Attention Blocks Improve White Matter Hyperintensity Semantic Segmentation using U-Nets

Kauê T. N. Duarte, Murilo C. Barros, Abhijot S. Sidhu, David G. Gobbi, Cheryl R. McCreary, Feryal Saad, Richard Camicioli, Eric E. Smith, Marco Carvalho, and Richard Frayne

***Abstract*—White matter hyperintensities (WMHs) are a common finding on magnetic resonance (MR) images in older individuals, appearing as high-signal intensity regions on fluid-attenuated inversion recovery (FLAIR) imaging. People with high WMH volume are at increased risk for dementia and stroke, controlling for vascular risk factors, but WMH burden is not reliably assessed in clinical practice. Manual segmentation of WMHs is accepted as the gold standard (or ground truth), however, it is a laborious and time-consuming method. Newer machine learning (ML)-based approaches are being proposed as alternatives to manual segmentation. Among these approaches, U-Net convolutional neural networks have demonstrated good WMH segmentation performance. However, even state-of-the-art ML models sometimes fail to correctly identify WMHs and their boundaries with sufficient accuracy. Attention blocks have emerged as a potential solution for improving the performance of U-Net models by enhancing the ability of the model to focus on relevant features in the data. We investigated the effectiveness of attention blocks in U-Net models for WMH segmentation compared to three other models (U-Net++, U-Net3+, and a standard U-Net). Attention blocks significantly improved the F-measure score for WMH segmentation (0.811 *vs* 0.789 for next best model, $p = 0.04$) in a diverse brain imaging dataset. This study demonstrates that attention blocks enhance U-Net models used for WMH identification and classification.**

Link to graphical and video abstracts, and to code: https://latamt.ieeer9.org/index.php/transactions/article/view/9615

***Index Terms*—attention blocks, artificial intelligence, alzheimer's disease, semantic segmentation, white matter hyperintensity (WMH)**

The associate editor coordinating the review of this manuscript and approving it for publication was Samuel Ortega (*Corresponding author: Kauê Tartarotti Nepomuceno Duarte*).

This study was supported by the Canadian Institutes for Health Research, the Canada Foundation for Innovation, and the University of Calgary Evolve2Innovate (E2I) program. The work of the research teams supporting the CNS, FAVR-I, and FAVR-II studies are acknowledged. KTND was an Eyes High Scholar, EES is the Katthy Taylor Chair, and RF is the Hopewell Professor, all held at the University of Calgary.

Kauê Tartarotti Nepomuceno Duarte, A. S. Sidhu, D. G. Gobbi, C. R. McCreary, F. Saad, E. E. Smith, and R. Frayne are with the Radiology and Clinical Neurosciences, Hotchkiss Brain Institute, University of Calgary; and the Seaman Family MR Centre and Calgary Image Processing and Analysis Centre, Foothills Medical Centre, Calgary, Alberta, Canada (e-mails: kaue.unicamp2011@gmail.com, sidhuas@ucalgary.ca, dgobbi@ucalgary.ca, crmcreary@ucalgary.ca, feryal.saad@ucalgary.ca, eesmith@ucalgary.ca, and rfrayne@ucalgary.ca).

M. C. Barros, and M. A. G. Carvalho are with the School of Technology, University of Campinas, Limeira, São Paulo, Brazil (e-mails: murilo.barros.sn@gmail.com, and magic@unicamp.br).

R. Camicioli is with the Department of Medicine (Neurology), Neuroscience and Mental Health Institute, University of Alberta, Edmonton, Alberta, Canada (e-mail: rcamicio@ualberta.ca).

I. Introduction

WHITE matter hyperintensities (WMHs) are radiologically defined regions of increased signal intensity observed on T2-weighted or fluid-attenuated inversion recovery (FLAIR) magnetic resonance (MR) imaging [1]. The number of WMHs, and their location and volume, reflect a combination of factors, such as age and the presence of active disease processes (*e.g.*, neurodegeneration, inflammation or demyelination). WMHs commonly develop in older individuals (>65 years) [2] and, at least initially, are asymptomatic. The exact cause of WMHs is not fully understood, though it is seen as likely multi-factorial. Factors that have been implicated in their development include age-related changes in blood vessels and chronic conditions, such as hypertension, diabetes, smoking, and cardiovascular disease [3]. Other potentially contributing factors include genetic susceptibility, inflammation, and the presence of other small vessel diseases [4], [5]. Studies have shown an association between the prevalence of WMH and extent of cognitive loss, dementia [6] or other neurodegenerative disorders [7], [8]. An increased WMH burden in the frontal, parietal, and periventricular brain regions is typical in elders.

The widely accepted ground truth for quantifying WMHs in clinical-research studies is manual segmentation of lesions on FLAIR images. However, this step is time-consuming and costly, because it requires highly trained specialists. Manual segmentation suffers from high inter-rater variation. Machine learning (ML)-based techniques hold much promise for overcoming these issues and potentially may provide low cost, fast and accurate WMH segmentation while eliminating inter-rater variation. [9] Deep learning-based techniques have advanced the field of WMH segmentation [10], in part, because of the increasing availability of higher quality, annotated data for model training. In particular, a variation of convolutional neural networks (CNN), known as the U-Net [11], has been shown to be effective in segmenting WMH [12]. Nevertheless, segmenting WMHs using U-Net models remains a challenging task as the total WMH burden is often $< 0.05\%$ of the entire brain volume and due to the difficulties encountered in correctly identifying WMH lesions and their boundaries.

Attention blocks are a newer technique that were proposed to emphasize important image features, while suppressing irrelevant ones [13], [14]. They may present a viable solution to the current challenges of WMH segmentation. Attention blocks in U-Nets employ additional convolutional layers and implement spatial/channel-wise attention mechanisms. The spatial attention module calculates attention weights for each

spatial location in the feature maps, enabling a more focused processing of specific regions of interest. In this study, we use attention blocks and implement, train, and evaluate a 2D U-Net attention model for segmenting WMHs. We expect to observe an increase in the accuracy compared to traditional U-Net models. The principle novelty of this work is the strategic integration of attention blocks into VGG16-based U-Nets to not only enhance segmentation of subtle WMH, but also to improve boundary delineation across diverse imaging protocols and disease stages.

Our goal is to explore the role of attention blocks in order to identify U-Net variants that improve WMH segmentation. Section II reviews the existing work on attention blocks and WMH segmentation. In Section III, we describe our methodology, expanding on the important details required to build and test our models. Section IV summarizes our results, and Section V discusses these findings and their broader implications. Finally, in Section VI we conclude this work and suggest areas for future study. Our major contributions are 1) identifying an improved U-Net variant for WMH segmentation, 2) assessing generalizability of this variant by testing on several datasets obtained with different acquisition protocols, acquired on multiple scanners located at different sites, and 3) establishing the benefit of attention blocks for segmenting WMHs.

II. Literature Review

Variants of U-Net models [12] have demonstrated much success in identifying and segmenting WHMs including in a generalized, clinical setting. Heinen *et al.* [15], for example, performed automated WMH segmentation in a multi-centre study using five different methods. They found that the kNN-TTP method had the best performance within and across sites and scanner vendors. However, they stressed the need for further improvements before applying the model in a clinical setting. This effort highlights a critical gap in generalizability that provides motivation for using attention mechanisms. Focusing on routine WMH segmentation, Zhu *et al.* [16] proposed the 2D VB-Net model, along with dedicated labeling protocols, for accurate segmentation of WMH and other intracranial lesions in large datasets obtained from a diverse population and acquired on multiple scanners. While effective for 2D imaging (with relatively thicker slices), their reliance on 2D models limit spatial integration across slices – a challenge we address here by combining 2D orientations. Liu *et al.* [17] proposed a U-Net model to identify and segment WMH in patients with acute ischemic lesions and showed that U-Net models outperformed other methods. Indeed, their focus on acute lesions can potentially diminish identification of chronic WMHs or those due to other neurodegenerative etiologies.

Segmenting WMHs is a challenging task because the lesions can vary in size, be irregularly shaped and occur in both deep white matter and periventricular areas. To address these issues, Sundaresan *et al.* [18] introduced TrUE-Net, a 2.5D U-Net segmentation method that averages the lesion probability map (outcome of the individual 2D U-Net models) over three processing orientations (axial, coronal, sagittal). This approach was evaluated using five MR datasets, including Alzheimer's disease and vascular disease data, along with training data obtained from MR and MR-PET examinations [18]. They demonstrated positive validations with the external data, suggesting the effectiveness of TrUE-NET across diverse datasets. 2D multi-orientation averaging (2.5D) has the advantage of being less complex and computational resource demanding than 3D volumes to process while preserving important spatial information.

Traditional U-Nets, however, can lack spatial specificity in deep layers because they focus on feature extraction. Attention U-Nets reduce this problem. Their application is becoming more prevalent because of their improved generalization capabilities for segmentation tasks. Attention blocks combine strong spatial resolution and feature characteristics through the implementation of skip connections. The capacity to generalize is essential as MR images are commonly acquired with different acquisition parameters from different sites, scanner vendors, and imaging techniques. Our work tests robustness by evaluating our results over heterogeneous cohorts and scanner vendors.

The application in medical image processing of attention blocks in ML models is expanding and several U-Net variants have been recently developed. Using another variant, Park *et al.* [19] combined a U-Net model with multi-scale foreground highlighting. Classification was performed based on three parameters: 1) WMH volume, 2) clinical variables (age, cardiovascular risk, education, gender, ApoE4 genotype, and race), and 3) the interaction between WMH volume and clinical variables. Their model exhibited high detection rates for small clusters and near WMH borders but was limited by the detection of only some WMH voxels. This result shows the need for architectures like ours that achieve voxel-level precision without non-imaging data. Other WMH segmentation methods include diverse architectural and training innovations. Lee et al. [20], for example, introduced AQUA, a 2D U-Net enhanced with bottleneck modules across encoder-decoder blocks to improve small lesion sensitivity. The *sysu_media* approach [21] employed an ensemble of three U-Net-like networks with distinct initializations, augmented by slice-based false-positive suppression. The *nlp_logix* method [21] implemented a multi-scale network trained across ten folds, averaging predictions from top checkpoints, while *cian* [21] utilized 3D MD-GRU with deformable patch augmentation to address spatial heterogeneity.

In other previous work [22], [23], the WMH segmentation was performed using different architectures and styles. However, the need to improve aspects such as the detection of WMH burden boundaries was still required to be implemented. Our work differs from these studies in the following aspects: 1) It explores the role of attention blocks in WMH segmentation tasks. 2) It evaluates the role of this approach over a range of subjects, raging from cognitively normal (or intact) through to those with dementia. 3) It compares distinct U-Net model variants, including one incorporating attention blocks, for WMH segmentation.

III. Materials and Methods

A. Dataset and Participant Demographics

Three local datasets ($N = 260$) and three publicly available, free-access datasets provided specifically for WMH segmentation training with manual ground truth labels ($N = 60$) were utilized in this work. These data were acquired on five scanner models from three scanner vendors and were located at five centres. Both the local and public datasets were used in training, validation, and testing steps. Details regarding the acquisition of each dataset are provided in [22]. Table S1 (see Supplementary Material) summarizes the number of individuals by scanner model and site, sex, age, and disease class for each dataset.

Local Datasets. The *Calgary Normative Study* (CNS) is an ongoing longitudinal MR study focused on quantitative imaging and analysis techniques in aging [24]. The CNS includes images from cognitively normal (CN) individuals residing in the community that were acquired on Scanner A (3 T Discovery MR750, General Electric (GE) Healthcare, Waukesha, WI). In total, we included 94 individuals from this study, divided into younger and older cohorts (CNS_1: $N = 74$ all ≤ 35 years and CNS_2: $N = 20$ all ≥ 40 years).

The *Functional Assessment of Vascular Reactivity I* (FAVR-I) study [25] contributed $N = 71$ participants and was a single-center observational study aimed at investigating the relationship between cerebral blood flow and cognitive status by disease class (CN; mild cognitive impairment, MCI; and Alzheimer's disease, AD). Participants had a neuropsychological assessment and were examined by a physician specialist in dementia. MCI and AD were diagnosed using consensus clinical criteria. [26] The FAVR-I data were also acquired using Scanner A.

FAVR-II, an extension of FAVR-I, is an ongoing two-center study from which we have data for $N = 95$ participants [27]. In FAVR-II, data were acquired from two scanners: Scanner A ($N = 65$, 68.4%) and Scanner B ($N = 30$, 31.6%; 3 T Siemens Prisma; Siemens Healthineers, Erlangen, Germany), located at a second site. For both studies, CN individuals with no history of central nervous system diseases were recruited by community advertising. The same criteria for MCI and AD were used in FAVR-II as in FAVR-I.

The local datasets were segmented using a semi-automated seed-based approach to create initial WMH masks (*Cerebra-Lesion Extractor* [28] for CNS and FAVR-II, and Quantamo [29] for FAVR-I). The initial WMH masks were reviewed and manually edited (as needed) to produce final ground truth WMH masks. Each voxel in the mask was marked as either *True* (containing WMH) or *False* (not containing WMH).

Public Datasets. We also included data from three public, free-access datasets (provided as part of the WMH Segmentation Challenge [21]). Specifically, we utilized the challenge training data from three additional sites: 1) Amsterdam (*AMS*, $N = 20$) – data acquired on Scanner C (3 T GE Signa HDxt), 2) *Utrecht* ($N = 20$) – data acquired on Scanner D (3 T Philips Achieva; Philips Healthcare, Eindhoven, the Netherlands), and 3) Singapore (*SIN*, $N = 20$) – data acquired on Scanner E (3 T Siemens Trio Tim). Ground truth segmentation masks were provided for the AMS, SIN, and Utrecht datasets [21].

B. Data Preparation

To improve the quality of the FLAIR scans, we applied the N4 bias-field correction [30]–[32] technique to reduce variations in intensity caused by inhomogeneity of the scanner. In general, all 2D FLAIR volumes were of size 256×256 with varying depth. To standardize the data volume size, we added zero-valued (*i.e.*, blank) images to achieve dimensions of $256 \times 256 \times 64$. Each image volume was then divided into sixteen equal-sized patches (dimensions: $64 \times 64 \times 64$). To address class imbalance, we followed the approach of [33] and used only patches that contained at least one labelled True WMH voxel in the training, validation, and test steps. Most data volumes were normalized by mapping the 0^{th} and 98^{th} percentile of the image intensity to $[0.0, 1.0]$. The AMS and SIN data exhibited an image contrast suggesting that fat suppression was used when acquiring the FLAIR images. For both of these datasets we used the 0^{th} and 100^{th} percentile (*i.e.*, min-max normalization).

C. Attention and Other U-Net Models

Adding attention blocks to a 2D U-Net model introduced specialized skip connections that combine high-resolution spatial information from the encoding layers with knowledge obtained from features extracted in the deeper decoding layers [34]. We included the attention blocks in our VGG16 feature extractor [35] before each max-pooling layer (refer to Figure S1). We chose this architecture based on previous studies that confirmed the benefits of VGG16 for effective WMH segmentation. [22], [23]

We also implemented three other U-Net variants (*Traditional U-Net*, [12] *U-Net++*, [36] *U-Net 3+* [37]) and evaluated them against our Attention U-Net model. The expected strengths of the *Attention U-Net* and three other models were:

1) *Attention U-Net*. Enhancement to the traditional U-Net architecture by incorporating attention mechanisms (blocks) [34] that selectively combine high spatial resolution with high feature information. Attention blocks improve the ability of the model to capture fine details at boundaries and therefore more accurately segment WMHs.
2) *Traditional U-Net* [12]. Has an encoder-decoder structure with skip connections.
3) *U-Net++* [36]. Builds on the traditional U-Net model by introducing a nested architecture that consists of multiple U-Net modules, capturing both local and global contextual cues for improved segmentation performance and better object boundary detection.
4) *U-Net 3+* [37]. Extends the U-Net++ model by integrating residual learning and densely connected pathways. Residual connections decrease optimization challenges, while dense connections facilitate information flow, leading to stronger feature representations and superior segmentation performance.

Figure S1 illustrates the VGG16-based architecture for each of the four models.

D. Model Training

Our models were trained for a maximum of 600 epochs at a learning rate of $l_0 = 5 \times 10^{-4}$. To ensure consistent activation, we employed a sigmoid-shaped activation function and dichotomized the output at 0.5. For each of the four implementations (Figure S1), three separate 2D U-Net models were trained using images derived by extracting data from the volume along the x (axial), y (coronal), and z (sagittal) directions.

Loss Function. We employed a loss function that balanced the impact of the Dice loss (DL) and the binary focal loss (FL). The DL is a commonly utilized loss function in the medical field. We chose this function because of the unequal number of *True* and *False* WMH mask elements:

$$DL = \frac{(1+\beta^2)\cdot TP}{(1+\beta^2)\cdot FP + \beta^2 \cdot FN + FP} \tag{1}$$

where the β is the balance coefficient, and TP, FP, and FN represent the true positive, false positive, and false negative voxels, respectively. The binary focal loss (FL) further addresses class imbalance by adjusting the cross-entropy criterion:

$$\begin{aligned} FL = &- GT\alpha(1-PT)^{\gamma}\log(PT) \\ &- (1-GT)\alpha PT^{\gamma}\log(1-PT) \end{aligned} \tag{2}$$

where GT refers to the ground truth and PT corresponds to the predicted truth. The hyperparameters $\alpha = 0.25$ and $\gamma = 2.0$ are values that were fine-tuned through a calibration process.

Performance Metrics. Because the number of WMH and non-WMH voxels were unequal, accuracy was not an appropriate performance metric due to the disproportionately high number of true negative (TN) counts. Instead, F-measure, intersection-over-union (IoU), and Hausdorff distance were employed to assess model effectiveness. F-measure (F) is a widely used performance metric in image segmentation

$$\mathrm{F} = 2 \times \frac{P \times R}{P + R} \tag{3}$$

that represents the harmonic mean of precision (P) and recall (R):

$$P = \frac{TP}{TP + FP} \quad \text{and} \quad R = \frac{TP}{TP + FN}.$$

IoU compares the predicted outcome to the ground truth, as follows:

$$\mathrm{IoU} = \frac{TP}{FP + TP + FN}. \tag{4}$$

During training, the model with the highest IoU metric was saved as the best model.

The Hausdorff distance measures the distance between predicted and ground truth WMH boundaries. Given that each boundary is defined by two sets of points, A and B, the Hausdorff distance is:

$$\begin{aligned} d_H(x,y) &= max\{d_{AB}, d_{BA}\} \\ &= max\{\underset{a\in A}{max}\{\{\underset{b\in B}{min}\{d(a,b)\}\}, \underset{b\in B}{max}\{ \\ &\qquad \underset{a\in A}{min}\{d(a,b)\}\}\} \end{aligned} \tag{5}$$

where a and b represent the elements of the sets A and B, respectively. The distance $d(a,b)$ is measured by the Euclidean distance between elements a and b. The 95^{th} percentile value of the Hausdorff distance distribution (d_{H95}) was used as the performance metric. Higher F-measure and IoU and lower d_{H95} values indicate better performance.

Hardware Platform. A computational cluster consisting of four nodes, each equipped with two Tesla V100-PCIE-16GB GPUs and a total memory capacity of 754 GBytes was used. Training of each brain projection, aligned with other U-Net variants for comparison, were conducted in parallel and independently, resulting in a significant reduction in the overall training time. The models were developed using Python 3.6 (Jupyter Notebook) and subsequently converted into Python scripts to facilitate execution on the cluster. The complete source code and the Keras-based models can be freely accessed at *https://github.com/KaueTND/WMH_CNS*.

E. Statistical Analysis

Five-fold cross-validation was used to evaluate the variability of our results. To assess generalizability across folds, we repeated this validation ten times. Details of the dataset division are provided in Section S2 of the Supplementary Material (Table S2 and Figure S2). We report mean and standard deviation summary statistics. Separate one-way analysis of variance (ANOVA) tests by U-Net variant, disease class and scanner were used to determine the statistical significant of differences in IoU, F-measure and d_{H95}. ANOVA tests were followed by *post hoc* two-sample t-tests with pooled variance (as appropriate). As we were interested only in a limited comparison of the Attention U-Net model against each of the three other models, only three pairwise comparisons were made with level of significance being appropriately adjusted for this number of multiple comparisons using the Holm-Bonferroni method. The level of statistical significance was set at $\alpha = 0.05$.

IV. Results

Fig. 1 shows example WMH segmentation masks obtained from each of the four U-Net variants by disease class. Both true and predicted lesion volume increased with worsening disease class (Table S3). The majority of voxels in each disease class were correctly classified by the four variants. FN voxels were more common than FP voxels (Table S5) indicating that the models tended to underestimate the true lesion. Overall, the Attention U-Net had the smallest error compared to the ground truth manual WMH segmentation than the other U-Net variants. Table I provides a comparison between our method and several published approaches, evaluated on the public dataset using the F-measure and Recall metrics. We report

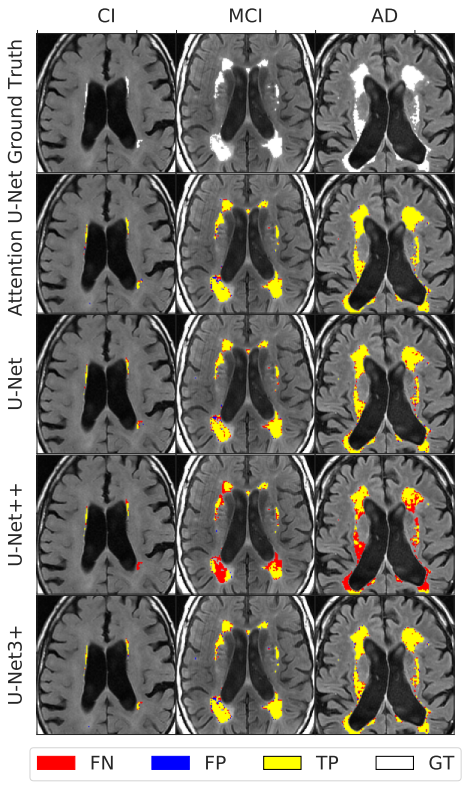


Fig. 1. Example segmentation results for each U-Net model at each disease stage (CI = cognitively normal, 74 year old female; MCI = mild cognitive impairment, 83 year old female; AD = Alzheimer's disease, 72 year old male). Results were obtained using a 2.5D implementation using the VGG16 feature extractor. Total ground truth and Attention U-Net predicted normalized WMH volume for these single images were CI: 0.27% *vs* 0.26% , MCI: 2.02% *vs* 2.06%, and AD: 3.35% *vs* 3.39%, respectively. Table S3 summarizes the ground truth-predicted nWMH averaged over all subjects by disease class.

the 10-times repeated 5-fold cross-validation, grouped by local and public dataset, in Table S8.

Introducing attention blocks before each pooling layer improved the F-measure compared to the other three U-Net variants evaluated (Fig. 2a). The U-Net++ model had the lowest F-measure scores, while the U-Net 3+ model performed second best in most individuals. These results when analyzed by image orientation, showed that sagittal extracted images performed better than axial or coronal images. However, the combined 2.5D model had the best F-measure score (Table S6). The F-measure Precision-Recall curve of both public and local datasets (Figure S5) showed the optimal F-measure score across different Precision and Recall thresholds. Similar performance patterns were found for IoU (Figure S3a and Table S7). The d_{H95} values were lowest for the Attention U-Net, followed by traditional U-Net and U-Net 3+ models (Figure S4a and Table S4). Notably, the U-Net++ model showed larger d_{H95} values with greater variability compared to other models.

Statistically significant differences in F-measure across the U-Net variants were found (One-way ANOVA: $F_{3,1276} = [61.38], p < 0.001$). Post-hoc t-tests confirmed that F-measure was significantly larger for the Attention U-Net compared to the three other U-Net variants (U-Net, $p_{corr} < 0.001$; U-Net++, $p_{corr} < 0.001$; and U-Net 3+, $p_{corr} < 0.001$). Similar statistically significant differences in IoU were found ($F_{3,1276} = [73.09], p < 0.001$). The mean IoU score was significantly higher in the Attention U-Net model compared to the other three models: U-Net ($p_{corr} < 0.001$), U-Net++ ($p_{corr} < 0.001$), and U-Net 3+ ($p_{corr} < 0.001$). Statistically significant differences in mean d_{H95} were also observed ($F_{3,1276} = [27.79], p < 0.001$). Significant differences in d_{H95} were only found between Attention U-Net and U-Net++ (t-test, $p_{corr} < 0.001$).

Fig. 2b shows F-measure values *vs* disease stage. F-measure, on average, increased with increasing disease severity for the majority of the U-Net variants. Attention U-Net models showed the best performance across all disease stages. The other U-Net variants were found to perform better in late disease stages. Similar findings by disease stage were found for IoU (Fig. S3b) and d_{H95} (Fig. S4b). No statistical differences were found by disease stage for F-measure ($F_{2,122} = [0.841]$, $p = 0.434$), IoU ($F_{2,122} = [1.355]$, $p = 0.262$), and d_{H95}, ($F_{2,122} = [0.471]$, $p = 0.625$).

Figs. 2c, S3c, and S4c show F-measure, IoU, and d_{H95} by scanner type, respectively. Significant differences in mean F-measure ($F_{4,284} = [5.44]$, $p < 0.001$), IoU score ($F_{4,284} = [5.586]$, $p < 0.001$) and d_{H95} ($F_{4,284} = [4.276]$, $p = 0.002$) were found across Scanners A-E. Holm-Bonferroni corrected pairwise t-test analysis found that the mean differences in all three metrics were localized between Scanner D and the other four scanners ($p_{corr} \leq 0.020$, IoU: $p_{corr} \leq 0.015$, and d_{H95}: $p_{corr} \leq 0.044$).

Fig. 3 summarizes graphically our main findings when using the Attention model and F-measure. Individuals with larger WMH burdens typically have more advanced disease class (MCI or AD). Within the CN, increasingly larger WMH lesion volumes were observed with older age. Larger WMH volumes had higher F-measure, suggesting that the Attention model performed better with larger disease burden.

TABLE I
COMPARISON OF PERFORMANCE BETWEEN ATTENTION U-NET AND EXISTING APPROACHES, EVALUATED ON THE PUBLIC DATASET [21]

Methods	F-Measure	Recall
Attention U-Net (ours)	0.82 ± 0.10	0.80 ± 0.11
PGS [19]	0.79	0.82
AQUA [20]	0.82 ± 0.02	0.75 ± 0.03
sysu media [21]	0.76 ± 0.03	0.84 ± 0.02
cian [21]	0.70 ± 0.02	0.83 ± 0.01
nlp logix [21]	0.78 ± 0.02	0.73 ± 0.03

V. DISCUSSION

Segmenting WMHs is a challenging task for several reasons including differences in the location of the WMH, their variable volume and irregular shape. Deep learning techniques have been proposed to overcome these problems and have sought to provide a more uniform and generalized response. In this study, we analyzed different U-Net variants including one that employed attention blocks [34] and demonstrated an effect on WMH segmentation performance.

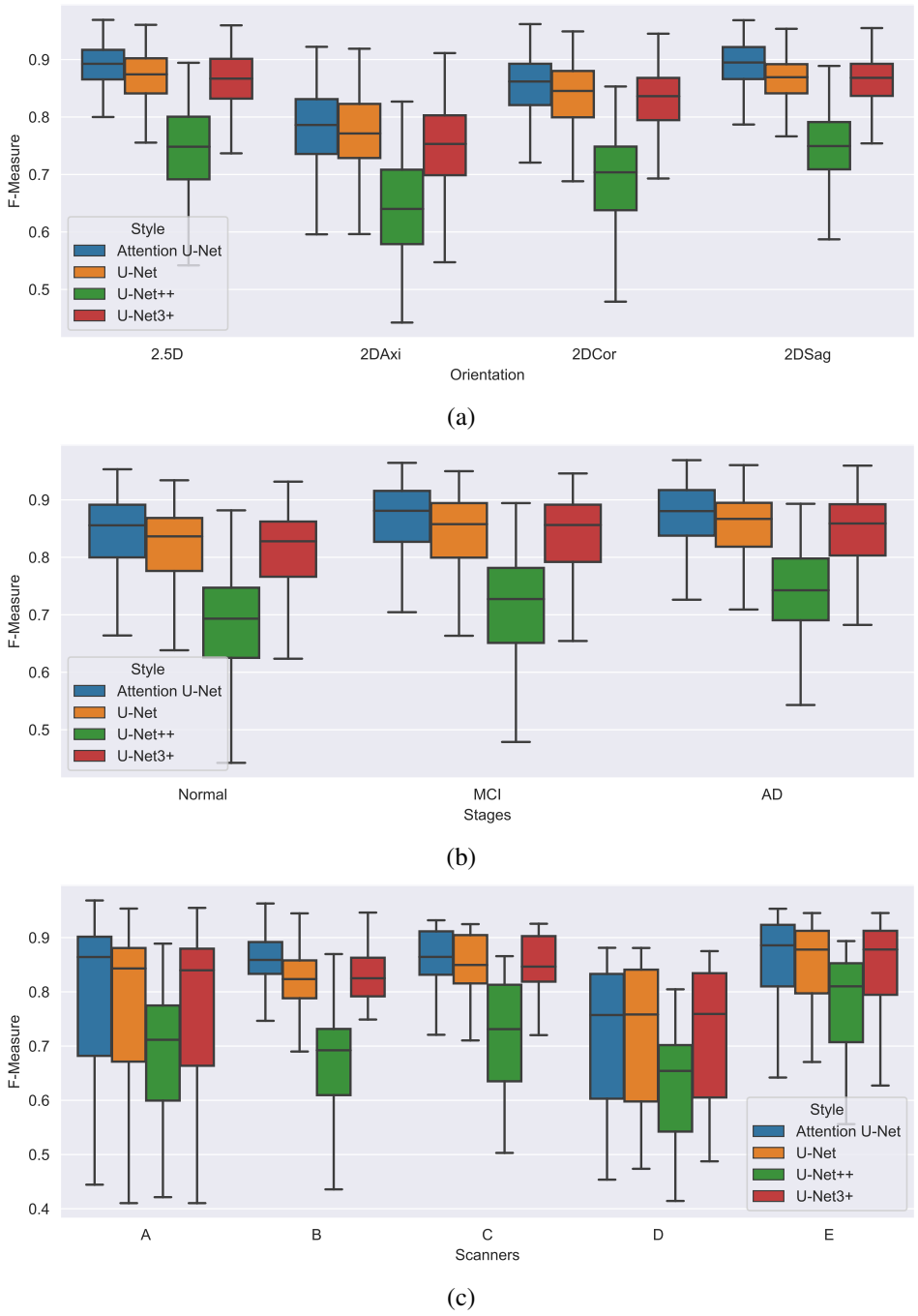


Fig. 2. Box plot of F-measure score across U-Net model by (a) input image orientation (Axi = axial, Cor = coronal, Sag = sagittal) and combined 2.5D model, (b) disease stage (CN = cognitively normal, MCI = mild cognitive impairment, AD = Alzheimer's disease), and (c) scanner (A-E, see description in text). Outliers have been suppressed to aid visualization.

Attention blocks allow combination of rich spatial features from the encoding part of the model and rich features from the decoding part of the model [34]. The traditional U-Net model performs a localized detection, but poorly delineates the boundaries of the WMHs [34], resulting in FP voxels (Fig. 1). Adding attention blocks to U-Net models reduced the false positive fraction (FPF) (Table S5) and improved the F-measure score (Fig. 2a and Table S6). The Attention U-Net model also performed significantly better with respect to IoU score (Fig. S3a, Table S7) and d_{H95} (Fig. S4a, Table S4).Generalization of deep learning model is a crucial aspect that evaluates the suitability for processing unseen data. Model developers need to consider robustness of the WMH prediction to other factors including disease stage (*e.g.*, CN *vs* MCI *vs* AD), the WMH size and location, and scanner and protocol used to acquire the FLAIR images.

In our analysis of the models by extracted image (axial, coronal, or sagittal), we noted a performance improvement for sagittal relative to the axial or coronal images across U-Net variations. However, it is important to recognize that this

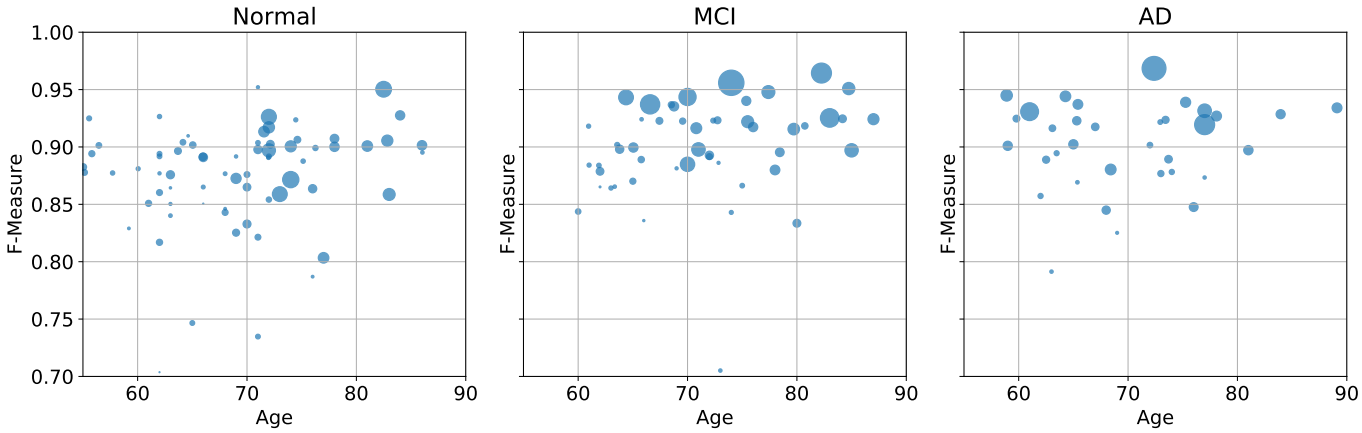


Fig. 3. Evaluation of F-measure score for the Attention U-Net model in 260 individual subjects over age grouped by disease stage: (a) CN = cognitively normal (148 individuals), (b) MCI = mild cognitive impairment (55 individuals), and (c) AD = Alzheimer's disease (37 individuals). Size of the filled circle in the plots reflects rank of the normalized WMH volume (*i.e.*, percentage of the inter-cranial volume). Plotted are data from the local datasets described in Table S1.

increase is perhaps explained by the increase in the number of sagittal slices compared to the two other orientations [18]. We adopted the best performing 2.5D model that combined the results from each of the three orthogonal projections. The 2.5D model performed better than the individual projection model (F-measure score, Fig. 2a). For this reason, we primarily evaluated our models using the 2.5D model.

Our results also found that the Attention U-Net model performance did not vary with disease class (*e.g.*, F-measure score, Fig. 2b). Other U-Net variants showed greater change in F-measure score when segmenting WMH in CN compared to MCI or AD subjects. These other models tended to perform better with increasing WMH burden, suggesting that their generalizability in healthy populations might be compromised.

While our results suggested that U-Net variants can achieve robustness to scanner differences when trained on sufficiently diverse multi-scanner datasets, this claim is not without caveats. For instance, Scanner D (contributing only 20/320, 6.3% of training data, Table S1) exhibited poorer performance across all models compared to other scanners (Figs. 2c, S3c, and S4c; Tables S6, S7, and S4). We expect that increasing the number of scanners and the number of example images per scanner dataset would lead to more consistent performance across different scanners/acquisition protocols. We found increased F-measure and Recall values when comparing our method with other published approaches (Table I), suggesting better performance for the attention block variant.

The main limitations of this study include the relatively small cohort size ($n = 320$ individuals) for training and validation, which may restrict the ability of the model to generalize to broader populations. Additionally, while our evaluation included multi-scanner data, the performance across distinct clinical imaging contexts (*e.g.*, WMHs due to multiple sclerosis *vs* due to stroke *vs* due to neurodegeneration, as studied here) remains poorly explored. Such differences in lesion characteristics and acquisition protocols across pathologies represent a common trade-off in artificial intelligence solutions in medical imaging and often require targeted validation in future applications. Another limitation was the reduced number of publicly available datasets with manually annotated ground truth data, despite the relative abundance of FLAIR scans.

VI. Conclusions

Attention blocks in U-Net models play a crucial role in improving the segmentation of WMH in FLAIR images by extracting additional information from the images to enhance the detection process. We showed that an Attention U-Net segmentation approach had statistically significant better performance (F-measure, $p_{corr} < 0.001$) compared to the other tested models. Similar findings were found with IoU metric ($p_{corr} < 0.001$). The comparison across U-Net variants showed that attention blocks reliably enhance the identification and segmentation of WMH by reducing FPF and FNF (Table S4). This improvement is likely due to the more precise delineation of WMH boundaries, where most FP and FN errors occur. These experiments demonstrated the advantages of incorporating attention mechanisms in improving identification and segmentation. The F-measure, IoU, and d_{H95} performance metrics of the Attention U-Net model were not influenced by disease class (Figs. 2b, 3, S3b, and S4b), in contrast to the results obtained from other U-Net variants.

Scanner differences (and the associated variations in acquisition protocol) were found to significantly affect performance with one scanner (Scanner D) performing worse than the other four scanners across all three performance measures (Figs. 2c, S3c, and S4c). This finding underscores a key challenge in deep learning-based WMH segmentation: Training models benefit from access to diverse datasets in order to seek both robustness and generalizability and, thus, potentially increase the likelihood of accurate segmentation on unseen data.

It is also important to recognize that WMH segmentation provides a complementary addition to other clinical information. Knowing the location and temporal progression of WMHs, for instance, can help form a more comprehensive understanding. Our future objective in utilizing Attention Block U-Net models is to harness their potential as a predictive

tool for conditions like AD, vascular dementia, and related neurodegenerative disorders. To achieve this goal, future work is need to expand dataset diversity, specifically including use of a range of scanners and acquisition protocols.

REFERENCES

[1] J. M. Wardlaw, M. C. V. Hernández, and S. Muñoz-Maniega, "What are white matter hyperintensities made of?" *Journal of the American Heart Association*, vol. 4, no. 6, p. e001140, 2015. [Online]. Available: https://doi.org/10.1161/JAHA.114.001140

[2] C. E. Bauer, V. Zachariou, E. Seago, and B. T. Gold, "White matter hyperintensity volume and location: Associations with wm microstructure, brain iron, and cerebral perfusion," *Frontiers in Aging Neuroscience*, vol. 13, 2021. [Online]. Available: https://doi.org/10.3389/fnagi.2021.617947

[3] J. M. Wardlaw, E. E. Smith, G. J. Biessels, C. Cordonnier, F. Fazekas, R. Frayne, R. I. Lindley, J. T. O'Brien, F. Barkhof, O. R. Benavente, S. E. Black, C. Brayne, M. Breteler, H. Chabriat, C. DeCarli, F.-E. de Leeuw, F. Doubal, M. Duering, N. C. Fox, S. Greenberg, V. Hachinski, I. Kilimann, V. Mok, R. van Oostenbrugge, L. Pantoni, O. Speck, B. C. M. Stephan, S. Teipel, A. Viswanathan, D. Werring, C. Chen, C. Smith, M. van Buchem, B. Norrving, P. B. Gorelick, and M. Dichgans, "Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration," *The Lancet Neurology*, vol. 12, no. 8, pp. 822–838, 2013. [Online]. Available: https://doi.org/10.1016/S1474-4422(13)70124-8

[4] C. Haffner, R. Malik, and M. Dichgans, "Genetic factors in cerebral small vessel disease and their impact on stroke and dementia," *Journal of Cerebral Blood Flow & Metabolism*, vol. 36, no. 1, pp. 158–171, 2016. [Online]. Available: https://doi.org/10.1038/jcbfm.2015.71

[5] K. T. N. Duarte, A. S. Sidhu, M. C. Barros, D. G. Gobbi, C. R. McCreary, F. Saad, R. Camicioli, E. E. Smith, M. P. Bento, and R. Frayne, "Multi-stage semi-supervised learning enhances white matter hyperintensity segmentation," *Frontiers in Computational Neuroscience*, vol. 18, 2024. [Online]. Available: https://doi.org/10.3389/fncom.2024.1487877

[6] K. Duarte, P. V. de Paiva, P. Martins, and M. Carvalho, "Predicting the early stages of the Alzheimer's disease via combined brain multi-projections and small datasets," in *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. SCITEPRESS - Science and Technology Publications, 2019. [Online]. Available: http://doi.org/10.5220/0007404705530560

[7] G. Tosto, M. E. Zimmerman, J. L. Hamilton, O. T. Carmichael, and A. M. Brickman, "The effect of white matter hyperintensities on neurodegeneration in mild cognitive impairment," *Alzheimer's and Dementia*, vol. 11, no. 12, pp. 1510–1519, Jun. 2015. [Online]. Available: https://doi.org/10.1016/j.jalz.2015.05.014

[8] S. Lee, F. Viqar, M. E. Zimmerman, A. Narkhede, G. Tosto, T. L. Benzinger, D. S. Marcus, A. M. Fagan, A. Goate, N. C. Fox, N. J. Cairns, D. M. Holtzman, V. Buckles, B. Ghetti, and E. M. et al, "White matter hyperintensities are a core feature of alzheimer's disease: Evidence from the dominantly inherited Alzheimer network," *Annals of Neurology*, vol. 79, no. 6, pp. 929–939, Apr. 2016. [Online]. Available: https://doi.org/10.1002/ana.24647

[9] E. E. Smith, M. O'Donnell, G. Dagenais, S. A. Lear, A. Wielgosz, M. Sharma, P. Poirier, G. Stotts, S. E. Black, S. Strother, M. D. Noseworthy, O. Benavente, J. Modi, M. Goyal, S. Batool, K. Sanchez, V. Hill, C. R. McCreary, R. Frayne, S. Islam, J. DeJesus, S. Rangarajan, K. Teo, S. Yusuf, and on behalf of the PURE Investigators, "Early cerebral small vessel disease and brain volume, cognition, and gait," *Annals of Neurology*, vol. 77, no. 2, pp. 251–261, 2015. [Online]. Available: https://doi.org/10.1002/ana.24320

[10] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2017. [Online]. Available: https://doi.org/10.1109/TPAMI.2016.2572683

[11] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-Net and its variants for medical image segmentation: A review of theory and applications," *IEEE Access*, vol. 9, pp. 82 031–82 057, 2021. [Online]. Available: https://doi.org/10.1109/ACCESS.2021.3086020

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241. [Online]. Available: https://doi.org/10.48550/arXiv.1505.04597

[13] B. Chen, Y. Liu, Z. Zhang, G. Lu, and A. W. K. Kong, "Transattunet: Multi-level attention-guided U-Net with transformer for medical image segmentation," 2022. [Online]. Available: https://doi.org/10.1109/TETCI.2023.3309626

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017. [Online]. Available: https://doi.org/10.48550/arXiv.1706.03762

[15] R. Heinen, M. D. Steenwijk, F. Barkhof, J. M. Biesbroek, W. M. van der Flier, H. J. Kuijf, N. D. Prins, H. Vrenken, G. J. Biessels, J. de Bresser, and TRACE-VCI study group, "Performance of five automated white matter hyperintensity segmentation methods in a multicenter dataset," *Scientific Reports*, vol. 9, no. 1, 2019. [Online]. Available: https://doi.org/10.1038/s41598-019-52966-0

[16] W. Zhu, H. Huang, Y. Zhou, F. Shi, H. Shen, R. Chen, R. Hua, W. Wang, S. Xu, and X. Luo, "Automatic segmentation of white matter hyperintensities in routine clinical brain MRI by 2D VB-Net: A large-scale study," *Frontiers in Aging Neuroscience*, vol. 14, 2022. [Online]. Available: https://doi.org/10.3389/fnagi.2022.915009

[17] S. Liu, X. Wu, S. He, X. Song, F. Shang, and X. Zhao, "Identification of white matter lesions in patients with acute ischemic lesions using U-Net," *Frontiers in Neurology*, vol. 11, 2020. [Online]. Available: https://doi.org/10.3389/fneur.2020.01008

[18] V. Sundaresan, G. Zamboni, P. M. Rothwell, M. Jenkinson, and L. Griffanti, "Triplanar ensemble U-Net model for white matter hyperintensities segmentation on MR images," *Medical Image Analysis*, vol. 73, p. 102184, 2021. [Online]. Available: https://doi.org/10.1016/j.media.2021.102184

[19] G. Park, J. Hong, B. A. Duffy, J.-M. Lee, and H. Kim, "White matter hyperintensities segmentation using the ensemble U-Net with multi-scale highlighting foregrounds," *NeuroImage*, vol. 237, p. 118140, 2021. [Online]. Available: https://doi.org/10.1016/j.neuroimage.2021.118140

[20] S. Lee, Z. Rieu, R. E. Kim, M. Lee, K. Yen, J. Yong, and D. Kim, "Automatic segmentation of white matter hyperintensities in T2-FLAIR with AQUA: A comparative validation study against conventional methods," *Brain Research Bulletin*, vol. 205, p. 110825, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0361923023002502

[21] H. J. Kuijf, A. Casamitjana, D. L. Collins, M. Dadar, A. Georgiou, and et al., "Standardized assessment of automatic segmentation of white matter hyperintensities and results of the WMH segmentation challenge," *IEEE Transactions on Medical Imaging*, vol. 38, pp. 2556–2568, 2019. [Online]. Available: https://doi.org/10.1109/tmi.2019.2905770

[22] K. T. Duarte, D. G. Gobbi, A. S. Sidhu, C. R. McCreary, F. Saad, R. Camicioli, E. E. Smith, and R. Frayne, "Segmenting white matter hyperintensities in brain magnetic resonance images using convolution neural networks," *Pattern Recognition Letters*, vol. 175, pp. 90–94, 2023. [Online]. Available: https://doi.org/10.1016/j.patrec.2023.07.014

[23] K. T. Duarte, D. G. Gobbi, A. S. Sidhu, C. R. McCreary, F. Saad, N. Das, E. E. Smith, and R. Frayne, "Segmenting white matter hyperintensity in Alzheimer's disease using U-Net cnns," in *2022 35th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, vol. 1, 2022, pp. 109–114. [Online]. Available: https://doi.org/10.1109/SIBGRAPI55357.2022.9991752

[24] C. R. McCreary, M. Salluzzi, L. B. Andersen, D. Gobbi, L. Lauzon, F. Saad, E. E. Smith, and R. Frayne, "Calgary Normative Study: Design of a prospective longitudinal study to characterise potential quantitative MR biomarkers of neurodegeneration over the adult lifespan," *BMJ Open*, vol. 10, no. 8, 2020. [Online]. Available: https://doi.org/10.1136/bmjopen-2020-038120

[25] S. Peca, C. R. McCreary, E. Donaldson, G. Kumarpillai, N. Shobha, and et al, "Neurovascular decoupling is associated with severity of cerebral amyloid angiopathy," *Neurology*, vol. 81, no. 19, pp. 1659–1665, 2013. [Online]. Available: https://doi.org/10.1212/01.wnl.0000435291.49598.54

[26] M. S. Albert, S. T. DeKosky, D. Dickson, B. Dubois, H. H. Feldman, N. C. Fox, A. Gamst, D. M. Holtzman, W. J. Jagust, R. C. Petersen, P. J. Snyder, M. C. Carrillo, B. Thies, and C. H. Phelps, "The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimer's & Dementia*, vol. 7, no. 3, pp. 270–279, 2011. [Online]. Available: https://doi.org/10.1016/j.jalz.2011.03.008

[27] A. Subotic, C. McCreary, F. Saad, A. Nguyen, A. Alvarez-Veronesi, and et al, "Cortical thickness and its association with clinical cognitive

and neuroimaging markers in cerebral amyloid angiopathy," *Journal of Alzheimer's Disease*, vol. 81, pp. 1–9, 05 2021. [Online]. Available: https://doi.org/10.3233/JAD-210138

[28] D. Gobbi, Q. Lu, R. Frayne, and M. Salluzzi, "Cerebra-WML: A rapid workflow for quantification of white matter hyperintensities," *Canadian Stroke Congress*, vol. 40, pp. E128–E129, 2012. [Online]. Available: https://www.researchgate.net/publication/278294844_Cerebra-WML_a_rapid_workflow_for_quantification_of_white_matter_hyperintensities

[29] J. C. Kosior, S. Idris, D. Dowlatshahi, M. Alzawahmah, M. Eesa, and et al, "Quantomo: Validation of a computer-assisted methodology for the volumetric analysis of intracerebral haemorrhage," *International Journal of Stroke*, vol. 6, no. 4, pp. 302–305, 2011. [Online]. Available: https://doi.org/10.1111/j.1747-4949.2010.00579.x

[30] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee, "N4ITK: improved N3 bias correction," *IEEE Trans. Medical Imaging*, vol. 29, no. 6, pp. 1310–1320, 2010. [Online]. Available: https://doi.org/10.1109/TMI.2010.2046908

[31] K. T. N. Duarte, M. C. de Barros, D. G. Gobbi, A. S. Sidhu, M. A. G. de Carvalho, and R. Frayne, "Changes in 3D radiomic texture descriptors in Alzheimer's disease stages," in *18th International Symposium on Medical Information Processing and Analysis*, J. Brieva, P. Guevara, N. Lepore, M. G. Linguraru, L. Rittner, and E. R. C. M.D., Eds., vol. 12567, International Society for Optics and Photonics. SPIE, 2023, p. 125670S. [Online]. Available: https://doi.org/10.1117/12.2670246

[32] K. T. N. Duarte, D. G. Gobbi, R. Frayne, and M. A. G. de Carvalho, "Detecting Alzheimer's disease based on structural region analysis using a 3D shape descriptor," in *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2020, pp. 180–187. [Online]. Available: https://doi.org/10.1109/SIBGRAPI51738.2020.00032

[33] R. Guerrero, C. Qin, O. Oktay, C. Bowles, L. Chen, and et al., "White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks," *NeuroImage: Clinical*, vol. 17, pp. 918–934, 2018. [Online]. Available: https://doi.org/10.1016/j.nicl.2017.12.022

[34] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning where to look for the pancreas," 2018. [Online]. Available: https://doi.org/10.48550/arXiv.1804.03999

[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv 1409.1556*, 09 2014. [Online]. Available: https://doi.org/10.48550/arXiv.1409.1556

[36] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, D. Stoyanov, Z. Taylor, G. Carneiro, T. Syeda-Mahmood, A. Martel, L. Maier-Hein, J. M. R. Tavares, A. Bradley, J. P. Papa, V. Belagiannis, J. C. Nascimento, Z. Lu, S. Conjeti, M. Moradi, H. Greenspan, and A. Madabhushi, Eds. Cham: Springer International Publishing, 2018, pp. 3–11. [Online]. Available: https://doi.org/10.48550/arXiv.1807.10165

[37] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "Unet 3+: A full-scale connected UNet for medical image segmentation," 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2004.08790

Kaue T. N. Duarte holds a B.Sc. degree in System Analysis and Technology from the University of Campinas in São Paulo, Brazil (2014). His M.Sc. (2017) in texture analysis of natural images and Ph.D. (2021) in Alzheimer's disease prognosis using image retrieval and AI were earned at the same university. He is currently a Postdoctoral Fellow at University of Calgary in Alberta, Canada and working on identifying potential biomarkers of WMH using AI algorithms.

Murilo C Barros holds a B.Sc. degree from the University of Campinas. His undergaduate thesis was on using X-band electromagnetic waves to image ceramic blocks. He holds a M.Sc. from the University of Campinas in Information and Communication Systems, for a disertain on predicting Tourette syndrome using traditional machine learning techniques. He is currently a Ph.D. candidate at the University of Campinas working on applying CNNs to the study of Tourette syndrome. He completed an internship at National Taiwan University.

Abhijot S Sidhu holds a Bachelor of Health Sciences Degree (Honours) from the University of Calgary, with emphasis in Biomedical Science and Nanoscience. He obtained his M.Sc. in Biomedical Engineering investigating physiological brain changes in healthyaging using functional magnetic resonance images. He is currently pursuing a Ph.D. in Biomedical Engineering at the Seaman Centre.

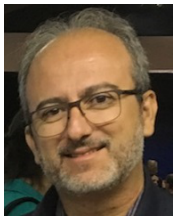
David G Gobbi is a research scientist at the University of Calgary with a diverse background that includes work in medical image registration, intra-operative imaging, quantitative MRI processing methods, and image database management. He currently leads software development at the Calgary Image Processing and Analysis Centre. David holds an MSc in Physics from Carleton University in Ottawa and a PhD in Medical Biophysics from Western University in London, Ontario.

Cheryl R McCreary has an extensive background in adopting imaging techniques to tissues, animal models and people. These techniques include imaging of white matter to evaluate myelin in a murine model of multiple sclerosis, and evaluating biomarkers of neurodegeneration associated with cerebral small vessel disease and aging. She is an imaging research scientist and MR research manager.

Feryal Saad earned her medical school degree at the Damascus University and pursued training in radiology at the University of Henri-Poincare and Nancy University Hospitals in France. She is a former director and consultant radiologist at Safita Medical Centre-a polyclinic in radiology, cardiology, and digestive specialties. She worked as a consultant radiologist at the Dammam Central Hospital and Dammam Medical Complex Tower in Saudi Arabia.

Richard Camicioli worked in engineering and medicinal chemistry prior to obtaining his MD CM, from McGill University where he also completed a neurologic residency. He completed a fellowship in geriatric neurology at Oregon Health and Sciences University (1994), joined the University of Alberta, Canada as an associate professor (2000) and became full professor (2008). Dr. Camicioli's research interests include cognitive dysfunction and motor dysfunction.

Eric E Smith is Professor of Radiology, Neurology, and Community Health Sciences at the University of Calgary. He is the Medical Director of the Cognitive Neurosciences Clinic and a member of the Calgary Stroke Program. He runs the Clinical and Research Fellowship program in Cognitive Neurosciences. He graduated from McGill University, trained in Neurology in teaching hospitals of Harvard Medical School, and was Assistant Professor of Neurology at Harvard University, before joining UofC.



Marco AG Carvalho holds a B.Sc. in Electrical Engineering (Universidade Federal do Rio Grande do Norte, Brazil, 1994), a M.Sc. in image processing at the School of Electrical and Computer Engineering (University of Campinas, Brazil, 1997) and a Ph.D. degree in image processing at School of Electrical and Computer Engineering (University of Campinas, Brazil, 2004). His main contributions are in areas of image processing, architectural segmentation, and image feature understanding.

Richard Frayne is a tenured Professor at University of Calgary (Departments of Radiology and Clinical Neurosciences), and the Deputy Director of the Hotchkiss Brain Institute (HBI). In addition, he is a member of the Libin Cardiovascular Institute, all at the Cumming School of Medicine. He directs the Vascular Imaging Laboratory group. He was the Scientific Director of the Seaman Centre (2010-7). Over 2003-13, he held a Canada Research Chair in Image Science, as well as over 2010-14, the Hopewell Professorship in Brain Imaging.

SUPPLEMENTARY MATERIAL

Supplementary material is provided to support the methodology and findings of the main manuscript. The supplementary figures provide detailed specifications of the architecture and summarize additional results of the IoU and d_{H95} metrics. Supplementary tables support the interpretation of the figures in the main paper, as well as summarizing individual demographics including by fold.

VII. SUPPLEMENTARY METHODOLOGY

Tabular data are reported by disease level (cognitively normal = CN, mild cognitive impairment = MCI and Alzheimer's disease = AD) and for the public dataset. Results shown in the Supplementary material are calculated over voxels containing brain tissue. Normalized WMH (nMWH) values report the fraction of the intracranial (or brain) volume. Provided are the mean $\pm$ standard deviation value calculated across each each subject.

To ensure robustness and generalizability, we adopted a stratified five-fold cross-validation across ten independent trials (see details in main paper). For each trial, the dataset was partitioned into 70% training, 10% validation, and 20% testing, preserving the distribution of scanners (Scanner A-E), disease classes (CN/MCI/AD), and cohorts (local/public) in all splits. This stratification strategy attempted to mitigate biases arising from imbalanced representation (*e.g.*, scanner type or vendor, disease stage, or dataset).

VIII. SUPPLEMENTARY RESULTS

Fig. 4 summarizes the four implemented U-Net architectures (Attention U-Net, Traditional U-Net, U-Net++, and U-Net3+) used in this work.

Table II describes the CN, MCI and AD cohorts within the local dataset ($n = 260$) as well as within the public dataset group ($n = 60$). The pooled data ($n = 320$) in each supplementary table includes both the local dataset and the public dataset subjects. Note: Sex, age and disease class (CN/MCI/AD) information was not provided for the public datasets.

Table III summarizes the number of individuals allocated to training, validation, and testing from each dataset. Participants were permutated across folds while maintaining the stratified proportions, ensuring that all models were evaluated on distinct subsets of the data in each trial. Fig. 5 visualizes the age distributions across the data splits and demonstrates consistent stratification patterns throughout the cross-validation process.

Table IV summarizes the ground truth nWMH volumes and the nWMH values measured by the four U-Net variants. Table VI summarizes the average false negative (FNF), false positive (FPF) and true positive (TPF) fractions by disease level and U-Net model variant.

Table VII summarizes the F-measure measurements by disease level and in the public and pooled data for each U-Net variant. Fig. 6 and Table VIII similarly summarize the IoU measurements. Fig. 7 and Table V summarize the d_{H95} metrics.

Table IX reports the average F-measure for each trial.

Fig. 8 presents the F-measure Precision-Recall plots for the public and private datasets.

TABLE II
PUBLIC AND PRIVATE DATASET DEMOGRAPHICS

Dataset [Ref]	***N***	**Scanner** (see text)	**Sex**	**Age** (years)	**Subject Breakdown** (CI-MCI-AD)
Local Dataset					
CNS_1	74	A	Male (50.%)	31.6 ± 4.5	
			Female (50.0%)	31.7 ± 4.4	(74-0-0)
			All (100.0%)	31.6 ± 4.4	
CNS_2	20	A	Male (55.0%)	44.0 ± 17.0	
			Female (45.0%)	43.2 ± 16.6	(20-0-0)
			All (100.0%)	43.7 ± 17.3	
FAVR-I	71	A	Male (52.1%)	70.9 ± 7.7	
			Female (47.9%)	68.4 ± 9.0	(24-29-18)
			All (100.0%)	69.7 ± 8.3	
FAVR-II	65	A	Male (57.1%)	70.3 ± 6.4	
	30	B	Female (42.9%)	71.0 ± 7.5	(50-26-19)
	95		All (100.0%)	70.7 ± 6.9	
Total	260		Male (53.5%)	57.4 ± 7.0	
			Female (46.5%)	57.0 ± 7.7	(148-55-37)
			All (100.0%)	57.2 ± 7.3	
Public Dataset					
AMS	20	C	N/A*	N/A*	N/A*
SIN	20	D	N/A*	N/A*	N/A*
Utrecht	20	E	N/A*	N/A*	N/A*
Total	60				

Reported are count (percentage) or mean $\pm$ standard deviation.
*Distribution of disease class, sex and age was not reported for the AMS, SIN and Utrecht datasets, though no significant differences by site for age ($p = 0.45$) or sex ($p = 0.87$) were reported. AD = Alzheimer's disease, CN = cognitively intact, CNS = Calgary Normative Study, FAVR = Functional Assessment of Vascular Reactivity. MCI = mild cognitive impairment, N/A = Not available.

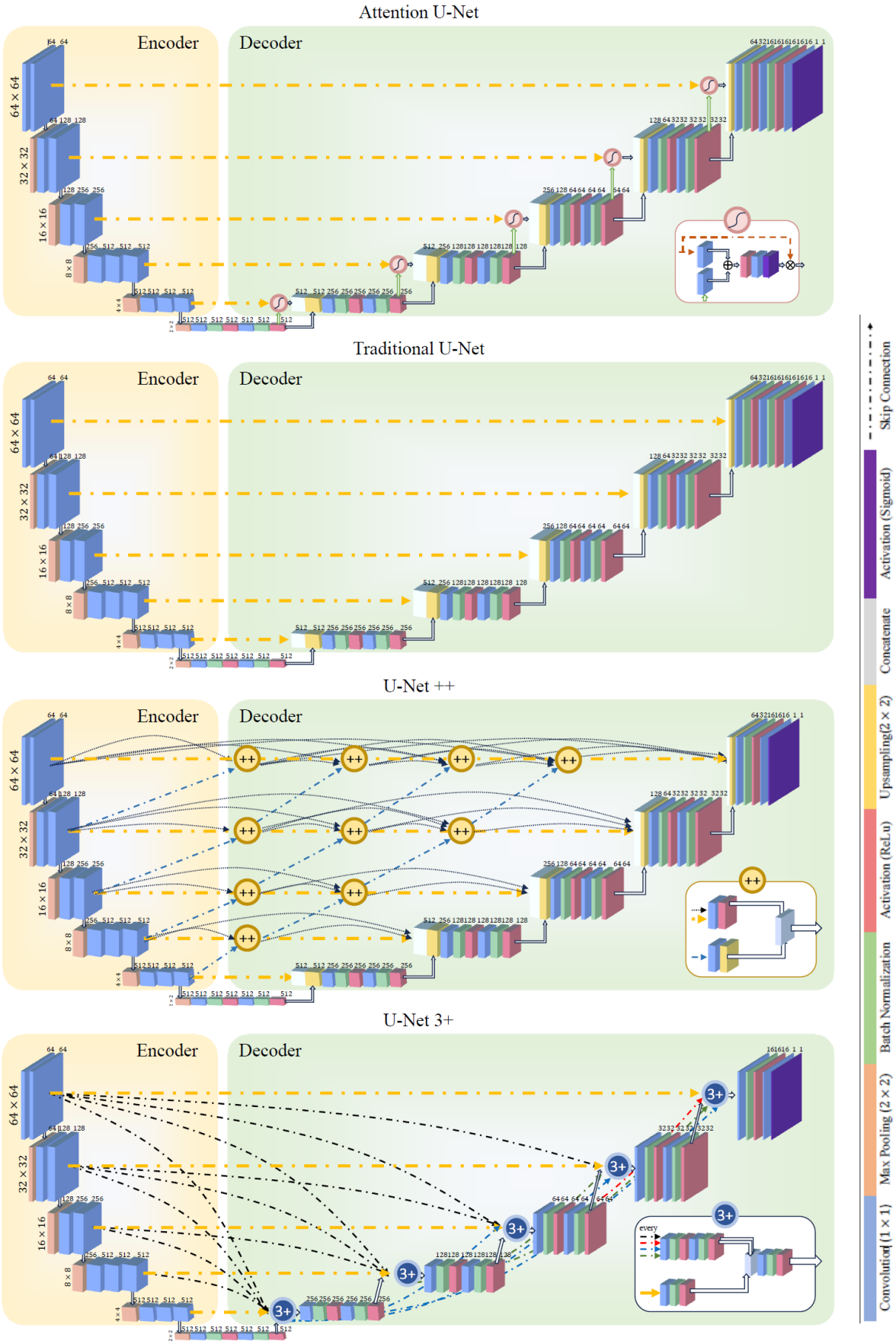
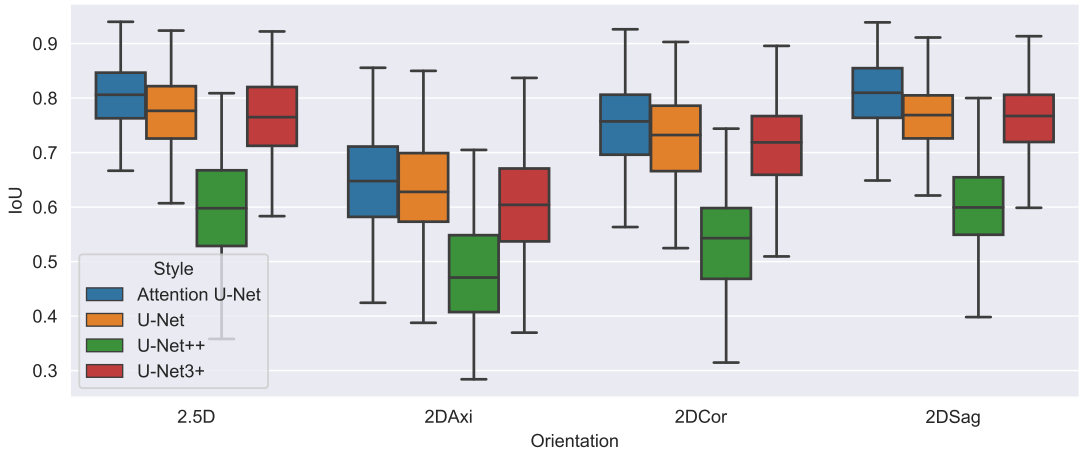
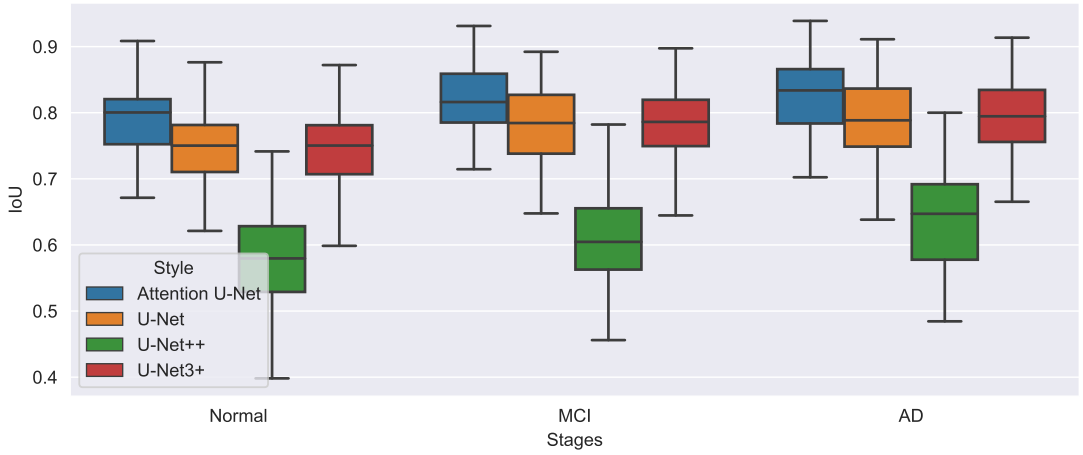


Fig. 4. Architecture of individual U-Net variations used in this work.

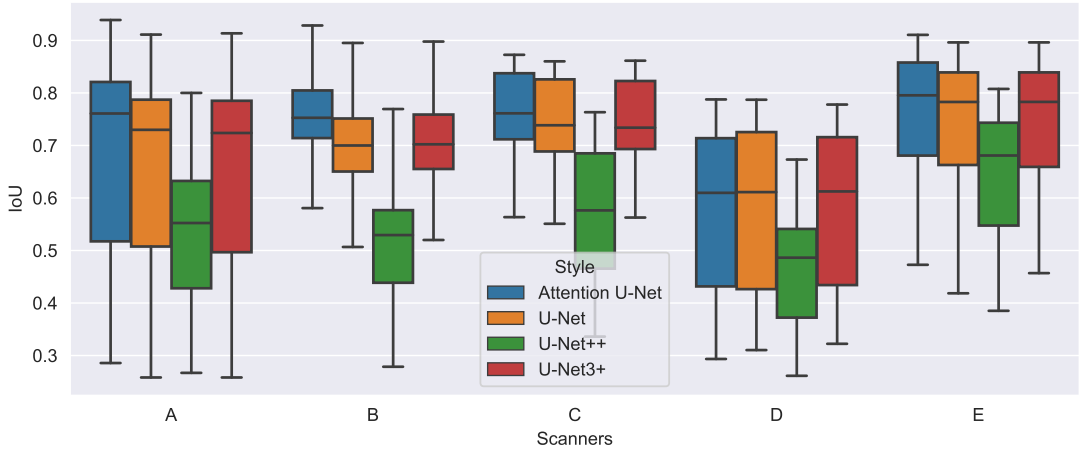
Fig. 5. Age distribution across cross-validation splits for local datasets. The grid of plots shows ten independent trials (rows) and five cross-validation folds (columns). For each fold, the test set corresponds to the held-out fold (20% of data), while the training (70%) and validation (10%) sets were derived from the remaining data. Boxplots show the distribution of participant ages within each split (colors indicate training (blue), validation (orange), and test (green) sets). Public dataset ages are omitted as the age data were not available.



(a)

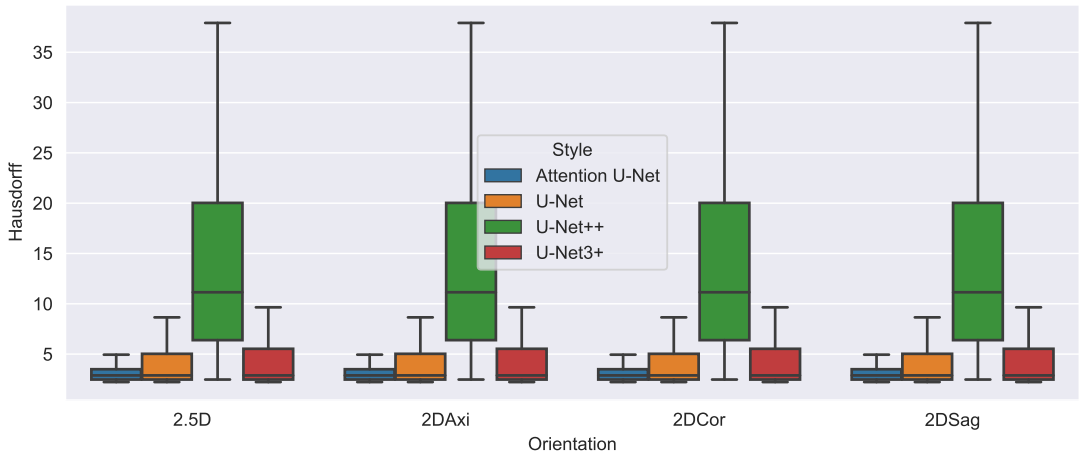


(b)

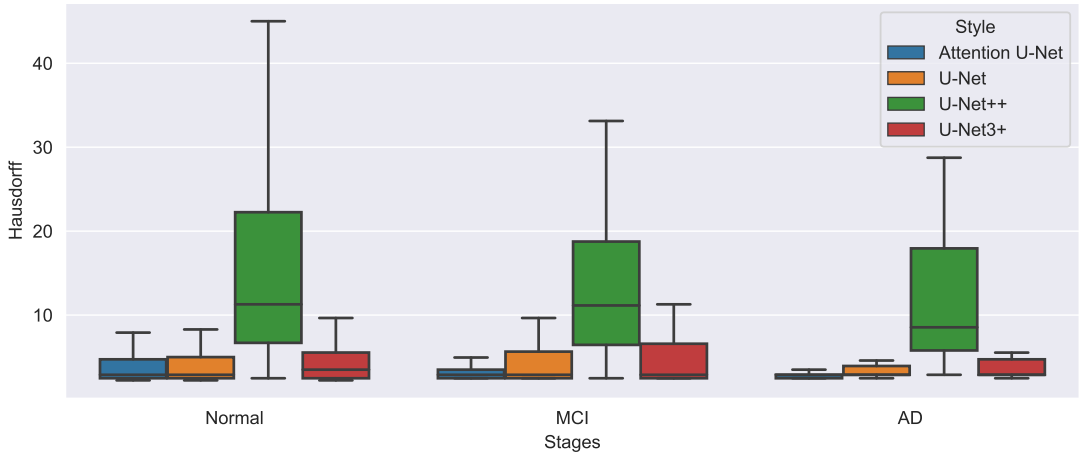


(c)

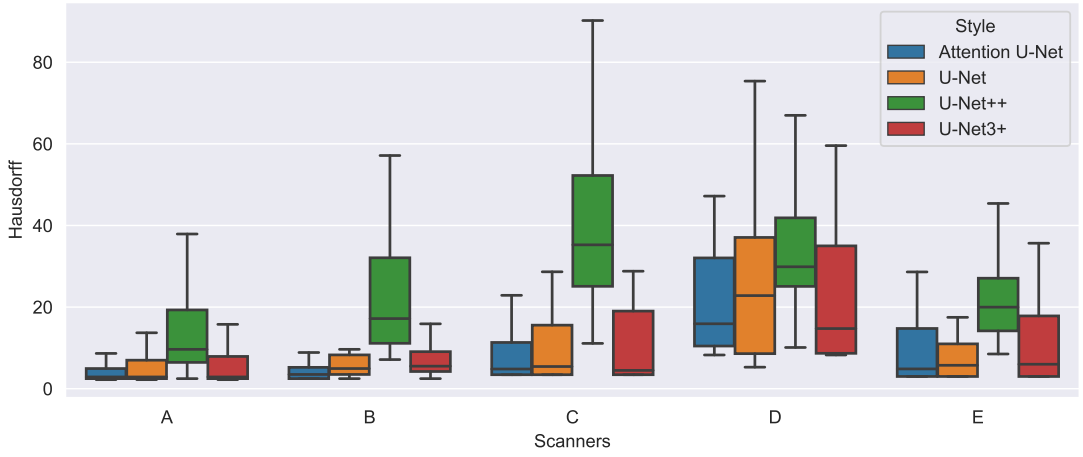
Fig. 6. Box plot of IoU score across U-Net variant, by (a) input image orientation (Axi = axial, Cor = coronal, Sag = sagittal) and combined 2.5D model, (b) disease stage (CN = cognitively intact, MCI = mild cognitive impairment, AD = Alzheimer's disease), and (c) scanner (A-E, see description in main paper). Outliers have been suppressed to aid visualization.



(a)



(b)



(c)

Fig. 7. Box plot of Hausdorff 95% percentile distance (d_{H95}) across U-Net variant, by (a) input image orientation (Axi = axial, Cor = coronal, Sag = sagittal) and combined 2.5D model, (b) disease stage (CN = cognitively intact, MCI = mild cognitive impairment, AD = Alzheimer's disease), and (c) scanner (A-E, see description in main paper). Outliers have been suppressed to aid visualization.

TABLE III
DATASET DIVISION FOR STRATIFIED FIVE-FOLD CROSS-VALIDATION. VALUES REPRESENT THE NUMBER OF SUBJECTS PER SPLIT. (TRAIN-70%,VALIDATION-10%,TEST-20%)

Dataset	N	Train/Val/Test
CNS (local)	94	66/9/19
FAVR-I (local)	71	50/7/14
FAVR-II (local)	95	67/9/19
AMS (public)	20	14/2/4
Utrecht (public)	20	14/2/4
SIN (public)	20	14/2/4
Total	320	225/31/64

TABLE IV
NORMALIZED WMH VOLUME (NWMH) (MEAN ± STANDARD DEVIATION) BY DISEASE LEVEL FOR THE GROUND TRUTH AND OBTAINED BY EACH U-NET VARIANT. ALSO SHOWN ARE RESULTS OF PUBLIC DATASET AND POOLED RESULTS

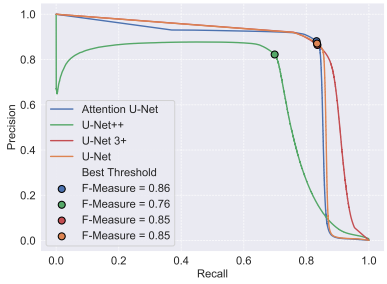
Level	Ground Truth (%)	Attention U-Net (%)
CN	0.35 ± 0.37	0.32 ± 0.34
MCI	0.59 ± 0.74	0.56 ± 0.72
AD	0.58 ± 0.70	0.56 ± 0.68
Public	1.16 ± 1.13	1.04 ± 0.97
Pooled	0.67 ± 0.84	0.61 ± 0.75

Level	Traditional U-Net (%)	U-Net++ (%)
CN	0.31 ± 0.33	0.24 ± 0.26
MCI	0.56 ± 0.72	0.47 ± 0.63
AD	0.55 ± 0.68	0.48 ± 0.62
Public	1.03 ± 0.95	0.87 ± 0.83
Pooled	0.61 ± 0.75	0.51 ± 0.65

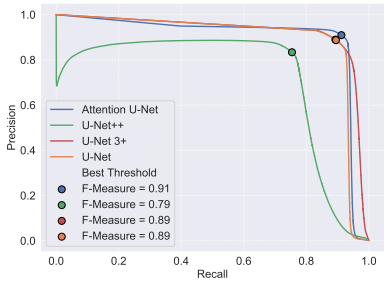
Level	U-Net3+ (%)	(%)
CN	0.31 ± 0.33	
MCI	0.56 ± 0.72	
AD	0.55 ± 0.68	
Public	1.02 ± 0.94	
Pooled	0.61 ± 0.74	

TABLE V
AVERAGE HAUSDORFF 95% DISTANCE (d_{H95}, MEAN ± STANDARD DEVIATION) BY DISEASE LEVEL FOR EACH U-NET VARIANT. ALSO SHOWN ARE RESULTS OF PUBLIC DATASET AND POOLED RESULTS

Level	Attention U-Net (mm)	U-Net (mm)	U-Net++ (mm)	U-Net 3+ (mm)
CN	4.6 ± 4.2	6.2 ± 10.5	16.4 ± 13.9	6.7 ± 9.6
MCI	4.4 ± 5.2	4.7 ± 3.2	13.8 ± 9.3	5.6 ± 6.3
AD	3.3 ± 1.6	3.9 ± 2.1	13.6 ± 12.3	4.0 ± 2.1
Public	13.7 ± 12.2	15.0 ± 15.2	32.1 ± 19.9	15.6 ± 15.8
Pooled	6.9 ± 8.4	7.9 ± 11.0	20.4 ± 17.8	8.5 ± 11.3



(a)



(b)

Fig. 8. Precision-Recall curve by U-Net variant for (a) public and (b) private datasets.

TABLE VI
AVERAGE FALSE NEGATIVE (FNF), FALSE POSITIVE (FPF) AND TRUE POSITIVE (TPF) FRACTIONS (MEAN ± STANDARD DEVIATION) BY DISEASE LEVEL FOR EACH U-NET VARIANT. ALSO SHOWN ARE RESULTS OF PUBLIC DATASET AND POOLED RESULTS. NOTE: THE DATASET IS IMBALANCED AND IN ALL CASES THE TRUE NEGATIVE FRACTION (TNF) EXCEEDS 98%

	Attention U-Net		
Level	FNF (%)	FPF (%)	TPF (%)
CN	0.05 ± 0.05	0.03 ± 0.03	0.30 ± 0.33
MCI	0.05 ± 0.05	0.04 ± 0.04	0.54 ± 0.70
AD	0.05 ± 0.05	0.04 ± 0.03	0.53 ± 0.66
Public	0.21 ± 0.28	0.11 ± 0.12	0.95 ± 0.92
Pooled	0.09 ± 0.17	0.06 ± 0.08	0.57 ± 0.72
	Traditional U-Net		
Level	FNF (%)	FPF (%)	TPF (%)
CN	0.05 ± 0.06	0.04 ± 0.04	0.29 ± 0.32
MCI	0.06 ± 0.06	0.05 ± 0.05	0.53 ± 0.69
AD	0.06 ± 0.06	0.05 ± 0.04	0.52 ± 0.65
Public	0.21 ± 0.27	0.12 ± 0.13	0.95 ± 0.92
Pooled	0.10 ± 0.17	0.07 ± 0.08	0.57 ± 0.71
	U-Net++		
Level	FNF (%)	FPF (%)	TPF (%)
Normal	0.12 ± 0.14	0.05 ± 0.04	0.23 ± 0.25
MCI	0.15 ± 0.18	0.07 ± 0.07	0.43 ± 0.59
AD	0.14 ± 0.14	0.07 ± 0.06	0.44 ± 0.57
Public	0.37 ± 0.42	0.15 ± 0.15	0.80 ± 0.79
Pooled	0.20 ± 0.27	0.08 ± 0.10	0.47 ± 0.61
	U-Net3+		
Level	FNF (%)	FPF (%)	TPF (%)
Normal	0.05 ± 0.06	0.04 ± 0.04	0.29 ± 0.32
MCI	0.06 ± 0.06	0.06 ± 0.06	0.53 ± 0.69
AD	0.06 ± 0.06	0.05 ± 0.05	0.52 ± 0.65
Public	0.21 ± 0.28	0.12 ± 0.12	0.94 ± 0.91
Pooled	0.10 ± 0.17	0.07 ± 0.08	0.57 ± 0.71

TABLE VII
AVERAGE F-MEASURE (MEAN ± STANDARD DEVIATION) BY DISEASE LEVEL FOR EACH U-NET VARIANT. ALSO SHOWN ARE RESULTS OF PUBLIC DATASET AND POOLED RESULTS

	Attention U-Net				U-Net			
Level	2D Axial	2D Coronal	2D Sagittal	2.5D	2D Axial	2D Coronal	2D Sagittal	2.5D
CI	0.76 ± 0.07	0.83 ± 0.06	0.88 ± 0.05	0.87 ± 0.04	0.75 ± 0.07	0.82 ± 0.06	0.85 ± 0.05	0.85 ± 0.05
MCI	0.79 ± 0.09	0.85 ± 0.09	0.89 ± 0.07	0.88 ± 0.08	0.77 ± 0.10	0.83 ± 0.09	0.87 ± 0.06	0.87 ± 0.09
AD	0.79 ± 0.06	0.87 ± 0.04	0.90 ± 0.04	0.90 ± 0.03	0.79 ± 0.06	0.86 ± 0.05	0.88 ± 0.04	0.89 ± 0.04
Public	0.78 ± 0.11	0.78 ± 0.11	0.81 ± 0.12	0.82 ± 0.10	0.77 ± 0.11	0.78 ± 0.11	0.81 ± 0.11	0.82 ± 0.10
Pooled	0.78 ± 0.09	0.83 ± 0.09	0.86 ± 0.08	0.87 ± 0.08	0.77 ± 0.09	0.82 ± 0.09	0.85 ± 0.08	0.85 ± 0.08
	U-Net++				U-Net3+			
Level	2D Axial	2D Coronal	2D Sagittal	2.5D	2D Axial	2D Coronal	2D Sagittal	2.5D
CN	0.62 ± 0.08	0.67 ± 0.08	0.72 ± 0.08	0.71 ± 0.09	0.73 ± 0.08	0.81 ± 0.06	0.85 ± 0.04	0.85 ± 0.05
MCI	0.65 ± 0.09	0.70 ± 0.09	0.76 ± 0.07	0.75 ± 0.09	0.76 ± 0.09	0.83 ± 0.09	0.86 ± 0.07	0.86 ± 0.08
AD	0.67 ± 0.08	0.72 ± 0.06	0.77 ± 0.06	0.78 ± 0.06	0.77 ± 0.07	0.85 ± 0.04	0.88 ± 0.04	0.88 ± 0.04
Public	0.68 ± 0.10	0.67 ± 0.11	0.71 ± 0.12	0.72 ± 0.11	0.76 ± 0.12	0.77 ± 0.11	0.80 ± 0.11	0.82 ± 0.09
Pooled	0.65 ± 0.09	0.68 ± 0.09	0.73 ± 0.09	0.73 ± 0.10	0.75 ± 0.09	0.81 ± 0.09	0.84 ± 0.08	0.85 ± 0.07

TABLE VIII
AVERAGE IOU (MEAN ± STANDARD DEVIATION) BY DISEASE LEVEL FOR EACH U-NET VARIANT. ALSO SHOWN ARE RESULTS OF PUBLIC DATASET AND POOLED RESULTS

	Attention U-Net				U-Net			
Level	2D Axial	2D Coronal	2D Sagittal	2.5D	2D Axial	2D Coronal	2D Sagittal	2.5D
CN	0.62 ± 0.09	0.72 ± 0.08	0.78 ± 0.07	0.78 ± 0.07	0.60 ± 0.09	0.69 ± 0.09	0.74 ± 0.07	0.75 ± 0.07
MCI	0.66 ± 0.12	0.75 ± 0.12	0.81 ± 0.10	0.80 ± 0.11	0.64 ± 0.12	0.73 ± 0.12	0.77 ± 0.09	0.77 ± 0.11
AD	0.66 ± 0.09	0.78 ± 0.06	0.82 ± 0.06	0.82 ± 0.06	0.65 ± 0.08	0.75 ± 0.07	0.78 ± 0.06	0.80 ± 0.06
Public	0.65 ± 0.14	0.65 ± 0.14	0.70 ± 0.15	0.71 ± 0.13	0.64 ± 0.14	0.65 ± 0.14	0.69 ± 0.15	0.70 ± 0.13
Pooled	0.64 ± 0.11	0.72 ± 0.12	0.77 ± 0.11	0.77 ± 0.11	0.63 ± 0.11	0.70 ± 0.11	0.74 ± 0.11	0.75 ± 0.10
	U-Net++				U-Net3+			
Level	2D Axial	2D Coronal	2D Sagittal	2.5D	2D Axial	2D Coronal	2D Sagittal	2.5D
CN	0.45 ± 0.09	0.51 ± 0.09	0.57 ± 0.09	0.56 ± 0.10	0.58 ± 0.09	0.68 ± 0.08	0.74 ± 0.07	0.74 ± 0.07
MCI	0.48 ± 0.10	0.54 ± 0.11	0.61 ± 0.09	0.61 ± 0.11	0.62 ± 0.11	0.71 ± 0.12	0.77 ± 0.09	0.76 ± 0.11
AD	0.51 ± 0.09	0.57 ± 0.07	0.63 ± 0.08	0.64 ± 0.08	0.63 ± 0.09	0.74 ± 0.07	0.78 ± 0.07	0.79 ± 0.06
Public	0.52 ± 0.11	0.51 ± 0.12	0.56 ± 0.14	0.57 ± 0.13	0.63 ± 0.14	0.64 ± 0.14	0.69 ± 0.15	0.70 ± 0.13
Pooled	0.49 ± 0.10	0.53 ± 0.10	0.59 ± 0.11	0.59 ± 0.11	0.61 ± 0.11	0.69 ± 0.11	0.74 ± 0.11	0.74 ± 0.10

TABLE IX
AVERAGE CLASSIFICATION F-MEASURE (%, MEAN ± STANDARD DEVIATION) FOR EACH OF TEN TRIALS OBTAINED USING TEN-TIMES REPEATED FIVE-FOLD CROSS-VALIDATION

	Local				Public			
Trial	Attention U-Net	U-Net	U-Net++	U-Net3+	Attention U-Net	U-Net	U-Net++	U-Net3+
1	0.84 ± 0.11	0.84 ± 0.11	0.78 ± 0.11	0.84 ± 0.11	0.83 ± 0.09	0.83 ± 0.09	0.76 ± 0.09	0.83 ± 0.09
2	0.83 ± 0.12	0.83 ± 0.12	0.76 ± 0.12	0.83 ± 0.13	0.82 ± 0.10	0.82 ± 0.10	0.75 ± 0.10	0.82 ± 0.11
3	0.84 ± 0.11	0.83 ± 0.11	0.77 ± 0.10	0.83 ± 0.11	0.82 ± 0.09	0.82 ± 0.09	0.74 ± 0.09	0.82 ± 0.09
4	0.85 ± 0.10	0.84 ± 0.11	0.78 ± 0.10	0.84 ± 0.11	0.83 ± 0.09	0.83 ± 0.09	0.76 ± 0.08	0.83 ± 0.09
5	0.82 ± 0.09	0.81 ± 0.09	0.74 ± 0.08	0.81 ± 0.09	0.80 ± 0.07	0.80 ± 0.07	0.73 ± 0.06	0.80 ± 0.08
6	0.83 ± 0.09	0.82 ± 0.09	0.75 ± 0.08	0.82 ± 0.09	0.81 ± 0.07	0.81 ± 0.07	0.74 ± 0.07	0.81 ± 0.07
7	0.81 ± 0.11	0.79 ± 0.11	0.70 ± 0.09	0.78 ± 0.11	0.80 ± 0.11	0.79 ± 0.10	0.70 ± 0.10	0.79 ± 0.10
8	0.83 ± 0.08	0.82 ± 0.09	0.75 ± 0.08	0.82 ± 0.09	0.81 ± 0.07	0.81 ± 0.07	0.74 ± 0.07	0.82 ± 0.07
9	0.84 ± 0.08	0.83 ± 0.09	0.76 ± 0.08	0.83 ± 0.09	0.83 ± 0.07	0.82 ± 0.07	0.75 ± 0.07	0.83 ± 0.07
10	0.84 ± 0.09	0.83 ± 0.09	0.76 ± 0.08	0.83 ± 0.09	0.83 ± 0.07	0.83 ± 0.06	0.75 ± 0.07	0.83 ± 0.07