Published February 11, 2021 as 10.3174/ajnr.A6970

A Stacked Generalization of 3D Orthogonal Deep Learning Convolutional Neural Networks for Improved Detection of White Matter Hyperintensities in 3D FLAIR Images

L. Umapathy, G.G. Perez-Carrillo, M.B. Keerthivasan, J.A. Rosado-Toro, M.I. Altbach, B. Winegar, C. Weinkauf, and A. Bilgin, for the Alzheimer's Disease Neuroimaging Initiative

ABSTRACT

BACKGROUND AND PURPOSE: Accurate and reliable detection of white matter hyperintensities and their volume quantification can provide valuable clinical information to assess neurologic disease progression. In this work, a stacked generalization ensemble of orthogonal 3D convolutional neural networks, StackGen-Net, is explored for improving automated detection of white matter hyperintensities in 3D T2-FLAIR images.

MATERIALS AND METHODS: Individual convolutional neural networks in StackGen-Net were trained on 2.5D patches from orthogonal reformatting of 3D-FLAIR ($n = 21$) to yield white matter hyperintensity posteriors. A meta convolutional neural network was trained to learn the functional mapping from orthogonal white matter hyperintensity posteriors to the final white matter hyperintensity prediction. The impact of training data and architecture choices on white matter hyperintensity segmentation performance was systematically evaluated on a test cohort ($n = 9$). The segmentation performance of StackGen-Net was compared with state-of-the-art convolutional neural network techniques on an independent test cohort from the Alzheimer's Disease Neuroimaging Initiative-3 ($n = 20$).

RESULTS: StackGen-Net outperformed individual convolutional neural networks in the ensemble and their combination using averaging or majority voting. In a comparison with state-of-the-art white matter hyperintensity segmentation techniques, StackGen-Net achieved a significantly higher Dice score (0.76 [SD, 0.08], F1-lesion (0.74 [SD, 0.13]), and area under precision-recall curve (0.84 [SD, 0.09]), and the lowest absolute volume difference (13.3% [SD, 9.1%]). StackGen-Net performance in Dice scores (median = 0.74) did not significantly differ ($P = .22$) from interobserver (median = 0.73) variability between 2 experienced neuroradiologists. We found no significant difference ($P = .15$) in white matter hyperintensity lesion volumes from StackGen-Net predictions and ground truth annotations.

CONCLUSIONS: A stacked generalization of convolutional neural networks, utilizing multiplanar lesion information using 2.5D spatial context, greatly improved the segmentation performance of StackGen-Net compared with traditional ensemble techniques and some state-of-the-art deep learning models for 3D-FLAIR.

ABBREVIATIONS: ADNI = Alzheimer's Disease Neuroimaging Initiative; AUC = area under curve; Ax = axial; CNN = convolutional neural network; E-A = ensemble average; E-MV = ensemble majority vote; F1-L = F1 lesion; HD = Hausdorff distance; VD = volume difference; WMH = white matter hyperintensity

Received May 13, 2020; accepted after revision October 26.

From the Departments of Electrical and Computer Engineering (L.U., A.B.), Medical Imaging (L.U., G.G.P.-C., M.B.K., J.A.R.-T., M.I.A., B.W., A.B.), Surgery (C.W.), and Biomedical Engineering (A.B.), University of Arizona, Tucson, Arizona.

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). Thus, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Paper previously was presented in part at: Annual Meeting and Exhibition of International Society for Magnetic Resonance in Medicine, Virtual Conference and Exhibition, October 8–14, 2020.

This work was supported, in part, by the Arizona Health Sciences Center Translational Imaging Program Project Stimulus Fund, BIO5 Team Scholar's Program, Technology and Research Initiative Fund Improving Health Initiative, and Arizona Alzheimer's Consortium. Data collection and sharing for this project were funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health grant U01 AG024904) and Department of Defense ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica; Biogen; Bristol-Myers Squibb; CereSpir; Cogstate; Eisai; Elan Pharmaceuticals; Eli Lilly; EuroImmun; F. Hoffmann-La Roche and its affiliated company Genentech; Fujirebio; GE Healthcare; IXICO; Janssen Alzheimer Immunotherapy Research and Development; Johnson & Johnson Pharmaceutical Research and Development; Lumosity; Lundbeck; Merck; Meso Scale Diagnostics; NeuroRx Research; Neurotrack; Novartis Pharmaceuticals; Pfizer; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study was coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Please address correspondence to Ali Bilgin, PhD, Department of Biomedical Engineering, Thomas R. Keating Building, University of Arizona, 1657 E. Helen Street, Tucson, AZ, 85721; e-mail: bilgin@email.arizona.edu

🔑 Indicates open access to non-subscribers at www.ajnr.org

▤ Indicates article with online supplemental data.

http://dx.doi.org/10.3174/ajnr.A6970

White matter hyperintensities (WMHs) correspond to pathologic features of axonal degeneration, demyelination, and gliosis observed within cerebral white matter.[1] Clinically, the extent of WMHs in the brain has been associated with cognitive impairment, Alzheimer's disease and vascular dementia, and increased risk of stroke.[2,3] The detection and quantification of WMH volumes to monitor lesion burden evolution and its correlation with clinical outcomes have been of interest in clinical research.[4,5] Although the extent of WMHs can be visually scored,[6] the categoric nature of such scoring systems makes quantitative evaluation of disease progression difficult. Manually segmenting WMHs is tedious, prone to inter- and intraobserver variability, and is, in most cases, impractical. Thus, there is an increased interest in developing fast, accurate, and reliable computer-aided automated techniques for WMH segmentation.

Convolutional neural network (CNN)-based approaches have been successful in several semantic segmentation tasks in medical imaging.[7] Recent works have proposed using deep learning–based methods for segmenting WMHs using 2D-FLAIR images.[8-11] More recently, a WMH segmentation challenge[12] was also organized (http://wmh.isi.uu.nl/) to facilitate comparison of automated segmentation of WMHs of presumed vascular origin in 2D multislice T2-FLAIR images. Architectures that used an ensemble of separately trained CNNs showed promising results in this challenge, with 3 of the top 5 winners using ensemble-based techniques.[12]

Conventional 2D-FLAIR images are typically acquired with thick slices (3–4 mm) and possible slice gaps. Partial volume effects from a thick slice are likely to affect the detection of smaller lesions, both in-plane and out-of-plane. 3D-FLAIR images, with isotropic resolution, have been shown to achieve higher resolution and contrast-to-noise ratio[13] and have shown promising results in MS lesion detection using 3D CNNs.[14] Additionally, the isotropic resolution enables viewing and evaluation of the images in multiple planes. This multiplanar reformatting of 3D-FLAIR without the use of interpolating kernels is only possible due to the isotropic nature of the acquisition. Network architectures that use information from the 3 orthogonal views have been explored in recent works for CNN-based segmentation of 3D MR imaging data.[15] The use of data from multiple planes allows more spatial context during training without the computational burden associated with full 3D training.[16] The use of 3 orthogonal views simultaneously mirrors how humans approach this segmentation task.

Ensembles of CNNs have been shown to average away the variances in the solution and the choice of model- and configuration-specific behaviors of CNNs.[17] Traditionally, the solutions from these separately trained CNNs are combined by averaging or using a majority consensus. In this work, we propose the use of a stacked generalization framework (StackGen-Net) for combining multiplanar lesion information from 3D CNN ensembles to improve the detection of WMH lesions in 3D-FLAIR. A stacked generalization[18] framework learns to combine solutions from individual CNNs in the ensemble. We systematically evaluated the performance of this framework and compared it with traditional ensemble techniques, such as averaging or majority voting, and state-of-the-art deep learning techniques.

MATERIALS AND METHODS

StackGen-Net CNN Architecture

Figure 1*A* shows an overview of the proposed StackGen-Net architecture. Our ensemble consists of 3 orthogonal 3D CNNs (DeepUNET3D), each trained on axial, sagittal, and coronal reformatting of 3D-FLAIR. This is followed by a stacked generalization[18] of the orthogonal CNNs using a Meta CNN. The proposed multiscale, fully-connected DeepUNET3D architecture is shown in Fig 1*B*. Compared with a UNET,[19] DeepUNET3D uses *convolutional blocks* instead of convolutional layers. These convolutional blocks consist of a sequence of convolutions with 3D kernels (3 × 3 × 3), batch normalization, and rectified linear activation layers separated by a dropout layer. A final convolution layer combines the feature maps in the native resolution space to generate posterior probabilities for WMHs.

The stacked generalization scheme attempts to maximize the overall accuracy of the ensemble by deducing the bias rate of the individual DeepUNET3D CNNs. If we consider $p_a,\ p_s,\ p_c$, and p_f to be the axial, sagittal, coronal, and final WMH posterior probabilities for a voxel, then the Meta CNN learns a new functional mapping $f(.)$ from $[0,1]^3$ to $[0,1]$ where $p_f = f(p_a,\ p_s,\ p_c)$ and $p_f,\ p_a,\ p_s,\ p_c \in [0,\ 1]$. In this work, we consider the following mapping:

$$p_f = f(p_a,\ p_s,\ p_c) = \sigma(w_a\ p_a + w_s\ p_s + w_c\ p_c + b),$$

where $w_a,\ w_s,\ w_c$ are the weights for axial, sagittal, and coronal posteriors, respectively; b is the bias term, and $\sigma(.)$ represents a softmax operation. These weights are learned during training of the Meta CNN, which consists of a single convolution layer with a 1 × 1 × 1 3D kernel.

Study Population and Image Acquisition

A cohort of 35 subjects was prospectively recruited (2016–2017) for a study on extracranial carotid artery disease with approval of the local institutional review board. Adults 50–85 years of age with extracranial carotid artery disease (50%–90% stenosis) based on duplex sonography criteria were recruited from outpatient clinics/inpatient hospitals. Exclusion criteria included depression, dementia, MS, and contraindications to MR imaging.

Sagittal 3D-FLAIR images were acquired using the 3D spatial and chemical-shift-encoded excitation inversion recovery sequence on a 3T MR imaging scanner (Magnetom Skyra). Five of the 35 subjects were excluded due to poor FLAIR image quality (motion artifacts). The remaining study cohort (67.7 [SD, 8.7] years of age; 22 men, 8 women) was randomly split into groups for training ($n = 20$), validation ($n = 1$), and testing ($n = 9$). These test subjects formed test cohort 1.

To test the generalizability of the framework, we also evaluated its performance on a multi-institutional and multi-scanner external cohort (test cohort 2) from the Alzheimer's Disease Neuroimaging Initiative (ADNI) data base (adni.loni.usc.edu). The primary goal of ADNI, a public-private partnership led by Principal Investigator Michael W. Weiner, is to test whether imaging and biologic markers, along with clinical and neuropsychological assessments, can be combined to measure progression of mild cognitive impairment and early Alzheimer's Disease.

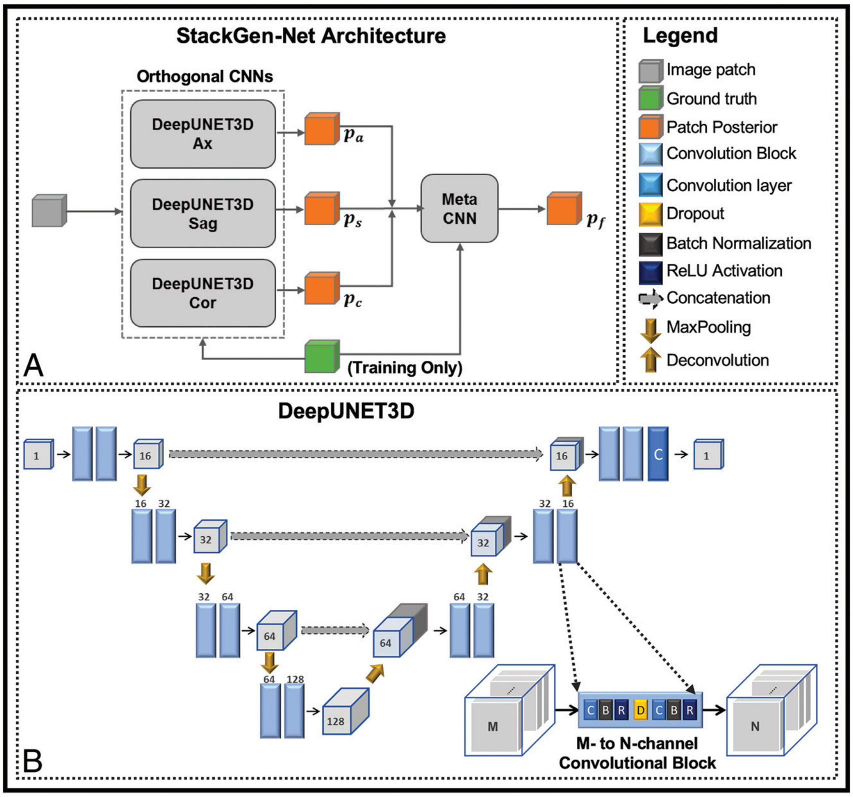

FIG 1. *A*, Overview of the proposed StackGen-Net. *B*, This consists of 3 DeepUNET3D CNNs, which are made up of convolutional blocks. The number of output feature maps is presented next to each convolutional block. Each DeepUNET3D predicts posterior probabilities for WMHs on orthogonal (axial, sagittal, and coronal) orientations of the 3D-FLAIR volumes. The Meta CNN combines axial, sagittal, and coronal posterior probabilities for a voxel to yield a final prediction for WMH. Sag indicates sagittal; Cor, coronal; ReLU, rectified linear unit.

Sagittal 3D-FLAIR volumes from 20 subjects (76.8 [SD, 9.3] years of age; 11 men, 9 women) were selected randomly from the cognitively normal and mild cognitive impairment groups. Additional information regarding the acquisition protocol and subject-selection criteria for ADNI3 is available in the Online Supplemental Data.

WMH Annotations

Two neuroradiologists, with certificate of added qualification, (observers 1 and 2) agreed on the following protocol to annotate WMH: 1) Deep WMHs should at least be 2 mm wide, spanning more than 1 imaging section; 2) periventricular WMHs should be >3 mm wide; 3) hyperintense regions due to partial volume effects near the ventricles or sulci or CSF flow artifacts should not be included; and 4) no deep gray matter lesions or cortical hyperintense lesions should be included.

Observer 1 manually annotated WMHs on 3D-FLAIR images for all the subjects. As part of an interobserver variability study to establish baseline human performance, observers 1 and 2 independently annotated 60 FLAIR images (12 imaging volumes, each with 5 consecutive images) from 3 subjects in test cohort 1. Observer 1 re-annotated these images after 5 months from the first annotations to avoid recall bias; these images were used to evaluate intraobserver variability. Observers 1 and 2 also annotated 10 additional subjects (1600 images) from the test cohort 2 to establish human performance on the external set. Both observers used an in-house Matlab (MathWorks)-based graphical user interface. The observers manually traced the WMH pixels on axial cross-sections of the 3D-FLAIR images. The observers were allowed to adjust the window width and level to improve WMH contrast and use 3D spatial context to annotate/edit/delete individual WMH masks.

Training Data

Image preprocessing consisted of skull stripping,[20] N4 bias correction,[21] total variation–based denoising, and contrast stretching. The image intensities were normalized, per subject, using a zero mean unit SD intensity normalization. Here, the mean signal was calculated from regions within the brain.

3D-FLAIR volumes in the training set were reformatted to axial, sagittal, and coronal orientations. From each orientation, overlapping 2.5D patches (64 × 64 × 7) were extracted using a sliding window over the entire brain to train the corresponding orthogonal CNNs. Patches with <30% brain voxels were discarded. Training data were generated by sampling the remaining patches to ensure an equal representation of patches with and without WMHs. Data augmentation was performed using the following schemes: in-plane flipping of patches, through-plane flipping of patches, and image filtering using Gaussian kernels.

Training data for the Meta CNN were generated by first predicting WMH posteriors for each subject in the training set using the trained orthogonal CNNs. After being reformatted to the axial orientation, 3D patches (16 × 16 × 16) were extracted from each of the posteriors and concatenated along the channel dimension.

During the test phase, 3D-FLAIR images were passed through the StackGen-Net framework to predict WMHs on the images in 1 pass through the network.

Experiments

Several ablation studies were conducted to systematically evaluate the choice of training data and architecture made in this study. A version of the DeepUNET3D architecture with 2D convolution kernels (DeepUNET2D) was trained on axially oriented 2D patches (64 × 64) to study the impact of additional spatial context in 2.5D patches on WMH segmentation performance.

We also trained an ensemble of 3 DeepUNET3D CNNs on axially oriented 2.5D training patches. The final prediction for

WMHs for this ensemble (DeepUNET3D-Ax E-A) was obtained by averaging posteriors from individual CNNs. For comparisons, the WMH posteriors from the orthogonal CNNs used in StackGen-Net were averaged (orthogonal E-A) or combined using a majority-voting scheme (orthogonal E-MV). Together, these experiments allowed us to determine whether a stacked generalization of orthogonal CNNs improves WMH segmentation performance.

We also explored the impact of the in-plane and through-plane spatial extent of the 2.5D training patches on WMH segmentation performance by training a series of DeepUNET3D CNNs with varying patch sizes.

All experiments were implemented in Python using Keras (https://keras.io) with TensorFlow (https://www.tensorflow.org/)[22] backend on a Linux system, with Titan P100 (NVIDIA) GPUs. The CNN implementation details, training parameters, and loss function definitions are available in the Online Supplemental Data. The StackGen-Net CNN used in this work will be available at https://github.com/spacl-ua/wmh-segmentation.

Comparisons with State-of-the-Art Techniques

We compared the performance of StackGen-Net with several state-of-the-art segmentation techniques. The UNET architecture[19] (UNET2D) was modified with zero-padded convolutions to yield input image–sized predictions. The multiclass cross-entropy loss used in the original article was modified to a weighted binary cross-entropy function. This CNN was trained using the same axial 2D patches as the DeepUNET2D.

DeepMedic[16] was trained with 3D-FLAIR images with the code and default training settings available at https://github.com/deepmedic/deepmedic. We also compared our performance with the ensemble technique[10] that achieved the highest Dice score, modified Hausdorff distance (95th percentile; HD95), and lesion-recall values in the recent WMH segmentation challenge.[12] This winning submission (UNET2D-WS-E), an averaging ensemble of three 2D UNETs, was trained using 2D-FLAIR images with the author's provided code and training settings (https://github.com/hongweilibran/wmh_ibbmTum). In contrast to the proposed DeepUNET3D, these reference architectures did not use dropout.

Additionally, we also used the lesion-prediction algorithm from the Lesion Segmentation Toolbox[23] (FMRIB Automated Segmentation Tool; https://www.applied-statistics.de/lst.html). Although some of the techniques compared here[10,16,23] can also use T1-weighted images, we made use of only FLAIR images to train and/or evaluate these techniques for a comparable assessment.

Evaluation Metrics

The WMH detection performance was evaluated using the metrics defined in the Online Supplemental Data: Dice score, precision, recall, F1, HD95, and absolute volume difference (VD). The precision, recall, and F1 metrics were evaluated at the pixel-level (Precision-P, Recall-P, F1-P) as well as lesion-level (Precision-L, Recall-L, F1-L). A connected component analysis was used to identify individual lesions in the predicted segmentations. We also generated precision-recall receiver operating curves to compare the areas under the curve (AUCs) for this heavily imbalanced class-detection problem.

Statistical Analysis

Two-sided paired t tests were used to determine whether StackGen-Net performance significantly differed from other state-of-the-art comparisons. When applicable, P values were Bonferroni-corrected for multiple comparisons. The total lesion volume was calculated on the ground truth annotations as well as WMH predictions from StackGen-Net. A Bland-Altman analysis was performed to assess the agreement in the number of detected lesions and lesion volume between ground truth and StackGen-Net predictions. The reproducibility coefficient, coefficient of variation, and correlation statistics were computed. A 2-sided paired t test was used to assess whether the WMH volumes significantly differed between the ground truth and StackGen-Net predictions.

Pair-wise Dice scores were calculated on the interobserver variability set from test cohorts 1 and 2 between human observers and StackGen-Net predictions. Repeated measures ANOVA was used to test whether the pair-wise Dice scores were significantly different between StackGen-Net and human observers. The value of α was set to .05 for all statistical comparisons.

RESULTS

The training of each DeepUNET3D CNN in StackGen-Net took approximately 40 hours, whereas the Meta CNN took approximately 1 hour. The end-to-end prediction time for a preprocessed 3D-FLAIR test image (240 × 270 × 176) was approximately 45 seconds on a GPU. The training and validation loss curves for a subset of DeepUNET3D architectures are presented in the Online Supplemental Data.

Figure 2 shows WMH predictions from StackGen-Net for representative multiplanar 3D-FLAIR images from a test subject, along with reference manual annotations. We see that StackGen-Net is able to identify smaller lesions, even when individual orthogonal CNNs miss them (Fig 2*B* and Online Supplemental Data). A comparison of StackGen-Net segmentation performance with variants of the DeepUNET3D CNN on test cohort 1 is presented in Table 1 and the Online Supplemental Data. StackGen-Net achieved a higher Dice score (0.76) compared with the individual orthogonal CNNs in the stacked generalization ensemble or their ensemble using averaging and majority voting (range, 0.72–0.75) (Table 1). StackGen-Net also yielded an absolute VD (12.36%) lower than the other CNNs. We also observed differences among performances of the orthogonal CNNs. On average, DeepUNET3D-Ax achieved the highest Dice score (0.74), whereas DeepUNET3D-Sagittal achieved the lowest absolute VD (16.9%), though these differences were not significant.

The introduction of convolutional blocks (DeepUNET2D-Ax versus UNET2D) significantly improved Dice scores ($P = .002$), F1-L ($P < .001$), and absolute VD ($P = .02$). Additional spatial information in the form of 2.5D patches (DeepUNET3D-Ax versus DeepUNET2D-Ax) significantly improved the performance in Dice scores, F1-L ($P < .005$), and absolute VD ($P = .03$). WMH segmentation performance of DeepUNET3D-

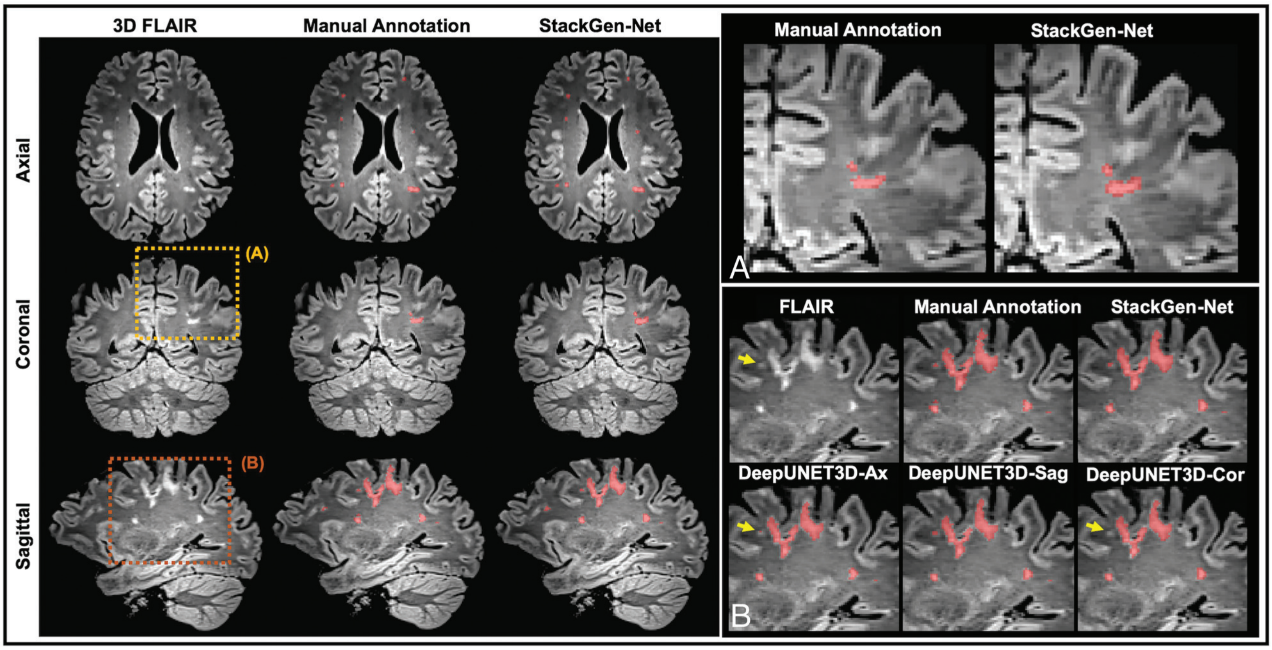

FIG 2. Qualitative evaluation of WMH detection performance by StackGen-Net. Representative axial, coronal, and sagittal slices from a test subject are shown in the left panel. Manual annotations and predictions from StackGen-Net are overlaid in red. *A*, The insets from coronal images are zoomed in for better comparison of the prediction with the ground truth. Compared with manual annotation, StackGen-Net slightly overestimates the lesion contour. *B*, A comparison of WMH predictions from the orthogonal CNNs (axial, sagittal, and coronal) is shown. The *yellow arrows* show WMHs that were missed by a majority of the CNNs in the ensemble. These lesions would have been missed by a simple averaging or majority voting of the orthogonal CNN predictions but are identified correctly by StackGen-Net. Sag indicates sagittal; Cor, coronal.

Table 1: Comparison[a] of StackGen-Net with variants of DeepUNET3D architecture

	DeepUNET3D				Orthogonal		
	Axial	Sagittal	Coronal	Axial (E-A)	(E-A)	(E-MV)	StackGen-Net
Dice (F1-P)	0.74 [SD, 0.06]	0.73 [SD, 0.08]	0.72 [SD, 0.02]	0.73 [SD, 0.07]	0.75 [SD, 0.08]	0.75 [SD, 0.08]	0.76 [SD, 0.07]
Precision-P	0.84 [SD, 0.08]	0.81 [SD, 0.07]	0.83 [SD, 0.08]	0.84 [SD, 0.08]	0.87 [SD, 0.06]	0.87 [SD, 0.06]	0.73 [SD, 0.11]
Recall-P	0.66 [SD, 0.08]	0.67 [SD, 0.10]	0.64 [SD, 0.12]	0.78 [SD, 0.09]	0.66 [SD, 0.10]	0.67 [SD, 0.10]	0.79 [SD, 0.1]
Precision-L	0.81 [SD, 0.10]	0.79 [SD, 0.09]	0.85 [SD, 0.11]	0.84 [SD, 0.09]	0.88 [SD, 0.09]	0.87 [SD, 0.09]	0.75 [SD, 0.11]
Recall-L	0.80 [SD, 0.15]	0.80 [SD, 0.10]	0.78 [SD, 0.11]	0.77 [SD, 0.14]	0.80 [SD, 0.13]	0.81 [SD, 0.13]	0.87 [SD, 0.08]
F1-L	0.80 [SD, 0.11]	0.79 [SD, 0.07]	0.80 [SD, 0.08]	0.80 [SD, 0.09]	0.83 [SD, 0.08]	0.83 [SD, 0.08]	0.80 [SD, 0.09]
|VD|(%)	21.2 [SD, 10.5]	16.9 [SD, 10.8]	23.5 [SD, 13.0]	22.7 [SD, 11.2]	24.3 [SD, 11.3]	22.3 [SD, 10.9]	12.3 [SD, 12.7]

Note:—|VD| indicates absolute volume difference; P, pixel; L, lesion.

[a] Mean [SD] on test cohort 1.

Ax with changes to the in-plane and through-plane spatial context of the training patch is shown in the Online Supplemental Data.

The predictions from StackGen-Net on test cohorts 1 and 2 are compared with state-of-the-art WMH segmentation techniques in Table 2 and the Online Supplemental Data. As expected, all deep learning CNNs outperformed the Lesion Segmentation Toolbox in all evaluation metrics. On average, StackGen-Net achieved significantly higher Dice scores (0.76 versus 0.33–0.66) ($P < .001$), F1-L (0.74 versus 0.40–0.65) ($P < .001$), and AUC (0.84 versus 0.53 - 0.62) ($P < .001$) in the test cohorts ($n = 29$). The absolute VD was significantly lower (13.3% versus 32.7%–64.1%) than UNET2D-WS-E ($P = .03$), DeepMedic ($P < .001$), and UNET2D ($P < .001$). The UNET2D-WS-E architecture had the next best performance across most of the evaluation metrics. The boxplots (Fig 3) show the median scores and interquartile ranges for these techniques over the test cohorts ($n = 29$). The correlation and Bland-Altman plots to assess agreement between ground truth and StackGen-Net WMH predictions in terms of the total

Table 2: Comparison of StackGen-Net with other WMH detection techniques

	Test Cohort 1				Test Cohort 2			
	UNET2D	DeepMedic	UNET2D WS-E	StackGen-Net	UNET2D	DeepMedic	UNET2D WS-E	StackGen-Net
Dice (F1-P)	0.43	0.62	0.67	0.76	0.27	0.58	0.66	0.76
	[SD, 0.17]	[SD, 0.09]	[SD, 0.09]	[SD, 0.07]	[SD, 0.20]	[SD, 0.15]	[SD, 0.17]	[SD, 0.09]
Precision-P	0.72	0.63	0.72	0.73	0.73	0.66	0.69	0.77
	[SD, 0.19]	[SD, 0.13]	[SD, 0.15]	[SD, 0.11]	[SD, 0.32]	[SD, 0.22]	[SD, 0.23]	[SD, 0.11]
Recall-P	0.32	0.63	0.64	0.79	0.18	0.53	0.67	0.75
	[SD, 0.19]	[SD, 0.18]	[SD, 0.07]	[SD, 0.1]	[SD, 0.16]	[SD, 0.13]	[SD, 0.12]	[SD, 0.09]
Precision-L	0.60	0.47	0.69	0.75	0.72	0.43	0.60	0.84
	[SD, 0.20]	[SD, 0.23]	[SD, 0.18]	[SD, 0.11]	[SD, 0.24]	[SD, 0.23]	[SD, 0.23]	[SD, 0.14]
Recall-L	0.37	0.86	0.79	0.87	0.26	0.71	0.74	0.67
	[SD, 0.09]	[SD, 0.09]	[SD, 0.15]	[SD, 0.08]	[SD, 0.14]	[SD, 0.10]	[SD, 0.13]	[SD, 0.13]
F1-L	0.44	0.54	0.71	0.80	0.37	0.50	0.63	0.73
	[SD, 0.10]	[SD, 0.11]	[SD, 0.09]	[SD, 0.09]	[SD, 0.16]	[SD, 0.20]	[SD, 0.15]	[SD, 0.11]
|VD|(%)	54.4	26.9	17.6	12.3	77.4	30.6	37.6	13.7
	[SD, 22.1]	[SD, 20.0]	[SD, 11.2]	[SD, 12.7]	[SD, 16.5]	[SD, 18.6]	[SD, 51.5]	[SD, 9.7]
HD95	19.5	15.9	10.8	5.27	30.6	21.8	19.5	17.1
	[SD, 8.6]	[SD, 16.1]	[SD, 6.7]	[SD, 3.15]	[SD, 20.9]	[SD, 22.9]	[SD, 18.8]	[SD, 21.0]
AUC	0.53	0.66	0.61	0.84	0.54	0.60	0.60	0.84
	[SD, 0.21]	[SD, 0.12]	[SD, 0.11]	[SD, 0.07]	[SD, 0.28]	[SD, 0.20]	[SD, 0.20]	[SD, 0.10]

Note:—HD95 indicates modified Hausdorff distance (mm); P, pixel; L, lesion; |VD| = absolute volume difference.

number of lesions (Online Supplemental Data) and their volumes (Fig 4) are also shown. The predicted WMH lesion volumes from StackGen-Net were highly correlated ($r = 0.99$) and were not significantly different from WMH volumes in ground truth ($P = .15$).

Table 3 compares human interobserver variability on the 2 test cohorts. The average intraobserver variability in Dice scores in observer 1 annotations on test cohort 1 was 0.70 (median = 0.71). The average pair-wise agreement in Dice scores between humans, calculated as an average of observer 1 versus observer 2, was 0.67 (median = 0.73) and 0.66 (median = 0.72) for test cohorts 1 and 2, respectively. The average agreement between human observers and StackGen-Net was 0.70 (median = 0.74) and 0.70 (median = 0.73) in these cohorts. Although the average pair-wise Dice scores for StackGen-Net were higher compared with human observers, we did not find this difference to be significant ($P = .22$).

DISCUSSION

Data and Architecture

In this work, we present the use of a stacked generalization of CNNs trained on 2.5D patches from orthogonal 3D-FLAIR orientations to improve WMH segmentation performance. The substantial improvement in performance as we move from UNET2D to DeepUNET2D illustrates the benefits of the convolutional blocks in the proposed architecture. The impact of additional spatial context provided by 2.5D training patches is evident in the superior performance of DeepUNET3D over its 2D counterpart.

The use of 2.5D training patches can be beneficial when working with a limited collection of annotated data or computational burden in optimally training a 3D network with 3D training patches.[24] Furthermore, in addition to random initialization in an ensemble framework, training each orthogonal CNN with 2.5D patches from a different orientation can provide training data diversity, a feature crucial to any ensemble-based training model.

The choice of CNN architecture, weights initialization, and hyperparameters has been shown to affect the task-specific performance of a CNN.[10,17] An ensemble of CNNs has been shown to average away the variances in the solution and model- and configuration-specific behaviors.[17] We also observed a similar trend (Table 1), in which the ensemble combination of CNNs performed better compared with individual CNNs in the ensemble.

The stacked generalization of orthogonal CNNs, with a higher Dice score and a lower absolute VD, outperformed individual DeepUNET3D CNNs or their ensemble combination using averaging or majority voting. An averaging ensemble assigns equal weights to WMH posteriors from individual CNNs, whereas majority voting prefers a majority consensus. Stacked generalization, on the other hand, learns a new functional mapping from individual CNN predictions in the ensemble to the target labels. This allows the Meta CNN to deduce the bias rate of individual DeepUNET3D CNNs in the ensemble and compensate for their flaws. In our experiments, we observed a difference in segmentation performance between the orthogonal CNNs, possibly due to learning different lesion characteristics that may depend on orientation. A stacked generalization framework is well-suited to learn and combine performance gains from the orthogonal CNNs. StackGen-Net is able to accurately detect WMHs, even when a majority in the ensemble predict a false-negative (Fig 2).

Human Observer Variability

The inter- and intraobserver variability in Dice scores between 2 experienced neuroradiologists reinforces the subjective nature of manual WMH annotations, even in the presence of pre-established annotation guidelines. The use of a trained CNN, with its deterministic framework, already eliminates intraobserver variability in predictions for a given FLAIR volume. Because we did not find the improvements in Dice scores between StackGen-Net and observers to be significant, we can say that StackGen-Net performance is comparable with human interobserver variability.

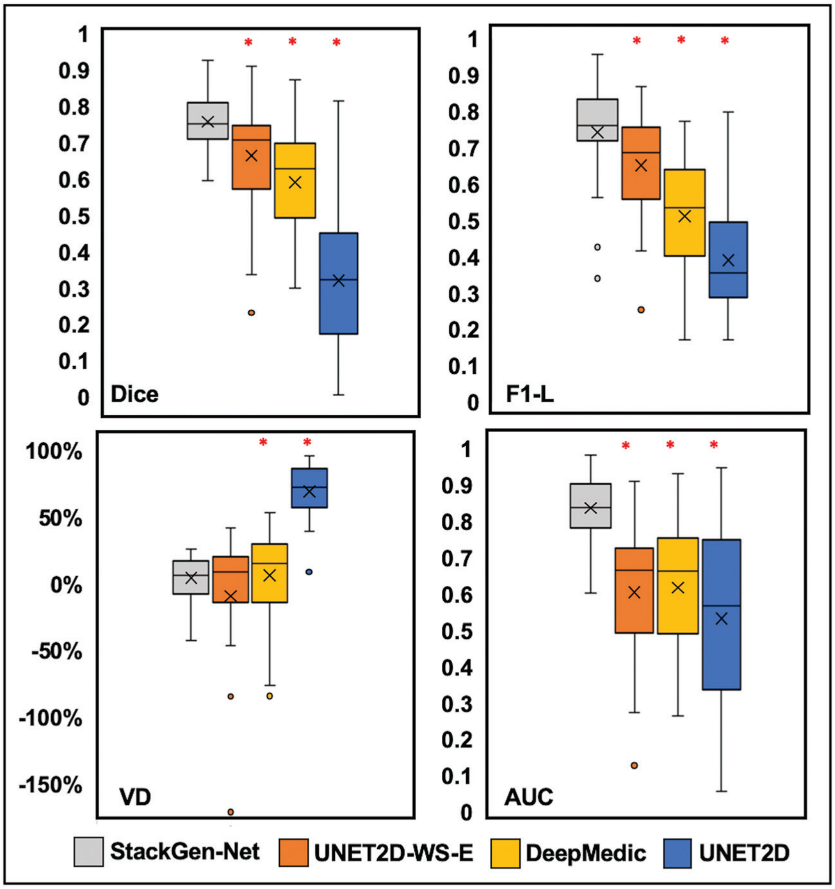

FIG 3. Boxplot comparison of Dice scores, lesion-based F1 (F1-L), volume difference (VD), and area under precision-recall curve (AUC) scores on the test set. We found a significant improvement in Dice scores, AUC, and F1-L in StackGen-Net compared with other WMH segmentation techniques compared here. The *asterisk* denotes $P < .001$ (2-sided paired *t* test, $n = 29$).

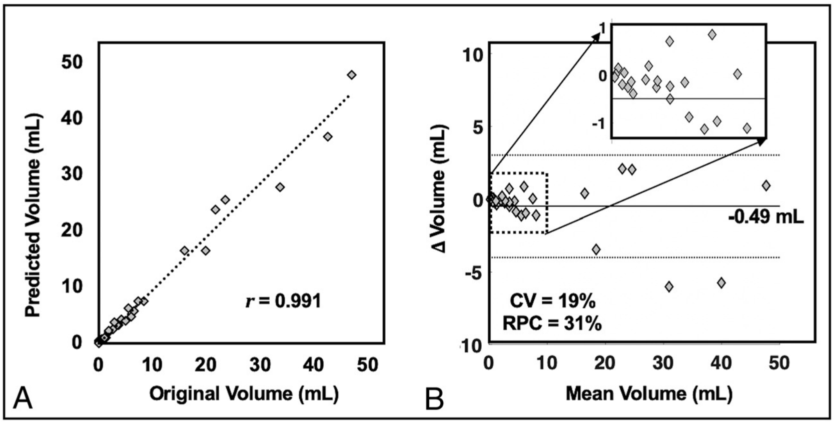

FIG 4. Correlation between WMH volumes (milliliters) in ground truth annotations and StackGen-Net predictions. *A*, We observed a strong correlation between the predictions and ground truth. *B*, Bland-Altman plot shows a good agreement in WMH volumes between the ground truth annotations and StackGen-Net predictions. We found no significant differences between the 2 volumes ($P = .15$, $n = 29$). The coefficient of variation (CV) and the repeatability coefficient (RPC) are also shown.

Comparison with Literature and Limitations

A wide range of Dice scores (0.51–0.80) have been reported in the literature for CNN-based WMH segmentation using 2D-FLAIR images.[8-10,25,26] In comparison with some state-of-the-art techniques evaluated on the 2 test cohorts in this study, we observed higher average Dice scores of 0.76 and 0.75, respectively. Although the Dice scores reported in this work are slightly lower than those in some of these earlier studies, the human interobserver variability baseline (0.67 compared with 0.77–0.79 reported in the literature) is also low in our study cohort.

The extent of WMH burden has been reported to affect evaluation metrics such as the Dice score.[8] The Online Supplemental Data show the histogram of WMH volumes in our study cohort ($n = 50$) and scatterplots of total WMH volumes and average WMH volumes versus Dice scores on the test cohorts ($n = 29$). We observed that most subjects in our study cohort had low WMH volumes (8.04 [SD, 11.3] mL), which are associated with lower Dice scores. For a comparable assessment, we trained and evaluated some of these state-of-the-art techniques on our 3D-FLAIR dataset.

The orthogonal CNNs in StackGen-Net exploit the 3D nature of FLAIR acquisition and combine WMH information from the 3 orthogonal planes for segmentation. The use of 3D convolutions may result in suboptimal performance when training on anisotropic 2D-FLAIR images with thick slices.[9,11] Interpolation to 3D space may affect the performance of 3D CNNs as a result of the blurring introduced along the slice direction. A similar observation was also made in Kuijf et al[12] regarding the results of the WMH segmentation challenge on 2D-FLAIR, in which most methods that used 3D convolutions appeared to perform poorly, ranking near the bottom.

Although 3D-FLAIR images are being widely used in research protocols such as ADNI, their clinical usage is not widespread. The clinical applicability of the proposed technique

Table 3: Interobserver variability in Dice scores[a]

	Test Cohort 1		Test Cohort 2	
	Observer 2	StackGen-Net	Observer 2	StackGen-Net
Observer 1	0.68 (0.72)	0.76 (0.74)	0.66 (0.72)	0.74 (0.75)
Observer 2		0.65 (0.66)		0.65 (0.72)

[a] Mean (median) pair-wise Dice scores.

needs to be further investigated on clinical 2D-FLAIR images to understand the impact of blurring on the detection of smaller lesions. Additionally, the CNNs in this study were all trained/evaluated on cohorts that excluded other pathologies that may produce hyperintensities on FLAIR images; the applicability of these CNNs requires further investigation on such images.

Although our study cohort is small compared with other published deep learning–based studies, the use of a 2.5D patch-based training framework, combined with data augmentation, has been useful to avoid the problem of limited annotated training data and overfitting. StackGen-Net, trained on images from a single scanner type, showed consistently improved performance on an independent cohort, demonstrating generalizability on images spanning multiple institutions and scanner manufacturers.

Clinical Outcomes

Results in the test cohorts show that StackGen-Net detects WMHs on 3D-FLAIR images with high Dice scores and lesion-wise F1. Fast and efficient 3D CNN architectures for WMH segmentation, such as StackGen-Net, can be used for the automatic, quantitative, and fast evaluation of WMH extent. With the demonstrated generalizability on a subset of ADNI data, the multiplanar StackGen-Net framework can be easily applied to larger 3D-FLAIR based longitudinal data repositories, including ADNI, to study the relationship between WMH burden and cognition. In conjunction with clinical visual rating scores, accurate WMH volume estimation can provide a better understanding of the relationship between lesion burden and clinical outcomes.

CONCLUSIONS

In this work, a stacked generalization of 3D orthogonal CNNs (StackGen-Net) was proposed to detect WMHs using multiplanar information from 3D-FLAIR images. We demonstrated that a stacked generalization ensemble outperforms traditional ensemble combinations as well as some state-of-the-art WMH detection frameworks. We also showed that we can reliably detect and quantify WMH in a time-efficient manner with performance comparable to human interobserver variability.

Disclosures: Gloria Guzman Perez-Carrillo—*UNRELATED*: *Patents (Planned, Pending or Issued)*: Quantitative Differentiation of Tumor Heterogeneity using Diffusion MR Imaging Data, Patent No. 016652-PRO1; *Stock/Stock Options*: Vanguard Mutual Funds. Mahesh Bharath Keerthivasan—*OTHER RELATIONSHIPS*: I am currently a paid employee of Siemens Healthineers.

REFERENCES

1. Fazekas F, Kleinert R, Offenbacher H, et al. **Pathologic correlates of incidental MRI white matter signal hyperintensities.** *Neurology* 1993;43:1683–83 CrossRef Medline
2. Brickman AM, Meier IB, Korgaonkar MS, et al. **Testing the white matter retrogenesis hypothesis of cognitive aging.** *Neurobiol Aging* 2012;33:1699–1715 CrossRef Medline
3. Chutinet A, Rost NS. **White matter disease as a biomarker for long-term cerebrovascular disease and dementia.** *Curr Treat Options Cardiovasc Med* 2014;16:392 CrossRef Medline
4. Carmichael O, Schwarz C, Drucker D, et al. **Longitudinal changes in white matter disease and cognition in the first year of the Alzheimer Disease Neuroimaging Initiative.** *Arch Neurol* 2010;67:1370– 78 CrossRef Medline
5. Bendfeldt K, Blumhagen JO, Egger H, et al. **Spatiotemporal distribution pattern of white matter lesion volumes and their association with regional grey matter volume reductions in relapsing-remitting multiple sclerosis.** *Hum Brain Mapp* 2010;31:1542–55 CrossRef Medline
6. Fazekas F, Chawluk JB, Alavi A, et al. **MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging.** *AJR Am J Roentgenol* 1987;149:351–56 CrossRef Medline
7. Litjens G, Kooi T, Bejnordi BE, et al. **A survey on deep learning in medical image analysis.** *Med Image Anal* 2017;42:60–88 CrossRef Medline
8. Rachmadi MF, Valdés-Hernández MD, Agan ML, et al; Alzheimer's Disease Neuroimaging Initiative. **Segmentation of white matter hyperintensities using convolutional neural networks with global spatial information in routine clinical brain MRI with none or mild vascular pathology.** *Comput Med Imaging Graph* 2018;66:28–43 CrossRef Medline
9. Guerrero R, Qin C, Oktay O, et al. **White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks.** *Neuroimage Clin* 2018;17:918–34 CrossRef Medline
10. Li H, Jiang G, Zhang J, et al. **Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images.** *Neuroimage* 2018;183:650–65 CrossRef Medline
11. Ghafoorian M, Karssemeijer N, Heskes T, et al. **Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities.** *Sci Rep* 2017;7:1–12 CrossRef Medline
12. Kuijf HJ, Casamitjana A, Collins DL, et al. **Standardized assessment of automatic segmentation of white matter hyperintensities and results of the WMH segmentation challenge.** *IEEE Trans Med Imaging* 2019;38:2556–68 CrossRef Medline
13. Bink A, Schmitt M, Gaa J, et al. **Detection of lesions in multiple sclerosis by 2D FLAIR and single-slab 3D FLAIR sequences at 3.0 T: initial results.** *Eur Radiol* 2006;16:1104–10 CrossRef Medline
14. Valverde S, Cabezas M, Roura E, et al. **Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach.** *Neuroimage* 2017;155:159–68 CrossRef Medline
15. Kushibar K, Valverde S, González-Villà S, et al. **Automated subcortical brain structure segmentation combining spatial and deep convolutional features.** *Med Image Anal* 2018;48:177–86 CrossRef Medline
16. Kamnitsas K, Ledig C, Newcombe VFJ, et al. **Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation.** *Med Image Anal* 2017;36:61–78 CrossRef Medline
17. Kamnitsas K, Bai W, Ferrante E, et al. **Ensembles of multiple models and architectures for robust brain tumour segmentation.** In: *International MICCAI Brainlesion Workshop:* Springer; 2017: 450–62 Accessed September 14, 2017

18. Wolpert DH. **Stacked generalization.** *Neural Networks* 1992;5:241–59 CrossRef

19. Ronneberger O, Fischer P, Brox T**. U-Net: convolutional networks for biomedical image segmentation.** May 2015. http://arxiv.org/abs/1505.04597. Accessed July 19, 2019

20. Smith SM. **Fast robust automated brain extraction.** *Hum Brain Mapp* 2002;17:143–55 CrossRef Medline

21. Tustison NJ, Avants BB, Cook PA, et al. **N4ITK: improved N3 bias correction.** *IEEE Trans Med Imaging* 2010;29:1310–20 CrossRef Medline

22. Abadi M, Agarwal A, Barham P, et al. **Tensorflow: large-scale machine learning on heterogeneous distributed systems.** February 28, 2016. https://arxiv.org/abs/1603.04467v2. Accessed April 25, 2020

23. Schmidt P, Gaser C, Arsic M, et al. **An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis.** *Neuroimage* 2012;59:3774–83 CrossRef Medline

24. Xing Y, Wang J, Chen X, et al. **2.5D convolution for RGB-D semantic segmentation.** *In: Proceedings of the International Conference on Image Processing (ICIP),* Taipei, Tiwan. September 22–25, 2019 CrossRef

25. Ghafoorian M, Karssemeijer N, Heskes T, et al. **Deep multi-scale location-aware 3D convolutional neural networks for automated detection of lacunes of presumed vascular origin.** *Neuroimage Clin* 2017;14:391–99 CrossRef Medline

26. Duong MT, Rudie JD, Wang J, et al. **Convolutional neural network for automated flair lesion segmentation on clinical brain MR imaging.** *AJNR Am J Neuroradiol* 2019;40:1282–90 CrossRef Medline