

Generic Contrast Agents

Our portfolio is growing to serve you better. Now you have a *choice*.



[VIEW CATALOG](#)

AJNR

This information is current as of May 15, 2025.

Automated Segmentation of MRI White Matter Hyperintensities in 8421 Patients with Acute Ischemic Stroke

Hosung Kim, Wi-Sun Ryu, Dawid Schellingerhout, Jonghyeok Park, Jinyong Chung, Sang-Wuk Jeong, Dong-Seok Gwak, Beom Joon Kim, Joon-Tae Kim, Keun-Sik Hong, Kyung Bok Lee, Tai Hwan Park, Jong-Moo Park, Kyusik Kang, Yong-Jin Cho, Byung-Chul Lee, Kyung-Ho Yu, Mi Sun Oh, Soo Joo Lee, Jae-Kwan Cha, Dae-Hyun Kim, Jun Lee, Man Seok Park, Hee-Joon Bae and Dong-Eog Kim

AJNR Am J Neuroradiol 2024, 45 (12) 1885-1894

doi: <https://doi.org/10.3174/ajnr.A8418>

<http://www.ajnr.org/content/45/12/1885>

Automated Segmentation of MRI White Matter Hyperintensities in 8421 Patients with Acute Ischemic Stroke

Hosung Kim,  Wi-Sun Ryu,  Dawid Schellingerhout, Jonghyeok Park,  Jinyong Chung,  Sang-Wuk Jeong, Dong-Seok Gwak,  Beom Joon Kim, Joon-Tae Kim, Keun-Sik Hong,  Kyung Bok Lee, Tai Hwan Park, Jong-Moo Park, Kyusik Kang, Yong-Jin Cho, Byung-Chul Lee, Kyung-Ho Yu, Mi Sun Oh, Soo Joo Lee, Jae-Kwan Cha, Dae-Hyun Kim, Jun Lee, Man Seok Park,  Hee-Joon Bae, and  Dong-Eog Kim



ABSTRACT

BACKGROUND AND PURPOSE: To date, only a few small studies have attempted deep learning–based automatic segmentation of white matter hyperintensity (WMH) lesions in patients with cerebral infarction; this issue is complicated because stroke-related lesions can obscure WMH borders. We developed and validated deep learning algorithms to segment WMH lesions accurately in patients with cerebral infarction using multisite data sets involving 8421 patients with acute ischemic stroke.

MATERIALS AND METHODS: We included 8421 patients with stroke from 9 centers in Korea. 2D UNet and squeeze-and-excitation (SE)-UNet models were trained using 2408 FLAIR MRIs from 3 hospitals and validated using 6013 FLAIR MRIs from 6 hospitals. WMH segmentation performance was assessed by calculating the Dice similarity coefficient (DSC), the correlation coefficient, and the concordance correlation coefficient compared with a human-segmented criterion standard. In addition, we obtained an uncertainty index that represents overall ambiguity in the voxel classification for WMH segmentation in each patient based on the Kullback-Leibler divergence.

RESULTS: In the training data set, the mean age was 67.4 (SD, 13.0) years, and 60.4% were men. The mean (95% CI) DSCs for UNet in internal testing and external validation were, respectively, 0.659 (0.649–0.669) and 0.710 (0.707–0.714), which were slightly lower than the reliability between humans (DSC = 0.744; 95% CI, 0.738–0.751; $P = .031$). Compared with the UNet, the SE-UNet demonstrated better performance, achieving a mean DSC of 0.675 (95% CI, 0.666–0.685; $P < .001$) in the internal testing and 0.722 (95% CI, 0.719–0.726; $P < .001$) in the external validation; moreover, it achieved high DSC values (ranging from 0.672 to 0.744) across multiple validation data sets. We observed a significant correlation between WMH volumes that were segmented automatically and manually for the UNet ($r = 0.917$, $P < .001$), and it was even stronger for the SE-UNet ($r = 0.933$, $P < .001$). The SE-UNet also attained a high concordance correlation coefficient (ranging from 0.841 to 0.956) in the external test data sets. In addition, the uncertainty indices in most patients (86%) in the external data sets were <0.35 , with an average DSC of 0.744 in these patients.

CONCLUSIONS: We developed and validated deep learning algorithms to segment WMH in patients with acute cerebral infarction using the largest-ever MRI data sets. In addition, we showed that the uncertainty index can be used to identify cases in which automatic WMH segmentation is less accurate and requires human review.

ABBREVIATIONS: C_b = bias correction factor; DSC = Dice similarity coefficient; HD = Hausdorff distance; KL = Kullback-Leibler; ReLU = rectified linear unit; SE = squeeze-and-excitation; WMH = white matter hyperintensity

White matter hyperintensities (WMHs), characterized by high signal intensity on T2-weighted MRI and FLAIR MRI, is commonly observed among the elderly.¹ WMHs are

associated with an increased risk of stroke,² adverse stroke outcomes,³ dementia,⁴ and depression.⁵ Manual segmentation of WMH lesions by humans is considered the criterion standard for

Received February 28, 2024; accepted after revision July 9.

From the USC Stevens Neuroimaging and Informatics Institute (H.K.), Keck School of Medicine of USC, University of Southern California, Los Angeles, California; Artificial Intelligence Research Center (W.-S.R., J.P.), J.L.K. Inc, Seoul, Republic of Korea; National Priority Research Center for Stroke and Department of Neurology (W.-S.R., J.C., S.-W.J., D.-S.G., D.-E.K.), Dongguk University Ilsan Hospital, Goyang, Republic of Korea; Department of Neuroradiology and Imaging Physics (D.S.), The University of Texas M.D. Anderson Cancer Center, Houston, Texas; Biomaging Data Curation Center (J.C., D.-S.G., D.-E.K.), KOREA-BioData Station, Daejeon, Republic of Korea; Department of Neurology (B.J.K., H.-J.B.), Seoul National University Bundang Hospital, Seongnam, Republic of Korea; Department of Neurology (J.-T.K., M.S.P.), Chonnam National University Hospital, Gwangju, Republic of Korea; Department of Neurology (K.-S.H., Y.-J.C.), Inje University Ilsan Paik Hospital, Goyang, Republic of Korea; Department of Neurology (K.B.L.),

Soonchunhyang University Hospital, Seoul, Republic of Korea; Department of Neurology (T.H.P.), Seoul Medical Center, Seoul, Republic of Korea; Department of Neurology (J.-M.P.), Uijeongbu Eulji Medical Center, Uijeongbu, Republic of Korea; Department of Neurology (K.K.), Nowon Eulji Medical Center, Eulji University School of Medicine, Seoul, Republic of Korea; Department of Neurology (B.-C.L., K.-H.Y., M.S.O.), Hallym University Sacred Heart Hospital, Anyang, Republic of Korea; Department of Neurology (S.J.L.), Eulji University Hospital, Daejeon, Republic of Korea; Department of Neurology (J.-K.C., D.-H.K.), Dong-A University Hospital, Busan, Republic of Korea; and Department of Neurology (J.L.), Yeungnam University Hospital, Daegu, Republic of Korea.

H. Kim and W.-S. Ryu equally contributed to this work.

This study was supported by the Multimimistry Grant for Medical Device Development (KMDF_PR_20200901_0098), the National Priority Research Center Program Grant

SUMMARY

PREVIOUS LITERATURE: WMH on FLAIR MRI is a quantifiable risk factor for stroke and dementia. Manual WMH segmentation is laborious, supporting automated segmentation methods, but these face difficulties in implementation. Specifically, cerebral infarcts (acute) can obscure the boundaries of WMH lesions (chronic), degrading performance. The few published studies in the current literature did not have sufficient data to undergo full external validation, limiting their ability to address the well-known “domain shift problem”: ie, an algorithm that performs well in its source domain but poorly when applied in the target domain.

KEY FINDINGS: We developed deep learning algorithms to segment WMH in patients with ischemic stroke using the largest-ever data set ($n = 8421$ patients), with full validation. Our SE-UNet algorithm achieved high segmentation performance, with DSC values from 0.672 to 0.744 across multiple validation data sets (compared with expert criterion standards), with low (<0.35) uncertainty indices in 86% of patients.

KNOWLEDGE ADVANCEMENT: Deep learning algorithms developed and using large MRI data sets from multicenter patients with stroke can accurately segment WMH lesions, distinguishing FLAIR lesions from DWI-positive infarcts without relying on DWIs. The uncertainty index can be used to identify those WMH segmentation cases that require human inspection.

volumetric assessment of WMHs.⁶ However, given the high prevalence of variously-sized scattered WMHs, such a manual procedure is laborious, time-consuming, and prone to rater-dependent bias and errors,⁷ particularly in large multicenter studies.

Approximately 40% of stroke survivors eventually have cognitive impairment.⁸ The extent and expansion of WMHs are intimately correlated with vascular cognitive impairment.^{4,9,10} Nevertheless, there is little evidence available to indicate that attenuating WMH progression can prevent functional decline, partly due to inconsistent volumetric measurement of WMHs.¹¹ Hence, to further advance our understanding of WMH in patients with ischemic stroke, precise and consistent WMH segmentation is required. Recent advances in deep learning approaches have markedly improved the automatic segmentation of brain lesions.^{12–14} Several studies on healthy elderly brains yielded promising results for fully automated WMH segmentation.^{12,15,16} However, patients with cerebrovascular lesions, such as acute or chronic infarcts, complicate these analyses because focal stroke-related lesions can obscure WMH borders.¹⁷ This area is understudied, representing a gap in our clinical armamentarium, which requires improved methods. Automated segmentation of WMH versus infarcts will contribute to more accurate quantification of acute-versus-chronic ischemia-related MR lesions. This contribution will help better predict poststroke outcomes such as WMH-related clinical worsening, neurologic recovery, and functional outcomes.³ It will also support research on WMH in stroke and dementia.

Only a few studies have attempted automatic segmentation of WMH lesions in patients with ischemic stroke.^{18,19} One study ($n = 250$) using convolutional neural networks achieved a Dice similarity coefficient (DSC) of near 70% in the test data set.¹⁸ A more recent study ($n = 429$) used the state-of-the-art UNet and UNet with squeeze-and-excitation (SE) blocks (SE-UNet),

achieving a higher DSC value of 74%–76%.¹⁹ These single-center studies with small sample sizes cannot address the notorious “domain shift problem,” in which an algorithm that performs well in the source domain proceeds to perform poorly in the target domain.²⁰ Moreover, because the acute and chronic infarcts obscure the boundaries of WMHs, the interrater and intrarater reliability of the manual segmentation of WMHs is reduced. This less reliable voxelwise labeling of WMHs reduces the quality of the training data and compounds the problem. The impact of this factor on the segmentation performance of deep learning algorithms has not yet been systematically investigated.

In this study, we first trained UNet and SE-UNet to segment WMH lesions using 3 different stroke center data sets involving 2408 patients with ischemic stroke and subsequently validated the deep learning algorithms using 6 separate stroke center data sets involving 6013 patients. Next, we used an uncertainty measure²¹ (based on the Kullback-Leibler divergence²²) from the UNet to investigate the possibility of predicting the accuracy of WMH segmentation. We also assessed lesion information that could affect the segmentation accuracy, such as WMH burden and infarct location and volume.

MATERIALS AND METHODS

Data Sets

The Korean Nationwide Image-Based Stroke Database project is a prospective multicenter study in Korea.^{2,3,17,23} From May 2011 to November 2013, we consecutively enrolled 10,423 patients with ischemic stroke who were admitted to the 9 participating centers within 7 days of symptom onset. We excluded the following patients: those with a contraindication to MRI ($n = 315$), poor quality or unavailability of FLAIR MRI or DWI ($n = 1632$), and MRI registration error ($n = 55$), leaving 8421 evaluable patients. The institutional review boards of all the participating centers approved the study. All patients or their legally authorized representatives provided a written informed consent for study participation. Brain MRI was performed on 1.5T ($n = 6583$) or 3T ($n = 1803$) MRI systems. FLAIR image protocols were TE = 76–187 ms, TR = 6000–11568 ms, voxel size = $1 \times 1 \times 3$ –7 mm³, spacing = 0.3–1.0 mm, slice thickness = 3–7 mm, FOV = 175–280 mm, and matrix size (row) = 256–768.

(NRF-2021RIA6A1A03038865), and the Basic Science Research Program Grant (NRF-2020RIA2C3008295) of National Research Foundation, funded by the Korean government.

Please address correspondence to Dong-Eog Kim, MD, PhD, Department of Neurology, Dongguk University Ilsan Hospital, 27, Dongguk-ro, Ilsandong-gu, Goyang, South Korea; e-mail: kdongeog@duh.org

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

📄 Indicates article with online supplemental data.

<http://dx.doi.org/10.3174/ajnr.A8418>

All scans were transferred to the Korean Brain MRI Data Center for central data storage and quantitative analysis. As previously reported,^{2,3,17,24-26} each patient's high signal intensity WM lesions on FLAIR images were manually segmented from scratch by 1 of 5 research assistants with at least 5 years' experience in WMH segmentation under careful supervision by an experienced vascular neurologist (W.-S. Ryu; Online Supplemental Data). When chronic WM lesions on FLAIR and acute infarct lesions on DWI overlapped or were adjacent, we determined the extent and distribution of FLAIR WMH on the basis of lesions in the hemisphere contralateral to the acute infarct location, because WMH symmetry of morphology and distribution is quite often observed between hemispheres. In preliminary investigations, we found that the performance of deep learning models reached a plateau at about 2000 patients' FLAIR MRIs. Additionally, we sought to test the robustness of our algorithms across various external data sets. Hence, a total of 2408 patients' FLAIR MRIs from 3 hospitals were designated as a training data set, and the remaining 6013 patients' FLAIR MRIs from 6 hospitals were designated as 6 external validation data sets. A training data set was divided into 6:2:2 ratios by random subsetting as training, validation, and internal test data sets.

Data Preprocessing

We applied 2D B-spline interpolation to resize the FLAIR slices to a dimension of 256×256 pixels. Next, we performed slice-wise intensity normalization in a uint8 format, ensuring that pixel intensities on each slice ranged from 0 to 255. We subsequently performed case-wise intensity equalization to ensure comparability in the imaging data across different MRI vendors using a histogram of 32 bins and shifting the highest peak of the histogram value to 150. Then, we generated a binary brain mask by including all voxels with signal intensities greater than a threshold value of 30 in the histogram domain. We filled tiny holes in this brain mask using `binary_fill_holes` (SciPy.ndimage).²⁷ Last, we performed the Gaussian normalization process to yield pixel intensities with a mean of zero and an SD of 1 for the brain area under the brain mask, thereby producing optimal arrays for training a deep learning model.

Deep Learning Algorithms: UNet versus SE-UNet Ensemble Network Architecture

We used the 2 neural network architectures from the literature,^{28,29} the 2D UNet and the 2D SE-UNet (Online Supplemental Data). The 2D UNet has an encoder path and a decoder path, each with 3 resolution steps. In the encoder path, each layer has three 3×3 convolutions and batch normalization, which is followed by a rectified linear unit (ReLU) activation function, and a 2×2 max pooling layer for downsampling. In the decoder path, each layer uses a deconvolution with a kernel size of 2×2 , followed by three 3×3 convolutions and batch normalization with ReLU. The network has shortcut connections from the layers in the encoder path to the corresponding layers with the same resolution in the decoder path. Finally, a fully connected layer with a 32-channel input was added to the end of the decoder path and activated using a sigmoid function, with random initial nonuniform weights. We used the Dice loss function. For the learning and optimization step, we

used the Adam optimizer with a 1e-3 learning rate and 1e-3/300 decay rate, 50 epochs. For the training, we did not use data augmentation techniques.

The 2D SE-UNet has a architecture similar to that of the 2D-UNet but with an additional SE-block and global average pooling applied after the convolution layer in the downsampling path. The SE block assigns weights to the network channel. Squeeze involves the portion of the layer that conducts global pooling, which embeds global information. Excitation recalibrates adaptively through ReLU and sigmoid activations.²⁸ The 2D-UNet and 2D SE-UNet models were trained using the same training data set. Each model was initialized with different random weights to ensure variety. We arbitrarily set the threshold of 0.5 to designate voxels with WMHs. The scripts for training deep learning models are available at GitHub (<https://github.com/jlk-jhpark/Whitematter-hyper-intensity-segmentation.git>).

The Uncertainty Index That Represents Overall Ambiguity in the Voxel Classification of WMH Segmentation in Each Patient

To calculate uncertainty, we used a deep ensemble of 5 different models. We used soft voting to combine the models. The individual networks share the same architecture but were trained on different 90% subsets of the training data set, each with different random initializations to ensure variability.^{30,31} The voxel-level Kullback-Leibler (KL) divergence values were calculated using the SciPy.stats.entropy module.³⁰ In the formula, the average of the predicted probabilities from the 5 models was used as input data, along with a prior probability of 0.5. This prior probability indicates that each voxel has an equal chance of being predicted as WMH. Because we assigned a prior probability of 0.5, the large KL divergence indicated low uncertainty. Next, we calculated the proportion of voxels with a substantial divergence value (<0.5) of all predicted voxels in each patient. Thus, the following patient-level uncertainty was calculated as

Uncertainty index =

$$\frac{\sum \text{predicted WMH pixels with divergence value} < 0.5}{\sum \text{predicted WMH pixels}}.$$

Segmentation Performance Evaluation Metrics

The following metrics were used to evaluate the automated WMH segmentation methods compared with manual WMH segmentation.

- DSC = $2 \times (\text{true-positive WMH voxels}) / (\text{true WMH voxels} + \text{predicted WMH voxels})$
- Hausdorff distance (HD) between 2 sets of points, representing segmented regions (true-positive WMH X and predicted WMH Y), is defined as follows: $HD(X, Y) = \max[HD(X, Y), HD(Y, X)]$, where the one-sided HD from X to Y is defined as $hd(X, Y) = \max_{x \in X} \min_{y \in Y} \|x - y\|$.

Correlation coefficient (R) =

$$\frac{n \times \sum xy - \sum x \sum y}{\sqrt{\left[n \sum x^2 - \left(\sum x \right)^2 \right] \left[n \sum y^2 - \left(\sum y \right)^2 \right]}}.$$

Where n indicates number of samples; x , automated WMH volume; y , manual WMH volume.

Experiment and Analysis

We implemented the networks in Python 3.6.10 using Keras 2.4.3. The baseline network (UNet, SE-UNet) for training was trained on a GeForce GTX 1080 Ti GPU with 11.0 CUDA Version, taking 13 minutes per epoch for 41,854, 14,303, 14,039 slices with the training/validation/test split of 60%, 20%, 20%. We empirically chose a batch size of 8.

To evaluate the interrater reliability of WMH segmentation, we computed the mean DSC and mean correlation coefficient among 5 research assistants using 135 randomly-sampled FLAIR images.

We validated deep learning models in 6 external data sets using DSC. A relationship between automated WMH volume and manual WMH volume was assessed using a correlation coefficient. After training on 100, 200, 500, 1000, 1500, and 2408 patients' FLAIR MRIs, the mean (SD) DSC in the combined external data set was calculated to evaluate the WMH segmentation performance as the amount of training data increases. In addition, we assessed the model performance after stratifying patients by infarction volume, infarction location, uncertainty, and WMH subregion. The infarction volume was categorized with cutoff points of 1.7 and 14 mL.¹⁵ WMH volumes were categorized into tertiles of the combined external validation data sets. The infarction location was categorized as the cortex, corona radiata, basal ganglia and internal capsule, thalamus, midbrain, pons, medulla, and cerebellum. Infarct location information was retrieved from a prospective stroke registry. In addition, the external validation data set was divided according to uncertainty levels of 0.2, 0.4, 0.6, and 0.8. To differentiate WMH subregions (periventricular versus deep), we trained a deep learning algorithm that automatically segments the ventricles in FLAIR images. A total of 145 patients' FLAIR images were used for the ventricle-segmentation learning using UNet, and the model showed a validation DSC of 0.9034. Then, a distance transform based on the OpenCV algorithm³² was used to define the periventricular region within 10 and 7.5 mm from the ventricular surface. WMH volumes observed in this region were classified as periventricular WMH, whereas WMH volumes observed outside this region were classified as deep WMH. In addition, we calculated the mean and 95% CI of the uncertainty index for the external validation data sets after stratification by DSCs (with the cutoff values of 0.60, 0.65, 0.70, 0.75, and 0.80) and predicted WMH volume (in quartiles).

Statistical Analysis

To compare subjects' characteristics between training and test data sets, we used the ANOVA or the Kruskal-Wallis test for the analysis of continuous variables and the χ^2 test for categorical variables as appropriate. Mean DSC and Hounsfield unit values were compared between the UNet and the SE-UNet using paired t tests. We compared the segmentation performance between automated methods that were applied to the test data set using ANOVA and Bonferroni post hoc analyses. We then compared the segmentation performance between test data subgroups stratified by infarct location, tertiles of WMH volume,

uncertainty, and WMH subregion. To evaluate the agreement between automated WMH segmentation volume by SE-UNet and manual segmentation volume, we used the concordance correlation coefficient, which is commonly used to assess agreement between 2 raters or 2 methods to measure a response when the data are measured on a continuous scale.³³ A concordance correlation coefficient of >0.8 has been suggested as an excellent level of agreement.³⁴ In addition, we measured a bias correction factor (C_b), which measures how far the best-fit line deviates from the 45° line (measure of accuracy).³³ When there is no deviation from the 45° line, the C_b is 1. The Pearson correlation coefficient ρ measures how far each observation deviates from the best-fit line (measure of precision).³³ Data were analyzed using Stata (StataCorp), and a 2-sided P value $< .05$ considered statistically significant.

RESULTS

Baseline Characteristics

Compared with the subjects in the combined data set for training and validation and internal testing (mean age, 67.4 [SD, 13.0] years; 60.4% men), those in the 6 data sets for external testing had significantly different clinical characteristics in terms of age, last-known-well time to admission, stroke subtype, and the frequencies of previous stroke history, coronary artery disease, and other cardiovascular risk factors (Table). Moreover, MRI vendors, magnetic field strength, and imaging parameters were variable across the data sets (Online Supplemental Data). Within the data sets for external testing, subjects' mean age and the prevalence of risk factors also varied across the data sets. These results show that we used distinct training and validation data sets, which should make the deep learning models more robust and effective on new, previously unseen data.

Algorithm Performance for Automatic Segmentation of WMH Lesions

Overall, the mean DSC values for the UNet segmentation on the internal test and external validation data sets were, respectively, 0.659 (95% CI, 0.649–0.669) and 0.710 (95% CI, 0.707–0.714), which were slightly lower than the reliability between human raters (mean, 0.744; 95% CI, 0.738–0.751; $P = .031$; Online Supplemental Data). The mean HD values for the UNet on the internal test and external validation data sets were 10.44 (95% CI, 9.75–11.14) and 10.44 (95% CI, 10.22–10.65), respectively. The mean values for the SE-UNet segmentation on the internal test and external validation data sets were, respectively, 0.675 (95% CI, 0.666–0.685) and 0.722 (95% CI, 0.719–0.726), exhibiting overall superior performance over the UNet (both $P < .001$). The mean HD for the SE-UNet on the internal test and external validation data sets were, respectively, 9.09 (95% CI, 8.43–9.74) and 8.98 (95% CI, 8.77–9.18), again demonstrating superior performance over the UNet (both $P < .001$). The mean (SD) processing times for the UNet and the SE-UNet on the overall external data sets were 67.0 (9.6) ms and 74.0 (6.6) ms, respectively ($P < .001$).

In addition, deep learning algorithms showed similarly high DSCs across multiple external validation data sets with the values

Baseline characteristics of the training, internal validation, and external validation data sets^a

	Training, Validation, and Internal Test (n = 2408)	External Validation 1 (n = 1105)	External Validation 2 (n = 838)	External Validation 3 (n = 2654)	External Validation 4 (n = 428)	External Validation 5 (n = 571)	External Validation 6 (n = 417)	P
Age (yr)	67.4 (13.0)	68.2 (12.8)	67.5 (13.3)	68.1 (12.5)	70.0 (11.7)	67.9 (13.1)	69.7 (12.7)	<.01
Sex, male	1419 (60.4)	568 (54.3)	491 (60.8)	1497 (57.7)	222 (52.7)	346 (63.0)	235 (58.6)	
LKW to admission (hr)	12.0 (3.5–37.2)	13.1 (3.5–32.0)	13.9 (3.2–39.0)	6.8 (2.6–24.0)	11.4 (3.1–34.8)	13.5 (3.3–39.4)	17.7 (5.7–51)	<.001 ^a
Prestroke mRS score >2	285 (12.1)	171 (16.4)	107 (13.2)	426 (16.4)	55 (13.1)	94 (17.1)	115 (28.7)	<.001
Admission NIHSS score	4 (2–8)	4 (2–7)	3 (1–8)	4 (2–9)	4 (2–10)	3 (2–8)	4 (2–8)	.11 ^a
Subtype								<.001
LAA	814 (35.4)	331 (32.3)	296 (38.0)	1030 (40.1)	179 (43.0)	197 (36.1)	162 (40.6)	
SVO	469 (20.4)	222 (21.6)	216 (27.7)	221 (8.6)	52 (12.5)	123 (22.5)	98 (24.6)	
CE	525 (22.8)	160 (15.6)	159 (20.4)	625 (24.3)	63 (15.1)	112 (20.5)	65 (16.3)	
Undetermined	435 (18.9)	289 (28.2)	93 (11.9)	654 (25.4)	120 (28.9)	103 (18.9)	64 (16.0)	
Other-determined	59 (2.6)	24 (2.3)	15 (1.9)	42 (1.6)	2 (0.5)	11 (2.0)	10 (2.5)	
Previous stroke	483 (20.5)	250 (23.9)	188 (23.3)	392 (15.1)	112 (26.6)	115 (21.0)	111 (27.7)	<.001
Coronary artery disease	391 (16.6)	95 (9.1)	95 (11.8)	88 (3.4)	37 (8.8)	66 (12.0)	33 (8.2)	<.001
Hypertension	1594 (67.8)	726 (69.4)	559 (69.2)	1585 (61.1)	322 (76.5)	409 (74.5)	311 (77.6)	<.001
Diabetes	791 (33.7)	338 (32.3)	280 (34.7)	730 (28.1)	164 (39.0)	199 (36.3)	158 (39.4)	<.001
Hyperlipidemia	1084 (46.1)	193 (18.5)	324 (40.1)	362 (14.0)	133 (31.6)	268 (48.8)	195 (48.6)	<.001
Smoking, current or quit ≤5 yr	950 (40.4)	391 (37.4)	330 (40.8)	989 (38.1)	162 (38.5)	259 (47.2)	166 (41.1)	<.001
Atrial fibrillation	496 (21.1)	180 (17.2)	158 (20.0)	621 (23.9)	86 (20.4)	114 (20.8)	66 (16.5)	.003
Revascularization	404 (17.2)	147 (14.1)	82 (10.2)	587 (22.6)	73 (17.3)	114 (20.8)	45 (11.2)	<.001
Infarct volume (mL)	1.7 (0.4–9.8)	1.1 (0.3–6.6)	2.0 (0.4–12.2)	3.6 (0.6–17.9)	1.2 (0.3–8.7)	2.2 (0.5–15.4)	1.4 (0.5–9.5)	
WMH volume (mL)	11.4 (5.4–24.7)	14.6 (6.6–30.0)	10.8 (4.2–25.1)	12.7 (6.9–24.3)	14.3 (7.0–28.7)	13.0 (6.7–25.5)	17.7 (8.9–40.0)	<.001 ^a

Note:—CE indicates cardioembolism; LAA, large-artery atherosclerosis; LKW, last known well; mRS, modified Rankin Scale; SVO, small-vessel occlusion.

Data are presented as mean ± SD, median (interquartile range), number (percentage).

^a Kruskal-Wallis test was used.

ranging from 0.672 to 0.744. In line with these findings, WMH volumes that were segmented automatically and manually showed a strong correlation for the UNet ($r = 0.917$, $P < .001$) and the SE-UNet ($r = 0.933$, $P < .001$) in the external validation data sets. Because the SE-UNet outperformed the UNet, the subsequent analyses were performed using the SE-UNet. The mean DSC for the external data sets increased from 0.659 to 0.699 when the training data set size was increased from 100 to 500 (Online Supplemental Data). However, with a larger sample size beyond 500 up to 2048, there was little but statistically significant improvement in DSC (from 0.699 to 0.722; $P < .001$), suggesting a saturation effect.

Impact of Volume and Location of Acute Infarct and WMH on the Accuracy of Automatic Lesion Segmentation

As shown in Fig 1A, the mean DSC (for the external data sets) was significantly lower in the large infarct group (>14 mL, DSC = 0.695; 95% CI, 0.688–0.703) and the moderate infarct group (1.7–14 mL, DSC = 0.721; 95% CI, 0.715–0.727), compared with the small infarct group (<1.7 mL, DSC = 0.737; 95% CI, 0.732–0.742; both $P < .001$). After stratification by infarct location (Online Supplemental Data), mean DSCs were comparable across the groups, except for the medulla infarct group with its DSC being <0.7 .

The mean DSC for the highest tertile of WMH volume was significantly greater than that for the lowest tertile (DSC = 0.617 versus 0.813, $P < .001$; Fig 1B). In addition, the segmentation accuracy for periventricular WMH was significantly higher than that for subcortical WMH (Online Supplemental Data), regardless of the definition of the periventricular region by the distance from the ventricular surface: 10 mm (DSC = 0.751

versus 0.617; $P < .001$) and 7.5-mm cutoff (DSC = 0.751 versus 0.628; $P < .001$).

Predictors of the Accuracy of Automatic WMH Segmentation: Uncertainty Index, Infarct Volume/Location, and WMH Volume

The mean patient-level uncertainty index for the external test data sets was 0.2214 (SD, 0.1184) (see Fig 2 for representative cases with low or high uncertainty indices). The uncertainty index was inversely correlated with the DSC in the combined external data set ($r = -0.753$, $P < .001$; Fig 3A). Regardless of the infarct size, the DSC was similar for subjects in the lowest uncertainty group (0–0.2). In higher uncertainty groups (0.4–1.0), however, large infarcts (>14 mL) were associated with a lower segmentation accuracy than small or moderate-sized infarcts (Fig 1C). Regardless of the infarct locations, the DSC decreased as the uncertainty index increased (Online Supplemental Data). For both automatic and manual WMH segmentations, the uncertainty index decreased as the WMH volume increased (Online Supplemental Data). In the combined external validation data sets, 86% of patients had an uncertainty index of <0.35 (Fig 3B). In these subjects, the mean DSC (0.746) was similar to the interrater reliability of manual segmentations (0.75). When patients were stratified by WMH quartiles and DSCs, we again observed an inverse relationship between the uncertainty index and WMH volume (Fig 1E). Of note, within each WMH volume stratum, there was an inverse relationship between the uncertainty index and the DSC. However, when WMH volumes were higher than ~ 10 mL, the uncertainty indices were lower than ~ 0.2 , regardless of their DSC values. In contrast, when WMH volumes were lower than ~ 4 mL, the uncertainty indexes were higher than ~ 0.3 , regardless of their DSC values.

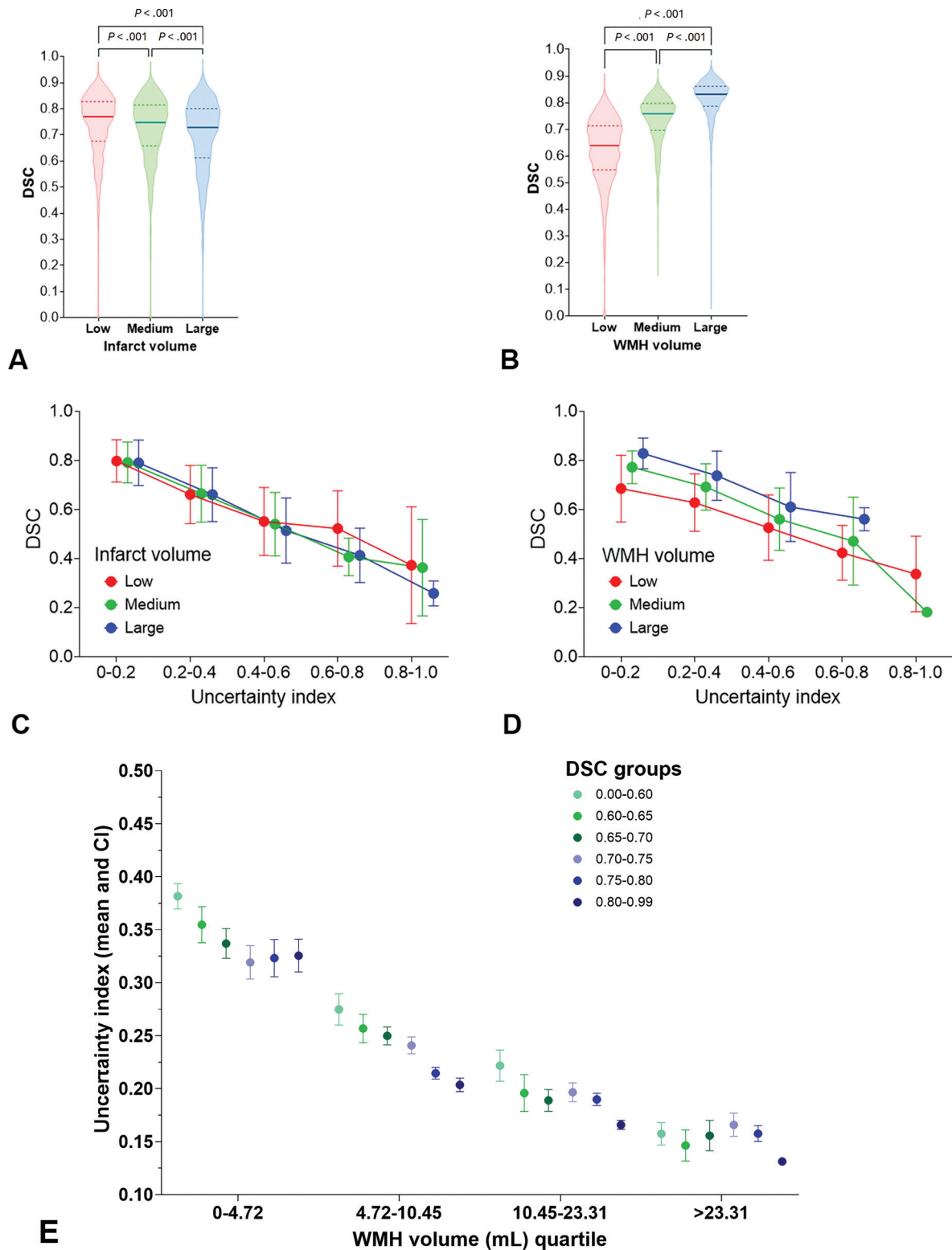


FIG 1. Segmentation performance of the SE-UNet, with stratification by infarct volume, white matter hyperintensity volume, and the uncertainty index in external test data sets. *A*, Violin plot for the performance of WMH segmentation, with stratification by infarct volume (low, <1.7 mL, $n = 2792$; medium, 1.7 – 14 mL, $n = 1785$; and large, >14 mL, $n = 1436$). *B*, Violin plot for the segmentation performance, with stratification by WMH volume (low, <7.6 mL, $n = 1984$; medium, 7.6 – 19.6 mL, $n = 1984$; and large, >19.7 mL, $n = 2045$). *C*, Segmentation performance stratified by the infarct volume and the uncertainty index. *D*, Segmentation performance, stratified by the WMH volume and the uncertainty index. Due to the association between the volume of WMHs and the uncertainty index, none of the patients in the large-volume group had an uncertainty index exceeding 0.8. *E*, The mean and 95% CI of the uncertainty index after stratification of DSCs and WMH volume (in quartiles). In the violin plots, lines and dotted lines indicate median and interquartile range, respectively. In the line graphs, dots and error bars indicate mean (SD), respectively.

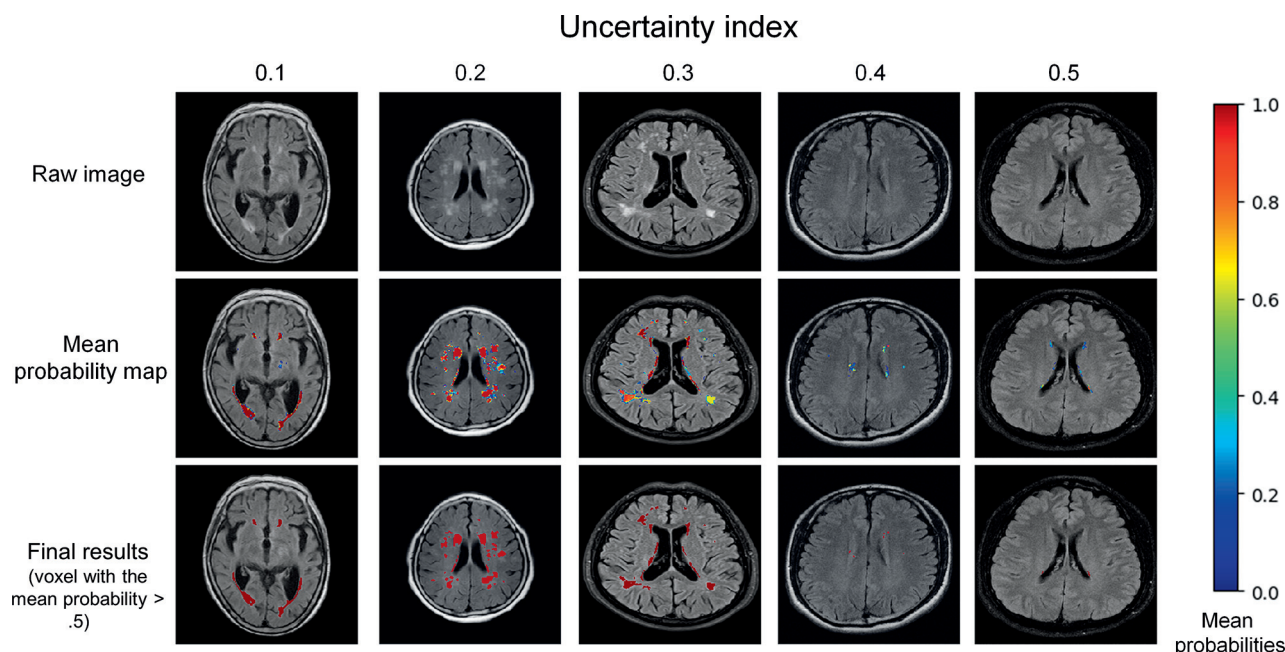


FIG 2. Representative cases of automatic white matter hyperintensity segmentation and their uncertainty indices. *Upper row:* raw images. *Middle row:* red and blue areas, respectively, show areas of agreement and disagreement between human segmentation and deep learning algorithm-based segmentation. *Lower row:* adjudicated final results, with uncertainties resolved.

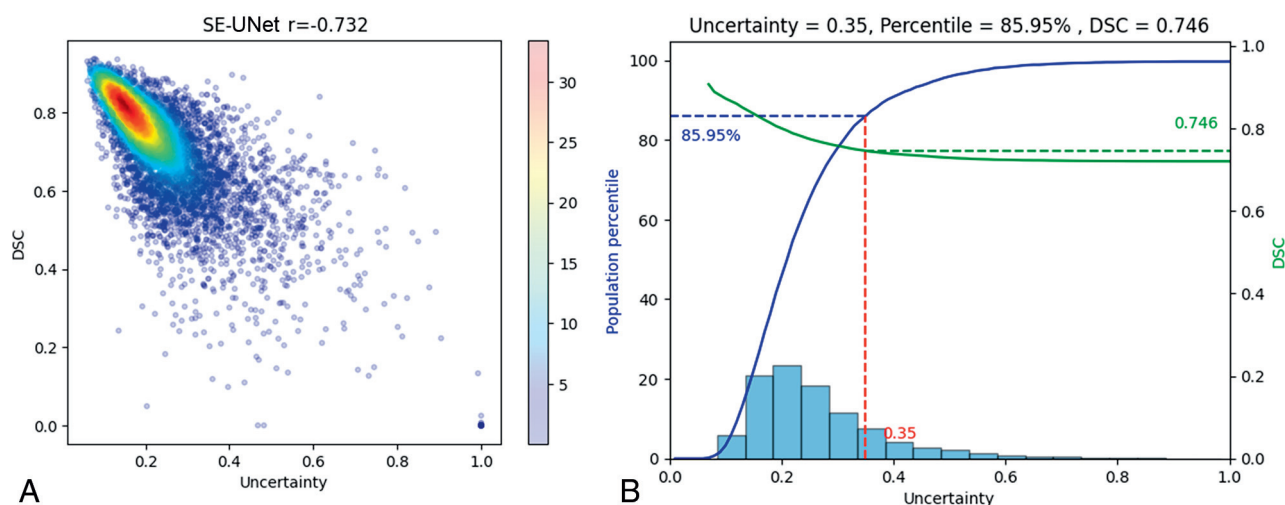


FIG 3. Relationship between the uncertainty index and the DSC. *A*, Density plot between the uncertainty index and the DSC. As expected, there is a negative correlation between the 2 values. *B*, Mean DSC, distribution of subjects, and the uncertainty index. The blue line represents the cumulative percentile of patients, while the green line indicates the mean DSC of cumulative patients. The blue dotted line represents the cumulative percentage of patients when the uncertainty index reaches 0.35 from 0 (red dotted line). The green dotted line shows interhuman DSC. At an uncertainty index of 0.35, accumulated patients comprise approximately 86% of the total population, with a mean DSC of 0.75.

Correlation between Volumes of WMH Segmented Automatically versus Manually

In the combined data set for external testing (6013 patients' MR images from 6 stroke centers), the median WMH volume estimated by the SE-UNet segmentation algorithm was 12.39 mL (interquartile range, 6.02–25.38 mL). There was an excellent correlation between volumes of WMH segmented automatically using the SE-UNet versus manual segmentation by experienced researchers (Fig 4), with the concordance correlation coefficient

and the C_b being 0.931 (95% CI, 0.928–0.934) and 0.990, respectively. When each external data set was assessed separately, the concordance correlation coefficient and C_b ranged from 0.841 to 0.956 and 0.973 to 0.994, respectively (Online Supplemental Data). In the combined external data set, 6 outlier cases that appeared to be far off the diagonal line were identified (Online Supplemental Data). In every case, we found that manual segmentation had erroneously designated acute or chronic infarcts as WMHs.

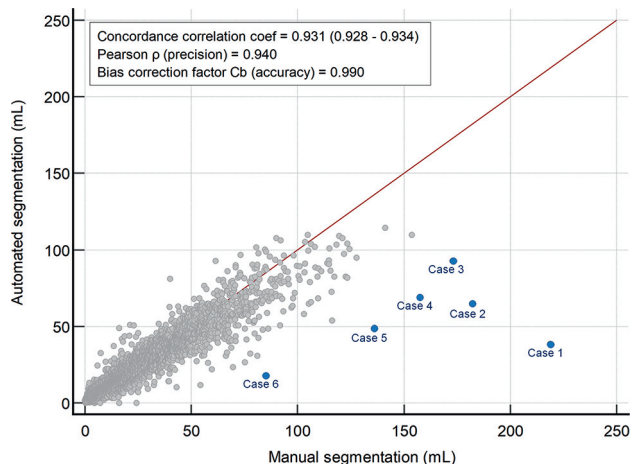


FIG 4. Scatterplot showing a strong correlation between the volumes of WMHs segmented automatically by SE-UNet and manually in external validation. *Gray dots* indicate each subject. The *red line* indicates a 45° line of identity. Outlier subjects (*blue dots*) are reviewed in the Online Supplemental Data. The given equation is $y = 2.1693 + 0.8311x$, where y represents the automatically segmented WMH volume and x represents the manually segmented WMH volume. After eliminating 6 outlier cases from the scatterplot, the slope showed a modest increase, resulting in the equation $y = 2.0671 + 0.8376x$. Coef indicates coefficient.

DISCUSSION

In the present study using multicenter data sets (MR images with voxelwise lesion [ground truth] annotation and prospectively collected clinical data) from a total of 8421 patients with ischemic stroke, we developed and validated deep learning algorithms for automatic WMH segmentation. This, to the best of our knowledge, is the largest-ever data set used for stroke research of this nature (Fig 5). The SE-UNet algorithm had a very high level of segmentation accuracy, with the concordance correlation coefficient and the C_b being, respectively, 0.919 and 0.987, compared with the ground-truth WMH volumes. As previously reported,⁷ however, even among the highly experienced raters, the level of voxelwise agreement was moderate (DSC-based interrater reliability of ~ 0.77). We found a close correlation between a high uncertainty index and a low DSC, suggesting that researchers can use the index to identify those cases that require human inspection following artificial intelligence-based automatic segmentation of WMH. The sample results and demonstration are available at the Web site https://stroke.medihub.ai/login/JBS_WMh_sample (Online Supplemental Data).

Because manual segmentation of WMH is labor-intensive and time-consuming, visual rating scales have been widely used in WMH research using large or multicenter data sets. However,

Previous studies vs. Current study

Year	Authors	Source of images used for deep learning training	Stroke population	Number of patients in the training and validation datasets	Model	DSC for internal validation	DSC for external validation	Source of images used for external validation	Number of patients in the external datasets
2018	Guerrero et al. ¹⁸	1 scanner at 1 hospital, UK	Yes ^a	167	uResNet	0.695	N/A	N/A	N/A
2020	Lee et al. ¹⁹	5 scanners (from 3 vendors) at 3 hospitals, Netherlands and Singapore (MICCAI 2017)	No	170	SEU-net	0.769	N/A	N/A	N/A
2021	Park et al. ¹²	5 scanners (from 3 vendors) at 3 hospitals, the Netherlands and Singapore (MICCAI 2017)	No	170	2D Ensemble U-net	0.81	N/A	N/A	N/A
2021	Sundaresan et al. ¹⁶	1 scanner, a Neurodegenerative cohort (NDGEN), UK	No	21	TrUE-Net	0.89	N/A	N/A	N/A
		1 scanner, Oxford vascular study (OXVASC), UK	Yes ^b	18		0.95			
		5 scanners (from 3 vendors) at 3 hospitals, Netherlands and Singapore (MICCAI 2017)	No	60		N/A			
2022	Li et al. ¹⁵	5 scanners (from 3 vendors) at 3 hospitals, Netherlands and Singapore (MICCAI 2017)	No	60	2D Ensemble SEU-net	0.833	0.783	2 different vendor scanners at 2 hospitals	60
		5 scanners (from 3 vendors) at 5 hospitals, Australia	No	40					
2024	Kim et al.: Current study	6 scanners ^c (from 3 vendors) at 3 hospitals, Korea	Yes	1927	SEU-net	0.675	0.722	7 scanners ^c (from 3 vendors) at 6 hospitals	6013

^anon-disabling lacunar or mild cortical ischemic stroke; ^bminor non-disabling stroke or transient ischemic attack; ^cScanners used for >5% of the study population were counted.

Automatic segmentation of WMHs by our deep learning algorithm, distinguishing the FLAIR lesions from DWI-positive infarcts without relying on DWIs

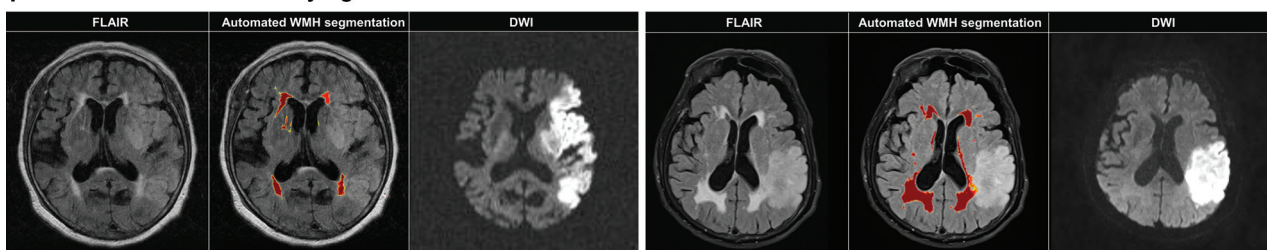


FIG 5. Comparison of previous studies and the current study on automated WMH segmentation using deep learning algorithms, along with representative cases demonstrating the performance of our algorithm. NA indicates not applicable.

visual rating systems showed limited correlation with quantitative assessments of WMH.³⁵ In addition, when examining longitudinal changes in white matter, volumetric measures of WMH provided a more reliable, sensitive, and objective substitute than visual rating scales.³⁶ Moreover, according to a recent study, expert agreement in manually segmenting WMH was low, with DSCs ranging from 0.56 to 0.62.¹¹ In our study, 5 experienced research assistants who have been segmenting WMH for >5 years under careful supervision (including regular audits and consensus discussions) of a vascular neurologist accomplished a DSC of ~ 0.77 . Taken together, these findings raise concerns about the accuracy and consistency of WMH quantifications in large multicenter investigations involving multiple raters, emphasizing the need for a verified automatic WMH segmentation method to serve as a labor saving-yet-accurate reference method. In the present study, multicenter validations demonstrated consistently high performance of the artificial intelligence algorithm, achieving a high DSC across various MRI vendors and parameters.

Because deep learning algorithms can have overfitting,³⁷ extensive external validation is necessary before they can be used in clinical practice. Unlike previous deep learning studies that have demonstrated similar¹⁸ or superior¹⁶ performance of WMH segmentation, our study rigorously validated our algorithms on 6 external data sets with various MRI vendors, acquisition protocols, and clinical features. This extensive validation underscores the robustness of our approach. As shown in our study, even among the highly experienced raters, the level of voxelwise agreement was moderate (DSC-based interrater reliability of ~ 0.77). Moreover, the performance of deep learning models reached a plateau at about 2000 FLAIR MRIs. Thus, even with a larger data set for supervised learning, it might be difficult to outperform experts in rigorous internal and external testing.

Prior studies on automatic WMH segmentation using deep learning did not include patients with acute ischemic stroke.^{12,15,16,18,19} These studies are difficult to apply to research in patients with stroke because it is challenging to distinguish between acute infarcts and WMH on FLAIR images. We demonstrated that our deep learning algorithm reliably differentiates WMHs from acute infarcts and quantifies WMH volume with excellent agreement with manual segmentation (Fig 5), thereby facilitating WMH research in patients with stroke with a large sample size. Future studies could validate our deep learning algorithms in outpatients with cognitive decline as well, expanding their clinical relevance. The general methodology we established in patients with stroke also should be applicable to many other populations.

Our study shows that manually or via deep learning approaches, segmenting WMHs may have an inherent element of ambiguity when the extent of WMH is small in the periventricular area and there are spatial overlaps between infarcts and WMHs. The uncertainty index that represents a patient-level ambiguity in voxel classification for WMH segmentation was negatively correlated with total WMH volume and DSC. Thus, the uncertainty index could be a useful stratification/triage tool to identify cases that require human inspection in research or clinical practice using WMH segmentation, enabling investigators to evaluate their own models and compare different study results.

Deep learning models typically perform worse on data that were not used for training.³⁸ In our study however, both the UNet and SE-UNet demonstrated higher DSCs in the external test data sets than in the internal test data set. This finding may be attributed to higher WMH volumes in the external validation data sets (ranging from 10.8 to 17.7 mL) compared with the internal data set (median 11.4 mL). As shown in Fig 1B, DSC values were higher when WMH volumes were higher.

Of note, while acknowledging the potential benefits of 3D models, we deliberately opted to use 2D UNets for several reasons: First, in typical clinical settings for screening patients with stroke, MR images mostly have high in-plane resolution (eg, 1×1 mm) but a relatively large slice thickness (3–7 mm), aligning well with the capabilities of 2D UNets. Preliminary investigations with 3D UNet³⁹ and nnUNet, a semantic segmentation method that automatically adapts to a given dataset,⁴⁰ yielded suboptimal segmentation performance compared with 2D methods (Online Supplemental Data). Furthermore, computational constraints, particularly with our large data sets, limited the feasibility of deploying and adjusting 3D approaches without sacrificing efficiency. Using 2D UNets, we achieved a balance between computational efficiency and model performance, ensuring that our experiments could be conducted within reasonable time and resource constraints.

Our study has limitations. First, T1-weighted images were unavailable because MRI was performed in the setting of acute stroke. The performance of deep learning algorithms was reported to improve when trained with both FLAIR and T1-weighted images, compared with the exclusive use of FLAIR images.¹² Second, given that most study participants were of Asian descent, the generalizability of the findings to other ethnic populations might be somewhat restricted. While a recent study found that racial differences in WMH burden are largely influenced by underlying vascular risk factors,⁴¹ further work with multinational cohorts is needed. Third, incorporating FLAIR and DWI into the algorithm may improve the differentiation of WMH from acute infarct. However, this step may limit the utility of the algorithm due to the requirements of both types of images. Finally, our algorithm was trained using FLAIR images acquired >10 years ago (from 2011 to 2013). However, when the algorithm was validated using a recent MRI data set from 30 patients with acute ischemic stroke (Online Supplemental Data), the validation showed a comparable level of DSC: 0.674 (95% CI, 0.619–0.729).

CONCLUSIONS

We developed and validated deep learning algorithms to segment WMH in patients with acute cerebral infarction using the largest-ever MRI data sets. In addition, we showed how the uncertainty index could assist researchers and physicians in identifying challenging cases that may necessitate human review after automatic WMH segmentation.

ACKNOWLEDGMENTS

We are grateful to Dr Sue Young Ha for her assistance with this research.

Disclosure forms provided by the authors are available with the full text and PDF of this article at www.ajnr.org.

REFERENCES

- Debette S, Markus HS. The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis. *BMJ* 2010;341:c3666 [CrossRef Medline](#)
- Ryu WS, Schellingerhout D, Hong KS, et al. White matter hyperintensity load on stroke recurrence and mortality at 1 year after ischemic stroke. *Neurology* 2019;93:e578–89 [CrossRef Medline](#)
- Ryu WS, Woo SH, Schellingerhout D, et al. Stroke outcomes are worse with larger leukoaraiosis volumes. *Brain* 2017;140:158–70 [CrossRef Medline](#)
- Prins ND, Scheltens P. White matter hyperintensities, cognitive impairment and dementia: an update. *Nat Rev Neurol* 2015;11:157–65 [CrossRef Medline](#)
- Herrmann LL, Le Masurier M, Ebmeier KP. White matter hyperintensities in late life depression: a systematic review. *J Neurol Neurosurg Psychiatry* 2007;79:619–24 [CrossRef](#)
- Wardlaw JM, Smith EE, Biessels GJ, et al. Neuroimaging; STandards for Reporting Vascular changes on nEuroimaging (STRIVE v1). standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *Lancet Neurol* 2013;12:822–38 [CrossRef Medline](#)
- Xiong Y, Yang J, Wong A, et al. Operational definitions improve reliability of the age-related white matter changes scale. *Eur J Neurol* 2011;18:744–49 [CrossRef Medline](#)
- Rost NS, Brodtmann A, Pase MP, et al. Post-stroke cognitive impairment and dementia. *Circ Res* 2022;130:1252–71 [CrossRef Medline](#)
- Alber J, Alladi S, Bae HJ, et al. White matter hyperintensities in vascular contributions to cognitive impairment and dementia (VCID): knowledge gaps and opportunities. *Alzheimers Dement (N Y)* 2019;5:107–17 [CrossRef Medline](#)
- Black S, Gao F, Bilbao J. Understanding white matter disease: imaging-pathological correlations in vascular cognitive impairment. *Stroke* 2009;40:S48–52 [CrossRef Medline](#)
- Carass A, Roy S, Gherman A, et al. Evaluating white matter lesion segmentations with refined Sorensen-Dice analysis. *Sci Rep* 2020; 10:8242 [CrossRef Medline](#)
- Park G, Hong J, Duffy BA, et al. White matter hyperintensities segmentation using the ensemble U-Net with multi-scale highlighting foregrounds. *Neuroimage* 2021;237:118140 [CrossRef Medline](#)
- Ryu WS, Kang YR, Noh YG, et al. Acute infarct segmentation on diffusion-weighted imaging using deep learning algorithm and RAPID MRI. *J Stroke* 2023;25:425–29 [CrossRef Medline](#)
- Noh YG, Ryu WS, Schellingerhout D, et al. Deep learning algorithms for automatic segmentation of acute cerebral infarcts on diffusion-weighted images: effects of training data sample size, transfer learning, and data features. 2023. *medRxiv* <https://www.medrxiv.org/content/medrxiv/early/2023/07/09/2023.07.02.23292150.full.pdf>. Accessed July 2, 2023
- Li X, Zhao Y, Jiang J, et al. White matter hyperintensities segmentation using an ensemble of neural networks. *Hum Brain Mapp* 2022;43:929–39 [CrossRef Medline](#)
- Sundaresan V, Zamboni G, Rothwell PM, et al. Triplanar ensemble U-Net model for white matter hyperintensities segmentation on MR images. *Med Image Anal* 2021;73:102184 [CrossRef Medline](#)
- Ryu WS, Woo SH, Schellingerhout D, et al. Grading and interpretation of white matter hyperintensities using statistical maps. *Stroke* 2014;45:3567–75 [CrossRef Medline](#)
- 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](#)
- Lee AR, Woo I, Kang DW, et al. Fully automated segmentation on brain ischemic and white matter hyperintensities lesions using semantic segmentation networks with squeeze-and-excitation blocks in MRI. *Informatics in Medicine Unlocked* 2020;21:100440 [CrossRef](#)
- Zhou K, Liu Z, Qiao Y, et al. Domain generalization: a survey. *IEEE Trans Pattern Anal Mach Intell* 2022;45:4396–415 [CrossRef Medline](#)
- Mehrtash A, Wells WM, Tempny CM, et al. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Trans Med Imaging* 2020;39:3868–78 [CrossRef Medline](#)
- van Erven T, Harremoës P. Rényi divergence and Kullback-Leibler divergence. *IEEE Trans Inform Theory* 2014;60:3797–820 [CrossRef](#)
- Ryu WS, Schellingerhout D, Hong KS, et al. Relation of pre-stroke aspirin use with cerebral infarct volume and functional outcomes. *Ann Neurol* 2021;90:763–76 [CrossRef Medline](#)
- Kim DE, Park JH, Schellingerhout D, et al. Mapping the supratentorial cerebral arterial territories using 1160 large artery infarcts. *JAMA Neurol* 2019;76:72–80 [CrossRef Medline](#)
- Kim DE, Ryu WS, Schellingerhout D, et al. Estimation of acute infarct volume with reference maps: a simple visual tool for decision making in thrombectomy cases. *J Stroke* 2019;21:69–77 [CrossRef Medline](#)
- Ryu WS, Schellingerhout D, Ahn HS, et al. Hemispheric asymmetry of white matter hyperintensity in association with lacunar infarction. *J Am Heart Assoc* 2018;7:e010653 [CrossRef Medline](#)
- Chitalya R, Pudipeddi S. *Image Processing and Acquisition using Python*. CRC Press; 2020; 89–107
- Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018:7132–41
- Huang H, Lin L, Tong R, et al. Unet 3+: a full-scale connected unet for medical image segmentation. In: *Proceedings of the 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. May 4–8, 2020. Barcelona, Spain; 10555–59
- Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. In: *Proceedings of the Annual Conference Neural Information Processing Systems*. December 4–8, 2017. Long Beach, California
- Jungo A, Reyes M. Assessing reliability and challenges of uncertainty estimations for medical image segmentation. *Medical Image Computing and Computer Assisted Intervention–MICCAI* 2019. Springer-Verlag; 2019:48–56
- Tuohy S, O’Cualain D, Jones E, et al. Distance determination for an automobile environment using inverse perspective mapping in OpenCV. In: *Proceedings of the IET Irish Signals and Systems Conference*. June 14, 2010. Belfast, Northern Ireland
- Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989;45:255–68 [CrossRef](#)
- Altman D. *Practical Statistics for Medical Research*. Chapman & Hall/ CRC Texts in Statistical Science. 1991; 396–439
- Kapeller P, Barber R, Vermeulen RJ, et al; European Task Force of Age-Related White Matter Changes. Visual rating of age-related white matter changes on magnetic resonance imaging: scale comparison, interrater agreement, and correlations with quantitative measurements. *Stroke* 2003;34:441–45 [CrossRef Medline](#)
- van den Heuvel DM, ten Dam VH, de Craen AJ, et al; PROSPER Study Group. Measuring longitudinal white matter changes: comparison of a visual rating scale with a volumetric measurement. *AJNR Am J Neuroradiol* 2006;27:875–78 [Medline](#)
- Seo H, Badiei Khuzani M, Vasudevan V, et al. Machine learning techniques for biomedical image segmentation: an overview of technical aspects and introduction to state-of-art applications. *Med Phys* 2020;47:e148–67 [CrossRef Medline](#)
- Zhang L, Wang X, Yang D, et al. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE Trans Med Imaging* 2020;39:2531–40 [CrossRef Medline](#)
- Çiçek Ö, Abdulkadir A, Lienkamp SS, et al. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. Springer-Verlag International Publishing; 2016:424–32
- Isensee F, Jaeger PF, Kohl SAA, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods* 2021;18:203–11 [CrossRef Medline](#)
- Morrison C, Dadar M, Manera AL, et al. Racial differences in white matter hyperintensity burden in older adults. *Neurobiol Aging* 2023;122:112–19 [CrossRef Medline](#)