

Supplementary Material

Multi-Stage Semi-Supervised Learning Enhances White Matter Hyperintensity Segmentation

Kauê TN Duarte^{1,2}, Abhijot S Sidhu^{3,4}, Murilo C Barros⁵, David G Gobbi^{1,2}, Cheryl R McCreary^{1,4}, Feryal Saad¹, Richard Camicioli⁶, Eric E Smith¹, Mariana P Bento³, Richard Frayne^{1,2,3,4}

¹Radiology and Clinical Neurosciences, Hotchkiss Brain Institute, University of Calgary, Calgary, Alberta T2N 1N4, Canada

²Calgary Image Processing and Analysis Centre, Foothills Medical Centre, Calgary, Alberta T2N 2T9, Canada

³Graduate Program in Biomedical Engineering, Hotchkiss Brain Institute, University of Calgary, Calgary, Alberta T2N 1N4, Canada

⁴Seaman Family MR Research Centre, Foothills Medical Centre, Calgary, Alberta T2N 2T9, Canada

⁵School of Technology, University of Campinas, Limeira, São Paulo, 13484-332, Brazil

⁶Medicine (Neurology), University of Alberta, Edmonton, Alberta, T6G 2R7, Canada

1 SOFTWARE DESIGN

The **Margarida White Matter Hyperintensities (WMH) Segmentation Toolbox** is a comprehensive tool designed to facilitate the segmentation of FLAIR images using variations of the U-Net architecture. The software provides an intuitive interface for researchers and clinicians to load and process medical images, customize segmentation settings, and obtain binary masks for analysis.

User Interface Overview

The main screen of the Margarida WMH Segmentation Toolbox (Figure S1) is divided into two primary areas:

- **Right Panel:** The right area features a 2×2 grid for displaying images. It allows users to navigate through different image orientations (axial, coronal, sagittal) and visualize the segmentation process in real-time. Users can scroll through the slices of the loaded FLAIR volume using this panel.
- **Left Panel:** The left area contains the control menu, which includes several key functions:
 - **Load FLAIR Volume:** This button allows the user to load the FLAIR volume into the software for segmentation.
 - **Load FLAIR Mask (Optional):** Users can load a corresponding manual mask, if available, to compare against the automated segmentation results.
 - **Image Orientation:** This dropdown menu offers different orientations for conducting the segmentation: 2D_{Axi} (axial), 2D_{Cor} (coronal), 2D_{Sag} (sagittal), and 2.5D (combined orientation).
 - **Model Variant:** This field allows users to select from various U-Net variants for segmentation. This field will be updated with new techniques as they are developed and evaluated, users are advised to check the latest options on the official GitHub repository.

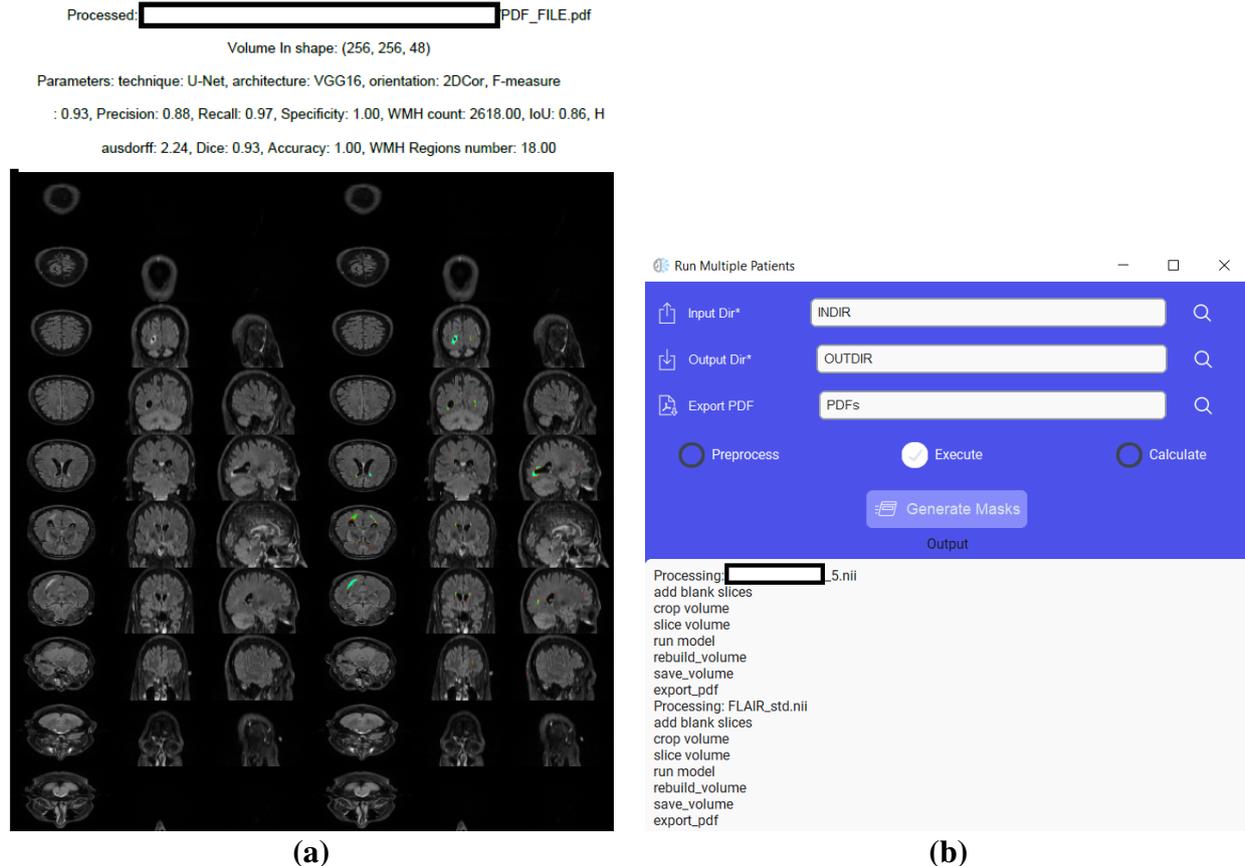


Figure S2. (a) The PDF report generated by the Margarida WMH Segmentation Toolbox. (b) The interface for running the segmentation model on multiple patients simultaneously.

ensure reproducibility and ease of use. The toolbox has been tested across multiple operating systems, including Windows, macOS, and Linux, to ensure broad compatibility.

Users can run the software either through its graphical user interface or via command line (Linux and macOS). Additionally, Docker and Singularity images are provided for streamlined deployment in different computing environments. For detailed guidance on installation and usage, a video tutorial is also available on the GitHub page.

This toolbox represents a significant step forward in automating the segmentation of white matter hyperintensities, offering a user-friendly interface and the flexibility to adapt to various research needs.

FLAIR Image Evaluation Dataset Availability

A subset of FLAIR images from the Calgary Normative Study (CNS) McCreary et al. (2020) is available for use with the Margarida WMH Segmentation Toolbox. This evaluation dataset is specifically prepared for testing and validating segmentation algorithms using the provided software.

The dataset includes a total of 140 images, with 20 images selected per decade from individuals aged 20 to 85 years old. Each decade is balanced for sex (10 male and 10 female individuals). The images were selected following a normal distribution within each age group to provide a representative sample of the population.

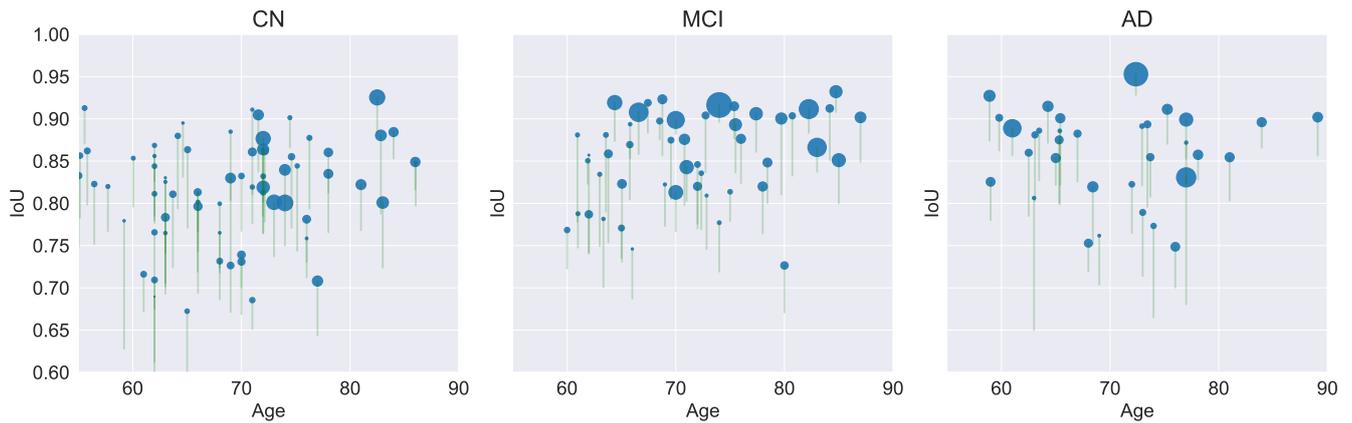


Figure S3. Evaluation of IoU score for the U-Net model across 260 individual subjects over age grouped by clinical stage: (a) CN = cognitively normal (148 individuals), (b) MCI = mild cognitive impairment (55 individuals), and (c) AD = Alzheimer's disease (37 individuals). Size of the filled circle in the plots reflects rank of the normalized WMH volume (*i.e.*, expressed as a percentage of the inter-cranial volume). Green vertical lines represent an increase in IoU values in the same individual, from baseline (bottom of the vertical line) to M3SL variant (top of the vertical line). Conversely, red vertical lines indicate a decrease in IoU for the M3SL compared to baseline variant. For all individuals (260/260 [100%]) IoU increased from baseline to M3SL model. This plot confirms the improved performance for M3SL compared to baseline model. Plotted are data from the local datasets described in Table 1.

Alongside the FLAIR images, corresponding manually annotated masks and demographic information (age and sex) are included for each subject. This dataset serves as a valuable resource for researchers and developers to evaluate the performance of their segmentation models.

These images and associated data will be available upon request (refer to Github). This Evaluation Dataset is derived from the Calgary Normative Study, as referenced in McCreary et al. (2020).

2 SUPPLEMENTARY METHODOLOGY

Supplementary Tables are provided to support interpretation of the Figures in the main paper (specifically Figures 2-5 in the main paper).

Supplementary Figures S3 and S4 extend the summary analysis of Figure 6.

Where available, data are report by clinical stage (cognitively normal = CN, mild cognitive impairment = MCI and Alzheimer's disease = AD). Table 1 (see main paper) describes the CN, MCI and AD cohorts within the three local datasets ($n = 260$), McCreary et al. (2020); Peca et al. (2013); Subotic et al. (2021) as well as over the annotated public datasets ($n = 60$). Kuijf et al. (2019)

The pooled group ($n = 320$) in each supplementary table includes both the local datasets and the annotated public dataset.

Results were calculated over voxels containing brain tissue. Normalized WMH (nMWH) reports the fraction of the intracranial (or brain) volume.

Reported are mean \pm standard deviation calculated across subjects.

3 SUPPLEMENTARY RESULTS

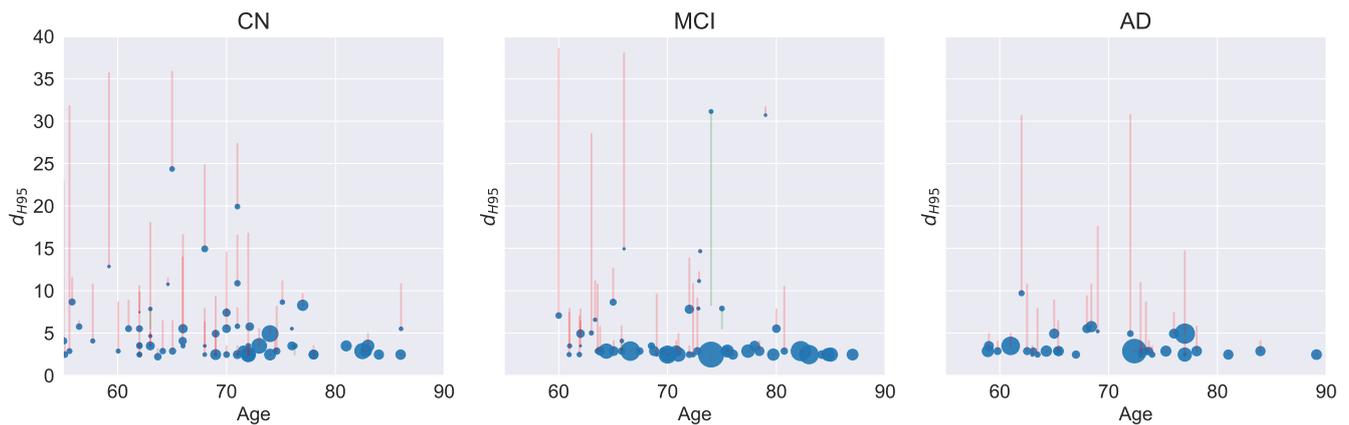


Figure S4. Evaluation of d_{H95} distance for the U-Net model across 260 individual subjects over age grouped by clinical stage: (a) CN = cognitively normal (148 individuals), (b) MCI = mild cognitive impairment (55 individuals), and (c) AD = Alzheimer's disease (37 individuals). Size of the filled circle in the plots reflects rank of the normalized WMH volume (*i.e.*, expressed as a percentage of the inter-cranial volume). Green vertical lines represent an increase in d_{H95} values in the same individual, from baseline (bottom of the vertical line) to M3SL variant (top of the vertical line). Conversely, red vertical lines indicate a decrease in d_{H95} for the M3SL compared to baseline variant. Only 7/260 (0.027%) of these lines were green. The decrease in d_{H95} (red lines) is desirable since shows an improvement in segmentation. This plot demonstrates better performance for M3SL compared to baseline model. Plotted are data from the local datasets described in Table 1.

Table S1. Normalized WMH volume (nWMH) (mean \pm standard deviation) by clinical stage for the ground truth and results obtained by each U-Net model (Multi-Stage Semi-Supervised Learning (M3SL), baseline, transfer learning (TL)). Also shown are results of annotated public dataset and pooled results.

Level	Ground Truth (%)	M3SL (%)	Baseline (%)	TL (%)
CN	0.35 ± 0.37	0.32 ± 0.34	0.31 ± 0.33	0.31 ± 0.33
MCI	0.59 ± 0.74	0.57 ± 0.72	0.56 ± 0.71	0.55 ± 0.71
AD	0.58 ± 0.70	0.56 ± 0.68	0.54 ± 0.65	0.53 ± 0.64
Annotated Public	1.16 ± 1.13	1.04 ± 1.01	0.77 ± 0.85	0.78 ± 0.86
Pooled	0.67 ± 0.84	0.61 ± 0.77	0.53 ± 0.67	0.53 ± 0.67

Table S2. Whole brain average false negative (FNF), false positive (FPF) and true positive (TPF) fractions (mean \pm standard deviation) by clinical stage for each U-Net model. Also shown are results of annotated public dataset and pooled results. Note: The dataset is imbalanced and in all cases the true negative fraction (TNF) exceeds 0.98

Level	M3SL			Baseline		
	FNF (%)	FPF (%)	TPF (%)	FNF (%)	FPF (%)	TPF (%)
CN	0.05 \pm 0.05	0.02 \pm 0.02	0.30 \pm 0.33	0.06 \pm 0.07	0.03 \pm 0.03	0.28 \pm 0.31
MCI	0.05 \pm 0.05	0.03 \pm 0.04	0.54 \pm 0.70	0.07 \pm 0.07	0.04 \pm 0.04	0.52 \pm 0.68
AD	0.05 \pm 0.06	0.02 \pm 0.02	0.53 \pm 0.66	0.08 \pm 0.12	0.04 \pm 0.03	0.50 \pm 0.63
Annotated Public	0.18 \pm 0.19	0.06 \pm 0.11	0.97 \pm 0.98	0.42 \pm 0.52	0.06 \pm 0.10	0.71 \pm 0.81
Pooled	0.08 \pm 0.12	0.03 \pm 0.07	0.58 \pm 0.74	0.16 \pm 0.32	0.04 \pm 0.06	0.49 \pm 0.64

Level	TL		
	FNF (%)	FPF (%)	TPF (%)
CN	0.07 \pm 0.07	0.03 \pm 0.03	0.28 \pm 0.31
MCI	0.07 \pm 0.08	0.04 \pm 0.04	0.51 \pm 0.67
AD	0.08 \pm 0.13	0.04 \pm 0.03	0.50 \pm 0.62
Public	0.40 \pm 0.52	0.06 \pm 0.10	0.72 \pm 0.82
Pooled	0.16 \pm 0.31	0.04 \pm 0.06	0.49 \pm 0.64

Table S3. Average F-Measure (mean \pm standard deviation) by clinical stage for each U-Net model and orientation. Also shown are results of annotated public dataset and pooled results. 2.5D represents the pooled average of the axial, coronal and sagittal predictions. Duarte et al. (2023)

Level	M3SL				Baseline			
	2D Axial	2D Coronal	2D Sagittal	2.5D	2D Axial	2D Coronal	2D Sagittal	2.5D
CN	0.82 \pm 0.06	0.87 \pm 0.04	0.85 \pm 0.05	0.89 \pm 0.04	0.74 \pm 0.06	0.84 \pm 0.04	0.80 \pm 0.05	0.85 \pm 0.05
MCI	0.84 \pm 0.07	0.89 \pm 0.05	0.87 \pm 0.07	0.90 \pm 0.07	0.77 \pm 0.07	0.86 \pm 0.06	0.83 \pm 0.07	0.86 \pm 0.07
AD	0.85 \pm 0.05	0.89 \pm 0.05	0.87 \pm 0.06	0.91 \pm 0.05	0.78 \pm 0.05	0.86 \pm 0.07	0.83 \pm 0.07	0.87 \pm 0.06
Public	0.82 \pm 0.07	0.82 \pm 0.07	0.80 \pm 0.09	0.84 \pm 0.07	0.70 \pm 0.10	0.72 \pm 0.13	0.70 \pm 0.14	0.72 \pm 0.14
Pooled	0.83 \pm 0.06	0.87 \pm 0.06	0.84 \pm 0.07	0.88 \pm 0.06	0.74 \pm 0.08	0.81 \pm 0.10	0.79 \pm 0.10	0.82 \pm 0.11

Level	TL			
	2D Axial	2D Coronal	2D Sagittal	2.5D
CN	0.74 \pm 0.06	0.83 \pm 0.05	0.80 \pm 0.06	0.84 \pm 0.04
MCI	0.76 \pm 0.07	0.85 \pm 0.05	0.82 \pm 0.07	0.86 \pm 0.06
AD	0.76 \pm 0.06	0.84 \pm 0.08	0.83 \pm 0.06	0.86 \pm 0.07
Annotated Public	0.71 \pm 0.10	0.71 \pm 0.15	0.69 \pm 0.14	0.72 \pm 0.14
Pooled	0.74 \pm 0.08	0.81 \pm 0.11	0.78 \pm 0.10	0.82 \pm 0.10

Table S4. Average IoU (mean \pm standard deviation) by clinical stage for each U-Net model and orientation. Also shown are results of annotated public dataset and pooled results. 2.5D represents the pooled average of the axial, coronal and sagittal predictions. Duarte et al. (2023)

Level	M3SL				Baseline			
	2D Axial	2D Coronal	2D Sagittal	2.5D				
CN	0.72 \pm 0.08	0.79 \pm 0.07	0.76 \pm 0.08	0.82 \pm 0.07	0.61 \pm 0.10	0.74 \pm 0.07	0.69 \pm 0.09	0.75 \pm 0.07
MCI	0.75 \pm 0.11	0.81 \pm 0.09	0.78 \pm 0.11	0.83 \pm 0.11	0.64 \pm 0.12	0.77 \pm 0.09	0.74 \pm 0.12	0.78 \pm 0.11
AD	0.77 \pm 0.07	0.82 \pm 0.08	0.80 \pm 0.08	0.85 \pm 0.07	0.66 \pm 0.08	0.76 \pm 0.12	0.73 \pm 0.13	0.78 \pm 0.12
Annotated Public	0.73 \pm 0.11	0.72 \pm 0.12	0.70 \pm 0.12	0.75 \pm 0.11	0.55 \pm 0.16	0.56 \pm 0.21	0.55 \pm 0.20	0.58 \pm 0.21
Pooled	0.74 \pm 0.10	0.78 \pm 0.10	0.75 \pm 0.11	0.81 \pm 0.10	0.61 \pm 0.12	0.70 \pm 0.16	0.67 \pm 0.16	0.72 \pm 0.16

Level	TL			
	2D Axial	2D Coronal	2D Sagittal	2.5D
CN	0.61 \pm 0.10	0.73 \pm 0.07	0.69 \pm 0.09	0.74 \pm 0.07
MCI	0.63 \pm 0.12	0.76 \pm 0.09	0.72 \pm 0.12	0.77 \pm 0.11
AD	0.65 \pm 0.09	0.75 \pm 0.12	0.72 \pm 0.13	0.77 \pm 0.11
Annotated Public	0.56 \pm 0.16	0.57 \pm 0.21	0.55 \pm 0.20	0.59 \pm 0.20
Pooled	0.61 \pm 0.12	0.70 \pm 0.15	0.66 \pm 0.15	0.71 \pm 0.15

Table S5. Average Hausdorff 95% distance (d_{H95} , mean \pm standard deviation) by clinical stage for each U-Net model and orientation. Also shown are results of annotated public dataset and pooled results. 2.5D represents the pooled average of the axial, coronal and sagittal predictions. Duarte et al. (2023)

Level	M3SL (mm)	Baseline (mm)	TL (mm)
CN	6.75 \pm 15.81	9.19 \pm 9.74	7.90 \pm 6.93
MCI	5.21 \pm 5.95	7.91 \pm 8.49	7.42 \pm 8.18
AD	3.48 \pm 1.50	7.28 \pm 7.03	7.06 \pm 8.46
Annotated Public	14.55 \pm 16.60	25.72 \pm 18.97	23.17 \pm 18.29
Pooled	8.03 \pm 13.50	13.11 \pm 14.62	11.89 \pm 13.54

Table S6. Average F-Measure (mean \pm standard deviation) by clinical stage for each U-Net model and scanner type (see description in Table 1). Also shown are results of annotated public dataset and pooled results.

Level	M3SL				
	A	B	C	D	E
CN	0.91 \pm 0.03	0.85 \pm 0.03	–	–	–
MCI	0.91 \pm 0.06	0.85 \pm 0.08	–	–	–
AD	0.91 \pm 0.06	0.89 \pm 0.03	–	–	–
Annotated Public	–	–	0.87 \pm 0.06	0.82 \pm 0.07	0.82 \pm 0.07
Pooled	0.91 \pm 0.05	0.86 \pm 0.05	0.87 \pm 0.06	0.82 \pm 0.07	0.82 \pm 0.07

Level	Baseline				
	A	B	C	D	E
CN	0.86 \pm 0.04	0.81 \pm 0.04	–	–	–
MCI	0.87 \pm 0.06	0.82 \pm 0.09	–	–	–
AD	0.87 \pm 0.06	0.85 \pm 0.03	–	–	–
Annotated Public	–	–	0.84 \pm 0.06	0.73 \pm 0.08	0.58 \pm 0.13
Pooled	0.87 \pm 0.05	0.82 \pm 0.05	0.84 \pm 0.06	0.73 \pm 0.08	0.58 \pm 0.13

Level	TL				
	A	B	C	D	E
CN	0.85 \pm 0.04	0.80 \pm 0.04	–	–	–
MCI	0.86 \pm 0.06	0.82 \pm 0.07	–	–	–
AD	0.86 \pm 0.07	0.84 \pm 0.03	–	–	–
Annotated Public	–	–	0.84 \pm 0.07	0.74 \pm 0.08	0.59 \pm 0.13
Pooled	0.86 \pm 0.05	0.81 \pm 0.05	0.84 \pm 0.07	0.74 \pm 0.08	0.59 \pm 0.13

Table S7. Average IoU (mean \pm standard deviation) by clinical stage for each U-Net model and scanner type (see description in Table 1). Also shown are results of annotated public dataset and pooled results.

M3SL					
Level	A	B	C	D	E
CN	0.84 ± 0.04	0.74 ± 0.07	–	–	–
MCI	0.84 ± 0.10	0.76 ± 0.14	–	–	–
AD	0.86 ± 0.07	0.81 ± 0.06	–	–	–
Annotated Public	–	–	0.79 ± 0.09	0.74 ± 0.13	0.71 ± 0.11
Pooled	0.85 ± 0.07	0.76 ± 0.09	0.79 ± 0.09	0.74 ± 0.13	0.71 ± 0.11
Baseline					
	A	B	C	D	E
CN	0.77 ± 0.06	0.68 ± 0.07	–	–	–
MCI	0.79 ± 0.10	0.72 ± 0.15	–	–	–
AD	0.79 ± 0.12	0.76 ± 0.08	–	–	–
Annotated Public	–	–	0.74 ± 0.10	0.61 ± 0.13	0.38 ± 0.19
Pooled	0.78 ± 0.09	0.70 ± 0.09	0.74 ± 0.10	0.61 ± 0.13	0.38 ± 0.19
TL					
	A	B	C	D	E
CN	0.77 ± 0.05	0.66 ± 0.08	–	–	–
MCI	0.78 ± 0.11	0.71 ± 0.13	–	–	–
AD	0.77 ± 0.12	0.74 ± 0.08	–	–	–
Annotated Public	–	–	0.74 ± 0.09	0.62 ± 0.13	0.39 ± 0.19
Pooled	0.77 ± 0.09	0.69 ± 0.10	0.74 ± 0.09	0.62 ± 0.13	0.39 ± 0.19

Table S8. Average Hausdorff 95% distance (d_{H95} , mean \pm standard deviation) by clinical stage for each U-Net model and scanner type (see description in Table 1). Also shown are results of annotated public dataset and pooled results.

M3SL					
	A	B	C	D	E
CN	6.33 \pm 17.70	8.17 \pm 6.33	–	–	–
MCI	4.62 \pm 5.01	8.94 \pm 9.89	–	–	–
AD	3.26 \pm 1.49	4.77 \pm 0.76	–	–	–
Annotated Public	–	–	9.65 \pm 8.52	17.36 \pm 19.74	16.65 \pm 18.83
Pooled	5.03 \pm 11.89	7.76 \pm 6.80	9.65 \pm 8.52	17.36 \pm 19.74	16.65 \pm 18.83
Baseline					
	A	B	C	D	E
CN	8.15 \pm 9.68	12.69 \pm 9.41	–	–	–
MCI	7.33 \pm 8.26	11.61 \pm 9.71	–	–	–
AD	6.97 \pm 7.47	9.06 \pm 3.58	–	–	–
Annotated Public	–	–	14.31 \pm 12.02	22.66 \pm 13.34	40.97 \pm 20.46
Pooled	7.60 \pm 8.67	11.77 \pm 8.60	14.31 \pm 12.02	22.66 \pm 13.34	40.97 \pm 20.46
TL					
	A	B	C	D	E
CN	6.29 \pm 4.91	13.30 \pm 9.75	–	–	–
MCI	6.67 \pm 7.91	12.25 \pm 8.92	–	–	–
AD	6.83 \pm 9.11	8.37 \pm 2.50	–	–	–
Annotated Public	–	–	12.16 \pm 9.42	21.02 \pm 13.72	36.29 \pm 22.32
Pooled	6.55 \pm 7.08	12.16 \pm 8.65	12.16 \pm 9.42	21.02 \pm 13.72	36.29 \pm 22.32

REFERENCES

- Duarte, K. T., Gobbi, D. G., Sidhu, A. S., McCreary, C. R., Saad, F., Camicioli, R., et al. (2023). Segmenting white matter hyperintensities in brain magnetic resonance images using convolution neural networks. *Pattern Recognition Letters* 175, 90–94. doi:10.1016/j.patrec.2023.07.014
- Kuijf, H. J., Casamitjana, A., Collins, D. L., Dadar, M., and et al., A. G. (2019). Standardized assessment of automatic segmentation of white matter hyperintensities and results of the WMH segmentation challenge. *IEEE Transactions on Medical Imaging* 38, 2556–2568. doi:10.1109/tmi.2019.2905770
- McCreary, C. R., Salluzzi, M., Andersen, L. B., Gobbi, D., Lauzon, L., Saad, F., et al. (2020). Calgary Normative Study: Design of a prospective longitudinal study to characterise potential quantitative MR biomarkers of neurodegeneration over the adult lifespan. *BMJ Open* 10. doi:10.1136/bmjopen-2020-038120
- Peca, S., McCreary, C. R., Donaldson, E., Kumarpillai, G., Shobha, N., and et al (2013). Neurovascular decoupling is associated with severity of cerebral amyloid angiopathy. *Neurology* 81, 1659–1665. doi:10.1212/01.wnl.0000435291.49598.54
- Subotic, A., McCreary, C., Saad, F., Nguyen, A., Alvarez-Veronesi, A., and et al (2021). Cortical thickness and its association with clinical cognitive and neuroimaging markers in cerebral amyloid angiopathy. *Journal of Alzheimer's Disease* 81, 1–9. doi:10.3233/JAD-210138