for data augmentation applied to deep-learning-based skull-stripping. In: *IEEE International Symposium on Biomedical Imaging*. IEEE; 2018, p 1114-1117.
11. Souza R, Lucena O, Garrafa J, et al. An open, multi-vendor, multi-field-strength brain MR dataset and analysis of publicly available skull stripping methods agreement. *Neuroimage* 2018;170:482-494.
12. Kalavathi P, Prasath VS. Methods on skull stripping of MRI head scan images—A review. *J Digit Imaging* 2016;29(3):365-379.
13. Zaitsev M, Maclaren J, Herbst M. Motion artifacts in MRI: A complex problem with many partial solutions. *J Magn Reson Imaging* 2015;42(4):887-901.
14. Lutkenhoff ES, Rosenberg M, Chiang J, et al. Optimized brain extraction for pathological brains (OPTIBET). *PLoS One* 2014;9(12):e115551.
15. Kleesiek J, Urban G, Hubert A, et al. Deep MRI brain extraction: A 3D convolutional neural network for skull stripping. *Neuroimage* 2016;129:460-469.
16. Lucena O, Souza R, Rittner L, Frayne R, Lotufo R. Convolutional neural networks for skull-stripping in brain MR imaging using silver standard masks. *Artif Intell Med* 2019;98:48-58.
17. Shattuck DW, Sandor-Leahy SR, Schaper KA, Rottenberg DA, Leahy RM. Magnetic resonance image tissue classification using a partial volume model. *Neuroimage* 2001;13(5):856-876.
18. Beare R, Chen J, Adamson CL, et al. Brain extraction using the watershed transform from markers. *Front Neuroinform* 2013;7:32.
19. Roura E, Oliver A, Cabezas M, et al. Multispectral adaptive region growing algorithm for brain extraction on axial MRI. *Comput Methods Programs Biomed* 2014;113(2):655-673.
20. Eskildsen SF, Coupé P, Fonov V, et al. BEaST: brain extraction based on nonlocal segmentation technique. *Neuroimage* 2012;59(3):2362-2373.
21. Jenkinson M. BET2: MR-based estimation of brain, skull and scalp surfaces. In: *Eleventh Annual Meeting of the Organization for Human Brain Mapping*; 2005. Available from: <https://ci.nii.ac.jp/naid/10030066593/en/>
22. Iglesias JE, Liu CY, Thompson PM, Tu Z. Robust brain extraction across datasets and comparison with publicly available methods. *IEEE Trans Med Imaging* 2011;30(9):1617-1634.
23. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer; 2015, p 234-241.
24. Çiçek O, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-net: Learning dense volumetric segmentation from sparse annotation. *CoRR*; 2016. Available from: <http://arxiv.org/abs/1606.06650>
25. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*; 2012, p 1097-1105.
26. Isensee F, Schell M, Pflueger I, et al. Automated brain extraction of multisequence MRI using artificial neural networks. *Human Brain Mapp* 2019;40(17):4952-4964. <https://doi.org/10.1002/hbm.24750>.
27. Salehi SSM, Erdogmus D, Gholipour A. Auto-context convolutional neural network (Auto-Net) for brain extraction in magnetic resonance imaging. *IEEE Trans Med Imaging* 2017;36(11):2319-2330.
28. Valverde, S., Salem, M., Cabezas, M., Pareto, D., Vilanova, J.C., Ramió-Torrentà, L., Rovira, A., Salvi, J., Oliver, A., Lladó, X.. One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks. *Neuroimage Clin* 2019;21:art 101638.
29. Marcus DS, Wang TH, Parker J, Csernansky JG, Morris JC, Buckner RL. Open access series of imaging studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *J Cogn Neurosci* 2007;19(9):1498-1507.
30. Shattuck DW, Mirza M, Adisetiyo V, et al. Construction of a 3D probabilistic atlas of human cortical structures. *Neuroimage* 2008;39(3):1064-1080.
31. Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (staple): An algorithm for the validation of image segmentation. *IEEE Trans Med Imaging* 2004;23(7):903-921.
32. Ioffe, S., Szegedy, C.. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *J Mach Learn Res* 2015;37:448-456. Available from: <http://arxiv.org/abs/1502.03167>
33. Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *Neuroimage* 2011;54:2033-2044. <https://doi.org/10.1016/j.neuroimage.2010.09.025>.
34. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *arXiv preprint arXiv:151203385*, 2015.
35. Guerrero, R., Qin, C., Oktay, O., Bowles, C., Chen, L., Joules, R., Wolz, R., Valdes-Hernandez, M.C., Dickie, D.A., Wardlaw, J., Rueckert, D.. White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks. *Neuroimage Clin* 2017;17:918-936. Available from: <http://arxiv.org/abs/1706.00935>
36. Maas AL, Hannun AY, Ng AY. Rectifier nonlinearities improve neural network acoustic models. In: *Proceedings of the 30th International Conference on Machine Learning*; 2013, p 6.
37. Zeiler MD. ADADELTA: An adaptive learning rate method. *arXiv preprint 12125701*, 2012, p 6. Available from: <http://arxiv.org/abs/1212.5701>
38. Diez Y, Oliver A, Cabezas M, et al. Intensity based methods for brain MRI longitudinal registration. A study on multiple sclerosis patients. *Neuroinformatics* 2014;12(3):365-379.