ONLINE-SECTION: DATA

Prospective 3D-FLAIR MRI acquisition was performed on a cohort of 35 subjects to study extracranial carotid artery disease and its impact on white matter hyperintensities (WMH) in the brain. The imaging parameters for the sequence are as follows: in-plane resolution=0.96mm, slice-thickness=0.99mm, TR=7000ms, TE=430ms, TI=2100ms, ETL=256, acquisition-matrix=270x240, refocusing flip angle=120°, FOV=258x230mm2, average-slices=176. Additionally, 3D-FLAIR volumes from 20 subjects were randomly selected from the third phase of publicly available Alzheimer's Disease Neuroimaging Initiative (ADNI) (http://adni.loni.usc.edu) repository to test the generalizability of WMH detection framework. ADNI is a longitudinal natural history study launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see *www.adni-info.org*. The sequence parameters for Sagittal 3D-FLAIR acquisition for ADNI3 are as follows: Effective TE=119ms, TR=4800ms, TI=1650ms, acquisition-matrix=256x256, average-slices=160. Additional details about ADNI3 MR protocols is available at http://adni.loni.usc.edu/wp-content/themes/freshnews-dev-v2/documents/mri/ADNI3-MRI-protocols.pdf.

These subjects were selected randomly from the cognitively normal (10 subjects) and mild cognitive impairment (10 subjects) using the following selection criteria: 1) Only FLAIR volumes from the initial (baseline) visit were considered. 2) Subjects with dementia diagnosis were excluded. 3) Subjects were amyloid beta negative as determined by standardized uptake value ratio threshold of 1.1 (18F-florbetapir PET scans) This resulted in 20 3D-FLAIR volumes from 16 unique imaging sites and multiple scanner manufacturers and systems. The scanner-wise subject breakdown are as follows: Siemens Prisma/Prisma Fit (11), Siemens Skyra (3), Siemens Verio (2), Siemens Trio/Tim (1), Philips Achieva (1), Philips Ingenia (1), and GE MR750: 1.

ONLINE-SECTION: METHODS

Loss Functions

The categorical cross entropy and weighted binary cross entropy loss functions used in this work is defined are defined as follows:

$$L_{cce} = -\sum_{n}(r_{1n}\log p_{1n}) + (r_{2n}\log p_{2n})$$

$$L_{wbce} = -\sum_{n} w_1(r_n \log p_n) + w_2((1-r_n)\log(1-p_n))$$

Here, r_n and p_n are the nth voxel in the ground truth annotations and the predicted white matter hyperintensities (WMH) mask, respectively. The weights w_1 and w_2 are the weights for the foreground and the background regions in the masks. This formulation assigns larger weights to the minority foreground class associated with WMH voxels and smaller weights to the majority foreground class, handling class-imbalance. The weights for a particular class are calculated from the ratio of total number of training samples to the number of samples belonging to that class i.e. $w_i = \frac{N_T}{C*N_i}$. Here, C refers to the number of classes.

CNN Implementation

The DeepUNET3D CNNs in StackGen-Net were trained with the following parameters: loss=weighted binary cross-entropy, epochs=120, batch size=10, optimizer=ADAM[21], learning rate=0.001, decay factor=0.1, and dropout probability=0.4. The learning rate was identified by a coarse grid search to ensure convergence of the training and validation loss curves. The same parameters were retained for training individual CNNs for the ablation studies. The training parameters for Meta-CNN were: loss=categorical cross entropy, epochs=400, batch size=64, optimizer=Stochastic Gradient Descent, learning rate=0.001, and decay factor=0.1.

Statistical Analysis

A Kolmogorov-Smirnov test was performed to determine if the sample distribution matched the characteristics of a normal distribution before performing parametric t-tests. For all the comparisons performed in the manuscript, the data did not significantly deviate from a normal distribution ($P>0.10$, $\alpha=0.05$).

Online-Table 1: Convolutional Neural Network training parameters

	UNET2D	UNET2D-WS	DeepMedic	DeepUNET2D	DeepUNET3D	Meta-CNN
Loss Function	WBCE	Dice	CCE	WBCE	WBCE	CCE
Training data	64x64	200x200	25x25x25	64x64	64x64x7	16x16x16
Optimizer	ADAM	ADAM	RMSProp	ADAM	ADAM	SGD
Learning Rate	1e-6	2e-4	1e-3	1e-5	1e-3	1e-3
Epochs	200	120	35	120	120	400
Trainable parameters (M)	~31.03	~8.74	~2.08	~0.562	~1.68	8

WBCE = weighted binary cross entropy; CCE = categorical cross entropy; SGD = stochastic gradient descent;

Online-Table 2: Definitions for segmentation evaluation metrics

Metric	Definition
Dice (F1-P)	$\frac{2\,\lvert P \cap GT\rvert}{\lvert P\rvert + \lvert GT\rvert}$
Precision-P	$\frac{\lvert P \cap GT\rvert}{\lvert P\rvert}$
Recall-P	$\frac{\lvert P \cap GT\rvert}{\lvert GT\rvert}$
Precision-L	$\frac{N_{(P \cap \mathrm{GT})}}{N_P}$
Recall-L	$\frac{N_{(P \cap \mathrm{GT})}}{N_{GT}}$
F1-L	$2 * \frac{(P\mathrm{L} * \mathrm{RL})}{\mathrm{PL} + RL}$
Volume Difference	$\frac{abs(\,V_P - V_{GT})}{V_{GT}}$
HD95	$\mathrm{P}_{95}\{\delta_H(G,P), \delta_H(P,G)\}$

GT = set of points constituting ground truth; P = set of points constituting predicted lesion mask;
N_p = number of lesions in predicted lesion mask; PL = precision-lesion; RL= recall-lesion;
V_{GT} = volume of ground truth mask; V_P = volume of predicted lesion mask; |A| = cardinality of set A;
δ_H = one-sided Hausdorff distance between two sets of points; P_{95} = 95th percentile

Online-Table 3: Comparison of architecture (Mean ± SD) choice on test cohort 1

	UNET2D	DeepUNET2D	DeepUNET3D
Dice (F1-P)	0.43 ± 0.17	0.54 ± 0.15*	0.74 ± 0.06††
Precision-P	0.72 ± 0.19	0.75 ± 0.15	0.84 ± 0.08
Recall-P	0.32 ± 0.19	0.45 ± 0.16*	0.66 ± 0.08†
Precision-L	0.60 ± 0.20	0.70 ± 0.16	0.81 ± 0.10
Recall-L	0.37 ± 0.09	0.64 ± 0.11**	0.80 ± 0.15†
F1-L	0.44 ± 0.10	0.65 ± 0.08**	0.80 ± 0.11†
|VD|(%)	54.4 ± 22.1	38.6 ± 25.1	21.2 ± 10.5

*P < .01 (two-sided paired t-test, UNET2D and DeepUNET2D)
**P < .001 (two-sided paired t-test, UNET2D and DeepUNET2D)
†P < .01 (two-sided paired t-test, DeepUNET2D and DeepUNET3D)
††P < .001 (two-sided paired t-test, DeepUNET2D and DeepUNET3D)

Online-Table 4: Comparison of StackGen-Net with state-of-the-art (Mean ± SD) on test cohort 1

	LST-LPA	UNET2D-WS*	StackGen-Net
Dice (F1-P)	0.23 ± 0.15	0.62 ± 0.10	0.76±0.07
Precision-P	0.49 ± 0.28	0.69 ± 0.16	0.73±0.11
Recall-P	0.21 ± 0.15	0.58 ± 0.08	0.79+0.1
Precision-L	0.28 ± 0.16	0.66 ± 0.17	0.75±0.11
Recall-L	0.25 ± 0.15	0.75 ± 0.14	0.87±0.08
F1-L	0.20 ± 0.07	0.68 ± 0.10	0.80±0.09
|VD|(%)	72.6 ± 47.1	19.1 ± 14.1	12.3±12.7
HD95	28.1 ± 12.8	12.4 ± 6.6	5.27±3.15
AUC	0.25 ± 0.18	0.44 ± 0.12	0.84±0.07

*Best CNN model in UNET2D-WS-E (Ensemble)

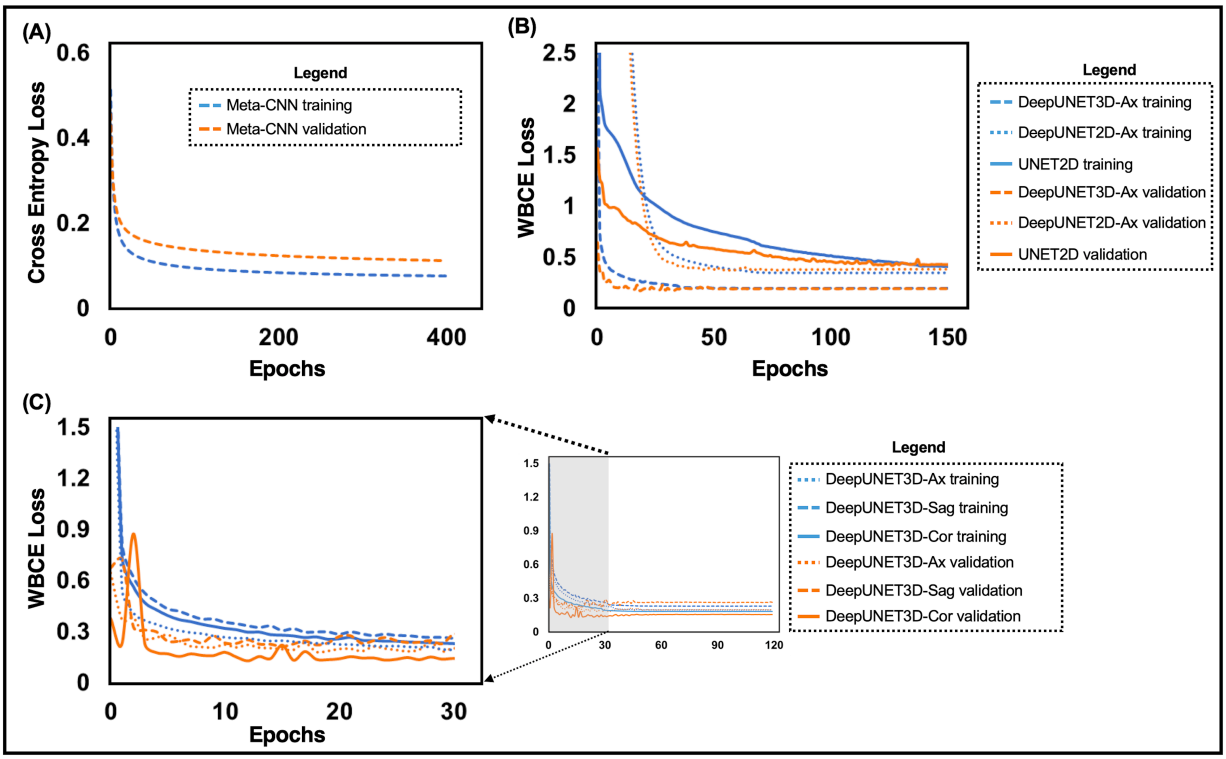
Online-Table 5: Comparison* of StackGen-Net (Mean ± SD) with DeepUNET3D trained with 3D cubes (64x64x64)

	DeepUNET3D-Cube	StackGen-Net
Dice (F1-P)	0.725 ± 0.041	0.76 ± 0.073
Precision-P	0.838 ± 0.067	0.729 ± 0.116
Recall-P	0.642 ± 0.053	0.797 + 0.1
Precision-L	0.848 ± 0.145	0.753 ± 0.115
Recall-L	0.703 ± 0.216	0.87 ± 0.08
F1-L	0.76 ± 0.187	0.798 ± 0.091
|VD|(%)	22.8 ± 9.2	12.3 ± 12.7

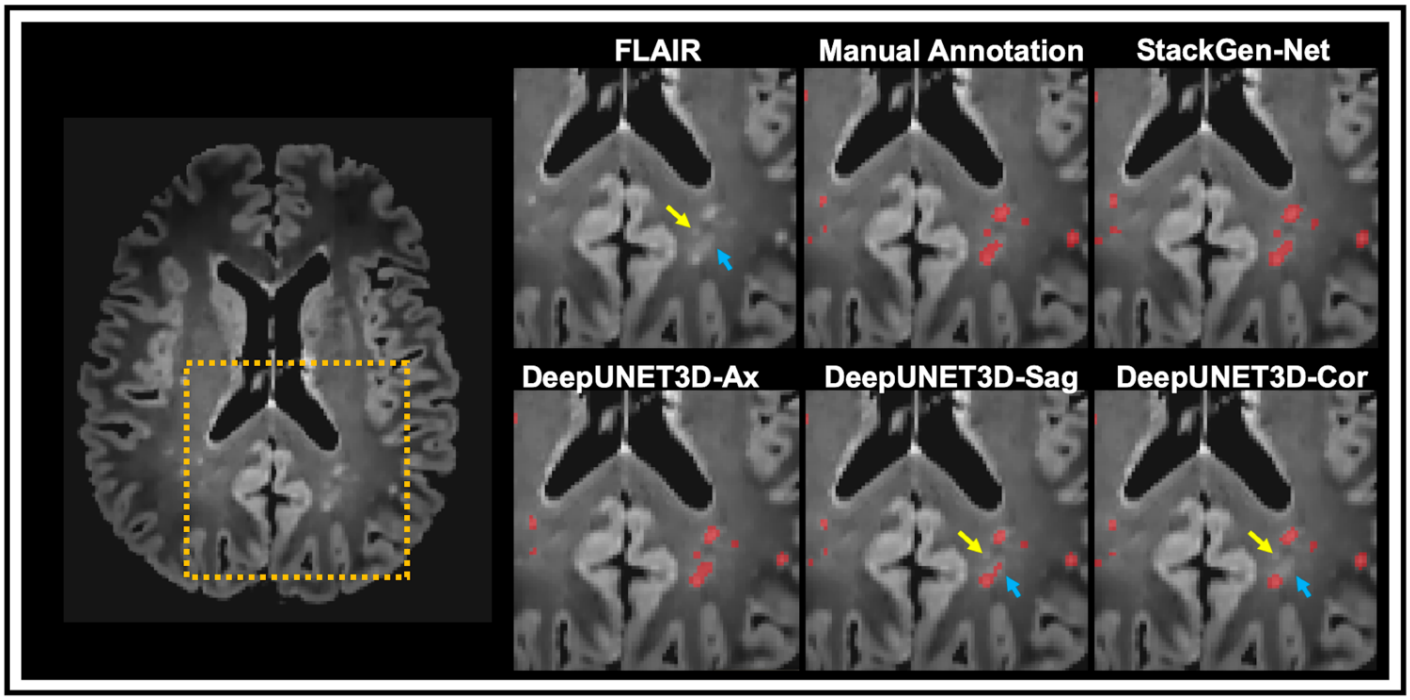
*Evaluated on test cohort 1

Online-Table 6: Pairwise F1-L (Mean, Median) on Test Cohort 2

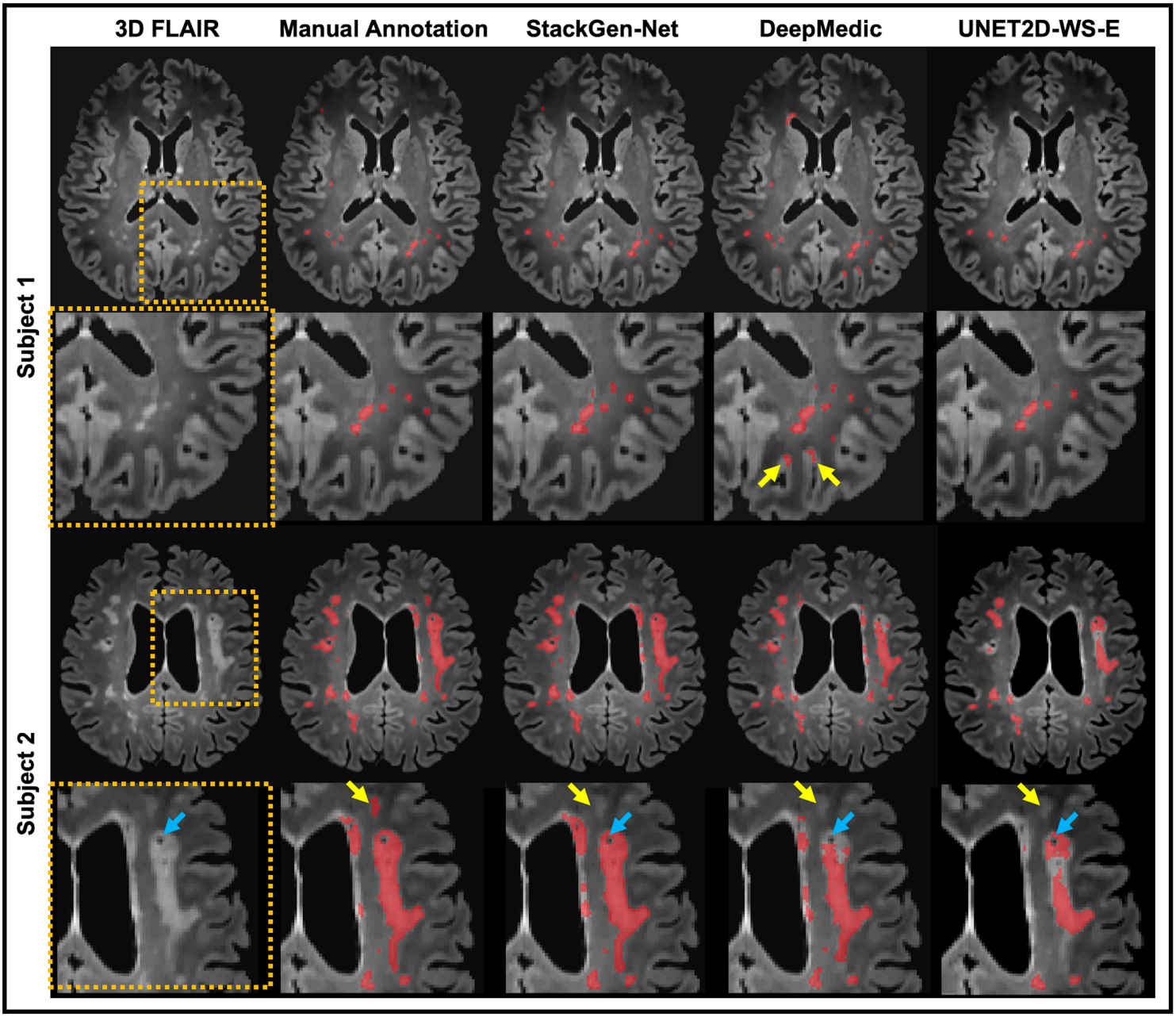
	Observer1	Observer2	StackGen-Net
Observer1	-	0.69, 0.65	0.70, 0.73
Observer2		-	0.62, 0.67
StackGen-Net			-



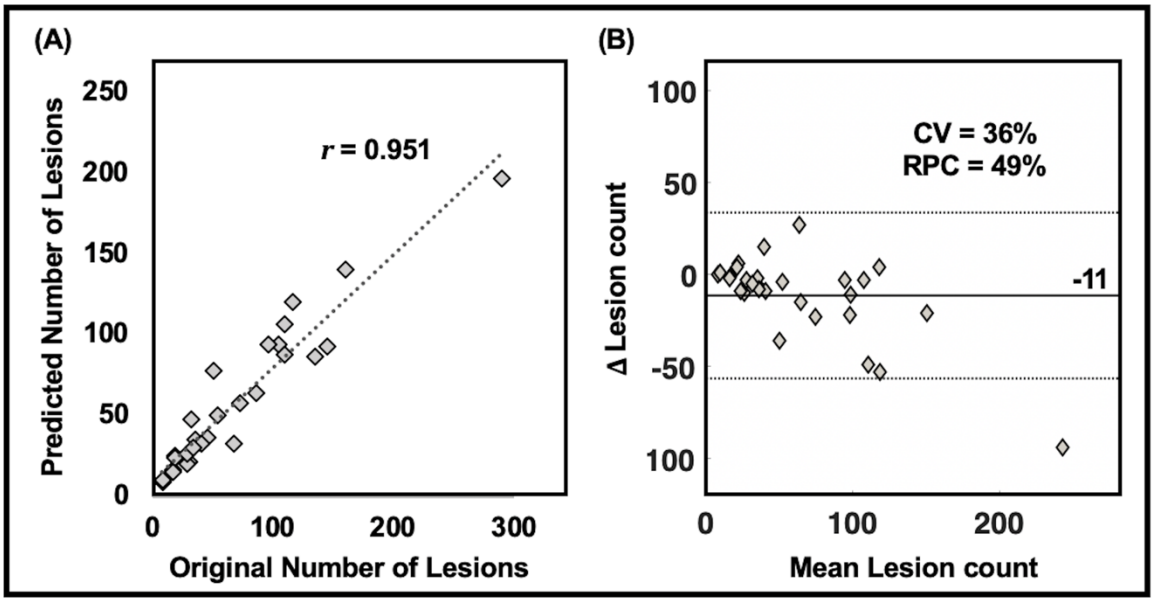
Online-Figure 1: Training (blue) and validation (orange) loss evolution curves for different convolutional neural networks (CNNs). The categorical cross entropy loss evolution curves for the Meta CNN over 400 epochs are shown in (A). A comparison of weighted binary cross entropy (WBCE) loss curves for DeepUNET3D, DeepUNET2D, and UNET2D is shown in (B). The three CNNs were trained on axially oriented training patches. The WBCE loss evolution curves for the orthogonal CNNs over the first 30 epochs (highlighted in gray) are shown in (C).



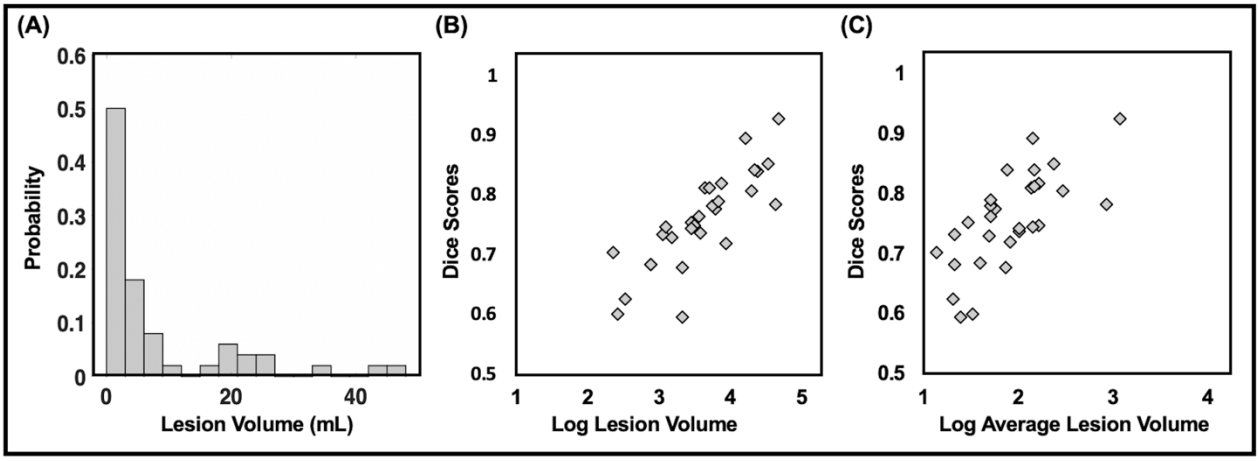
Online-Figure 2: WMH predictions from DeepUNET3D-Axial, DeepUNET3D-Sagittal, DeepUNET3D-Coronal, and StackGen-Net overlaid on a representative axial slice from a test subject. Manual annotations are shown for comparison. The arrows show WMH that were missed (yellow) or whose contours were mis-identified (blue) by a majority of the CNNs in the stacked-generalization ensemble. These WMH would have been missed by a simple averaging or majority voting of the orthogonal CNN predictions but are identified correctly by StackGen-Net.



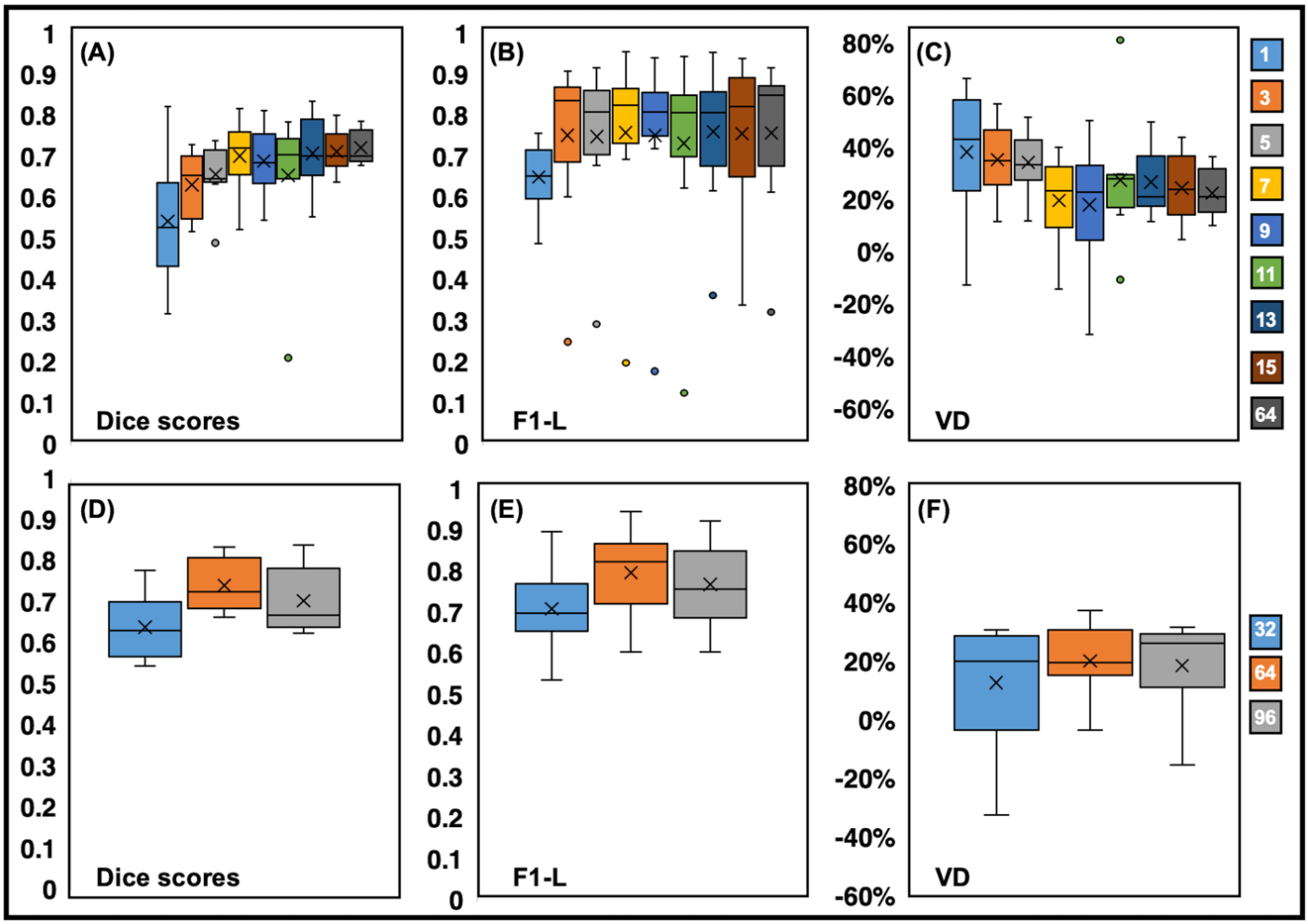
Online-Figure 3: Comparison of WMH predictions from StackGen-Net, DeepMedic, and UNET2D-WS-E. The manual annotations and predictions are overlaid in red on representative axial FLAIR images from two subjects. For each subject, the bottom row shows the inset zoomed. The yellow arrows in Subject 1 point to false positives predicted by DeepMedic at the juxtacortical margin. The blue arrow in Subject 2 shows WMH at left anterior centrum semiovale. StackGen-Net correctly identifies that the central area of encephalomalacia is not a white matter lesion. The yellow arrow points to a false negative region (diffuse WMH).



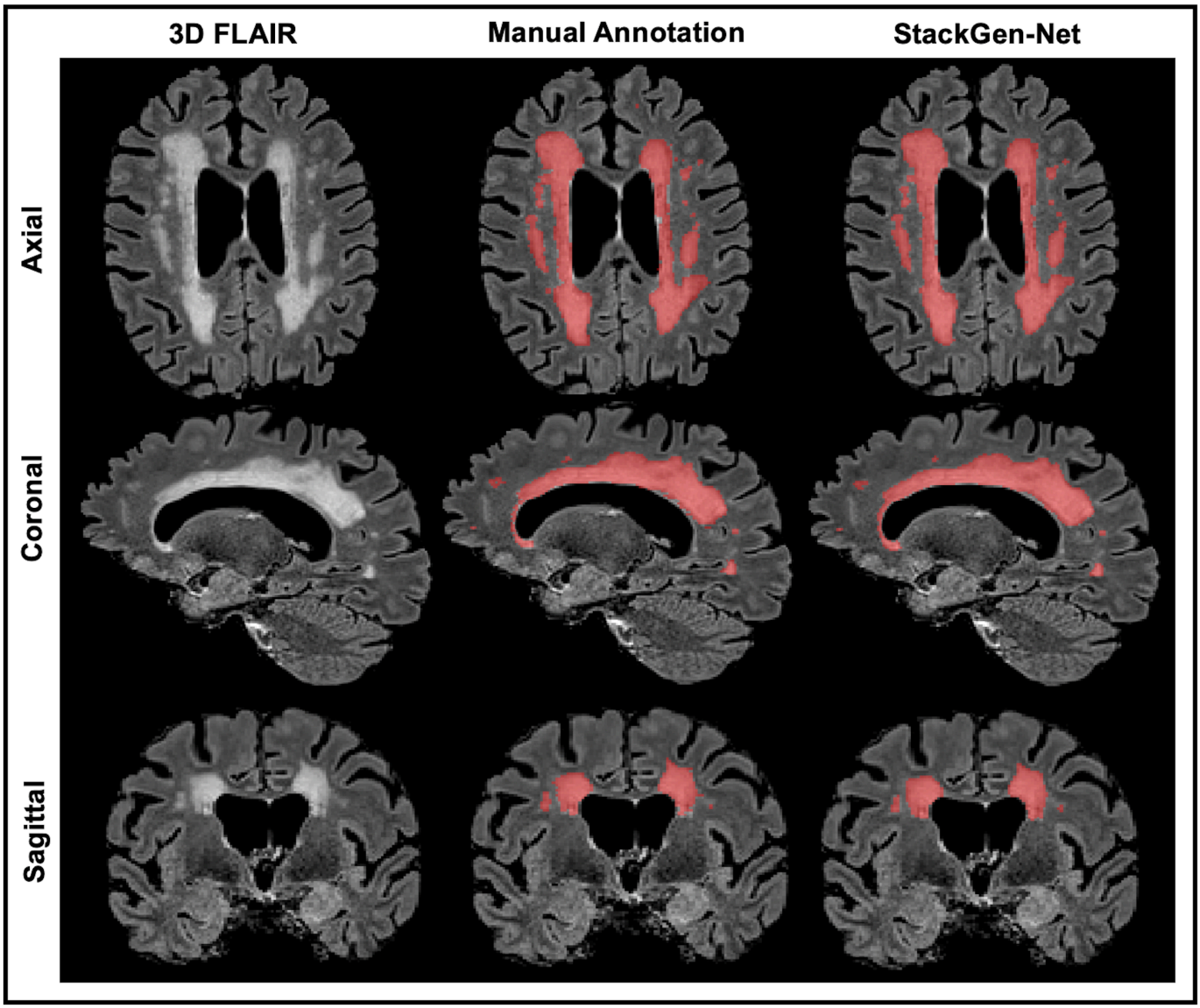
Online-Figure 4: Correlation plot and Bland-Altman analysis for agreement in the number of WMH lesion between the ground truth annotations and StackGen-Net predictions. The coefficient of variation (CV) and the repeatability coefficient (RPC) also shown. A connected component analysis was used to identify individual lesions in the WMH masks.



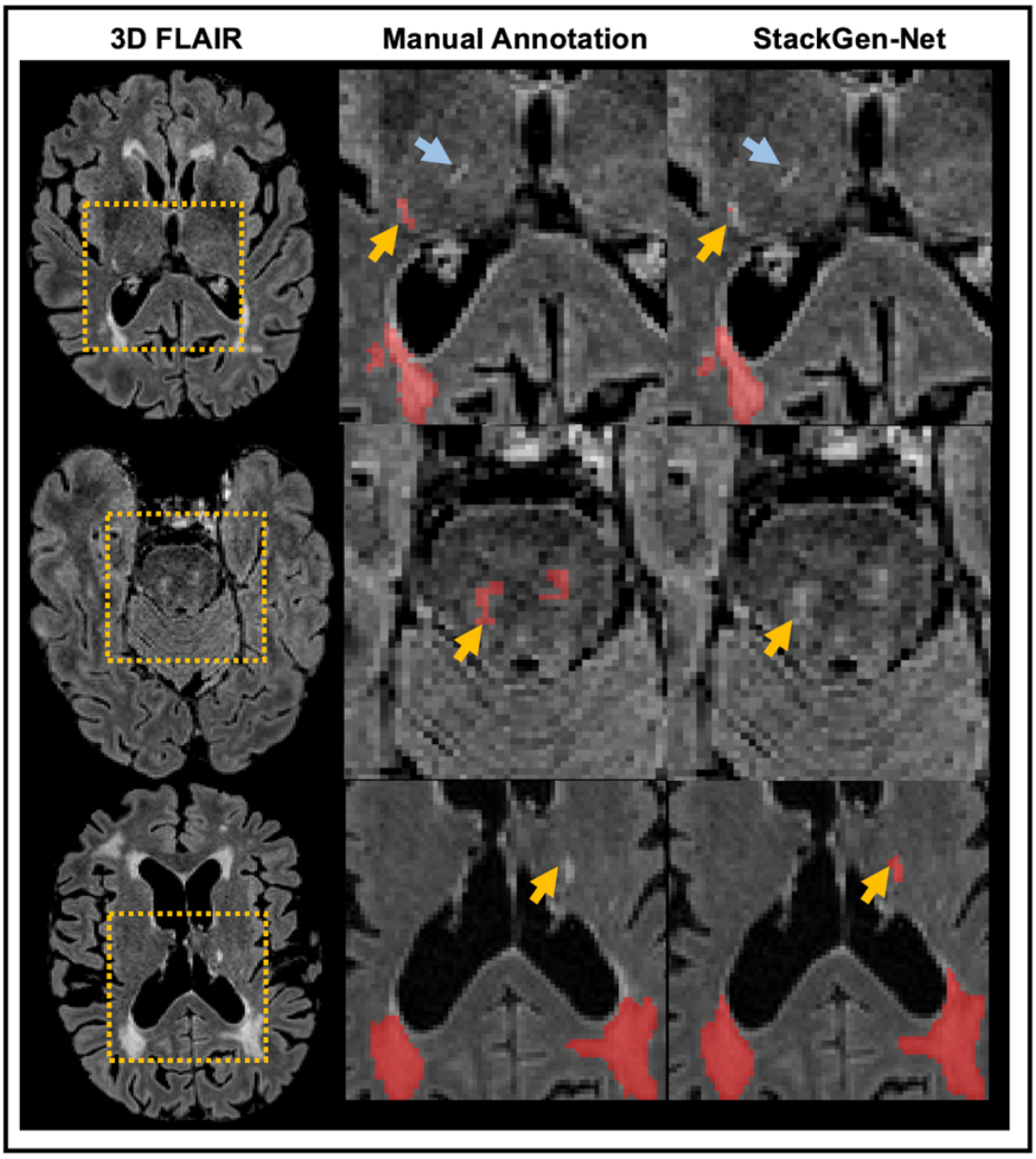
Online-Figure 5: WMH Lesion volumes (mL) in study cohort. (A) A histogram of white matter hyperintensities volume (mL) per subject in the study cohort (n=50, mean ± standard deviation=8.1±11.3 mL, median=2.9 mL) is shown. A majority of the cases in the study cohort have a lesion burden less than 15mL. (B and C) Scatter plot of white matter hyperintensities volume vs StackGen-Net performance measured in Dice metric on the test cohorts (n=29). The volumes are in logarithmic scale for display. As expected, lower lesion burden affects dice scores.



Online-Figure 6: Effect of varying through-plane and in-plane spatial extent on WMH segmentation performance of DeepUNET3D CNN on test cohort 1. (A-C) The number of slices in training patches of size 64x64xN were increased as indicated by the legend. With increasing spatial extent along the slice direction, there is an increase in Dice scores and F1-L values and a decrease in volume difference. However, after a certain through-plane patch size, there are no considerable changes. (D-F) The in-plane spatial extent of the training patch of size NxNx7 were varied as indicated by the legend.



Online-Figure 7: WMH predictions overlaid on multi-planar views from a subject in test cohort 2 (ADNI3). Manual annotations are presented for reference. StackGen-Net is able to accurately predict lesions on an independent test cohort with performance comparable to human observers.



Online-Figure 8: Representative examples of false-negatives and false-positives (top-row) in WMH predictions by StackGen-Net. The blue and yellow arrows in top row show hyperintensities within the thalamus and in the extracapsular region. StackGen-Net accurately ignores the WMH within the thalamus (blue arrow) but has trouble accurately detecting the lesion at the extracapsular region (yellow arrow). The second row shows false negatives in the midbrain region where StackGen-Net misses the WMH due to microvascular ischemic changes. The last row shows a false positive WMH. These are likely due to the absence of adequate training examples for WMH in the midbrain region or lesions in the gray matter region.