ON-LINE APPENDIX: METHODS

MR Imaging

All diffusion MR imaging data were obtained on 1.5T systems (GE Healthcare) using an axial acquisition. The FOV ranged from 200 to 300 mm, with the majority and median being 220 mm, with an acquisition matrix of 128 × 128, and most cases ($n = 233$) upsampled to a 256 × 256 reconstruction matrix, resulting in a median in-plane resolution of 0.86×0.86 mm^2 (0.86×0.86 to 0.94×0.94 mm^2). Slice thickness ranged from 5 to 7 mm with gap of 0–1 mm; median, 5 mm (5–5 mm) thickness; and 1 mm (1–1 mm) gap. Slice coverage ranged from 18 to 30 slices, with a median of 24 slices (23–26 slices). The median TR was 5000 ms (5000–5000 ms), and the median TE was 89 ms (85–96 ms). The number of diffusion gradient directions was 3 ($n = 13$), 6 ($n = 156$), 15 ($n = 2$), 21 ($n = 9$), and 25 ($n = 87$), with 1 average (for 21 and 25 direction acquisitions) to 5 averages (6 directions). Diffusion-weighting (b-value) of the high-b-value volume ranged from 1000 to 1221 s/mm^2, but most cases had b-values of 1000 s/mm^2 ($n = 259$). For the 8 cases with b-value = 1221 s/mm^2, the low b-value was 3.1 s/mm^2. All other data had a low b-value of 0 s/mm^2.

Convolutional Neural Network Training

DeepMedic is a 3D convolutional neural network that operates on 2 multiresolution pathways to allow efficient and accurate supervised segmentation.[1] This framework was chosen over other approaches such as multispectral support vector machines[2] and random forests because it was shown to perform best in the ISLES 2015 trial.[3] Other studies have also shown better or comparable performance of DeepMedic compared with other neural network architectures.[4-9] In brief, the DeepMedic framework includes a high- and low-resolution pathway (isotropic subsampling by factor 3) with an equal number of 8 convolutional layers consisting of 30, 30, 40, 40, 40, 40, 50, and 50 feature maps. The convolutional kernels were the same size (3, 3, 3) for all layers. The outputs of layers 2, 4, and 6 were connected to the outputs of layers 4, 6, and 8, respectively.[10] The outputs of the high- and low-resolution paths were concatenated and linked to 2 convolutional layers with isotropic kernels of 1 × 1 × 1 (each with 150 neurons). On-line Fig *1* shows the architecture. To avoid overfitting,[11] we applied a drop-out rate of 50% to the final convolutional and the classification layers.

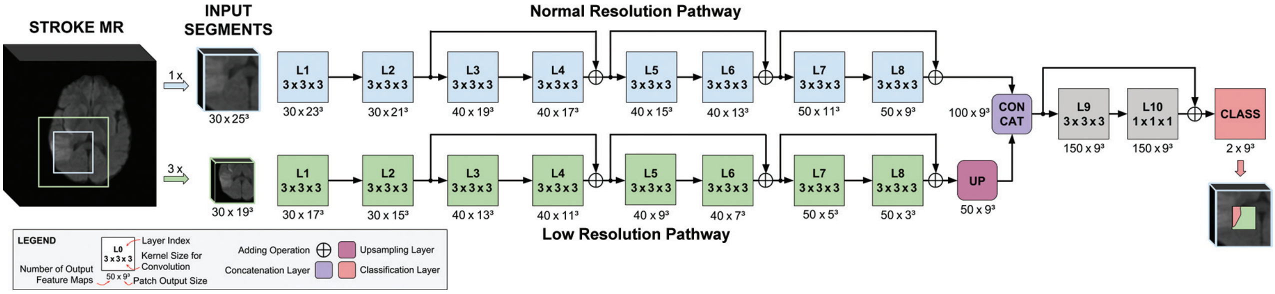
RESULTS

Differences in MR imaging acquisition parameters between the Training Cohort and Evaluation Cohort are shown in On-line Table 1.

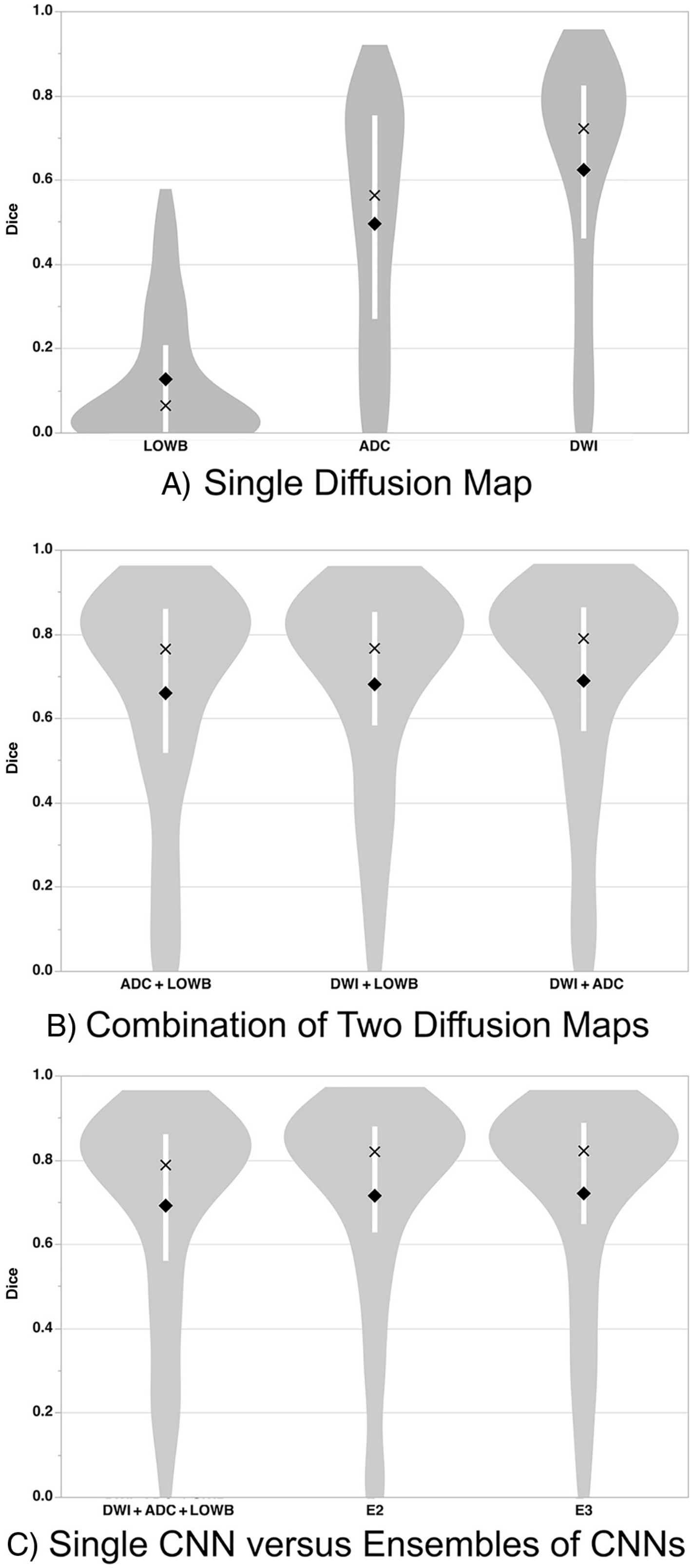
Five different CNNs were trained on 2 (DWI+ADC) or all 3 (DWI+ADC+LOWB) diffusion maps (On-line Tables 2 and 3, respectively). The performances were consistent across the CNNs, with marginal fluctuations and no measurable differences. Creating ensembles of each of the 5 CNNs improved the Dice scores and precision significantly compared with each individual CNN ($P < .001$). The sensitivity of E2 followed this trend; however, it was only significantly higher than 2 of the individual CNNs (On-line Table 2, CNN 2 and CNN 5, $P < .05$). E3 was significantly more sensitive than 2 CNNs also trained on 3 diffusion maps (On-line Table 3, CNN 2 and CNN 3, $P < .01$).

REFERENCES

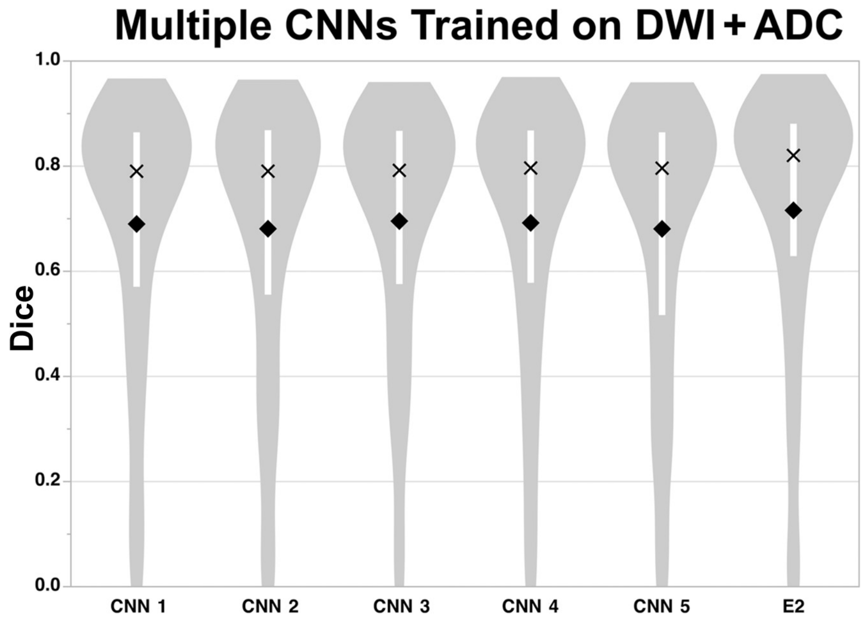
1. Kamnitsas K, Ledig C, Newcombe VF, et al. **Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation.** *Med Image Anal* 2017;36:61–78 CrossRef Medline
2. Maier O, Wilms M, Gablentz JV, et al. **Ischemic stroke lesion segmentation in multi-spectral MR images with support vector machine classifiers.** In: *Proceedings of Society of Photo-Optical Instrumentation Engineers Medical Imaging*, San Diego, California. February 15–20, 2014:12
3. Maier O, Menze BH, von der Gablentz J, et al. **ISLES 2015: a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI.** *Med Image Anal* 2017;35:250–69 CrossRef Medline
4. Wang G, Li W, Zuluaga MA, et al. **Interactive medical image segmentation using deep learning with image-specific fine tuning.** *IEEE Trans Med Imaging* 2018;37:1562–73 CrossRef Medline
5. Li W, Wang G, Fidon L, et al. **On the compactness, efficiency, and representation of 3D convolutional networks: brain parcellation as a pretext task**. In: Niethammer M, Styner M, Aylward S, et al, eds. *Information Processing in Medical Imaging.* 2017. Cham: Springer-Verlag; 2017:348–60
6. Casamitjana A, Puch S, Aduriz A, et al. **3D Convolutional neural networks for brain tumor segmentation: a comparison of multi-resolution architectures** In: Crimi A. Menzer B, Maier O, et al, eds. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries.* Cham: Springer-Verlag; 2016:150–61
7. Xue Y, Xu T, Zhang H, et al. **SegAN: adversarial network with multi-scale L1 loss for medical image segmentation.** *Neuroinformatics* 2018;16:383–92 CrossRef Medline
8. Nie D, Wang L, Adeli E, et al. **3-D fully convolutional networks for multimodal isointense infant brain image segmentation.** *IEEE Trans Cybern* 2019;49:1123–36 CrossRef Medline
9. Brosch T, Saalbach A. **Foveal fully convolutional nets for multi-organ segmentation.** In: *Medical Imaging* 2018: *Image Processing*, March 2, 2018;10574:105740U. *International Society for Optics and Photonics*
10. He K, Zhang X, Ren S, et al. **Deep Residual Learning for Image Recognition.** In: *Proceedings of the Institute of Electrical and Electronics Engineers Conference on Computer Vision and Pattern Recognition*, Las Vegas, Nevada; June 27–30, 2016:770–78
11. Srivastava N, Hinton G, Krizhevsky A, et al. **Dropout: a simple way to prevent neural networks from overfitting.** *J Mach Learn Res* 2014;15: 1929–58



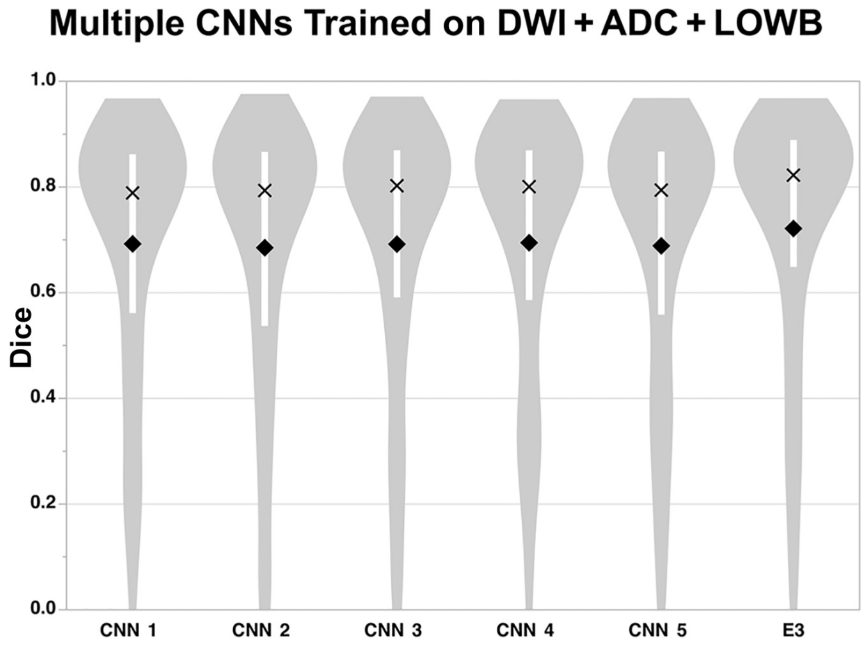
ON-LINE FIG 1. The DeepMedic architecture used operates on 2 different receptive fields, one with original resolution and one isotropically downsampled by a factor of 3. Each receptive field was processed by individual but equally constructed pathways. Each included 8 convolutional layers (L1–L8) with 3 × 3 × 3 kernel size and 3 residual connections between outputs of layers 2 and 4, 4 and 6, as well as 6 and 8 (+ signs). The final output of the low-resolution pathway was upsampled (UP) to match the output of the normal resolution pathway (ie, 9 × 9 × 9). Both outputs were then concatenated (CONCAT) and processed by 2 further convolutional (L9 and L10) layers with 3 × 3 × 3 and 1 × 1 × 1 kernel sizes, respectively, and 1 residual connection. The final classification layer (CLASS) provided the lesion prediction. Although the figure shows only the DWI channel, multiple channels can be easily used.



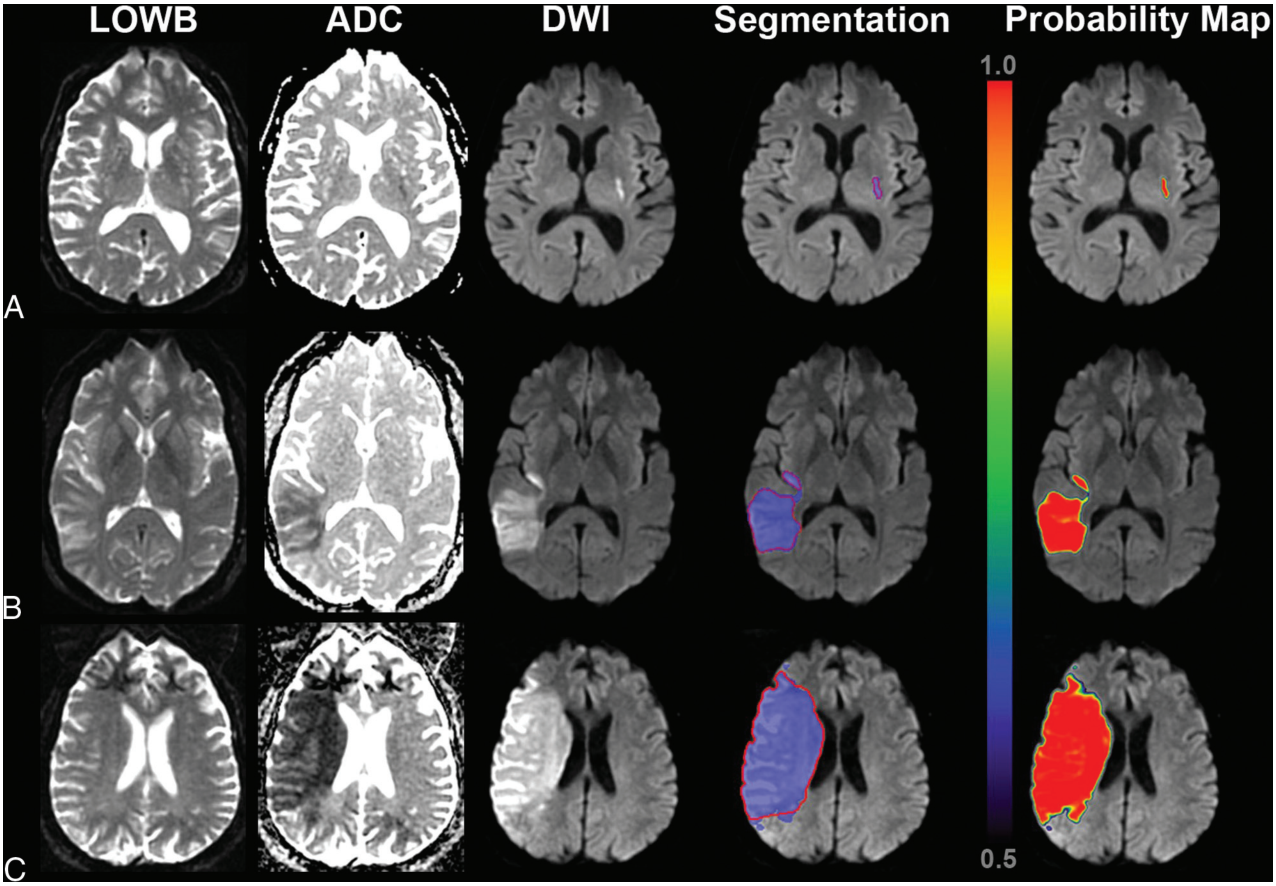
ON-LINE FIG 2. Distribution of Dice scores of all models. The models trained on single diffusion parametric maps are shown in the *first row* (*A*); 2 parametric maps, in the *second row* (*B*); and all 3 parametric maps along with ensemble results from 5 separate convolutional neural networks in the *third row* (*C*). Of the individual models, the one based on DWI did best, while the individual ADC model performed moderately well and the LOWB model performed worst. Dice scores using 2 parametric maps did better, with DWI+ADC yielding the best performance ($P < .05$). Adding LOWB maps to the CNN (DWI+ADC+LOWB) did not improve the Dice scores over the DWI+ADC model ($P = .49$). Both ensembles (*lowest row*), each consisting of 5 CNNs trained either on DWI+ADC (E2) or DWI+ADC+LOWB (E3), outperformed all other models ($P < .001$) but offered a similar performance compared with each other ($P = .66$). The *white bar* in the violin plot shows the IQR, mean is a *diamond*, and median is an *X* for all plots (*A*, *B*, *C*).



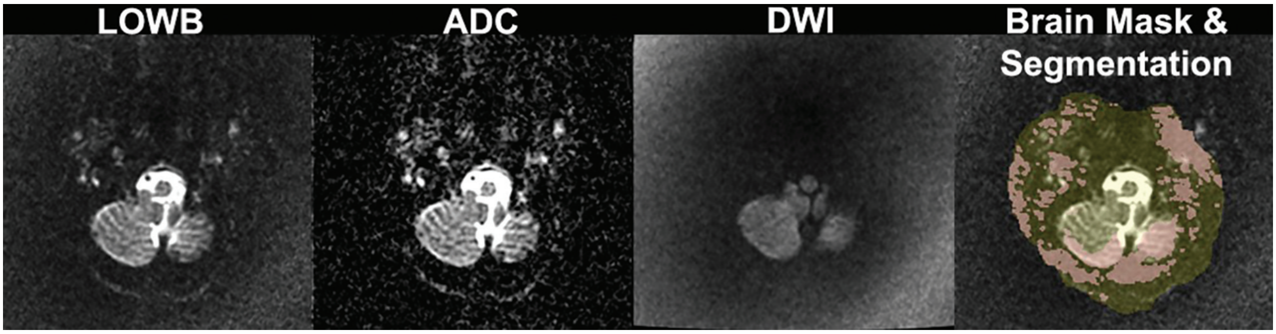
ON-LINE FIG 3. Distribution of Dice scores for all 5 CNNs trained on DWI+ADC maps and their ensemble (E2). The *white bar* within the violin plot shows the IQR, mean is a *diamond*, and median is an *X*.



ON-LINE FIG 4. Distribution of Dice scores for all 5 CNNs trained on DWI+ADC+LOWB maps and their ensemble (E3). The *white bar* within the violin plot shows the IQR, mean is a *diamond*, and median is an *X*.



ON-LINE FIG 5. Sample segmentation results of the ensemble of DWI+ADC+LOWB (*blue regions*) on sample subjects along with manual outlines (*red outlines*) and probability-of-infarction maps for the same patients shown in Fig 2.



ON-LINE FIG 6. Example of poor segmentation results. LOWB, ADC, and DWI maps are shown along with an automatically extracted brain mask (yellow) overlaid on the LOWB image from an 86-year-old woman presenting with an admission NIHSS score of 12 and scanned approximately 6 hours from when she was last known to be well. For this slice, there is no evident lesion on the DWI; however, the segmented lesion (*pink overlay* in "Brain Mask & Segmentation" panel) grossly encompasses normal tissue and background. This artifact is due to the poor automated brain extraction as a result of scanner inhomogeneity artifacts. The measured lesion volume is 14.1 cm^3, while the automated lesion volume is 133.7 cm^3.

On-line Table 1: Diffusion-weighted MRI acquisition parameters[a]

Characteristic	Training (*n* = 116)	Evaluation (*n* = 151)	*P* Value
TR (ms)	5000 (5000–5000)	5000 (5000–6000)	<.001
TE (ms)	88.9 (85.9–94.5)	89.7 (85.3–99.2)	.35
FOV (mm)	220 (220–220)	220 (220–220)	.28
Reconstructed matrix	256 × 256	128 × 128 (*n* = 34), 256 × 256 (*n* = 117)	<.001
Slices	24 (23–26)	24 (23–26)	.70
Slice spacing (mm)	6 (6–6)	6 (6–6)	.60
Directions	3 (*n* = 11), 6 (*n* = 84), 25 (*n* = 21)	3 (*n* = 2), 6 (*n* = 72), 15 (*n* = 2), 21 (*n* = 9), 25 (*n* = 66)	<.001

[a] Shown are median (IQR) values and statistical significance of differences between the training and Evaluation Cohort.

On-line Table 2: Five different CNNs trained on 2 diffusion maps (DWI+ADC) and their ensemble (E2)[a]

	CNN 1	CNN 2	CNN 3	CNN 4	CNN 5	E2
Dice	79.0 (57.1–86.4)	79.0 (55.6–86.9)	79.2 (57.6–86.7)	79.7 (57.8–86.8)	79.6 (51.7–86.5)	82.0 (62.9–88.1)
Precision	79.0 (62.1–90.5)	75.4 (54.5–89.6)	78.3 (55.4–90.2)	79.0 (60.3–88.3)	77.9 (47.2–90.3)	82.0[b] (65.1–92.6)
Sensitivity	82.6 (68.4–91.4)	83.9 (73.1–92.0)	85.4 (70.4–92.8)	83.6 (68.4–91.2)	83.1 (72.5–91.4)	84.1 (71.0–92.6)

[a] All performance metrics in median (IQR). Performance was robust across all single CNNs (CNNs 1–5). The ensemble had significantly better Dice performance than all single models ($P < .001$).

[b] Excludes 1 subject with automatically segmented lesion volume of zero because precision is undefined in this circumstance.

On-line Table 3: Five different CNNs trained on 3 diffusion maps (DWI+ADC+LOWB) and their ensemble (E3)[a]

	CNN 1	CNN 2	CNN 3	CNN 4	CNN 5	E3
Dice	78.9 (56.2–86.2)	79.3 (53.7–86.6)	80.2 (59.1–86.9)	80.1 (58.6–86.9)	79.4 (55.8–86.7)	82.2 (64.9–88.9)
Precision	77.4 (55.0–89.8)	77.2 (52.7–88.6)	78.2 (57.8–91.4)	76.9 (58.2–89.8)	77.7 (55.7–90.4)	83.2 (67.7–93.3)
Sensitivity	83.4 (71.3–91.8)	84.1 (72.1–93.4)	81.8 (69.1–90.9)	84.2 (70.3–92.2)	84.6 (71.6–91.4)	83.9 (71.9–92.4)

[a] All metrics are denoted in percentages as median (IQR). Performance was robust across all single CNNs (CNNs 1–5). The ensemble (E3) outperformed all single models ($P < .001$) in terms of Dice.

On-line Table 4: Ensemble results as a function of lesion location for the Evaluation Cohort[a]

	Cortical (*n*= 104)	Subcortical (*n* = 30)	Multiple (*n* = 4)	Cerebellum (*n* = 8)	Brain Stem (*n* = 5)	*P* Value
MLV (cm^3)	20.5 (5.7–57.2)	1.8 (0.6–4.9)	11.5 (9.6–13.5)	5.5 (1.3–11.3)	0.6 (0.3–0.9)	<.001
ALV (cm^3)	25.1 (7.0–55.9)	2.8 (1.0–9.3)	9.9 (5.8–10.9)	4.9 (0.2–9.7)	0.3 (0.1–0.6)	<.001
Dice	84.9 (71.1–90.5)	73.3 (48.5–84.6)	79.3 (54.4–92.3)	64.5 (14.1–83.7)	42.8 (0–71.6)	.002
Precision	85.6 (68.1–93.9)	66.3[b] (38.6–82.3)	93.3 (76.1–96.9)	67.9 (12.2–86.0)	75.9[b] (18.5–94.5)	.01
Sensitivity	86.9 (75.7–93.2)	90.2 (68.6–94.6)	76.3 (42.3–92.5)	71.9 (10.9–89.5)	30.1 (0–60.0)	<.001

Note:—ALV indicates automatically segmented lesion volume.

[a] All metrics are denoted in percentages as median (IQR).

[b] Excludes 1 subject with automatically segmented lesion volume of zero because precision is undefined in this circumstance.