Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Distance-to-Ventricle Guided Deep Learning for Periventricular and Deep Multiple Sclerosis Lesion Segmentation

Preethi Ramesh
Vellore Institute of Technology University
Nitharsana C
Vellore Institute of Technology University
Jayanandhitha P
Vellore Institute of Technology University
Appalaraju Muralidhar
muralidhar.a@vit.ac.in
Vellore Institute of Technology University

Article

Keywords: Multiple sclerosis, lesion segmentation, white matter hyperintensities, periventricular lesions, distance transform, anatomical prior, U-Net, MS3SEG

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11438101/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Distance-to-Ventricle Guided Deep Learning for Periventricular and Deep Multiple Sclerosis Lesion Segmentation

Preethi Ramesh[1], Nitharsana C[1], Jayanandhitha P[1] & Appalaraju Muralidhar[1]

[1]School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India

*Corresponding author: [Appalaraju Muralidhar[1], muralidhar.a@vit.ac.in]

Abstract

Periventricular lesions are central to the diagnosis of multiple sclerosis (MS), but automatic segmentation is the least dependable. Whether a lesion is periventricular (PV) or deep is said by its three-dimensional distance to the ventricles. We show that on a thick-slice clinical MRI this decision depends on the ventricles in adjacent slices and on the slice spacing, so that an in-plane distance would show 30% of PV lesion voxels as deep. We therefore give the geometry explicitly. A 3D distance-to-ventricle map, computed in millimetres on the full volume, is added in as one of the input channel to a 2.5D U-Net that segments ventricles, normal white matter hyperintensities (WMH), PV lesions and deep lesions. On the 100-patient MS3SEG dataset, against an identically trained baseline, the distance model raised held-out Dice for PV lesions from 0.731 to 0.758 and for deep lesions from 0.702 to 0.738 (Holm-corrected $p < 0.001$) and reduced PV/deep label confusions by 71%, while the pooled lesion Dice was unchanged. Ablation and counterfactual analyses shows that the prior improves the structural assignment of lesions rather than their detection itself.

Keywords: Multiple sclerosis, lesion segmentation, white matter hyperintensities, periventricular lesions, distance transform, anatomical prior, U-Net, MS3SEG

Introduction

Multiple sclerosis (MS) is a chronic immune-mediated disease of the central nervous system that produces focal demyelinating lesions in the white matter [1]. Magnetic resonance imaging (MRI) is the main tool for detecting these type of lesions. Diagnostic criteria use their location as well as their number and volume. Periventricular lesions are one of the defining topographies for dissemination in space [2,3]. Manual finding is slow and observer dependent [4], while automating could make it faster.

Lesions are classified as periventricular (PV) or deep by their three-dimensional distance to the ventricles, conventionally 10 mm [5]. Clinical MRI often provides limited through plane context, which makes it difficult for slice-based networks to determine a lesion's 3D proximity to the ventricles. Models that completely rely on image intensities must therefore use this structural relationship indirectly. In the MS3SEG data used here, the ventricles appear in only 56% of slices, and 23% of lesion voxels lie in slices that contain no ventricle at all ('Locality of the 10 mm rule on thick-slice MRI' in Methods). The MS3SEG dataset [6] marks the ventricles, normal white matter

hyperintensities (WMH) and abnormal WMH separately, so we can understand the relationship exactly and can be given to the network as an input.

Our contributions are:

- A geometric analysis of the 10 mm rule on thick-slice MRI, showing that the PV/deep class depends on the adjacent slices and the slice spacing, and a 3D distance-to-ventricle map that gives this geometry to a 2.5D U-Net as one additional input channel.
- A region-resolved formulation that segments PV and deep lesions as separate classes, so that location errors hidden by a pooled lesion score become measurable.
- A controlled comparison in which the only difference between the two models is the distance channel, evaluated once on a locked test set with paired, multiplicity-corrected statistics.
- An analysis of how the prior is used, based on confusion counts, distance bands, channel ablation, counterfactual inputs and Integrated Gradients.

Through-plane context in MS lesion segmentation

Different segmentation techniques for MS lesions vary with respect to the amount of context in three dimensions they utilize. On the one side of the spectrum, Papadopoulos et al. [7] demonstrated the suitability of simple 2D U-Net structure applied to FLAIR images, providing evidence for the Dice value of 0.729 in the analysis involving 30 patients in the clinical study. At the same time, the authors' approach to data splitting has not been considered stratified by the subject. In their study, Rondinella et al. [8] accounted for the context of the adjacent slices by combining squeeze attention and convolutional LSTM across three adjacent slices. However, as a result, the researchers managed to achieve the Dice index of only 0.46 on data from a different scanner (1.5 T vs 3 T), while their result on the ISBI-2015 subset reached 0.84. Overall, it is worth noting that the methods making the most extensive use of three-dimensional context, as well as those based on transformers, demonstrated good results, particularly LST-AI [9], which consists of three 3D U-Nets and showed the best periventricular-region Dice among the tools it was compared with: 0.78. Moreover, CNN-transformers demonstrated high coefficients of above 0.90 [10,11]. The figures mentioned in this context derive from various datasets and so they cannot be compared directly. Also, larger models involve their own costs, like SwinUNETR-v2 [11] which requires 349 GFLOPs. Here, acquisition type also plays a prominent role. MindGlide [12] was less sensitive on 2D acquisitions than on 3D ones, which is also important in routine clinical protocols with 6–8 mm thick slices typical for MS3SEG. Some works deal with robustness against distribution shift [13,12], some focus on efficiency [14], and some analyze confluent lesions [15]. In all papers, image intensity is the single input and is used to detect any relation between lesions and ventricles.

Separating normal from pathological hyperintensities

Our research is based on the work presented in the paper by Bashiri Bawil et al. [16]. While instead of varying the topology of the neural network, they worked on altering the label space — the normal periventricular changes receive its label in their own category rather than being included into the background. They worked with four different architectures and 2,750 FLAIR slices from 115 patients. With their three-class classification, the pathological WMH Dice for the U-Net increased from 0.497 while working on the binary classification to 0.768 and the authors attributed most of the improvement to achievement of fewer number of false positive results. Later, they proposed a joint model which included ventricles, normal WMH and abnormal WMH [17] and released the MS3SEG dataset [6] that was utilized in this work. They also mentioned the asymmetry of data coming from the two centers (87% vs 13%) and the unstable performance of the TransUNet as limitations, and named regional spatial analysis as future work. Our work is the one which takes up that regional analysis. Jeong et al. [18] approached the problem of periventricular changes from an alternative view point as they combined WMH segmentation with Fazekas grading [19] which would be based on the periventricular/deep distinction. Their cohort was not specific to MS, and they argued that there was no consensus about the cutoff of 10 mm. Analyzing all studies together, we can conclude that in every case the network is reasoning about the lesion position based only on the image features.

Explicit location information

Table 1 is a summary of the researches that provide a clear location data. Ghafoorian et al. [20] incorporated eight location characteristics in patch-based CNN for WMH in small-vessel disease:

- the positions of the voxels in the MNI atlas space;
- the distances from the left and right ventricles, the cortex, and the midsagittal line;
- the prior probability of WMH in that certain place.

The best way to include those characteristics is at the first fully-connected layer. This confirms that location information is crucial for WMH recognition. Still, the coordinates and the prior depend on non-linear registration with atlas. Dealing with distances implies making measurements inside the slice only, which is not good for thick-slice MRI as it may lead to inaccurate location across 10 mm border (‘Locality of the 10 mm rule on thick-slice MRI’ in Methods).

According to Liang et al. [21], the subject are mapped with the help of the regional atlas through the non-linear registration process, which is important for defining region-wise probabilities of WMH occurrence and joining them with MRI intensity data as inputs of U-Net. Regions were allocated to lesions only subsequent to segmentation. Both approaches do not produce PV and deep lesions as distinct classes. Griffanti et al. [5] and Chen et al. [22] do designate PV and deep areas by distance (Chen et al. determine the distance to the ventricles and the cortex). However, their criteria are applied

to an already existing segmentation, hence cannot enhance the segmentation itself. Distance maps are also employed in boundary losses [23], but in this case they are calculated from the ground truth and applied only in the training process. It is known that location is helpful and distance determines the clinical PV/deep classes. However, we do not know of any work in which a full 3D distance map, calculated without registration, is put into the network for it to analyze PV and deep lesions. This is the gap filled in by this research.

Table 1. Studies using explicit location information

Study	Location information	How it is used	PV/deep
Ghafoorian [20]	atlas coordinates (registration), in-plane distances	features in classifier layer	no
Liang [21]	registered WM atlas	separate encoder + attention	no
Griffanti, Chen [5,22]	distance to ventricles (and cortex)	rule after segmentation	rule
Kervadec [23]	distance to ground-truth boundary	training loss only	no
Ours	3D distance to ventricles, mm, native space	dense input channel	learned

Results

Backbone selection: comparison of base models

Before the distance prior was added, two candidate backbones were compared on the intensity-only, nine-channel 2.5D task: the 2D U-Net [24] and a 2D Swin UNETR [25]. Both were trained with the protocol of 'Implementation and model selection', with the same folds, loss, augmentation, model-selection rule and test-time augmentation, so that the differences reflect the architecture alone. Supplementary Table S1 reports pooled lesion Dice in five-fold CV and on the locked test set.

The U-Net exceeded Swin UNETR by 0.065 pooled lesion Dice in CV and by 0.067 on the test set. The U-Net was selected as the backbone of the Distance U-Net for five reasons.

Accuracy: The U-Net had the best pooled lesion DSC score in CV and this same score was true on the held-out test set.

Data size and lesion scale: The development set contains 70 patients (1,396 slices), and each cross-validation fold trains on only 56 of them. Transformer encoders are not as locality and translation equivariant as convolutions, and need more data or pre-training to reach their full potential. The number of voxels with MS lesions is 0.17% and lesions are frequently narrow in plane (few voxels). While a convolutional encoder with full-resolution skip connections retains this fine detail, Swin UNETR creates its transformer tokens by processing 2 × 2 patches and only uses a single convolutional encoder block at full resolution. This is in line with the reduced Swin UNETR rating.

Thick slices: With 18–20 slices of 6.0–8.1 mm, we selected a 2D backbone with 2.5D input. In 'Locality of the 10 mm rule on thick-slice MRI' in Methods, the current and adjacent slices are used to explain how they can be used to obtain the information for the 10 mm rule.

Controlled attribution: This study is to measure the effect of one input channel. The distance channel is only modified in the first convolution (288 additional weights), and there are no attention, deep-supervision or dense-skip components that can interact with the new input. The setup in the training phase was kept identical, thus eliminating the effect of adding the distance before training.

Efficiency and comparability: The U-Net consists of 7.76M parameters and is trained using batch size 8 on a 4 GB gpu in laptop. It is also the architecture used by Bashiri Bawil et al. in MS3SEG [16,17] which can be compared with their results. A gradient accumulation (micro-batches with 2 samples) was required for Swin UNETR and resulted in a time of 149 s per training epoch, compared to 37 s for the U-Net.

Table 2. Test-set results (30 patients). Dice: mean ± SD over patients. Δ: mean paired difference (Distance U-Net − baseline U-Net) with 95% bootstrap CI. *p*: Wilcoxon signed-rank test, Holm-corrected over six classes. n_+/n_-: patients improved / worsened. HD95: mean (median) in mm. Best value per row in bold.

	Dice ↑					HD95 (mm) ↓	
Class	Baseline U-Net	Distance U-Net	Δ [95% CI]	*p*	n_+/n_-	Baseline U-Net	Distance U-Net
PV lesion (≤10 mm)	0.731 ± 0.103	**0.758** ± 0.107	+0.027 [+0.016, +0.039]	<0.001	28/2	12.38 (7.28)	**11.00** (6.80)
Deep lesion (>10 mm)	0.702 ± 0.142	**0.738** ± 0.137	+0.036 [+0.022, +0.053]	<0.001	26/3	16.87 (12.78)	**14.65** (8.52)
Pooled lesion (PV + deep)	0.760 ± 0.127	**0.763** ± 0.122	+0.004 [−0.002, +0.010]	0.49	15/15	**10.40** (7.30)	10.72 (6.94)
Normal WMH	0.669 ± 0.062	**0.684** ± 0.058	+0.015 [+0.010, +0.021]	<0.001	23/7	6.38 (5.69)	**5.72** (4.24)
All WMH	0.769 ± 0.074	**0.776** ± 0.071	+0.007 [+0.004, +0.010]	<0.001	25/5	3.37 (2.83)	**3.28** (2.24)
Ventricle	0.894 ± 0.024	**0.942 ± 0.012**	+0.048 [+0.043, +0.054]	<0.001	30/0	1.12 (1.00)	**0.97** (1.00)

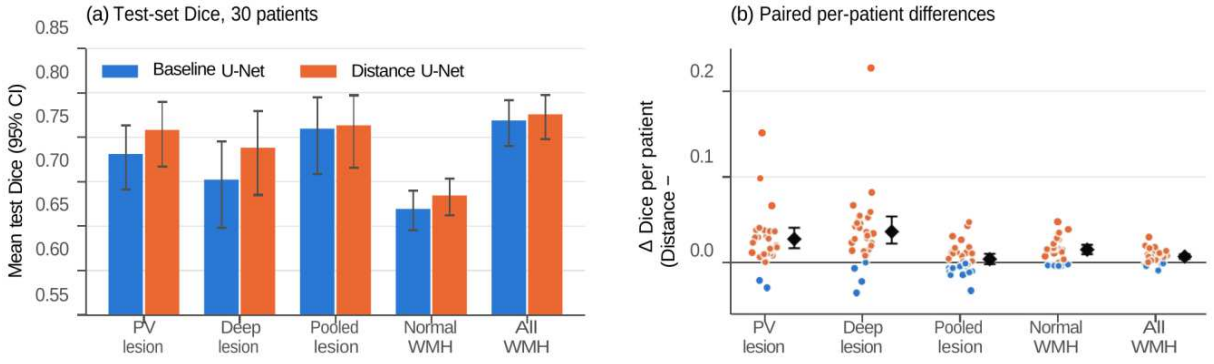


Figure 1. Test-set comparison of the baseline U-Net and the Distance U-Net (30 patients). (a) Mean Dice per class with 95% bootstrap CIs. (b) Per-patient Dice difference, Distance U-Net minus baseline U-Net; each dot is one patient (orange: improved, blue: worsened), and diamonds show the mean with 95% CI.

Overall comparison

Table 2 and Fig. 1 summarise the test results. The Dice scores of the Distance U-Net for PV and deep lesions were 0.758 and 0.738, significantly higher than those of the baseline U-Net (0.731 and 0.702; $p < 0.001$ for both). PV Dice improved in 28 of 30 patients and deep Dice in 26 of 30. Normal WMH (+0.015), all WMH (+0.007) and ventricle Dice (+0.048) also improved significantly ($p < 0.001$), whereas pooled lesion Dice did not change significantly (+0.004, $p = 0.49$). Median HD95 was lower for the Distance U-Net in all five WMH targets. Cross-validation showed the same pattern for the lesion classes (Supplementary Table S2): PV Dice was higher in all five folds and deep Dice in four of five. The effect size (matched-pairs rank-biserial r) was large for PV lesions (0.86), deep lesions (0.89) and normal WMH (0.84), and small for pooled lesions (0.15).

PV and deep lesions: location assignment

The PV/deep gains may be attributed to an increase in the number of lesions detected, or to the correct labelling of detected lesions. There are three results that support the second explanation.

Confusion counts. The number of lesion voxels assigned to the wrong location class decreased from 1609 to 463 (−71%; Fig. 2). PV voxels that were labelled deep decreased from 902 to 198, and deep voxels that were labelled PV decreased from 707 to 265. PV-to-deep errors decreased by 704 (80% of the net gain in PV true-positive voxels, 880), while deep-to-PV errors decreased by 442 (70% of the net gain in deep true-positive voxels, 628). Correcting a PV/deep misclassification increases the true positives of the correct class and reduces the false positives of the incorrect class. This is consistent with the simultaneous rise in PV sensitivity (0.777 to 0.819), deep sensitivity (0.735 to 0.773) and PV precision (0.717 to 0.735, $p = 0.014$). For example, as seen in Fig. 3, both models identify almost all the lesions of test patient 046, with the Distance U-Net assigning the deep lesions to the deep class.

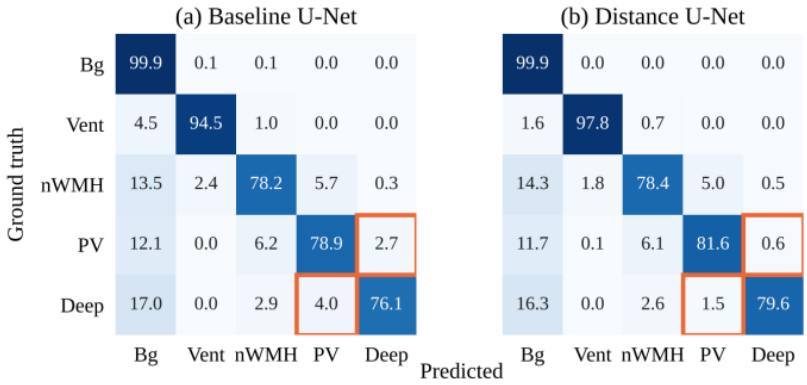


Figure 2. Test-set confusion (30 patients): percentage of each ground-truth class (rows) assigned to each predicted class. (a) Baseline U-Net; (b) Distance U-Net. Bg: background; Vent: ventricle; nWMH: normal WMH. Orange boxes: PV/deep swaps.

Distance bands. In Fig. 4a, the accuracy of PV/deep labels is displayed as a function of distance from the ventricles. The models were almost always right below 5 mm and above 20 mm. The models

differed near the 10 mm boundary. For the 5–10 mm band, the accuracy for the baseline U-Net was 0.943 and for the Distance U-Net was 0.989. In the 10–20 mm band, it was 0.931 and 0.974 (both $p < 0.001$). The majority of test lesion voxels (66%) are within these two bands (Table 3), suggesting that most lesion voxels are within 20 mm of the ventricles.

Adjacent slices. Baseline PV Dice decreased by 0.132 and deep Dice decreased by 0.153 when slices $z \pm 1$ were removed (Supplementary Table S3). The percentage of baseline lesion labels that matched the 10 mm geometry dropped from 0.956 to 0.824. The Distance U-Net changed little (PV −0.003, deep −0.014; consistency 0.991 to 0.989). As predicted by equation (4), the consistency of the baseline without neighbouring slices is close to the 0.805 achieved by the in-plane distance rule on the test lesions (Table 3). This is consistent with the view that the baseline uses neighbouring slices as an indirect indication of ventricular proximity, while the distance channel provides this information directly.

Counterfactual test. Replacing the true distance map with a donor map changed the 10 mm assignment for nearly 19,250 brain voxels per patient. The true map yielded 99.1% agreement between the predicted lesion labels and the true geometry. With the donor map, 97.1% of predicted lesion labels agreed with the donor geometry (Fig. 4b). Of the voxels that were predicted as lesions in both runs, 87.3% changed class in the direction of the donor map. The results of this experiment strongly indicate that the supplied distance map affects PV/deep location predictions.

Deep versus PV lesions

Deep lesions gained more than PV lesions (Dice +0.036 vs. +0.027). Three factors are consistent with this. The baseline U-Net mislabelled a larger share of deep voxels (3.98%) than PV voxels (2.74%). Deep lesions are also the smaller class, with 17,764 test voxels compared with 32,951 PV voxels. A correction in this smaller class can have a greater effect on its Dice score. Finally, a deep lesion labelled PV forms a distant false-positive component, which inflates HD95. Median deep HD95 fell from 12.78 to 8.52 mm. The mean HD95 reductions were not significant after correction ($p = 0.16$ for PV, 0.15 for deep), as a few patients with isolated distant errors dominate the mean.

Pooled lesion detection

Several results indicate that lesion detection depends on image contrast rather than on the prior. Without FLAIR, neither model detected any lesions. In the counterfactual test, only 0.5% of affected voxels changed between lesion and non-lesion, although many changed location class. Lesion sensitivity per distance band was similar for both models, for example 0.840 and 0.852 at 5–10 mm. IG assigned the largest attribution share to the centre FLAIR slice in both models (0.34–0.39; Fig. 4c), and DIST received 0.061 for PV and 0.114 for deep predictions. At the voxel level, missed lesion voxels fell by 3.8% and Counting every non-lesion voxel predicted as lesion (background, ventricle and normal WMH) we found the rise to be +2.0%, so pooled Dice was unchanged.

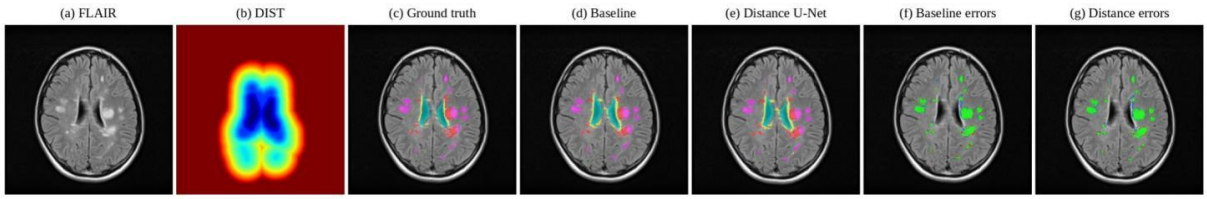


Figure 3. Test patient 046, axial slice 12. (a) FLAIR; (b) normalised distance channel $\tilde{d}$; (c) ground truth; (d) baseline U-Net and (e) Distance U-Net predictions (cyan: ventricle, yellow: normal WMH, red: PV lesion, magenta: deep lesion); (f, g) lesion detection maps of the two models (green: true positive, red: false positive, blue: false negative).

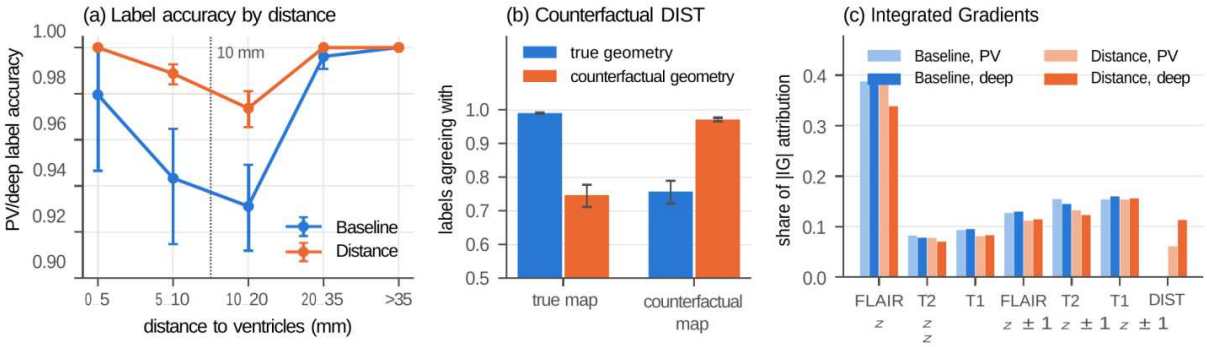


Figure 4. Analysis of the two models on the test set. (a) PV/deep label accuracy of correctly detected lesion voxels in each distance band (mean over patients with 95% CI); the dotted line marks the 10 mm boundary. (b) Share of predicted lesion labels that agree with the true or the donor (counterfactual) ventricle geometry, when the Distance U-Net receives the true or the donor DIST map. (c) Share of absolute Integrated Gradients attribution per input group; z: centre slice, $z \pm 1$: adjacent slices.

Normal WMH

The majority of normal WMH in MS3SEG is periventricular caps and rims: 90.4% of voxels in test normal-WMH are within 5 mm of the ventricles. The improvement came from precision (0.592 to 0.612, $p < 0.001$); sensitivity did not change significantly (0.791 to 0.796, $p = 0.47$). The predicted voxels of background WMH decreased by 7.2% and the predicted voxels of ventricle WMH decreased by 33%. The most significant difference was at 5–10 mm, just beyond the normal cap. There, precision increased from 0.396 to 0.488 and the volume of false-positives decreased from 0.38 to 0.24 mL per patient. The donor map reduced the normal-WMH Dice from 0.684 to 0.559. The results indicate that the Distance U-Net is able to reject the prediction of normal-WMHs in implausible poses, as seen in the distance channel.

Stability and robustness

The standard deviation of CV Dice decreased from 0.024 to 0.008 for PV lesions and from 0.026 to 0.015 for deep lesions (Supplementary Table S2). Pooled lesion Dice dropped to 0.239 and 0.363 when T2 or T1 was removed, while only to 0.490 and 0.617 when the Distance U-Net was applied (Supplementary Table S3).

Discussion

The results in the Results indicate that the influence of the prior is uneven across the two components of the task. It helps in location assignment, but no statistically significant improvement in pooled lesion Dice can be seen. This leads to two conclusions. First, in terms of evaluation, using a pooled lesion Dice metric will not show clinically important mistakes, the method will mark the same lesions but assign them the wrong location. The region-wise metrics should thus be used in case of methods aimed at the periventricular region. In terms of model development, the two components require different information. The detection step is mainly based on the tissue contrast, so the way to improve it is either to have better image features or more data. The location component, on the other hand, is a deterministic function of the ventricle mask and voxel spacing, as defined in equation (4).

The results address the three issues that Bashiri Bawil et al. left open (the Introduction). Boundary errors fell most in the 5–20 mm range. The distance prior made the adjacent slices nearly redundant for location, which indicates that it supplies the inter-slice information their 2D models lacked. Region-wise evaluation exposed an error that the pooled score concealed. The normal-WMH result complements their finding: they suppressed false positives by adding a class, and the position cue suppresses them further. Absolute values are not directly comparable because the cohorts and splits differ. Our baseline achieved a pooled lesion Dice of 0.760 and normal-WMH Dice of 0.669, compared with 0.825 and 0.677, respectively, in their joint model [17]. Compared with earlier location priors (Table 1), the main difference is that the distance is measured in 3D in the patient's own space and is used densely by the network. The 3D measurement matters: on MS3SEG, the in-plane distances used by Ghafoorian et al. [20] would mislabel 19% of lesion voxels (Table 3).

MS3SEG, to our knowledge, only public dataset with separate normal and abnormal WMH masks, so we used 100 patients. Small test set limits the detection of small differences. The data come from a single 1.5 T scanner with thick slices, so generalisation to other scanners and to 3D acquisitions is untested. The distance maps were computed from the ventricle masks supplied with MS3SEG, than from automatically segmented ventricles. Limited local hardware ruled out 3D architectures and higher epoch rates. Finally, evaluation was voxel-based rather than lesion-wise.

A 3D distance-to-ventricle map, added as one input channel to a 2.5D U-Net, improved the segmentation of PV and deep MS lesions on MS3SEG primarily by correcting their location assignment, while pooled lesion Dice showed no statistically significant improvement. The gain follows from the geometry of the 10 mm rule, which on thick-slice MRI depends on the adjacent slices and the slice spacing. Future work will derive the prior from automatically segmented ventricles, add lesion-wise metrics and multiple seeds, and validate the approach on multi-centre data.

Methods

Fig. 5 shows the framework. The proposed Distance U-Net and the baseline U-Net are identical except for the distance channel.

Problem formulation

Each patient has registered T1, T2 and FLAIR volumes and masks of the ventricles V, normal WMH and abnormal WMH. Every voxel is assigned to one of five classes: background, ventricle, normal WMH, PV lesion and deep lesion. Following the 10 mm convention [5], an abnormal-WMH voxel x is a PV lesion if $d(x) \leq 10$ mm and a deep lesion if not. Lesion labels take precedence over normal WMH and ventricle labels. Only seven abnormal-WMH voxels in the cohort overlap the ventricle mask. Normal WMH were held as a single class rather than separating them into PV and deep categories since deep normal WMH were not present in 33 of 100 patients, which would introduce an additional sparsely represented class. The classes are highly imbalanced: background accounts for 99.3% of voxels, PV lesions for 0.106% and deep lesions for 0.064%. This imbalance is addressed by the class-weighted loss described in 'Training objective' in Methods.

Distance-to-ventricle prior

For a patient with ventricle mask V on image grid Ω, the distance map [5,26] is

$$d(x) = \min_{v \in V} \| S(x - v) \|_2, \; x \in \Omega \tag{1}$$

where $S = \mathrm{diag}(s_x, s_y, s_z)$ is the voxel spacing, so that equation (1) gives the shortest physical distance from voxel x to any ventricle voxel v. Distances are therefore in millimetres and account for the 6.0–8.1 mm slice spacing; $d = 0$ inside the ventricles. We compute equation (1) with an exact Euclidean distance transform [26], as implemented in SciPy [27], on the full 3D volume before slicing. A per-slice transform would be wrong for slices that do not contain the ventricles, because it ignores ventricles a few millimetres above or below ('Locality of the 10 mm rule on thick-slice MRI' in Methods). The network receives the clipped, normalised map [28]

$$\hat{d}(x) = \frac{\min(d(x), D_{max})}{D_{max}} \in [0, 1] \tag{2}$$

in which the min(·) operation caps every distance at $D_{\max}$ and the division by $D_{\max}$ maps the result to a common [0, 1] scale for all patients. Clipping makes removes dependent of brain size and ensures that lesions are resolved. $D_{\max}$ is determined from the patient sample used in the initial development. The longest distance among their total of 128,433 voxels of abnormal-WMH is ranked in the 99th percentile at a distance of 32.1 mm. When rounded up to the nearest multiple of five mm this gives $D_{\max} = 35$ mm. Note that $D_{\max}$ scales only the input while the class labels use raw d at 10 mm. Because the map captures the geometry rather than intensity thus it is excluded from the process of intensity normalization and intensity augmentation. However, it is subjected to the same spatial transformations as the images.

The map has information that intensity values lack, the distance in 3D space defining the boundary between the PV and the deep structures, the surface of the ventricles where $\tilde{d} = 0$, and the structure that specifies normal position of the periventricular caps as well as providing the spatial references independent from the patient information. The only condition to obtain the map is to have available a mask of the ventricle in its proper place in space, it does not require having atlases or registration.

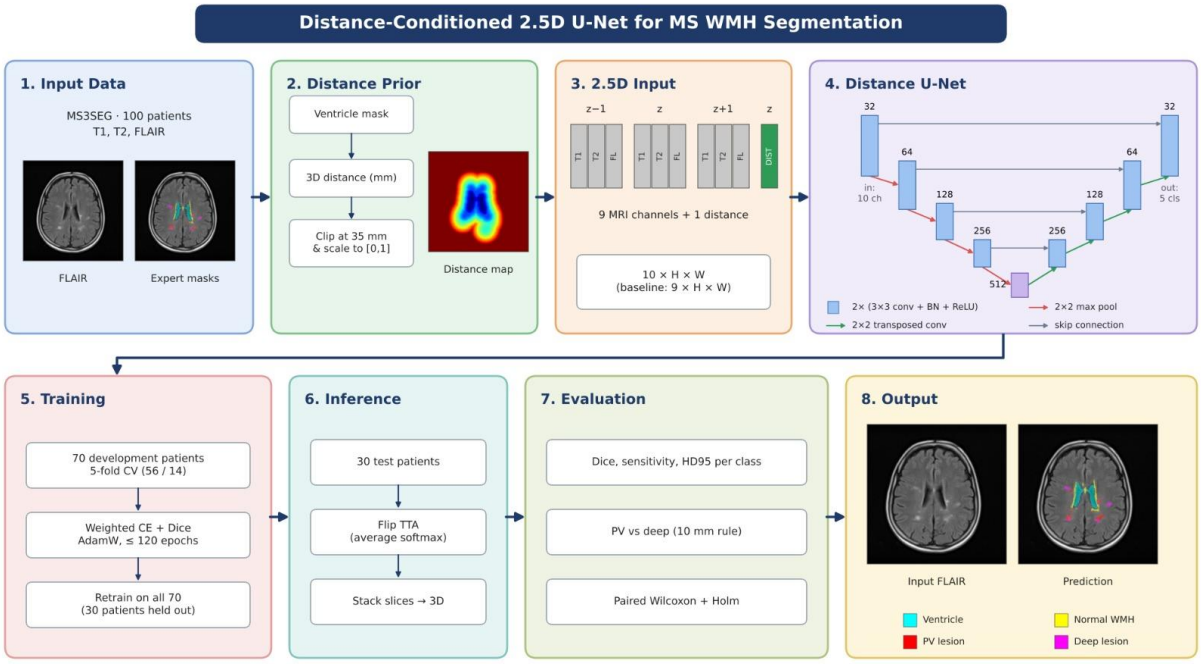


Figure 5. Architecture of the proposed Distance-Conditioned 2.5D U-Net for MS WMH segmentation. (1) Input data: T1, T2 and FLAIR with expert masks from MS3SEG. (2) Distance prior: 3D Euclidean distance (mm) from the ventricle mask, clipped at 35 mm and scaled to [0, 1]. (3) 2.5D input: T1, T2 and FLAIR of slices $z - 1$, z, $z + 1$ (9 channels) plus the distance channel at slice z (10 channels; the baseline uses 9). (4) Distance U-Net: 2D encoder–decoder with skip connections and a five-class output. (5) Training with five-fold cross-validation and retraining on the 70 development patients. (6) Inference on the 30 test patients with flip TTA. (7) Evaluation per class and PV/deep region. (8) Output: ventricle, normal WMH, PV lesion and deep lesion.

Locality of the 10 mm rule on thick-slice MRI

Let x lie in slice z, and let $d_k(x)$ be the in-plane distance in millimetres from x to the ventricle voxels of slice k, with $d_k = \infty$ if slice k contains none. Because S is diagonal, equation (1) separates into in-plane and through-plane parts [26]:

$$d(x) = \min_k \sqrt{d_k(x)^2 + ((z-k)s_z)^2} \tag{3}$$

Here the first term under the root is the in-plane distance within slice k, and $(z - k)s_z$ is the through-plane offset between slices z and k; the minimum over k selects the nearest ventricle voxel in any slice. Since $s_z \geq 6.0$ mm in MS3SEG, a ventricle voxel two or more slices away is always more than 10 mm from x. The PV/deep class therefore depends only on the current slice and its two neighbours [5,22]:

$$PV(x) \Leftrightarrow d_z(x) \leq 10 \;\; or \;\; \min_{k=z\pm1} d_k(x) \leq \sqrt{10^2 - s_z^2} \tag{4}$$

The first condition covers ventricle voxels in the current slice; the second covers the two neighbouring slices, where the through-plane offset s_z leaves an in-plane allowance of $\sqrt{(10^2 - s_z^2)}$. Equation (4) reproduced the stored PV/deep label of every lesion voxel in all 100 patients. The threshold in the adjacent slices, $\sqrt{(10^2 - s_z^2)}$, ranges from 8.0 mm at s_z = 6.0 mm to 5.9 mm at s_z = 8.1 mm. Table 3 quantifies three consequences on the ground truth.

The in-plane distance is not sufficient. The in-plane distance d_z is an upper bound for d. Employed alone, as in [20], it misclassifies 30.3% of PV lesion voxels as deep, and all patients are affected since 23.5% of lesion voxels belong to slices that lack ventricles and many other voxels are closest to a ventricle in the adjacent slice.

The 2.5D network should learn equation (4) implicitly. The 2.5D input includes all slices that equation (4) requires, so the rule is learnable in theory. However, the network should segment the ventricles in the adjacent slices, combine the in-plane and through-plane distances non-linearly, and know s_z, which is absent in the images; changing s_z within the range of values in the cohort affects the class of 5.4% of lesion voxels. The distance map supplies equation (3) in closed form with the spacing of the patient.

Sensitivity to the ventricle boundary. Moving the ventricle boundary by 1 mm in-plane (dilation or erosion with a 4-connected, cross-shaped in-plane structuring element) changes the class of 3.4% (dilation) or 6.1% (erosion) of lesion voxels, so the ventricle mask used in the prior probability must be accurate to about 1 mm.

Table 3. Geometry of the 10 mm rule in the MS3SEG ground truth (share of voxels)

Quantity	All 100	Test 30
Slices containing ventricle	55.9%	56.7%
Lesion voxels in slices without ventricle	23.5%	23.7%
PV voxels labelled deep by in-plane distance	30.3%	30.0%
Lesion voxels mislabelled by in-plane distance	18.9%	19.5%
Lesion voxels whose class depends on s_z	5.4%	5.0%
Lesion voxels changed by 1 mm dilation / erosion	3.4 / 6.1%	3.2 / 5.6%
Lesion voxels 5–20 mm from the ventricles	68.4%	66.4%
Lesion labels reproduced by (4)	100%	100%

2.5D input and network

Nine channels are produced from the input of slice z stacks, which are T1, T2, and FLAIR of slices $z - 1$, z, $z + 1$. The distance U-Net adds $\tilde{d}$ to slice z as the tenth channel. Both the models have been created using the same 2D U-Net [24] having 4 down-sampling blocks. In total, 7.76 million parameters are present in the system; the additional channel has added 288 more weights. 3D networks would have added no more information through planes at a much higher expense, as thick volumes mostly consist of either 18 or 20 slices. 2.5D input offers isolation and also helps get the distance map.

Training objective

The loss combines a class-weighted cross-entropy and a soft Dice loss [29,30]:

$$L = L_{CE}^{w} + 1 - \frac{1}{C-1}\sum_{c=1}^{C-1}\frac{2\sum_i p_{i,c}g_{i,c}+1}{\sum_i p_{i,c}+\sum_i g_{i,c}+1} \quad (5)$$

where $p_{i,c}$ and $g_{i,c}$ are predicted probabilities and one-hot labels, and $C = 5$. The first term, $L_{CE}^{w} = -(1/N)\,\Sigma_i\,\Sigma_c\, w_c\, g_{i,c} \log p_{i,c}$ over the N voxels of a batch, penalises wrong class probabilities voxel by voxel; the second term is one minus the mean soft Dice of the foreground classes, which rewards overlap regardless of class size, and the constant 1 keeps it defined when a class is absent. The Dice term excludes the background and is computed with the whole batch, which avoids unstable terms on lesion-free slices. The cross-entropy weights are $w_c \propto f_c^{-1/2}$, normalised to mean one, where f_c is the training frequency of class c. This counteracts class imbalance without over-weighting the rarest classes.

Dataset

MS3SEG [6] contains 100 MS patients scanned on a 1.5 T Toshiba scanner. Each patient has T1-weighted, T2-weighted and axial T2-FLAIR images with expert masks of the ventricles, normal WMH and abnormal WMH. The scans are clinical thick-slice acquisitions: 256 × 256 matrix, 0.898 mm pixels and 18–20 slices with 6.0–8.1 mm spacing. An audit confirmed that all images and masks matched the FLAIR image in shape, spacing, orientation and affine.

Pre-processing and data split

T1, T2 and the masks were resampled to a FLAIR-based grid with 1.0 mm in-plane spacing and the original slice spacing, using linear interpolation for images and nearest neighbour for masks. This gives volumes of 230 × 230 × (18–20) voxels. Each image was clipped to its 0.5th–99.5th percentiles and z-score normalised over non-zero voxels. The distance transform was applied to the resampled mask, since interpolating a distance map distorts it.

Patients were split once into 70 development and 30 test patients, stratified by tertiles of abnormal-WMH volume. The development set was divided into five folds of 56 training and 14 validation patients. All 1,985 slices were used. The test set was locked and evaluated once per model, after all design choices had been fixed on development data.

Implementation and model selection

The same parameters are used in both models and are given in Supplementary Table S4. The best epoch for cross-validation (CV) was determined using the average patient level 3D Dice of pooled lesions data for validation patients. Finally, retraining was done on 70 patients, and this final model was created by taking the median of the top 5 epochs, resulting in 86 for baseline U-Net and 109 for Distance U-Net. Inference: Flipped test-time augmentation (TTA) was performed as the average of the

softmax output from 4 flips of the inputs. The minimum lesion component size was tuned to match the CV predictions, and was finally set to 0 voxels for both models. The justification of using U-Net as the backbone was done by comparing it with Swin UNETR [25] with the same protocol as discussed in 'Backbone selection' in Results.

Evaluation and statistics

Metrics were computed per patient on 3D volumes. We report the Dice similarity coefficient, DSC = $2|P \cap G|/(|P| + |G|)$, where P and G are the predicted and ground-truth voxel sets of a class [31,32], and the 95th-percentile Hausdorff distance (HD95) in millimetres, the 95th percentile of the distances between the surfaces of P and G [33]. PV and deep lesions were also merged into a pooled lesion target, and all hyperintensities into an all-WMH target. Every class was present in all 30 test patients. Paired differences were tested with the two-sided Wilcoxon signed-rank test, with Holm correction over the six classes, applied separately to each metric; distance-band comparisons were Holm-corrected over the bands. All p-values reported in this paper are Holm-adjusted. We also report the mean paired difference with a 95% bootstrap confidence interval (CI; 5,000 resamples) and the number of patients improved or worsened.

To understand the results, we analysed the two final models on the test set without retraining. The analyses were: (i) PV/deep confusion counts; (ii) PV/deep label accuracy of correctly detected lesion voxels in five distance bands; (iii) channel ablation, setting one input group to its mean value; (iv) a counterfactual test in which each patient's distance map was replaced by one computed from the ventricles of the most similar development patient, so that the geometry changes while the MRI does not; and (v) Integrated Gradients (IG) [34] with 64 steps. For (iv), donor ventricles were aligned with an axis-aligned affine transform and accepted only if brain-overlap Dice was at least 0.85, scaling was 0.75–1.33 and the ventricle-volume ratio was 1/3–3; a valid donor was found for all 30 patients.

Ethics declarations

This study is a secondary analysis of the publicly available MS3SEG dataset [6] and did not involve the recruitment of participants or the collection of new data. The MRI data were acquired at Golgasht Medical Imaging Center, Tabriz, Iran, and the original data collection was approved by the ethics committee of Tabriz University of Medical Sciences (approval code IR.TBZMED.REC.1402.902). The dataset is distributed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence, which permits reuse with appropriate attribution. All methods were carried out in accordance with relevant guidelines and regulations.

Data availability

The MS3SEG dataset analysed in this study is publicly available at https://doi.org/10.6084/m9.figshare.30393475 [6].

References

1. Reich, D. S., Lucchinetti, C. F. & Calabresi, P. A. Multiple sclerosis. *N. Engl. J. Med.* **378**, 169–180 (2018).
2. Thompson, A. J. et al. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *Lancet Neurol.* **17**, 162–173 (2018).
3. Filippi, M. et al. MRI criteria for the diagnosis of multiple sclerosis: MAGNIMS consensus guidelines. *Lancet Neurol.* **15**, 292–303 (2016).
4. Commowick, O. et al. Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Sci. Rep.* **8**, 13650 (2018).
5. Griffanti, L. et al. Classification and characterization of periventricular and deep white matter hyperintensities on MRI: a study in older adults. *NeuroImage* **170**, 174–181 (2018).
6. Bashiri Bawil, M., Shamsi, M., Ghalehasadi, A., Jafargholkhanloo, A. F. & Shakeri Bavil, A. A multiple sclerosis MRI dataset with tri-mask annotations for lesion segmentation. *Sci. Data* (2026). Dataset: https://doi.org/10.6084/m9.figshare.30393475.
7. Papadopoulos, T. G. et al. White matter lesion segmentation for multiple sclerosis patients implementing deep learning. In *Proc. IEEE Eng. Med. Biol. Soc. Conf. (EMBC)* (2022).
8. Rondinella, A. et al. Boosting multiple sclerosis lesion segmentation through attention mechanism. *Comput. Biol. Med.* **161**, 107021 (2023).
9. Wiltgen, T. et al. LST-AI: a deep learning ensemble for accurate MS lesion segmentation. *NeuroImage Clin.* **42**, 103611 (2024).
10. Saadh, M. J. et al. A multi-center study of transformer-based CNNs for multiple sclerosis lesion segmentation on 3D FLAIR MRI. *Neuroradiology* **67**, 3637–3653 (2025).
11. Thabet, R. M., Shedeed, H. A., Al-Berry, M. & Khattab, D. A multi-sequence MRI integration framework using SwinUNETR-v2 for multiple sclerosis lesion segmentation. *BMC Med. Inform. Decis. Mak.* **26**, 195 (2026).
12. Goebl, P. et al. Enabling new insights from old scans by repurposing clinical MRI archives for multiple sclerosis research. *Nat. Commun.* (2025).
13. Coppola, E., Savardi, M. & Signoroni, A. Towards robust and reliable multi-modal 3D segmentation of multiple sclerosis lesions. *Pattern Recognit. Lett.* **200**, 115–122 (2026).
14. Proietto Salanitri, F. et al. SAM-guided prompt learning for multiple sclerosis lesion segmentation. *Pattern Recognit. Lett.* **199**, 205–211 (2026).
15. Wynen, M. et al. ConfLUNet: multiple sclerosis lesion instance segmentation in presence of confluent lesions. *Comput. Biol. Med.* **197**, 110919 (2025).
16. Bashiri Bawil, M., Shamsi, M., Jafargholkhanloo, A. F. & Shakeri Bavil, A. Incorporating normal periventricular changes for enhanced pathological white matter hyperintensity segmentation: on multiclass deep learning approaches. *BioMed. Eng. OnLine* **25**, 69 (2026).
17. Bashiri Bawil, M., Shamsi, M. & Shakeri Bavil, A. Simultaneous segmentation of ventricles and normal/abnormal white matter hyperintensities in clinical MRI using deep learning. Preprint at https://arxiv.org/abs/2506.07123 (2025).
18. Jeong, S. Y. et al. Development and validation of a deep learning-based automatic segmentation and classification of cerebral white matter hyperintensities. *Eur. Radiol.* **36**, 4812–4823 (2026).
19. Fazekas, F., Chawluk, J. B., Alavi, A., Hurtig, H. I. & Zimmerman, R. A. MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging. *Am. J. Roentgenol.* **149**, 351–356 (1987).
20. Ghafoorian, M. et al. Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. *Sci. Rep.* **7**, 5110 (2017).

21. Liang, L. et al. An anatomical knowledge-based MRI deep learning pipeline for white matter hyperintensity quantification associated with cognitive impairment. *Comput. Med. Imaging Graph.* **89**, 101873 (2021).
22. Chen, J. et al. Bilateral distance partition of periventricular and deep white matter hyperintensities: performance of the method in the aging brain. *Acad. Radiol.* **28**, 1699–1708 (2021).
23. Kervadec, H. et al. Boundary loss for highly unbalanced segmentation. *Med. Image Anal.* **67**, 101851 (2021).
24. Ronneberger, O., Fischer, P. & Brox, T. U-Net: convolutional networks for biomedical image segmentation. In *Proc. MICCAI*, LNCS 9351, 234–241 (2015).
25. Hatamizadeh, A. et al. Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In *Proc. BrainLes (MICCAI Workshop)*, LNCS 12962, 272–284 (2022).
26. Maurer, C. R., Qi, R. & Raghavan, V. A linear time algorithm for computing exact Euclidean distance transforms of binary images in arbitrary dimensions. *IEEE Trans. Pattern Anal. Mach. Intell.* **25**, 265–270 (2003).
27. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
28. Ma, J. et al. How distance transform maps boost segmentation CNNs: an empirical study. In *Proc. Medical Imaging with Deep Learning (MIDL)*, PMLR 121, 479–492 (2020).
29. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203–211 (2021).
30. Ma, J. et al. Loss odyssey in medical image segmentation. *Med. Image Anal.* **71**, 102035 (2021).
31. Kuijf, H. J. et al. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the WMH segmentation challenge. *IEEE Trans. Med. Imaging* **38**, 2556–2568 (2019).
32. Müller, D., Soto-Rey, I. & Kramer, F. Towards a guideline for evaluation metrics in medical image segmentation. *BMC Res. Notes* **15**, 210 (2022).
33. Taha, A. A. & Hanbury, A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med. Imaging* **15**, 29 (2015).
34. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. In *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, PMLR 70, 3319–3328 (2017).

Acknowledgements

The authors thank the creators of the MS3SEG dataset for making their MRI data and expert annotations publicly available, and the School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, for supporting this work

Author contributions

P.R., N.C. and J.P. contributed to this work. P.R. conceived the distance-to-ventricle prior, carried out the geometric analysis of the 10 mm rule and implemented the network and training pipeline. N.C. performed data pre-processing, the backbone comparison, model training and the statistical evaluation. J.P. performed the explainable AI analyses, including Integrated Gradients, channel ablation, the counterfactual distance-map test and the PV/deep confusion and distance-band analyses. P.R., N.C.

and J.P. wrote the main manuscript text and prepared the figures and tables. A.M. supervised the study. All authors reviewed the manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests

The authors declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.pdf