frontiers in Neuroinformatics

ORIGINAL RESEARCH
published: 18 June 2021
doi: 10.3389/fninf.2021.641600

3D Segmentation of Perivascular Spaces on T1-Weighted 3 Tesla MR Images With a Convolutional Autoencoder and a U-Shaped Neural Network

*Philippe Boutinaud[1,2], Ami Tsuchida[3,4,5], Alexandre Laurent[3,4,5], Filipa Adonias[3,4,5], Zahra Hanifehlou[1,6], Victor Nozais[1,3,4,5], Violaine Verrecchia[1,3,4,5], Leonie Lampe[7], Junyi Zhang[8], Yi-Cheng Zhu[8], Christophe Tzourio[9,10], Bernard Mazoyer[1,3,4,5,10] and Marc Joliot[1,3,4,5]**

[1] Genesislab, Bordeaux, France, [2] Fealinx, Lyon, France, [3] UMR 5293, GIN, IMN, Univ. Bordeaux, Bordeaux, France, [4] UMR 5293, GIN, IMN, CNRS, Bordeaux, France, [5] UMR 5293, GIN, IMN, CEA, Bordeaux, France, [6] Faculty of Computer Engineering, University of Isfahan, Isfahan, Iran, [7] Integrative Model-based Cognitive Neuroscience Research Unit Universiteit van Amsterdam, Amsterdam, Netherlands & Max Planck Institute for Human Cognitive and Brain Sciences Leipzig, Germany, [8] Department of Neurology, Peking Union Medical College Hospital, Beijing, China, [9] U1219, INSERM, Bordeaux Population Health, University Bordeaux, Bordeaux, France, [10] Centre Hospitalier Universitaire de Bordeaux, Bordeaux, France

OPEN ACCESS

Edited by:
Sean L. Hill,
Krembil Centre for Neuroinformatics, Centre for Addiction and Mental Health, Canada

Reviewed by:
Xiaoying Tang,
Southern University of Science and Technology, China
Florian Dubost,
Erasmus Medical Center, Netherlands

***Correspondence:**
Marc Joliot
marc.joliot@u-bordeaux.fr

Received: 14 December 2020
Accepted: 24 May 2021
Published: 18 June 2021

Citation:
Boutinaud P, Tsuchida A, Laurent A, Adonias F, Hanifehlou Z, Nozais V, Verrecchia V, Lampe L, Zhang J, Zhu Y-C, Tzourio C, Mazoyer B and Joliot M (2021) 3D Segmentation of Perivascular Spaces on T1-Weighted 3 Tesla MR Images With a Convolutional Autoencoder and a U-Shaped Neural Network. Front. Neuroinform. 15:641600. doi: 10.3389/fninf.2021.641600

We implemented a deep learning (DL) algorithm for the 3-dimensional segmentation of perivascular spaces (PVSs) in deep white matter (DWM) and basal ganglia (BG). This algorithm is based on an autoencoder and a U-shaped network (U-net), and was trained and tested using T1-weighted magnetic resonance imaging (MRI) data from a large database of 1,832 healthy young adults. An important feature of this approach is the ability to learn from relatively sparse data, which gives the present algorithm a major advantage over other DL algorithms. Here, we trained the algorithm with 40 T1-weighted MRI datasets in which all "visible" PVSs were manually annotated by an experienced operator. After learning, performance was assessed using another set of 10 MRI scans from the same database in which PVSs were also traced by the same operator and were checked by consensus with another experienced operator. The Sorensen-Dice coefficients for PVS voxel detection in DWM (resp. BG) were 0.51 (resp. 0.66), and 0.64 (resp. 0.71) for PVS cluster detection (volume threshold of 0.5 within a range of 0 to 1). Dice values above 0.90 could be reached for detecting PVSs larger than 10 mm^3 and 0.95 for PVSs larger than 15 mm^3. We then applied the trained algorithm to the rest of the database (1,782 individuals). The individual PVS load provided by the algorithm showed a high agreement with a semi-quantitative visual rating done by an independent expert rater, both for DWM and for BG. Finally, we applied the trained algorithm to an age-matched sample from another MRI database acquired using a different scanner. We obtained a very similar distribution of PVS load, demonstrating the interoperability of this algorithm.

Keywords: perivascular space, deep learning, U-net, MRI, brain cohort, segmentation

INTRODUCTION

With increasing longevity, cognitive impairment, stroke, and dementia are currently major causes of disability and dependence in elderly individuals (Seshadri and Wolf, 2007), representing a huge societal burden. Cognitive impairment and dementia are mainly the result of a mix of vascular brain injury and neurodegeneration. Most often, vascular brain injury does not result in stroke, but manifest as covert small vessel disease (cSVD), a pathology highly prevalent in elderly individuals (Schmidt et al., 1999). Small vascular brain injuries of various types (white matter lesions, lacunes, microbleeds, and enlarged perivascular spaces) can be detected with brain magnetic resonance imaging (MRI; Wardlaw et al., 2013) in older community individuals. These lesions are associated with a risk of cognitive decline and dementia (review in Debette et al., 2019) and are now considered imaging markers of cSVD. Thus, there is considerable interest in assessing the burden of these lesions from an epidemiological point of view and for the early identification of individuals at risk of developing severe cognitive deterioration and/or dementia.

Among these lesions, perivascular spaces (PVSs; Lawrence and Kubie, 1927), also referred to as Virchow-Robin spaces, are of particular interest. PVSs are extracellular spaces containing interstitial fluid or cerebrospinal fluid surrounding cerebral small penetrating arteries and veins (Patek, 1941). PVSs are formed during development, accompanying brain angiogenesis (Marin-Padilla and Knopman, 2011). As such, they are physiological spaces that, when large enough, can be visible on brain MRI scans of healthy individuals (Maclullich et al., 2004; Zhu et al., 2011; Yakushiji et al., 2014). Animal models have shown evidence that both degenerative and vascular mechanisms can lead to enlarged PVSs (ePVSs), and several studies have shown that ePVS is associated with age and the risk of both cognitive deterioration (Passiak et al., 2019) and dementia (Zhu et al., 2010b; Francis et al., 2019); ePVS is also associated with the presence of other cSVD imaging markers in elderly individuals (Doubal et al., 2010; Zhu et al., 2010a; Potter et al., 2015; Ramirez et al., 2016). Overall, ePVS is now considered a hallmark and a very early anomaly of cSVD, so its assessment has recently become a major area of interest.

ePVS burden is commonly assessed on T1- or T2-weighted MR images using visual ratings with semi-quantitative rating scales (Zhu et al., 2011; Adams et al., 2013; Potter et al., 2015). However, such visual reading lacks precision and reproducibility, which limits its usability for longitudinal studies, and leads to overall loss of analytic power. Therefore, there is a strong need for quantitative volumetric segmentation methods that could ideally identify every PVS in each individual as a 3-dimensional (3D) object in a perfectly reproducible manner. In fact, in the past few years, there have been several attempts at developing such automated PVS segmentation methods using broadly two different approaches, one based primarily on image processing (Wang et al., 2016; Gonzalez-Castro et al., 2017; Zhang et al., 2017; Ballerini et al., 2018; Boespflug et al., 2018; Schwartz et al., 2019; Sepehrband et al., 2019) and the other mainly based on deep learning (DL) (Lian et al., 2018; Dubost et al., 2019; Jung et al., 2019; Sudre et al., 2019; Dubost et al., 2020). The former approach is based on signal enhancement/noise reduction and/or specifically tailored morphological filters derived from the precise analysis of a few PVSs. The latter approach is based on a large set of convolutional filters (at different spatial resolutions) that extract features relevant for segmenting target objects, such as PVSs in the image. Such algorithms require a priori knowledge of the PVS number or locations on a subset of the input data. Both approaches have drawbacks: image processing methods are hampered by the large variance of PVS shapes and the signal-to-noise ratio (SNR), making it difficult to design a filter that will be optimal for the detection of all kinds of PVSs. Conversely, DL methods are sensitive to the quality and amount of a priori knowledge available: in particular, having a sufficiently large and reliable learning set of PVSs may be very difficult and cumbersome as it will require having one or several human operators manually tracing multiple PVSs on thousands of subjects. Both types of methods suffer from limited interoperability, as algorithms are usually tuned for the type of images they are designed or trained with.

Thus, despite several interesting attempts, there is still a need for an interoperable, and validated algorithm for the detection and quantification of PVSs in the entire brain volume. In the present study, we investigated the possibility of implementing such an approach using a class of DL methods based on autoencoders (Kingma and Welling, 2013) and U-shaped networks (U-nets, Ronneberger et al., 2015); a key feature of this approach is that the algorithm is able to learn from relatively sparse data, a major advantage over other DL algorithms. Others have used similar approaches for the PVS detection (Lian et al., 2018; Dubost et al., 2020). Dubost et al. (2020) presented a weakly supervised detection method based on U-net architecture that could be optimized with the PVS count, and applied it to a large dataset (2,200 subjects) of T2-weighted scans. Lian et al. (2018) proposed a multi-channel (T2-weighted, enhanced T2-weighted, and probability map) fully convolutional network and applied it to a small set (20 subjects) of 3D patches of scans acquired at 7T. Here, we report a simpler U-net implementation, based on the T1-weighted (T1w) single-channel input, on a large database of 3D T1w whole-brain volumes acquired at 3T from 1,832 young adults in the MRi-Share database for Magnetic Resonance of i-Share that is the Internet-based Student Health Research enterprise ([1]Tsuchida et al., 2020). The learning set was composed of 40 T1w MRI scans from this database in which all "visible" PVSs were manually annotated by an experienced operator. After learning, algorithm performance was assessed using another set of 10 T1w MRI scans from the same database in which PVSs were also traced by the same operator. Next, we applied the algorithm to the rest of the MRi-Share database (1,782 individuals) and compared its output to a visual rating given by a trained rater based on a validated scale. Finally, we applied the algorithm to the T1w MRI images of age-matched subjects from the BIL&GIN database (for Brain Imaging of Lateralization by the Groupe d'Imagerie Fonctionnelle; Mazoyer et al., 2016), acquired

[1]www.i-share.fr

from a different scanner; subsequently, we compared the PVS distribution of both databases.

METHODS

The brain MRI data were taken from the MRi-Share database (Tsuchida et al., 2020), a subcomponent of i-Share (internet-based Student Health Research enterprise, www.i-share.fr), a large prospective cohort study aiming to investigate French university student health. The MRi-Share database was designed to allow the investigation of structural and functional brain phenotypes in a sample of approximately 2,000 young adults in the post-adolescence period. In the present study, we included 1,832 i-Share participants who completed the full MRi-Share brain imaging examination and did not have any incidental findings on their brain MRI. The study sample age was 22.1 $\pm$ 2.3 years (mean $\pm$ SD, range: [18-35], median = 21.7 years), with a high proportion of women (72%, 1,320 women) as was the case with the rest of the i-Share cohort. The study was approved by the local ethics committee (Bordeaux, France).

For the purpose of the present study, we used the T1w scans acquired from each participant on the same Siemens Prisma 3-Tesla MRI scanner using a three-dimensional high-resolution MPRAGE sequence (TR = 2000 ms; TE = 2.03 ms; flip angle = 8°; inversion time = 880 ms; field of view = 256 $\times$ 256 $\times$ 192 mm^3; isotropic voxel size = 1 $\times$ 1 $\times$ 1 mm^3, and in-plane acceleration = 2).

Manual Segmentation and Visual Rating of PVS

Manual Segmentation of PVS in a Subsample of 50 Participants

Supratentorial PVSs are usually classified based on their location, either in the basal ganglia (BG) along lenticulostriate arteries or in the deep white matter (DWM) of the brain. These two types of PVS are usually rated separately and/or quantified since they were demonstrated to be differentially associated with SVD and dementia (Ding et al., 2017), as well as to have different genetic determinants (Duperron et al., 2018). Accordingly, to train and evaluate the performance of our detection algorithm, a subset of 50 individuals exhibiting varying amounts of visible PVS either in the DWM or in the BG were selected from the study sample by a neuroradiologist (BM) who reviewed the raw T1w images of the entire dataset. A trained investigator (AT) performed a voxelwise manual delineation of each PVS on the raw T1w images of each of these 50 individuals. Manual annotation of each PVS was performed using Medical Image Processing, Analysis and Visualization (MIPAV, v7.4.0). Specifically, each axial, coronal, and sagittal slice from the raw T1w scans from each individual was reviewed using the 3D view setting of MIPAV to detect PVSs in the DWM and BG regions. DWM PVSs are typically visible as tubular shapes, often running perpendicular to the cortical surface following the orientation of perforating vessels, whereas those in the BG are visible at the base of the basal ganglia along lenticulostriate arteries. Based on these shape and location characteristics, each visible PVS was segmented as best as possible using the MIPAV pen tool and occasionally expanded using the MIPAV paint grow tool that can automatically "paint" every neighboring voxel that has a lower intensity level than the selected voxel; the distance limitation set for the paint grow tool was 3 mm. The PVS segmented volume was saved as a binary mask in the individual native acquisition space. The PVS annotation procedure was first optimized by having the first ten MRI datasets reviewed by a second expert (LL) who recorded all potential disagreements she had, whether false positive or false negative according to her own opinion, for every dataset. These discordances were then jointly checked one by one by the two experts and were resolved by consensus between them. Subsequently, the remaining 40 MRI datasets were manually annotated for PVS by the first expert only.

Visual Rating of PVS Burden in All Participants

The global PVS burden estimated with our algorithm was also compared to a classical visual semi-quantitative assessment. For this, another investigator (JZ) visually rated the global PVS burden for each of the 1,832 individuals of the sample using a previously validated protocol and rating scale (Zhu et al., 2011). Briefly, for each individual, all axial slices of the T1w images were first examined to identify the slice containing the largest amount of PVS (one for DWM and one for the BG). The selected slice was then used to rate the burden of PVS by the number of spaces observed on a 4-level severity score as follows: for the BG, degree 1 when there were < 5 PVSs, degree 2 when there were ≥ 5 and ≤ 10 PVSs, degree 3 when there were > 10 PVSs but they were countable, and degree 4 when there were innumerable PVS; for cerebral DWM, degree 1 when there were < 10 PVSS in the entire cerebral white matter, degree 2 when there were ≥ 10 PVSs in the total cerebral white matter but < 10 in the slice with the largest number of PVSs, degree 3 when there were ≥ 10 and ≤ 20 PVSs in the slice with the most PVSs, and degree 4 when there were > 20 PVSs in the slice with the most PVSs. The reliability of this visual rating was assessed by having the same investigator blindly rate two subsets of 60 individuals (one for the BG and one for DWM, with 40% of individuals common to both subsets) twice. The kappa concordance coefficients between the two ratings were 0.81 and 0.77 for DWM and the BG, respectively (both were significantly different from 0 at $p < 10^{-4}$), and all discrepancies between the two ratings were minor, i.e., consisting of a difference of one scale level.

Segmentation Model Training, Validation, and Testing

This section details the methodology for the PVS segmentation model architecture, training, and testing. KNIME 4.0 (Berthold et al., 2009) was used for the data management workflows, and Python-based Keras 2.2.4[2], Scikit-learn (Pedregosa et al., 2011) and TensorFlow 2.1 (Abadi et al., 2016) were used for implementing our segmentation model. The algorithms were run on a Centos computer with a Xeon ES2640, 40 cores, 256 Gb RAM and two Tesla P100 GPUs with 16 Gb RAM.

[2]https://keras.io

Methodology of PVS Segmentation

Defining data subsets for training and testing the PVS segmentation algorithm

The full dataset of 1,832 T1w volumes was split into 3 non-overlapping subsets:

- An autoencoder subset (ENCOD), including the 1,782 volumes without manual annotations of PVSs; the T1w volumes of ENCOD subset was used to train the 3D convolutional autoencoder (see "Visual Rating of PVS Burden in All Participants"), and visual rating of this subset (see "Visual Rating of PVS Burden in All Participants" above) was also used to assess the model-predicted PVS load against visual rating (see "Testing the PVS Segmentation Model of a Large Subset: ENCOD Subset Analysis").
- A training subset (TRAIN), including 40 of the volumes with manually annotated PVSs; the TRAIN subset was used to train and validate the segmentation model.
- A testing subset (EVAL), including the remaining 10 volumes with annotated PVS; the EVAL subset was used to test the segmentation model.

The 10 T1w volumes of the EVAL subset were selected by an expert neuroradiologist (BM) to be representative of the full set of 50 volumes of manually annotated PVSs. This was checked by comparing the distribution of manually annotated PVSs in the two subsets.

T1-weighted MRI preprocessing

Prior to training the DL, T1w MRI volumes (N = 1,832) were preprocessed following a 5-step procedure: 1- tissue segmentation with FreeSurfer, 2- creation of an intracranial volume mask (ICV), 3- voxel intensity rescaling, 4- creation of a brain volume bounding box, and 5- creation of a BG mask.

- First, each T1w volume was segmented using FreeSurfer v6.0[3], and the different tissue components were identified.
- Second, an ICV mask was defined as the union of the gray matter (GM), WM, and cerebrospinal fluid (CSF) tissue voxels.
- Third, voxel intensity values were linearly rescaled between 0 and 1 by setting the 99th percentile of each subject's sample as the maximum. The values greater than 1 were set back to 1.
- Fourth, for each individual, we computed the minimal bounding box (oriented along the 3 axes of the T1w acquisition) that included his/her brain volume. In doing so, we eliminated the neck and some of the background air signals. The union of the 1,832 individual bounding boxes (registered using their centers) was then computed and used to crop each T1w volume. Note that in the process, T1w volumes were not interpolated but were only translated by an integer number of voxels since all individual boxes had the same orientation. This cropping process led to a 52% data size reduction (from 256 × 256 × 192 voxels to 160 × 214 × 176 voxels), resulting in a gain of a factor of approximately 2 in computational burden.
- Fifth, since PVS distribution is usually separately examined when localized in the BG or in the DWM, we created a basal ganglia mask (BG-mask) for each individual including the tissue classes identified by FreeSurfer as thalamus, thalamus-proper, caudate, putamen, pallidum and accumbens regions. The labels were used to determine whether a PVS belonged to the BG or the DWM.

Autoencoder and U-net architecture

Our segmentation model used a U-net architecture similar to the one described in Ronneberger et al. (2015). The main constraints in our application were (1) the small size of the annotated datasets available for training and validation (TRAIN set) and (2) the large volume of data: the model parameters and a batch of volumes had to fit into the 16 Gb memory of the GPU.

To train the U-net, an autoencoder was first trained on the large ENCOD set. The convolutional autoencoder and the U-net share a similar architecture to transfer the weights learned by the former to the latter. The difference resides in the addition of "skip connections" in the U-net between the corresponding encoding and decoding blocks (**Figure 1A**); the output of the encoding block is concatenated to the input of the decoding block. We used a U-net architecture denoted by its main hyperparameters: the initial number of kernels for the first stage of convolutions (*nb_kernel_init = 8)*, the number of stages (NStages = 7) and the number of 3D convolutions for a stage (*nConvolutions = 2*). This configuration is referred in the following as the 8.7.2 autoencoder/U-net architecture.

Figure 1B shows the encoding and decoding levels for stage i of the encoding/decoding architecture. On the encoding side, the tensors went through 2 convolution blocks. Each convolution block consisted a 3D convolution with a kernel size of 3 × 3 × 3 that produced 2i kernels, followed by batch normalization (Ioffe and Szegedy, 2015), and activation using a rectified linear unit (ReLU) (Glorot and Bengio, 2010). In the autoencoder, the weights were initialized randomly using a Glorot uniform initializer (Glorot and Bengio, 2010); in the U-net the weights were initialized using the trained autoencoder. ReLUs were used for all activation functions except for the last step of the decoding block, where a sigmoid was used to return to the initial volume (for the autoencoder training) or manually annotated mask (for the U-net training). After the convolution blocks, there was a 2 × 2 × 2 max pooling (Ciresan et al., 2011) and a dropout layer (Srivastava et al., 2014). We tested an architecture with strided convolutions instead of max pooling layers (as in Milletari et al., 2016), but the added parameters produced a model that did not fit into the GPU without reducing the number of kernels used at each stage, and the resulting model did not perform better.

On the decoding side, the first layer was a 2 × 2 × 2 nearest neighbor upsampling. Then, a padding/cropping layer was used to align the spatial dimensions of the tensor to those in the encoding block at the same stage. It allowed the subsequent concatenation of the tensors coming from the encoding block to those from the decoding layers in the U-net configuration, and eliminated the need for initial padding or cropping when

[3]https://surfer.nmr.mgh.harvard.edu/

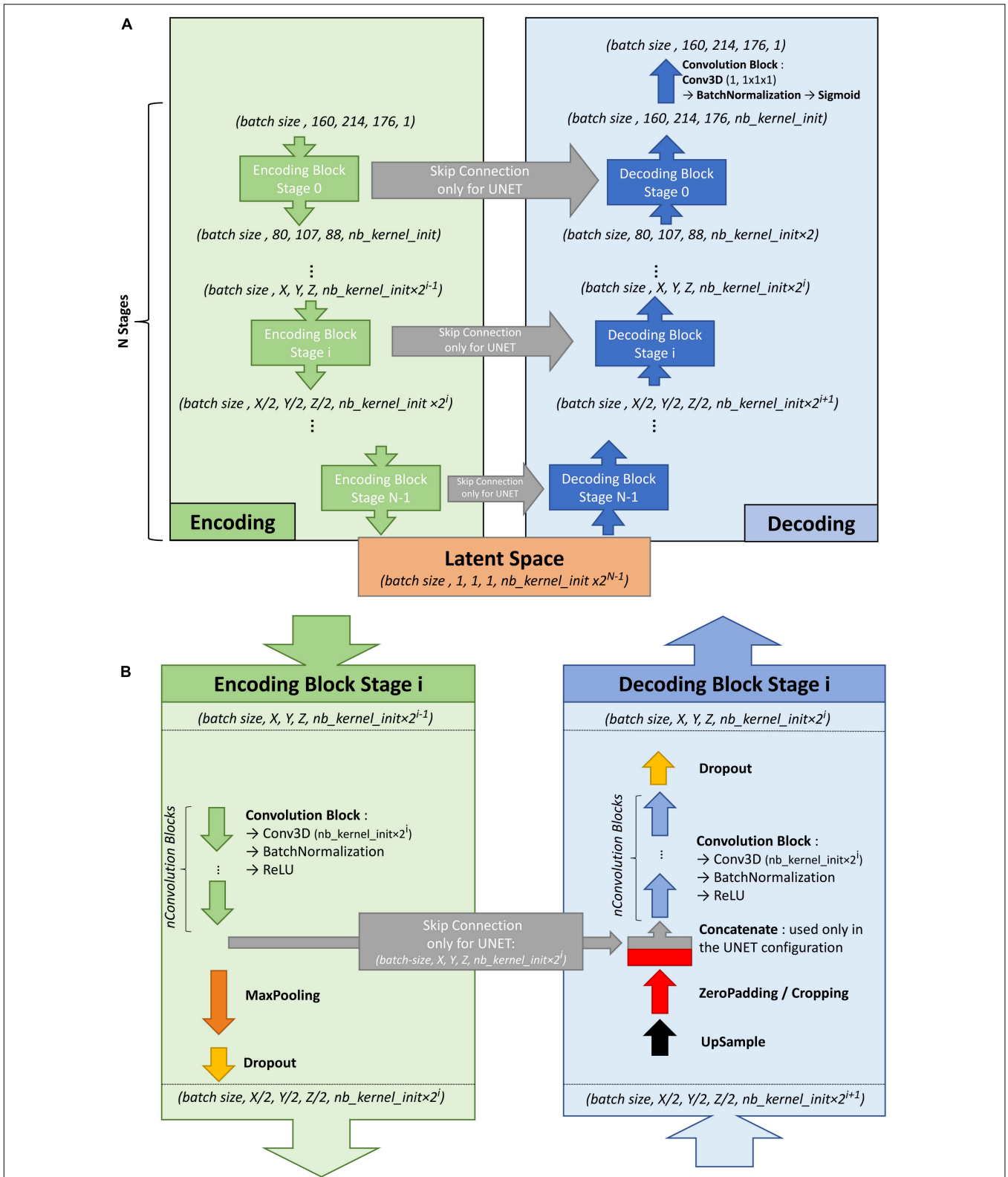


FIGURE 1 | (A) Overall architecture for the convolutional autoencoder (without skip connections) and U-net (with skip connections). On the left, the encoding blocks are shown. Each block used 3D convolutions to create new features and max pooling operations to downsize the image. On the right are the decoding blocks that used upsampling and convolutions to reconstruct the image (for the autoencoder configuration) or to create the segmentation probability map (for U-net configuration with skip connections). **(B)** Details of stage i of the model architecture.

the volume had resolutions that were not a power of 2. It was followed by a symmetric number of convolution blocks as the encoding side (i.e., 2 in our case), and finally, by a dropout layer. Unlike the encoding convolutions, the decoding convolutions were initialized using a Glorot uniform initializer for both the autoencoder and the U-net.

With our configuration of the 8.7.2 autoencoder/U-net architecture, the final model had 44 million trainable parameters. The resulting latent space had 512 dimensions. The Adam algorithm (Kingma and Welling, 2013) was used with the default parameters recommended by the authors: beta1 = 0.9, beta2 = 0.999, and decay = 0.0. The dropout rate used for both encoding and decoding dropout layers was 0.1. In the autoencoder, the mean square error (MSE) was used as the loss function, whereas Keras implementation of the Dice loss was used as the loss function in the U-net. Note that we tested different loss functions such as binary cross entropy, focal loss, tversky loss without seeing any significant performance gain upon using the dice loss. As we did not have enough computing resources to systematically explore the hyperparameter space for each possible configuration, the different hyperparameter values were chosen based on previous experiences. The most sensitive hyperparameter was the initial kernel size (larger was better), and we tried several configurations to balance it with the batch size while allowing the model and the batch of volumes to fit in GPU memory, resulting in a batch size of 2.

PVS Segmentation Model Training and Testing

We first trained the autoencoder by randomly splitting the ENCOD dataset into 80% training and 20% validation subsets. We then initialized the U-net segmentation model with the weights of the first encoding stages and trained it using the TRAIN set with a 5-fold cross-validation scheme. The input for the U-Net model was the T1 image, and the output was the manually annotated mask. Each fold consisted of 80% training (32 volumes) and 20% (8 volumes) validation, and each subject's data appeared only once in the fold validation set and four times in the fold training set. The repartition in the 5-fold validation set was manually defined to include a similar pseudo-uniform distribution of the individual PVS load in each fold. For each fold, the model with the lowest validation loss was selected. The five resulting models were used to predict PVS maps for the EVAL subset, and the five output maps for each input image were averaged to create the final PVS map called thereafter the *consensus* (of the 5 folds) segmentation. Note that each predicted map was coded in values in the 0 to 1 range. Four metrics were computed to quantify segmentation algorithm performance on $N = 8$ validation set for each of the 5 folds of TRAIN and $N = 10$ of the EVAL sets. For this quantification, the prediction maps were binarized with a 0.5 threshold and compared to the manually traced PVS masks using *Dice-loss* scores (see "Visual Rating of PVS Burden in All Participants" above), the *true positive rate (*TPR*)* and the *positive predictive value (*PPV*),* as well as their harmonic mean known as the *(Sorensen-) Dice coefficient*. The *TPR* is defined as the number of predicted PVS voxels that were correctly identified (true positive, TP), i.e., overlapping with manually annotated PVS voxels, divided by the number of manually annotated PVS voxels (i.e., TP plus false negative (FN) voxels). The *PPV* is defined as the number of TPs divided by the sum of the number of TPs plus the number of false positive (FP) voxels, i.e., voxels that were wrongly predicted as PVSs. Note that in the context of machine learning, the *TPR* and *PPV* are also named *sensitivity* and *precision,* respectively. The Sorensen-Dice coefficient is then computed as $1/\text{Dice} = (1/\text{TPR} + 1/\text{PPV})/2$. The 3 metrics will be referred to as voxel-level metrics (VL) in the following sections.

In the testing phase, only the EVAL consensus prediction was used, and VL metrics were computed with 9 amplitude thresholds applied to the prediction map (*PredMap-Thr*, 0.1 to 0.9 in 0.1 step). Additionally, cluster-level (CL) TPR and PPV were computed as following: Clusters were defined using a 26-neighbor connectivity rule (i.e., 1 voxel is connected to its 26 surrounding voxels) both for the manually annotated and for the predicted PVS masks. A cluster of the predicted PVS volume was assigned to the TP class if it included at least one voxel of the manually annotated PVS volume. Similarly, a cluster of the predicted PVS volume was assigned to the FP class if it included no voxels of the manually annotated PVS volume. Conversely, a cluster of the manually annotated PVS volume was assigned to the FN class if it included no voxels of the predicted PVS volume. We will refer them as TPR-CL and PPV-CL, whereas their harmonic mean will be referred as Dice-CL. The final metrics were computed by averaging the values of the 10 EVAL subjects for each *PredMap-Thr* and each metric. Note that the EVAL dataset (10 volumes) was not used in the training/validation step; thus, independent scores were generated. In addition, for each subject the Hausdorff Distance (HD, Huttenlocher et al., 1993) between each manual traced cluster and its predicted counterpart was computed. The subject's metric was defined as the 95th percentile (HD95) of the HD distribution.

PVS Segmentation Model Robustness

In order to test the reproducibility of the model, we computed the segmentation model 4 more times. In addition, we assessed the robustness of the model with regard to the size of the training dataset by training the model with the reduced training set of 20 or 30 volumes. In each case, we present graphs of VL and CL TPR vs PPV for the 9 *PredMap-Thr*.

Testing the PVS Segmentation Model Against the Manual Annotation: EVAL Subset

This section reports the results obtained using the first segmentation model (based on 40 TRAIN participants' data) with the EVAL subset.

PVS Segmentation Model Performance

Algorithm performance was assessed in the BG or in the DWM defined individually (see "Methodology of PVS Segmentation"). For each tissue type and for each subject of the EVAL set, the 4 quantification metrics

(TPR-VL, PPV-VL, TPR-CL, and PPV-CL) and their harmonic means (Dice-VL and Dice-CL) (see "PVS segmentation model training and testing") were computed independently for each of the 9 *PredMap-Thr* (0.1 to 0.9 in 0.1 step).

PVS Segmentation Model Performance for Varying PVS Cluster Sizes

Since enlarged PVS is of primary interest, we investigated how TPR-CL and PPV-CL were modified if we focus on PVSs that were larger than a given threshold. Accordingly, TPR-CL vs PPV-CL plots with the 9 *PredMap-Thr* were generated for cluster size thresholds varying from 0 to 15 voxels with a step size of one voxel. For this analysis, the metric values were computed using all the clusters of the EVAL set, not the average value for each subject.

To compute the TPR-CL, in this approach, we removed all clusters with a size smaller than a given threshold from the manually annotated map. Accordingly, TPs were clusters of the predicted map that had at least one voxel in common with a cluster of the thresholded annotated map. Meanwhile, to compute the PPV-CL, we removed the clusters smaller than a given threshold in the predicted map and compared that map with the annotated map.

PVS Segmentation Model Performance at Predicting PVS Cluster Sizes

Finally, to assess the ability of the model to estimate the PVS size, we compared the size of manually annotated PVS clusters to that of the model-predicted clusters (i.e., TP clusters). For each cluster in the EVAL set, DWM or BG, we computed the linear fit with the model-predicted cluster size as the independent and manually annotated cluster size as the dependent variables, applying 3 of the *PredMap-Thrs* (0.1, 0.5, and 0.9). We report both the slope and the R^2 of the fit. Note that predicted clusters encompassed more than one manually traced cluster were not considered in this analysis. In percentage of the total number of TP clusters, the number of clusters removed was 7.6%, 3.6% and 2% for *PredMap-Thrs* of 0.1, 0.5, and 0.9, respectively.

Testing the PVS Segmentation Model of a Large Subset: ENCOD Subset Analysis

We used the first segmentation model (trained on 40 TRAIN participants' data) to estimate the PVS load in the larger ENCOD subsets, in which the visual rating of PVSs was available for both DWM and BG. Note that no subjects of the ENCOD set were part of the EVAL U-net training set. For each participant, we computed both the numbers of PVS voxels and of PVS clusters at three values of the *PredMap-Thrs* (0.1, 0.5, 0.9). We examined the relationship between the two numbers, searching for the best polynomial fit (up to 3rd degree).

We then used the number of clusters as the proxy for the PVS load and compared it to the visual rating score, since the visual rating was based primarily on the evaluation of a number of PVSs. A logistic regression analysis was used to predict the visual rating score from the PVS cluster number for the whole, ENCOD set individuals (N = 1,782), separately for DWM and for BG.

Assessment of the Prediction Algorithm Interoperability

To indirectly assess the interoperability of our model, we predicted the PVS load in the T1w images from the BIL&GIN database (Mazoyer et al., 2016) that were acquired using a different MRI scanner (Philips Achieva 3T) 10 years prior to the acquisition of the data from the MRi-Share cohort, using the segmentation model trained on the MRi-Share dataset. T1w images were acquired using a different sequence (3D-FFE-TFE; TR = 20 ms, TE = 4.6 ms, flip angle = 10°, inversion time = 800 ms, turbo field echo factor = 65, sense factor = 2, matrix size = 256 x 256 x 180 mm^3, and 1 mm^3 isotropic voxel size). From the 453 subjects included in the BIL&GIN database, 354 were selected for being aged between 18 and 35 years to match the age range in the MRi-Share cohort. We assessed the similarity of PVS distributions in these age-matched cohorts using a QQ plot and tested using the Kolmogorov-Smirnov test. The tests were also computed using cluster size thresholds varying from 0 to 15 voxels by a step size of one voxel, and distributions were compared using the Kolmogorov-Smirnov test.

RESULTS

Manual Segmentation of PVS

The distribution of the number of voxels manually identified as DWM PVS (**Figure 2A**) in 50 participants showed a bimodal shape with 34 participants below 1500 (1.5 cm^3), 14 participants between 1500 and 3000 (1.5 and 3.0 cm^3) and 2 outliers above 4000 (4 cm^3). The distribution of the number of clusters in the DWM of the same participants (**Figure 2C**) shows the same shape than the voxel distribution. For the PVS in the BG (**Figure 2B**), the distribution was more homogeneous, with an average of 180 voxels (or mm^3) and 2 outliers above 500 voxels (or mm^3). The distribution of the number of clusters (**Figure 2D**) had the same shape as the voxel distributions. **Figure 2E** shows representative PVS of both categories. On average, in the whole set, a PVS in the BG was 2 times larger than a PVS in DWM.

Segmentation Model Training and Testing

The most time-consuming step was the autoencoder training using the ENCOD dataset; one epoch run required ∼2000 seconds for 1,425 volumes with a validation subset of 357 volumes. **Figure 3A** shows the evolution of the loss function while training the autoencoder using the ENCOD dataset. The training was stopped after 23 epochs, and previous experiments showed that further training the autoencoder do not improved significantly the training speed of the *U-net*. **Figure 3B** shows the loss evolution on the TRAIN dataset with 80/20 split training/validation for one-fold cross-validation. In this fold, the model with the lowest validation loss (0.4894) at epoch 119 was selected. Training of the autoencoder took ∼12.6 h (∼2000s per epoch) and training of the Unet took ∼12 h (∼72s per epoch). Prediction for one subject is linear in complexity and takes less than one second. **Table 1** shows the average of the 10

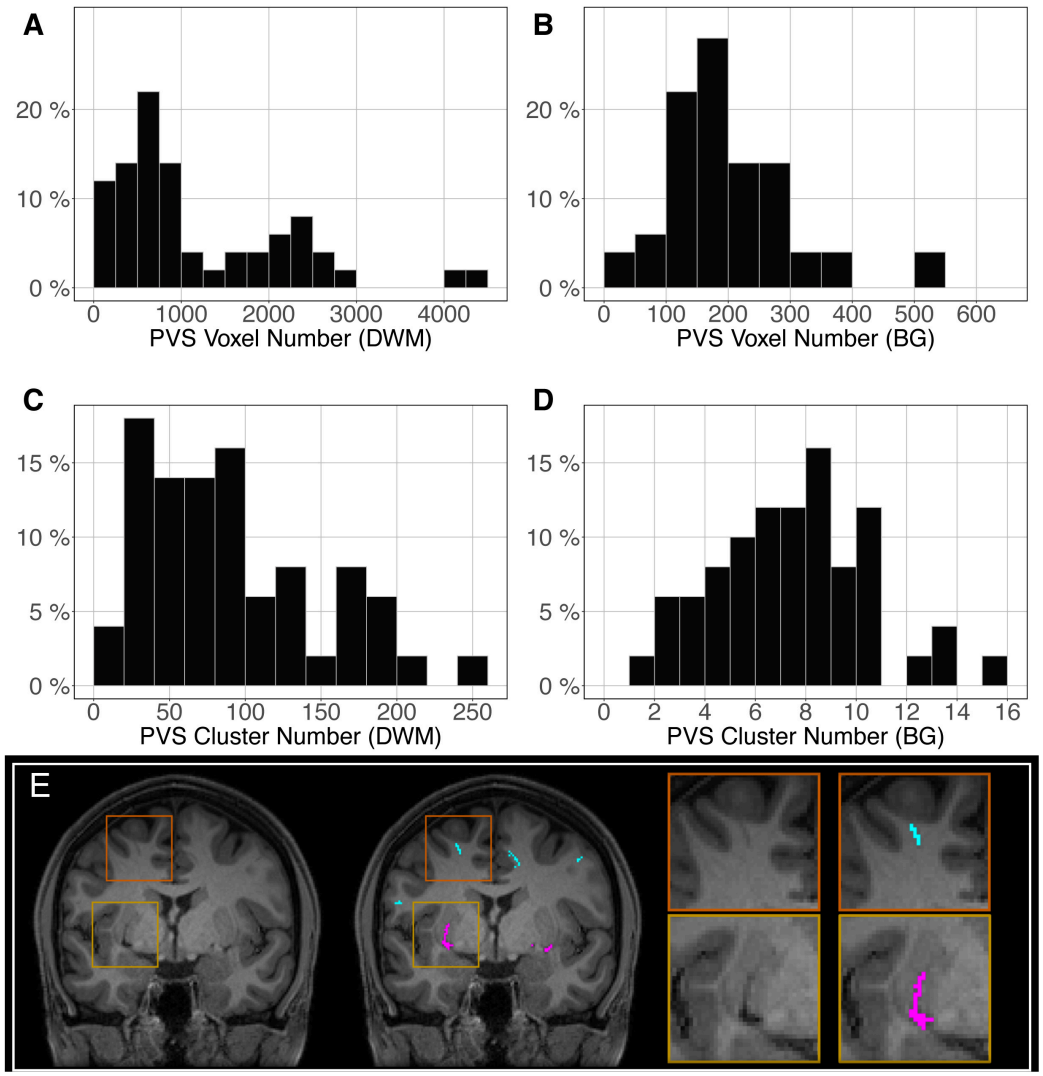


FIGURE 2 | Distribution of the manually delineated PVS voxel load in the deep white matter **(A)** and basal ganglia **(B)** of 50 participants. Distribution of the manually delineated PVS number of clusters in the deep white matter **(C)** and basal ganglia **(D)** of 50 participants. **(E)** Example of delineated PVSs displayed over an MRI slice in one participant in the deep white matter (cyan) and basal ganglia (magenta). DWM, deep white matter; BG, basal ganglia.

EVAL volume result scores for each fold and for the consensus segmentation of the full model. Using the 40 subjects training set, we tested the training with and without the autoencoder. The use of an autoencoder improved both the lesion prediction and the training speed. The DICE-VL metric (**Table 1**) computed with a prediction map threshold of 0.5 increased by 7% on the 5-fold average value; the average gain for all thresholds was 13%. With the autoencoder the U-net training speed was 30% faster than without.

PVS Segmentation Model Robustness Analysis

Figure 4A illustrates the good reproducibility of the model predictions across the 5 repetitions. **Figure 4B** shows that even when the TRAIN dataset size was reduced to 30 individuals, performance was only slightly degraded compared to that obtained with the 40-subject training. While trying to train the algorithm using 20 datasets only, the algorithm performance visibly deteriorated, especially when analyzing the lowest *PredMap-Thrs* (between 0.1 and 0.4). Qualitatively, the degraded performance manifested as some predicted PVS clusters encompassing a large portion of the WM in some subjects of the TRAIN dataset, producing numerous FP voxels.

Testing the PVS Segmentation Model: EVAL-Based Subset Analysis

PVS Segmentation Model Performance

Figure 5 shows for one subject of the EVAL subset, the visual display centered on one of the clusters identified as a TP. Such pictures were computed for each of the TP/FP/FN clusters.

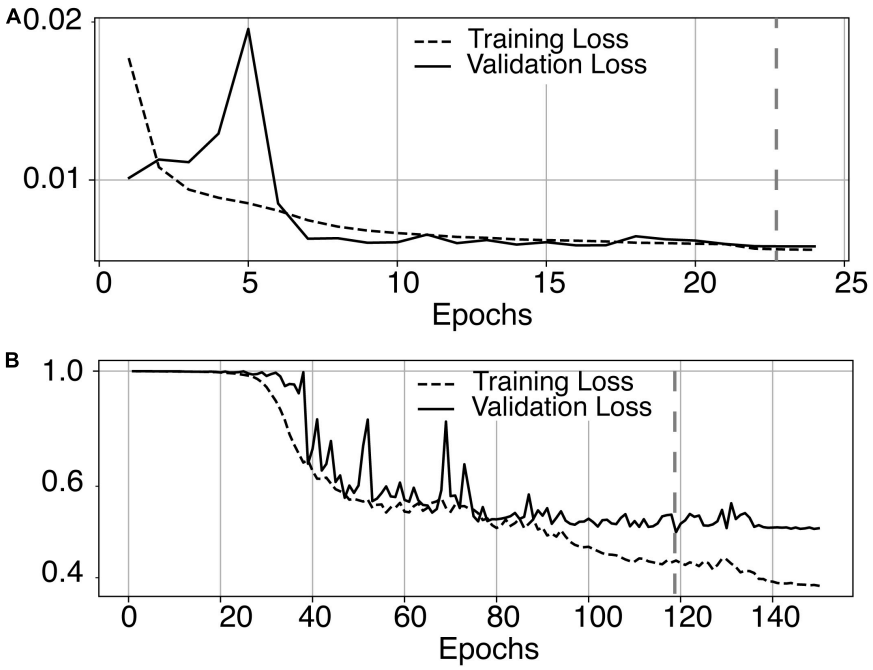


FIGURE 3 | (A) Training loss (MSE, dashed line) and validation loss (continuous line) evolution for the 8.7.2 autoencoder training. The model with the lowest validation loss (0.0075, epoch 23, dashed line) was selected. **(B)** One-fold training loss (dice loss, dashed line) and validation loss (continuous line) evolution of 8.7.2. U-net with weights initialized from the autoencoder. The model with the best validation loss (0.49, epoch 119, dashed line) was selected. Logarithm scale is used for loss.

Figure 6 shows plots of TPRs vs PPVs of segmented PVSs for the 9 *PredMap-Thrs* at both the voxel and cluster levels. **Tables 2**, **3** show the metric values with Dice values. As expected, TPRs decreased and PPVs increased with increasing *PredMap-Thrs*, in both DWM and the BG, and at both VL and CL. While the PPV was similar at both VL and CL, the TPR was markedly higher at the cluster level. Regardless of the type of tissue (DWM/BG) and the level (VL/CL), the Dice coefficients were maximal for *PredMap-Thrs* between 0.4 and 0.6. This is unsurprising, considering that Dice coefficient is the harmonic mean of the TPR and PPV. More precisely, the best Dice values were for an amplitude of 0.6 in the DWM and for an amplitude of 0.4 in the BG. When comparing Dice coefficients on both levels, the impact of the prediction map amplitude thresholding seemed to be higher at the CL than at the VL.

TABLE 1 | Dice loss, true positive rate (TPR-VL), positive predictive value (PPV-VL), and Dice (Dice-VL) metric for each fold and for the 5-fold consensus of the EVAL dataset (10 volumes).

Fold	TPR-VL	PPV-VL	Dice-VL
0	0.526	0.578	0.551
1	0.541	0.579	0.559
2	0.509	0.637*	0.566*
3	0.561*	0.520ˆ	0.540ˆ
4	0.503ˆ	0.633	0.561
5-fold consensus	0.526	0.634	0.575

Note that those metrics were computed without differencing the DWM and BG located VRS.
TPR: true positive rate, PP: positive predictive value, V: voxel level, CL: cluster level.
**Best values of the 5 folds, ˆWorst values of the 5 folds.*

PVS Segmentation Model Performance for Varying PVS Cluster Sizes

Figure 7 shows grid plots of TPR-CL vs. PPV-CL (**Figures 7A,B** for DWM and BG PVSs, respectively) and surface plots of the Dice-CL (**Figures 7C,D** for DWM and BG PVSs, respectively) at different intensity (0.1 to 0.9) and cluster size (0 to 15 voxels) thresholds. The values of each of these metrics are provided in the **Supplementary Material**.

As expected, better performance for both metrics was observed when the cluster size threshold was increased. At lower values (1 to 5 voxels), the cluster size threshold had a major impact on PPVs regardless of the intensity threshold. For instance, for DWM PVSs, ignoring clusters made of only one voxel increased the PPV by approximately 15% while increasing TPRs by only a few percentage points. Furthermore, Dice scores above 0.9 could be obtained regardless of the threshold when removing clusters of size 15 or less (**Figure 7C**). Data above a cluster threshold of 15 are not shown, as less than half of the EVAL volumes had at least one cluster in the FP or FN classes at this threshold.

For the BG PVSs the grid plot (**Figure 7B**) is less smooth than the DWM PVSs grid plot (**Figure 7A**), due to a lower number of PVSs in the BG (101 clusters on the whole EVAL dataset) compared to the number in DWM (882 clusters). Nevertheless, a

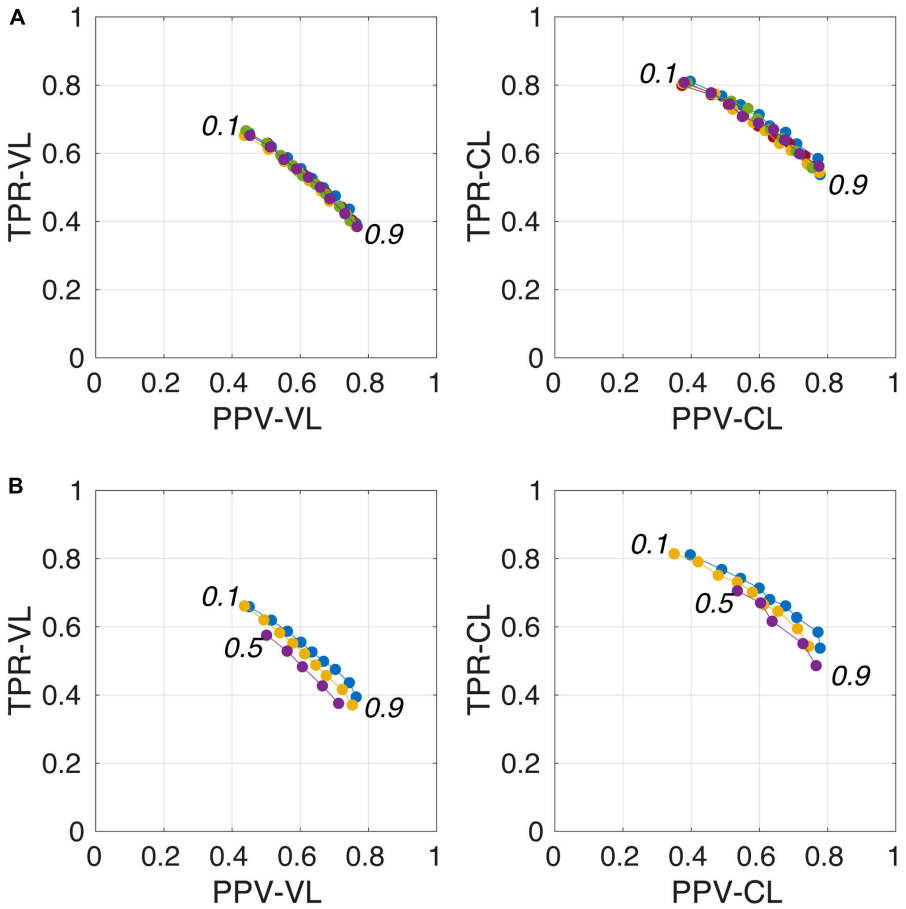


FIGURE 4 | Plot of true positive rates (TPRs) vs. positive predictive values (PPVs) for the 9-amplitude prediction map thresholds (PredMap-Thrs, 0.1 to 0.9 in step sizes of 0.1) computed for replication analysis based on 40 TRAIN subjects (each of the 5 replications is shown with a different color) **(A)** and for training the model with 20 (purple), 30 (yellow), and 40 (blue) subjects **(B)** using voxel level (VL) (left) and cluster level (CL) (right).

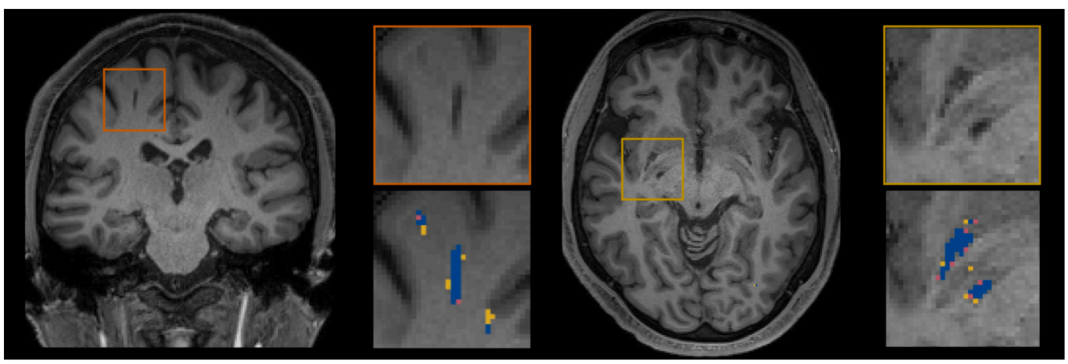


FIGURE 5 | Example of true positive clusters detected and displayed in coronal and axial orientations. Blue indicates TP voxels, orange indicates FP voxels, and yellow indicates FN voxels.

marked increase was observed in both the TPR and the PPV after removing the PVS of one voxel. The progression toward optimal values of the PPV was faster in BG PVSs than in DWM PVSs. The Dice metric reached 0.9 (regardless of the *PredMap-Thr*) using a cluster threshold size of 7 voxels.

PVS Segmentation Model Performance at Predicting PVS Cluster Sizes

Figure 8 shows the model-predicted PVS size vs. the size of its corresponding manually annotated PVS. At the *PredMap-Thr* of 0.5, the linear regression slopes were 0.58 (R^2 = 0.72) and

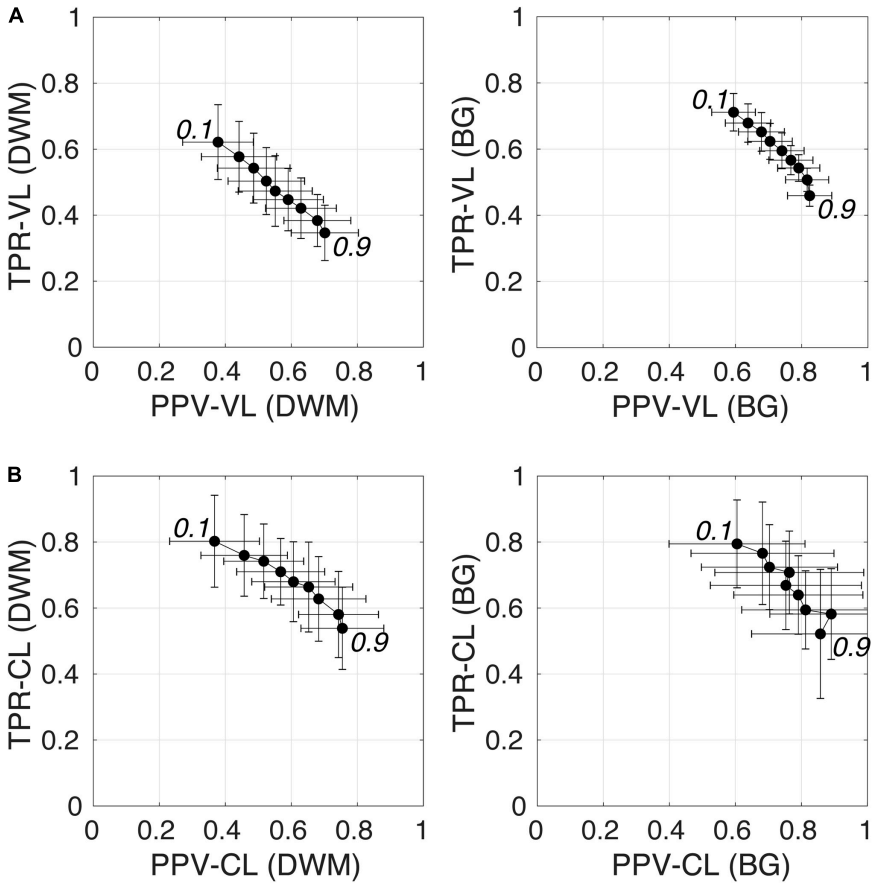


FIGURE 6 | (A) Plot of the true positive rate (TPR) vs the positive predictive value (PPV) for the 9 amplitude prediction map thresholds (PredMap-Thrs, 0.1 to 0.9 in steps of 0.1) in the deep white matter (DWM, left) and basal ganglia (BG, right) at the voxel level (VL) **(A)** and cluster level (CL) **(B)**. Each point gives the average and standard deviation of metrics across the 10 individuals of the EVAL dataset.

0.87 (R^2 = 0.93) for the DWM and BG PVSs, respectively. With a 0.1 *PredMap-Thr*, the slopes were closer to 1 (0.83 and 1.2 for DWM and BG PVS, respectively) and were markedly lower with a *PredMap-Thr* of 0.9 (0.39 and 0.63). In both cases, the uncertainty analysis demonstrated that there was no bias in the model regardless of the values.

Testing the PVS Segmentation Model of a Large Subset: ENCOD Subset Analysis

Relation Between the Voxel and Cluster Charge

Figure 9 shows that the relation between the PVS voxel load and the number of clusters can be modeled using a second-order polynomial fit. The adjusted squares of the correlation were 0.94, 0.96 and 0.96 for *PredMap-Thrs* of 0.1, 0.5 and 0.9, respectively.

Testing the Segmentation Model Against Visual Rating Results

The logistic regression between the number of DWM PVS clusters (**Figure 10A**) and the visual grading rating (**Figure 10B**) in the ENCOD set was highly significant for all 3 *PredMap-Thrs* ($R^2_{0.1}$ = 0.38, $R^2_{0.5}$ = 0.45 see **Figure 10C**, $R^2_{0.9}$ = 0.47, $p < 0.001$, N = 1,782).

Figure 11 shows the comparison of the number of BG PVS clusters (**Figure 11A**) and the first 3 levels of visual grading (**Figure 11B**) in the ENCOD set. Again, the logistic regression showed significant correspondence between the two ($R^2_{0.1}$ = 0.02, $R^2_{0.5}$ = 0.05 see **Figure 11C**, $R^2_{0.9}$ = 0.04, $p < 0.001$, N = 1782). Note that there were no subjects with BG PVS category of level 4 in the visual rating scale, as such a level is more routinely observed in elderly subjects.

Assessment of the Prediction Database Interoperability

Figures 12A,B shows the overlap of the number of cluster distributions in the MRi-Share and BIL&GIN datasets. Although the datasets were age-matched, more PVSs were observed in the BIL&GIN subjects when not filtering small clusters (**Figure 12A**). Note that this result can also be clearly seen on the cumulative plot of both distributions, also called the QQ plot, presented in **Figure 12C**. The difference was quantified by a Kolmogorov-Smirnov test, which showed a significant difference in the distributions (d = 0.23, p-Value $< 10^{-4}$). As we showed in the previous sections that filtering out the small cluster improves the reliability of the algorithm, we aimed to find the level of

TABLE 2 | Voxel level (VL) detection true positive rate (TPR), positive predictive value (PPV), and the harmonic mean (Dice) for the 9 amplitude prediction map thresholds (PredMap-Thrs) in deep white matter (DWM) and the basal ganglia (BG).

Tissue	*PredMap-Thr*	TPR-VL	PPV-VL	Dice-VL
DWM	0.1	0.62*	0.38	0.47
	0.2	0.58	0.44	0.50
	0.3	0.54	0.49	0.51
	0.4	0.50	0.52	0.51
	0.5	0.47	0.55	0.51*
	0.6	0.45	0.59	0.51
	0.7	0.42	0.63	0.50
	0.8	0.38	0.68	0.49
	0.9	0.35	0.70*	0.46
BG	0.1	0.71*	0.59	0.65
	0.2	0.68	0.64	0.66
	0.3	0.65	0.68	0.66
	0.4	0.62	0.70	0.66*
	0.5	0.59	0.74	0.66
	0.6	0.57	0.77	0.65
	0.7	0.54	0.79	0.64
	0.8	0.51	0.82	0.63
	0.9	0.46	0.82*	0.59

*Best values.

TABLE 3 | Cluster level (CL) detection true positive rate (TPR), positive predictive value (PPV), harmonic mean (Dice) metric, and Hausdorff distance (HD95) for the 9 amplitude prediction map thresholds (PredMap-Thrs) in deep white matter (DWM) and the basal ganglia (BG).

Tissue	*PredMap-Thr*	TPR-CL	PPV-CL	Dice-CL	HD95
DWM	0.1	0.80*	0.37	0.50	2.05
	0.2	0.76	0.46	0.57	2.02
	0.3	0.74	0.52	0.61	1.99
	0.4	0.71	0.57	0.63	1.96
	0.5	0.68	0.61	0.64	2.01
	0.6	0.66	0.65	0.66*	2.13
	0.7	0.63	0.68	0.65	2.21
	0.8	0.58	0.74	0.65	2.29
	0.9	0.54	0.75*	0.63	2.33
BG	0.1	0.79*	0.60	0.69	1.88
	0.2	0.77	0.68	0.72	1.84
	0.3	0.72	0.70	0.71	1.90
	0.4	0.71	0.76	0.73*	2.15
	0.5	0.67	0.75	0.71	2.17
	0.6	0.64	0.79	0.71	2.31
	0.7	0.59	0.81	0.69	2.45
	0.8	0.58	0.89*	0.70	2.73
	0.9	0.52	0.86	0.65	3.32

*Best values.

cluster filtering that made the distributions more comparable. After removing clusters below 5 voxels, both distributions (see **Figure 12B** and the QQ plot in **Figure 12D**) were not different according to the Kolmogorov-Smirnov test (d = 0.077, p-Value = 0.058). The discrepancies remained (see **Figure 12D**) mainly for the few subjects showing the highest number of PVSs, with more subjects in the MRi-Share dataset having the higher number of PVS clusters than in the BIL&GIN dataset.

DISCUSSION

Methodological Issues

Model Building

The U-net architecture for our machine learning model was chosen for its segmentation performance with a small training set. We initially tried to implement our algorithm with a training set made of only 10 images, but such a small training set did not allow us to obtain stable and reproducible results; in some cases, the training phase did not converge on a solution or took a very long time (more than 1000 epochs). We then observed that initializing the U-net model with the weights determined by training the autoencoder on the ENCOD set helped provide stable results with fewer epochs (usually under 200 epochs).

The main hyperparameters chosen for the U-net used in our work were based on the U-net topology described in Ronneberger et al. (2015) adapted for 3D images and constrained by the available RAM in a GPU where the model parameters together with a batch of images should be stored. The initial number of kernels for the first stage of convolutions is important since it provides the basis for the number of features that may be extracted at each level of resolution: the larger the number is, the more features it can extract; however, the resolution has a large impact on the size of the model. The number of stages is also important since it adds information for each successive resolution, whereas the number of convolutions for each stage, when greater than 1, seems to be less important. As stated above, 7 stages are needed for the model for going to the bottom stage with an image size of (1, 1, 1), which only allows for an initial number of kernels of 8 with the GPU we used. Since training the model took approximately 10 h, it was difficult to perform a full grid search or even to use the kind of optimization described in Bergstra et al. (2013). Obvious tests, such as decreasing the number of stages to increase the number of kernels, were performed without significant gains. Image cropping during the preprocessing phase provided a 52% reduction in data volume size, which authorized two larger batches and thus increased the training speed. We also tested the segmentation using data registered in MNI stereotaxic space with the same sampling as the acquisition (1 mm^3); however, both VL TPR and PPV metrics were lower than without normalization (15 and 25%, respectively, for an amplitude PredMap-Thr of 0.5). Visual analysis of the prediction showed that it could be attributed to smoothing due to interpolation. In fact, smoothing small elongated structures such as PVS makes it more difficult for them to be detected because of the induced partial volume effect. We also tested the increase in the training set (both with and without data stereotaxic normalization) using flipping with respect to the interhemispheric plane, but neither case provided any significant increases in the TPR and PPV metrics.

Finally, we tested a more complex U-net topology with U-net++ (Zhou et al., 2020), but it proved to be a failure for the

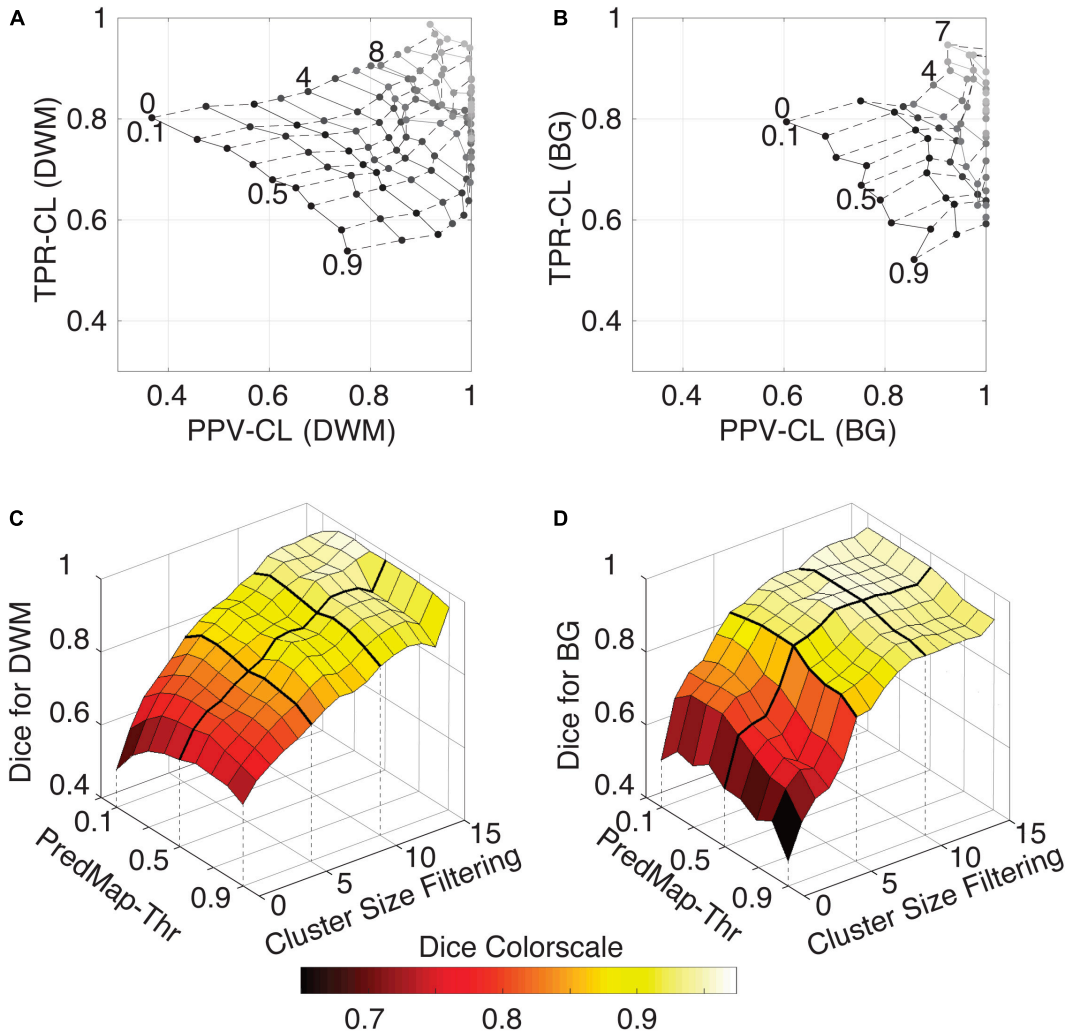


FIGURE 7 | (A) Cluster-level true positive rate (TPR-CL) vs. the positive predictive value (PPV-CL) of PVS detection in deep white matter (DWM) for the EVAL dataset (N = 10). Metric values are shown with masking clusters with increasing size (dashed lines, 0 to 15 voxels) and 9 amplitude prediction map thresholds (PredMap-Thrs, solid lines, 0.1 to 0.9 in steps of 0.1). **(B)** Same for detection of basal ganglia (BG) PVS, increasing the cluster size threshold from 0 to 7 voxels. **(C)** Dice scores of PVS detection in deep white matter for the EVAL dataset (N = 10) at the 9 PredMap-Thr and masking clusters with increasing size (0 to 15 voxels). **(D)** Same as C but for basal ganglia PVS detection.

small amplitude PredMap-Thr (0.1 and 0.2) with TPR-VL near 1 and PPV-VL near 0 and marginally better at the other amplitude thresholds for the TPR and worst for the PPV. For the sake of parsimony, we chose to use the more basic U-net topology.

Model Reproducibility and Robustness

The random nature of the initialization was tested and proved not to be an issue; thus, we did not implement a double-level procedure, such as repeating the training for each fold. This would lengthen the training process (×5) and require the management of 5 times the number of parameters.

The size of the training set proved to be an issue when it was reduced to 20. At this size, using an amplitude threshold below 0.5, the segmentation failed for some of the folds. With a weaker reduction of the training set at 30 data points, the only visible effect was a decrease in the TPR (and not the PPV) of a few percentage points—values that were well below the uncertainties of the measured values (see **Figure 6**). Such good results were expected when using *U-net* technology. However, to both maintain robustness and limit overlearning, the training dataset size should be as large as possible.

Algorithm Performance

We presented our algorithm performance at both the VL and CL. Regardless of the amplitude threshold, both TPR values and PPVs were higher at the CL than at the VL. This discrepancy was related to an imbalance in FNs, possibly due to the small difference in shape between manually traced and predicted clusters. While it was difficult to obtain a definite answer, it is probable that the full extent of the PVS was not included in the manual tracing, leading

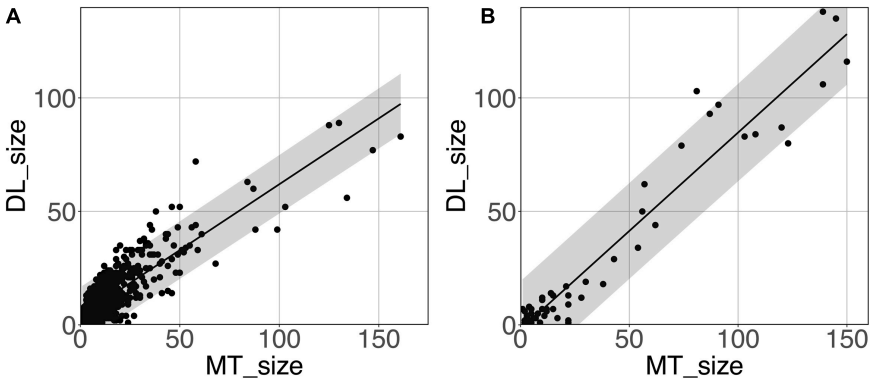


FIGURE 8 | Predicted cluster size (DL_size) vs. the manual tracing size (MT_size) for deep white matter **(A)** and the basal ganglia **(B)**. Both were computed with the amplitude prediction map threshold (PredMap-Thr) of 0.5. Black line: linear regression curve; shaded area: 95% confidence interval of individual values.

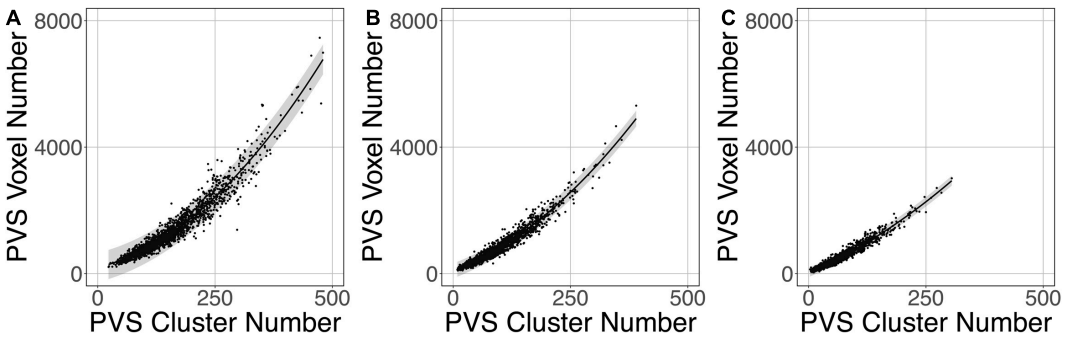


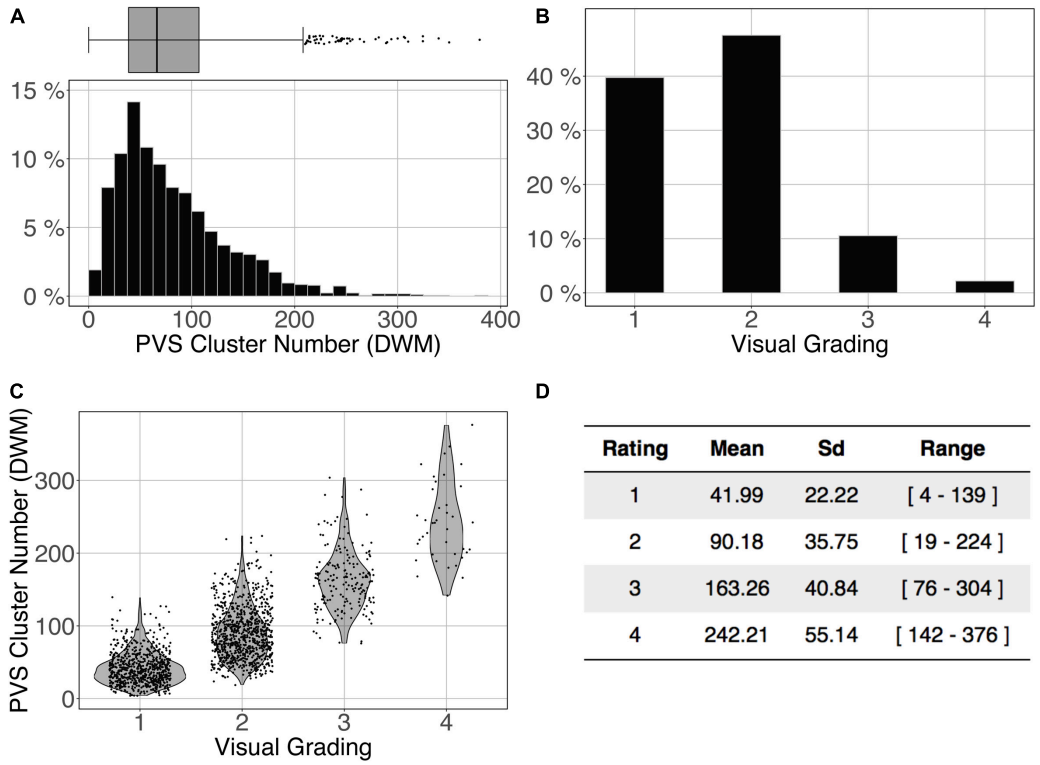
FIGURE 9 | Relationship between the number of clusters and the PVS charge load at 3 amplitude prediction map thresholds (PredMap-Thrs) of 0.1 **(A)**, 0.5 **(B)**, and 0.9 **(C)**. Black line: best second-order polynomial regression curve; shaded area: 95% confidence interval of individual values.

to the observed discrepancies. Nevertheless, metrics measured at both levels exhibited a strong correlation with each other through a monotonic second-order polynomial relationship. Whatever the chosen metric level (voxel or cluster), TPR (resp. PPV) decreased (resp. increased) when the amplitude threshold increased, which led to the best Sorensen-Dice coefficient value with a medium threshold. We proposed that the 0.5 threshold should be used if one does not have any specific reason to favor one of the TPR or PPV metrics.

The algorithm was trained without taking the PVS location into account, namely, whether the PVS was located in the DWM or in the BG. However, for the reasons explained above (see section 1.2), the algorithm predictions on the EVAL and ENCOD sets were analyzed independently for the two locations. From manual tracing, we observed in this dataset of young subjects that the deep WM-located PVS was more numerous and smaller than the BG PVS. Averaging cluster-level metrics across the EVAL dataset demonstrated that, regardless of the amplitude PredMap-Thr, TPR values were equivalent for the 2 locations, whereas PPVs were higher for BG than for DWM. As PPVs are dependent on the FP rate, we investigated their behavior when filtering out clusters according to their size. For both locations, filtering out clusters of 1 voxel size led to a large increase in both metric values. This was expected since no PVS of one voxel size was manually traced. Nevertheless, increasing the size of filtered out clusters led to further increases in values of both metric values with the increase being larger for PPVs than for TPRs. To summarize, the good performance of our DL algorithm could be improved by considering only PVSs of larger sizes, a feature that could be very interesting when the goal is to detect and quantify dilated PVS.

Likewise, we showed that the PVS sizes were linearly related to the true sizes, albeit underestimated, and the degree of underestimation depended on the chosen amplitude PredMap-Thr. The best estimation of PVS size was obtained with the lowest threshold, albeit at the expense of lowered PPVs and Sorensen-Dice values.

Comparing the PVS prediction by the algorithm with their visual rating on the ENCOD dataset provided important information regarding the clinical utility of the algorithm. Although a visual rating of the PVS burden was made on only one slice for each location (DWM or BG), we observed a very significant agreement between PVS global burdens estimated by the algorithm and by visual rating ($p < 0.0001$) in both locations. For the best results, nearly 50% of the visual rating variance explained by the algorithm predictions was observed for the number of DWM clusters estimated with an amplitude PredMap-Thr above 0.5. Though also significant, the BG PVS explained-variance was 5 times lower, due in part to very



Rating	Mean	Sd	Range
1	41.99	22.22	[4 - 139]
2	90.18	35.75	[19 - 224]
3	163.26	40.84	[76 - 304]
4	242.21	55.14	[142 - 376]

FIGURE 10 | Cluster analysis of the deep white matter (DWM) PVS of the ENCOD set (N = 1,782). The prediction maps were thresholded at 0.5. **(A)** Histogram distribution of the number of clusters per subject. **(B)** visual grading. **(C)** Number of clusters vs. visual grading ($R^2_{0.5}$ = 0.45, p-Value $< 10^{-4}$). **(D)** Summary of PVS load values within the 4 groups.

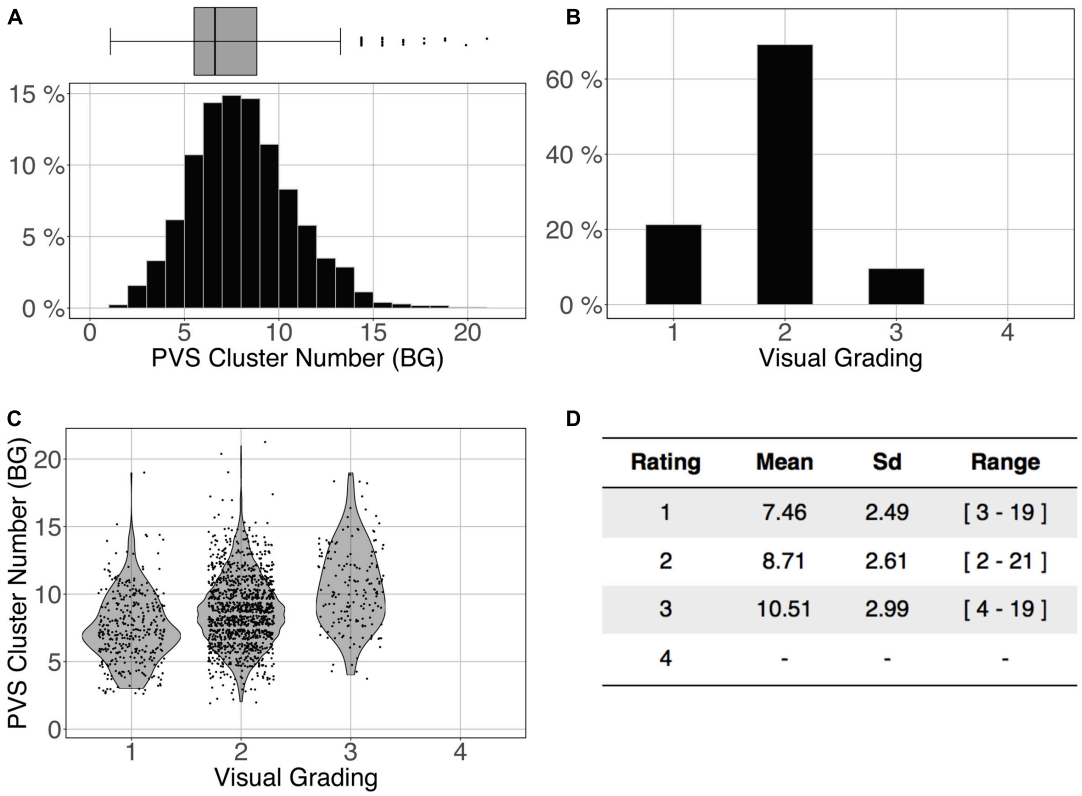
unbalanced frequencies of rating for the visual scale degrees; 65% of individuals were rated as degree 3 and none were rated as degree 4.

Interoperability of imaging marker detection algorithms is crucial, especially in the context of population neuroimaging in which multiple databases must be jointly analyzed. This can be achieved either by retraining the neural network for each database or by applying the neural network previously trained on one dataset to the other datasets. The latter approach is preferable since the former would require manual tracing of at least 40 subjects for each dataset (30 subjects for the training validation and 10 subjects for the test); manual tracing for this number of subjects is time consuming and could introduce some bias if the operators are not the same. Here, we applied the neural network trained on the MRi-Share dataset to the detection of PVSs in images acquired on a different scanner in a different sample of individuals having the same age range. The PVS number distributions were found to be very similar for the two datasets and not significantly different when removing clusters with sizes less than 5 voxels.

Comparison to Other Segmentation Methodologies

As stated in the introduction, the different segmentation methods proposed in the literature fall into two broad categories: those based mainly on image processing designed to enhance PVS visibility on the image and those that emphasize machine learning classifications and increasingly, DL-based approaches. In fact, this subdivision is not as clear-cut as some methods of the former category often use machine learning classifiers after image enhancement (support vector machine (SVM) (Gonzalez-Castro et al., 2017), or random forests (Zhang et al., 2017)), while some of the latter category used enhanced images as input for the neural networks (Lian et al., 2018). Regardless of the category, the performance is typically evaluated with several different metrics, and the choice of evaluation method is dictated by what is available as the ground truth. Briefly, the TRP (also known as *sensitivity*) and PPV (also known as *precision*) are often reported whenever the ground truth is based on PVS voxel manual tracing, whereas Pearson correlation or Lin's coefficient is used when only the number of PVSs is available. We computed similar metrics in order to facilitate the comparison of the performance of our algorithm to those in the literature.

When reviewing and comparing existing PVS detection algorithms, several other factors should be taken into account. First is the quality and type of the input image used for the PVS detection, such as the strength of the acquisition MR scanner and image resolution. Some studies used data acquired either at 1.5 T (Gonzalez-Castro et al., 2017; Dubost et al., 2020), 3T ((Ballerini et al., 2018; Boespflug et al., 2018; Sepehrband et al., 2019; Sudre et al., 2019) and our data) or 7T (Lian et al., 2018;



Rating	Mean	Sd	Range
1	7.46	2.49	[3 - 19]
2	8.71	2.61	[2 - 21]
3	10.51	2.99	[4 - 19]
4	-	-	-

FIGURE 11 | Cluster analysis of the basal ganglia (BG) PVS of the ENCOD set (N = 1,782). The prediction maps were thresholded at 0.5. **(A)** Histogram distribution of the number of clusters per subject. **(B)** Visual grading. **(C)** Number of clusters vs. visual grading results ($R^2_{0.5}$ = 0.05, p-Value $< 10^{-4}$). **(D)** Summary of PVS load values within the 4 groups.

Jung et al., 2019). Up to 3T, the data are most commonly acquired with approximately the same 1 mm^3 sampling size, whereas with 7T scanners, sampling is usually 8 times higher, providing a crucial advantage for detecting small DWM PVS. Notably, the detection of thin features like PVS, especially in DWM, is much improved at high resolution imaging; thus, comparison of the present work with the literature is only meaningful for data acquired on 3T and 1.5T scanners. The imaging sequence is another important factor, as T2-weighted sequences provide better contrast (Zong et al., 2016) for PVS, whereas we worked with T1w images. However, the T1w has the advantage of reduced potential confusion between PVS and WM hyperintensities, which is a common problem when working with T2-weighted images. Another important advantage of an algorithm trained with T1w is the fact that the T1w images are the most commonly available 3D images with millimeter resolution, and it opens the possibility to quantify PVS in datasets that were not originally designed to detect PVS.

The type and size of the evaluation set is also a determining factor in the comparisons. When the gold standard was a visual rating the size of the evaluation set in previous studies ranges from 20 (Ballerini et al., 2018) to 28 (Boespflug et al., 2018) to 100 (Sepehrband et al., 2019). With 1782 subjects we are far above those values. When the gold standard was manual tracing (like our EVAL set) it ranges from 2 (Sudre et al., 2019), to 6 (Lian et al., 2018) to 19 (Zhang et al., 2017). However, it should be noted that this last study used the same set to train and test. If only considering studies using an independent estimation set, our EVAL set with 10 subjects is larger than any other published studies. One other study reported tracing from 1000 subjects for evaluation (Dubost et al., 2020). Their manual tracing consisted of tagging the PVSs in two slices (one in the BG and one in the DWM), in the complete volumes of the hippocampus and the midbrain.

Overall, our algorithm exhibited better performance when compared to the performance measures reported in previous studies that used T2-weighted images acquired at 3T or less. In the "image processing" category, Ballerini et al. (Ballerini et al., 2018), using a Frangi filter on a dataset of 20 subjects, reported a Pearson correlation between visual rating and either the total PVS volume burden (r = 0.53) or the total PVS number (r = 0.67). The corresponding values were 0.79 and 0.77, respectively, based on the number of clusters derived from our algorithm (0.5 amplitude prediction map threshold) and the visual rating in the ENCOD set (N = 1782 subjects). Using filtering techniques and morphological constraints, Boespflug et al. (Boespflug et al., 2018) reported correlations of 0.58 (N = 28 subjects) and 0.76 for the PVS volume and number of clusters; these values indicated worse performance for the total volume burden and equivalent performance for the number of PVS when compared

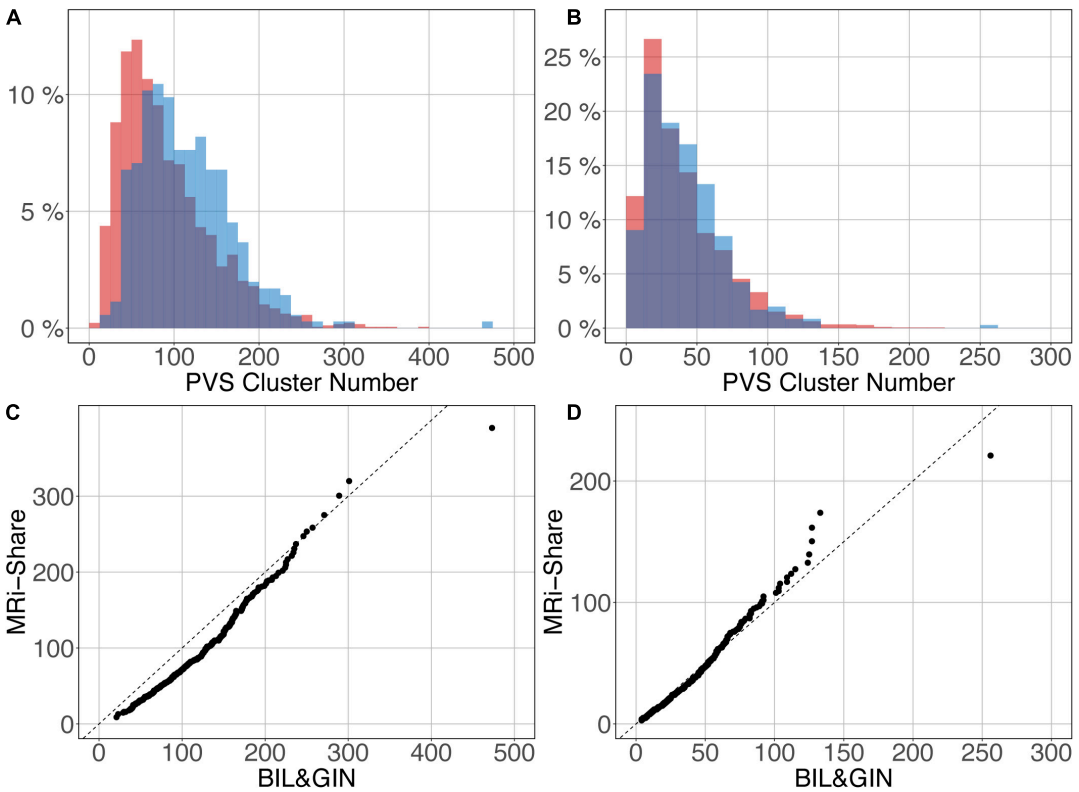


FIGURE 12 | Overlap of the distribution of the number of clusters per subject for MRi-Share (red) and BIL&GIN (blue) cohorts computed at an amplitude threshold of 0.5 and without **(A)** and with **(B)** filtering clusters that were ≤5 in size. **(C,D)** QQ plots of the distributions shown in panels **(A,B)**, respectively.

to ours. Finally, using the Frangi filter to create "vesselness" images and nonlocal mean filtering techniques, Sepehrband et al. (Sepehrband et al., 2019) reported a visual rating (N = 100 subjects) to cluster number correlation of 0.61 (filtering out the less than 5 voxel clusters), again showing lower correlation than ours. In the 2 studies using a DL methodology, Dubost et al. (Dubost et al., 2020) used a convolutional neural network (CNN) weakly supervised detection approach with attention maps to create class activation maps and reported a cluster-like level sensitivity (called TPR-CL in our study) of 0.51 and 0.57 for DWM and BG PVS, respectively, whereas we reported values of 0.68 and 0.67 for the PVS in the two regions. The PPV metrics were in the same range between our study and Dubost et al study; they reported values equivalent to our PPV-CL index to be 0.69 and 0.7 for DWM and BG PVS, respectively, compared to 0.61 and 0.75 in our study. Note that the comparison is only approximate, since we used a full 3D PVS manual tracing on 10 subjects for the evaluation of PPV, while they used PVS tagging on two slices in 1000 subjects, in which only the center of mass of each PVS was traced. While we quantified the prediction performance based on the overlap of manually traced and predicted PVS across the whole brain, they quantified the matching between the annotated and predicted PVS in one slice based on a proximity distance of mass centers computed using the Hungarian algorithm. Sudre et al. (Sudre et al., 2019), using region-based CNN (R-CNN), based on a test set of 2 subjects, reported a sensitivity of 0.73 (N = 2) for PVSs above 5 voxels in size, whereas we reported a value of 0.76 and 0.86 (N = 10, for DWM and BG, respectively) under the same conditions (see **Figure 7**).

Even when comparing with studies using T2-weighted images acquired at 7T, our results indicate competitive performance of our algorithm: Zhang et al. (2017), using a structured random field on extracted vascular features on 19 subjects, using the same 19 subjects to test the algorithm they reported a Dice coefficient of 0.66, which was identical to what we obtained for BG PVS, whereas for DWM PVS, it was lower (0.51). However, the slightly better performance in Zhang study for DWM PVS is to be expected, as a higher sampling rate increases the detectability of small PVSs that are mainly located in the DWM. Similarly, Lian et al. (Lian et al., 2018), who used a U-net approach called M2EDN on higher resolution (0.5^3 mm^3) T2-weighted images, testing the algorithm on 6 manual traced subjects they reported 0.77 ± 0.04 (N = 6) Dice values (at the VL regardless of the localization) compared to 0.66 and 0.51 (N = 10) for the BG and DWM VRS, respectively, in our study. Without the same resolution in the data while searching small objects, it is difficult to interpret those quantitative differences; thus, we will discuss the differences in the methodology. First, our implementation allows us to process the whole-brain volume simultaneously as opposed to patches of the volume as done in M2EDN. By doing so, we avoid artificially cutting PVSs in 2 or more parts and thus

make it possible to compute the number of PVS and not solely the number of PVS voxels. This choice impacts the architecture of the CNN; while Lian et al. (Lian et al., 2018) choose 3 stages with 64 features per level to accommodate the local nature of the patches with PVS, we choose 7 levels with an increasing number of features (from 8 for the first level to 512 at the deepest level) to accommodate both the local and global patterns of PVS repartition in the volume (PVSs are not located everywhere in the volume). A second difference relies on using or not using the multiscale feature, which is equivalent to the U-net++ topology (Zhou et al., 2020) that we tested in the selection of the best topology (see 3.1.1). We did not select it for further analysis because of failures for the small amplitude PredMap-Thr (0.1 and 0.2) and no univocal amelioration of the evaluation metrics at the other thresholds. Note that Lian et (Lian et al., 2018) did not report such failure, but they also did not report the amplitude PredMap-Thr used to compute their algorithm evaluation metrics. Finally, it must be emphasized that M2EDN is based on multichannel MRI data, including raw T2-weighted images, enhanced T2-weighted images and probability maps. Using the probability map in what Lian et al. (Lian et al., 2018) called "autocontext" did not demonstrate a decisive advantage, and the procedure was limited to one iteration. In addition, we believe that having a portable PVS detection algorithm operating on 3D-T1w images only constitutes an important advantage in favor of our approach since this algorithm could be easily applied to most of the existing MRI databases and/or clinical brain MRI protocols that often include a high-resolution 3D T1w acquisition.

Additionally we tested two promising methods applied in medical imaging but not yet to the segmentation of PVS. Adding the Generic Autodidactic model (Zhou et al., 2021) during the training phase did not improve neither the DICE-VL (−1,8%) nor DICE-CL (−1,5%) metrics. We also tested the nnU_Net based segmentation (Isensee et al., 2021). Compared to our method, decreases in performance were observed for the DICE-VL (−12%) and DICE-CL (−10%).

Limitations and Potential Solutions

The first limitation of the work is the number of subjects included in the training set. While 40 fully traced subjects is more than what is typically used in the previously published methods, a higher number could improve the prediction. The 3-Dimensional tracing is a very complex and time consuming task and the PVS are very small objects thus sometimes difficult to detect. One solution proposed by Lutnick et al. (2019) consists of integrating an iterative annotation technique in the training loop. At the first iteration the prediction map of a set subject's data not included in the training set is reviewed by an expert and cluster detected tagged either true PVS or artefacts. At the second iteration those data are added to the initial training set and the neural network retrained.

Similarly 10 subjects for the testing set is again above what is used in most studies. Nonetheless, performance index values can vary significantly across subjects, partly depending on their PVS load, making it difficult to assess the performance. Having a larger testing set would definitely make it easier to evaluate improvements in the model. However, it would require more time-consuming manual tracing of subjects by an experienced rater.

Another potential limitation is the applicability of this algorithm trained on young subjects to the PVS prediction in older subjects. Unlike other small vessel disease imaging markers (such as white matter lesions, lacunes, or microbleeds), the number of PVS can be very large even in young subjects (see distribution in the MRi-Share database shown in the article), since PVS is a physiological space that appears and develops with the growth of brain vessels during the fetal life (see Introduction section). What may be dependent on age and the age-related brain disorders is the occurrence and number of enlarged PVS. Since we observed better performance of the algorithm for predicting larger PVS clusters (both better sensitivity and precision), we expect it to perform reasonably well in older subjects. However, the presence of other small vessel diseases that result in hypointense T1, in particular lacunes that are rare in young subjects, may hamper with the precision. Future work should test the performance of the algorithm in scans from older subjects to evaluate these issues.

It should also be noted that the present algorithm was trained using 1 cubic millimeter sampling size voxel, and it cannot accommodate other sample sizes without retraining. Thus, this algorithm cannot be applied to scans obtained at a lower resolution. For higher resolution images with sub millimetric voxels, either the data will have to be resample to 1mm3 or the neural network needs to be trained on a new set of manual traced data.

CONCLUSION

We implemented a U-net-based DL algorithm for the 3D detection of PVS on T1w images both in the DWM and the BG area. Overall, when considering images of comparable resolution, our U-net-based DL PVS segmentation algorithm exhibited better performance than that of previously published methods working with T2-weighted images, whether based on signal processing or DL methods. The algorithm performance and its interoperability for 3T T1w data are important features in the context of both routine clinical analysis and mega- or meta-analysis of PVS across databases, as 3D millimeter T1 images are available for many existing neuroimaging databases.

DATA AVAILABILITY STATEMENT

The BIL&GIN is available through a data-sharing model based on collaborative research agreements. Request for joint research projects can be made through the BIL&GIN website (http://www.gin.cnrs.fr/BIL&GIN). To access i-Share and MRi-Share de-identified data (including the labeled data), a request can be submitted to the i-Share Scientific Collaborations Coordinator (ilaria.montagni@u-bordeaux.fr) with a letter of intent (explaining the rationale and objectives of the research proposal), and a brief summary of the planned means and options for funding. The i-Share Steering Committee will assess this request, and provide a response (principle agreement, request to

reformulate the application or for further information, refusal with reasons). If positive, applicants will have to complete and return an application package which will be reviewed by the principal investigator, the Steering Committee, and the operational staff. Reviews will be based on criteria such as the regulatory framework and adherence to regulations (access to data, confidentiality), the scientific and methodological quality of the project, the relevance of the project in relation to the overall consistency of the cohort in the long term, the complementarity/competition with projects planned or currently underway, ethical aspects. De-identified data (and data dictionaries) will be shared after (i) final approval of the application, and (ii) formalization of the specifics of the collaboration.

ETHICS STATEMENT

The MRi-Share protocol has been approved by the CPP-SOM III ethics committee and the BIL&GIN protocol by the Basse-Normandie ethics committee.

AUTHOR CONTRIBUTIONS

PB: methodology, software, validation, writing – original draft, writing – review and editing, supervision, project administration, and funding acquisition. AT: data curation, writing – original draft, and writing – review and editing. AL: software, writing – original draft, validation, writing – review and editing, and visualization. FA: validation. ZH: software. VN: software, writing – original draft, and writing – review and editing. VV: software, writing – original draft, and writing – review and editing. LL: data curation. JZ: data curation. Y-CZ: data curation. CT: investigation, writing – review and editing, and funding acquisition. BM: conceptualization, formal analysis, investigation, writing – original draft, writing – review and editing, supervision, project administration, and funding acquisition. MJ: conceptualization, formal analysis, investigation, writing – original draft, writing – review and editing, visualization, supervision, project administration, and funding acquisition.

FUNDING

This work has been supported by a grant from "La Fondation pour la Recherche Médicale" (DIC202161236446 WAIMEA, BM PI, AL, LL, and AT) and by a grant overseen by the French National Research Agency (ANR) as part of the "Investissements d'Avenir" Program ANR-18-RHUS-002. This work was supported by a grant from the French National Research Agency (ANR-16-LCV2-0006-01, LABCOM Ginesislab MJ and PB PIs, ZH, VN, VV). The i-Share study has received funding from the ANR (Agence Nationale de la Recherche) via the 'Investissements d'Avenir' programme (grant ANR-10-COHO-05, CT PI). The MRi-Share cohort was supported by grant ANR-10-LABX-57 (BM PI) and supplementary funding was received from the Conseil Régional of Nouvelle Aquitaine (ref. 4370420). The work was also supported by the "France Investissements d'Avenir" program (ANR–10–IDEX-03-0, CT PI) and (ANR-18-RHUS-0002, S. Debette).

ACKNOWLEDGMENTS

We thank Pierre-Louis Bazin for providing the medical image processing, analysis, and visualization software. University of Bordeaux and CNRS provided infrastructural support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fninf.2021.641600/full#supplementary-material

"Cluster_amp_size.xls" file: TPR, PPV, and Dice-Sorensen metrics for PVSs located in the DWM and the BG are provided at 9 amplitude thresholds of the prediction map threshold (PredMap-Thr, 0.1 to 0.9 in steps of 0.1) and cluster size thresholds f0 (no filtering) to f15 (PVS with volume greater than 15 mm^3) in steps of 1 mm^3.

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (2016). "TensorFlow: a system for large-scale machine learning," in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16)* (Savannah, GA), 265–283.

Adams, H. H., Cavalieri, M., Verhaaren, B. F., Bos, D., van der Lugt, A., Enzinger, C., et al. (2013). Rating method for dilated virchow-robin spaces on magnetic resonance imaging. *Stroke* 44, 1732–1735. doi: 10.1161/STROKEAHA.111.000620

Ballerini, L., Lovreglio, R., Valdes Hernandez, M. D. C., Ramirez, J., MacIntosh, B. J., Black, S. E., et al. (2018). Perivascular spaces segmentation in brain MRI using optimal 3D filtering. *Sci. Rep.* 8:2132. doi: 10.1038/s41598-018-19781-5

Bergstra, J., Yamins, D., and Cox, D. D. (2013). "Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures," in *Proceedings of the 30th International Conference on International Conference on Machine Learning* (Atlanta), 115–123.

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., et al. (2009). KNIME - the konstanz information miner: version 2.0 and beyond. *SIGKDD Explor. Newsl.* 11, 26–31. doi: 10.1145/1656274.1656280

Boespflug, E. L., Schwartz, D. L., Lahna, D., Pollock, J., Iliff, J. J., Kaye, J. A., et al. (2018). MR Imaging-based multimodal autoidentification of perivascular spaces (mMAPS): automated morphologic segmentation of enlarged perivascular spaces at clinical field strength. *Radiology* 286, 632–642. doi: 10.1148/radiol.2017170205

Ciresan, D. C., Meier, U., Masci, J., Gambardella, L. M., and Schmidhuber, J. (2011). "Flexible, high performance convolutional neural networks for image classification," in *Proceedings of the 22nd International Joint Conference on Artificial Intelligence* (Barcelona), 1237–1242.

Debette, S., Schilling, S., Duperron, M. G., Larsson, S. C., and Markus, H. S. (2019). Clinical Significance of magnetic resonance imaging markers of vascular

brain injury: a systematic review and meta-analysis. *JAMA Neurol.* 76, 81–94. doi: 10.1001/jamaneurol.2018.3122

Ding, J., Sigurosson, S., Jonsson, P. V., Eiriksdottir, G., Charidimou, A., Lopez, O., et al. (2017). Large perivascular spaces visible on magnetic resonance imaging, cerebral small vessel disease progression, and risk of dementia. The Age, Gene/Environment Susceptibility-Reykjavik Study. *JAMA Neurol.* 74, 1105–1112. doi: 10.1001/jamaneurol.2017.1397

Doubal, F. N., MacLullich, A. M., Ferguson, K. J., Dennis, M. S., and Wardlaw, J. M. (2010). Enlarged perivascular spaces on MRI are a feature of cerebral small vessel disease. *Stroke* 41, 450–454. doi: 10.1161/STROKEAHA.109.564914

Dubost, F., Adams, H., Yilmaz, P., Bortsova, G., Tulder, G. V., Ikram, M. A., et al. (2020). Weakly supervised object detection with 2D and 3D regression neural networks. *Med. Image Anal.* 65:101767. doi: 10.1016/j.media.2020.101767

Dubost, F., Yilmaz, P., Adams, H., Bortsova, G., Ikram, M. A., Niessen, W., et al. (2019). Enlarged perivascular spaces in brain MRI: automated quantification in four regions. *Neuroimage* 185, 534–544. doi: 10.1016/j.neuroimage.2018.10.026

Duperron, M. G., Tzourio, C., Sargurupremraj, M., Mazoyer, B., Soumare, A., Schilling, S., et al. (2018). Burden of dilated perivascular spaces, an emerging marker of cerebral small vessel disease, is highly heritable. *Stroke* 49, 282–287. doi: 10.1161/STROKEAHA.117.019309

Francis, F., Ballerini, L., and Wardlaw, J. M. (2019). Perivascular spaces and their associations with risk factors, clinical disorders and neuroimaging features: a systematic review and meta-analysis. *Int. J. Stroke* 14, 359–371. doi: 10.1177/1747493019830321

Glorot, X., and Bengio, Y. (2010). "Understanding the difficulty of training deep forward neural networks," in *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics* (Sardinia).

Gonzalez-Castro, V., Valdes Hernandez, M. D. C., Chappell, F. M., Armitage, P. A., Makin, S., and Wardlaw, J. M. (2017). Reliability of an automatic classifier for brain enlarged perivascular spaces burden and comparison with human performance. *Clin Sci (Lond)* 131, 1465–1481. doi: 10.1042/CS20170051

Huttenlocher, D. P., Klanderman, G. A., and Rucklidge, W. J. (1993). Comparing images using the Hausdorff distance. *IEEE Trans. Pattern Anal. Mach. Intell.* 15, 850–863. doi: 10.1109/34.232073

Ioffe, S., and Szegedy, C. (2015). "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning* (Lille).

Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., and Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* 18, 203–211. doi: 10.1038/s41592-020-01008-z

Jung, E., Chikontwe, P., Zong, X., Lin, W., Shen, D., and Park, S. H. (2019). Enhancement of perivascular spaces using densely connected deep convolutional neural network. *IEEE Access* 7, 18382–18391. doi: 10.1109/ACCESS.2019.2896911

Kingma, D. P., and Welling, M. (2013). "Auto-encoding variational bayes," in *Proceedings of the 2nd International Conference on Learning Representations (ICLR)* (Banff, AB).

Lawrence, S., and Kubie, M. D. (1927). A study of the perivascular tissues of the central nervous system with the supravital technique. *J. Exp. Med.* 46, 615–626.

Lian, C., Zhang, J., Liu, M., Zong, X., Hung, S. C., Lin, W., et al. (2018). Multi-channel multi-scale fully convolutional network for 3D perivascular spaces segmentation in 7T MR images. *Med. Image Anal.* 46, 106–117. doi: 10.1016/j.media.2018.02.009

Lutnick, B., Ginley, B., Govind, D., McGarry, S. D., LaViolette, P. S., Yacoub, R., et al. (2019). An integrated iterative annotation technique for easing neural network training in medical image analysis. *Nat. Mach. Intell.* 1, 112–119. doi: 10.1038/s42256-019-0018-3

Maclullich, A. M., Wardlaw, J. M., Ferguson, K. J., Starr, J. M., Seckl, J. R., and Deary, I. J. (2004). Enlarged perivascular spaces are associated with cognitive function in healthy elderly men. *J. Neurol. Neurosurg. Psychiatry* 75, 1519–1523. doi: 10.1136/jnnp.2003.030858

Marin-Padilla, M., and Knopman, D. S. (2011). Developmental aspects of the intracerebral microvasculature and perivascular spaces: insights into brain response to late-life diseases. *J. Neuropathol. Exp. Neurol.* 70, 1060–1069. doi: 10.1097/NEN.0b013e31823ac627

Mazoyer, B., Mellet, E., Perchey, G., Zago, L., Crivello, F., Jobard, G., et al. (2016). BIL&GIN: a neuroimaging, cognitive, behavioral, and genetic database for the study of human brain lateralization. *Neuroimage* 124, 1225–1231. doi: 10.1016/j.neuroimage.2015.02.071

Milletari, F., Nassir, N., and Seyed-Ahmad, S. A. (2016). "V-net: fully convolutional neural networks for volumetricmedical image segmentation," in *Proceedings of the 4th International Conference on 3D Vision* (Stanford, CA).

Passiak, B. S., Liu, D., Kresge, H. A., Cambronero, F. E., Pechman, K. R., Osborn, K. E., et al. (2019). Perivascular spaces contribute to cognition beyond other small vessel disease markers. *Neurology* 92, e1309–e1321. doi: 10.1212/WNL.0000000000007124

Patek, P. (1941). The perivascular spaces of the mammalian brain. *Anat. Rec.* 38, 1–24.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830.

Potter, G. M., Chappell, F. M., Morris, Z., and Wardlaw, J. M. (2015). Cerebral perivascular spaces visible on magnetic resonance imaging: development of a qualitative rating scale and its observer reliability. *Cerebrovasc. Dis.* 39, 224–231. doi: 10.1159/000375153

Ramirez, J., Berezuk, C., McNeely, A. A., Gao, F., McLaurin, J., and Black, S. E. (2016). Imaging the perivascular space as a potential biomarker of neurovascular and neurodegenerative diseases. *Cell Mol. Neurobiol.* 36, 289–299. doi: 10.1007/s10571-016-0343-6

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Med. Image Comput. Comp. Assist. Interv.* 9351, 234–241.

Schmidt, R., Fazekas, F., Kapeller, P., Schmidt, H., and Hartung, H. P. (1999). MRI white matter hyperintensities: three-year follow-up of the austrian stroke prevention study. *Neurology* 53, 132–139. doi: 10.1212/wnl.53.1.132

Schwartz, D. L., Boespflug, E. L., Lahna, D. L., Pollock, J., Roese, N. E., and Silbert, L. C. (2019). Autoidentification of perivascular spaces in white matter using clinical field strength T1 and FLAIR MR imaging. *Neuroimage* 202:116126. doi: 10.1016/j.neuroimage.2019.116126

Sepehrband, F., Barisano, G., Sheikh-Bahaei, N., Cabeen, R. P., Choupan, J., Law, M., et al. (2019). Image processing approaches to enhance perivascular space visibility and quantification using MRI. *Sci. Rep.* 9:12351. doi: 10.1038/s41598-019-48910-x

Seshadri, S., and Wolf, P. A. (2007). Lifetime risk of stroke and dementia: current concepts, and estimates from the Framingham Study. *Lancet Neurol.* 6, 1106–1114. doi: 10.1016/S1474-4422(07)70291-0

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15, 1929–1958.

Sudre, C. H., Gomez-Anson, B., Silvia, I., Lane, C. D., Jimenez, D., Haider, L., et al. (2019). 3D multirater RCNN for multimodal multiclass detection and characterisation of extremely small objects. *Procc. Mach. Learn. Res.* 102, 447–456.

Tsuchida, A., Laurent, A., Crivello, F., Petit, L., Joliot, M., Pepe, A., et al. (2020). The MRi-Share database: brain imaging in a cross-sectional cohort of 1,870 university students. *bioRxiv*[Preprint] doi: 10.1101/2020.06.17.154666

Wang, X., Valdes Hernandez, Mdel, C., Doubal, F., Chappell, F. M., Piper, R. J., et al. (2016). Development and initial evaluation of a semi-automatic approach to assess perivascular spaces on conventional magnetic resonance images. *J. Neurosci. Methods* 257, 34–44. doi: 10.1016/j.jneumeth.2015.09.010

Wardlaw, J. M., Smith, E. E., Biessels, G. J., Cordonnier, C., Fazekas, F., Frayne, R., et al. (2013). Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *Lancet Neurol.* 12, 822–838. doi: 10.1016/S1474-4422(13)70124-8

Yakushiji, Y., Charidimou, A., Hara, M., Noguchi, T., Nishihara, M., Eriguchi, M., et al. (2014). Topography and associations of perivascular spaces in healthy adults: the Kashima scan study. *Neurology* 83, 2116–2123. doi: 10.1212/WNL.0000000000001054

Zhang, J., Gao, Y., Park, S. H., Zong, X., Lin, W., and Shen, D. (2017). Structured learning for 3-D perivascular space segmentation using vascular features. *IEEE Trans. Biomed. Eng.* 64, 2803–2812. doi: 10.1109/TBME.2016.2638918

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., and Liang, J. (2020). U-net++: redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med. Imaging* 39, 1856–1867. doi: 10.1109/TMI.2019.2959609

Zhou, Z., Sodha, V., Pang, J., Gotway, M. B., and Liang, J. (2021). Models genesis. *Med. Image Anal.* 67:101840. doi: 10.1016/j.media.2020.101840

Zhu, Y. C., Dufouil, C., Mazoyer, B., Soumare, A., Ricolfi, F., Tzourio, C., et al. (2011). Frequency and location of dilated virchow-robin spaces in elderly people: a population-based 3D MR imaging study. *AJNR Am. J. Neuroradiol.* 32, 709–713. doi: 10.3174/ajnr.A2366

Zhu, Y. C., Dufouil, C., Soumare, A., Mazoyer, B., Chabriat, H., and Tzourio, C. (2010a). High degree of dilated virchow-robin spaces on MRI is associated with increased risk of dementia. *J. Alzheimers Dis* 22, 663–672. doi: 10.3233/JAD-2010-100378

Zhu, Y. C., Tzourio, C., Soumare, A., Mazoyer, B., Dufouil, C., and Chabriat, H. (2010b). Severity of dilated virchow-robin spaces is associated with age, blood pressure, and MRI markers of small vessel disease: a population-based study. *Stroke* 41, 2483–2490. doi: 10.1161/STROKEAHA.110.591586

Zong, X., Park, S. H., Shen, D., and Lin, S. (2016). Visualization of perivascular spaces in the human brain at 7 T: sequence optimization and morphology characterization. *Neuroimage* 125, 895–902. doi: 1016/j.neuroimage.2015.10.078

Conflict of Interest: PB was employed by Fealinx.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.