SUPPLEMENTAL MATERIALS

METHODS

MRI

All diffusion MRI data were obtained on 1.5T systems (General Electric Medical Systems) using an axial acquisition. The field-of-view ranged from 200 to 300 mm, with the majority/median being 220 mm, with acquisition matrix of 128x128, and majority of cases (N=233) up-sampled to 256x256 reconstruction matrix, resulting in median in-plane resolution of 0.86x0.86 [0.86x0.86-0.94x0.94] mm^2. Slice thickness ranged from 5 mm to 7mm with gap of 0-1 mm, median 5 mm [5-5] thickness, and 1 [1-1] mm gap. Slice coverage ranged from 18 to 30 slices, median 24 [23-26]. Median TR was 5000 [5000-5000] ms and median TE was 89 [85-96] ms. The number of diffusion gradient directions were 3 (n=13), 6 (N=156), 15 (N=2), 21 (N=9) and 25 (N=87), with 1 average (for 21 and 25 direction acquisition) to 5 averages (6 directions. Diffusion-weighting (b-value) of the high-b-value volume ranged from 1000 to 1221 s/mm^2, but the majority of the cases had b-value of 1000 s/mm^2 (N=259). For the 8 cases with b-value=1221 s/mm^2, the low b-value=3.1 s/mm^2. All other data had low b-value=0 s/mm^2.

Convolutional Neural Network (CNN) training

DeepMedic is a 3D convolutional neural network that operates on two multi-resolution pathways to allow efficient and accurate supervised segmentation.[1] This framework was chosen over other approaches such as multi-spectral support vector machines[2] and random forests since it had been shown to perform best in the Ischemic Stroke Lesion Segmentation Challenge (ISLES) 2015.[3] Other studies have also shown better or comparable performance of DeepMedic compared to other neural network architectures.[4-9] In brief, the Deep Medic framework includes a high and a low resolution pathway (isotropic sub-sampling by factor 3) with equal number of 8

convolutional layers consisting of 30, 30, 40, 40, 40, 40, 50, 50 feature maps. The convolutional kernels were the same size [3, 3, 3] for all layers. The output of layers 2, 4, and 6 were connected to the output of layers 4, 6 and 8 respectively.[10] The outputs of the high and low resolution paths were concatenated and linked to two convolutional layers with isotropic kernel 1x1x1 (each with 150 neurons). Supplementary Figure A1 shows the architecture. To avoid overfitting[11], a dropout rate of 50% was applied to the final convolutional layer and the classification layer.

RESULTS

Differences in MRI acquisition parameters between the training cohort and evaluation cohort are shown in Table A1.

Five different CNNs were trained on two (DWI+ADC) or all three (DWI+ADC+LOWB) diffusion maps (Tables A2 and A3 respectively). The performances were consistent across the CNNs with marginal fluctuations and no measurable differences. Creating ensembles of each of the five CNNs improved Dice scores and precision significantly compared to each individual CNN ($p<0.001$). The sensitivity of ensemble E2 followed this trend, however was only significantly higher than two of the individual CNNs (Table A2. CNN #2 and CNN #5, $p<0.05$). Ensemble E3 was significantly more sensitive than two CNNs trained also on three diffusion maps (Table A3. CNN #2 and CNN#3, $p<0.01$).

REFERENCES

1. 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

2. Maier O, Wilms M, Gablentz Jvd, Krämer U, Handels H. Ischemic stroke lesion segmentation in multi-spectral MR images with support vector machine classifiers. *SPIE Medical Imaging*: SPIE; 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;35:250-269

4. Wang G, Li W, Zuluaga MA, et al. Interactive Medical Image Segmentation Using Deep Learning With Image-Specific Fine Tuning. *IEEE Transactions on Medical Imaging* 2018;37:1562-1573

5. Li W, Wang G, Fidon L, Ourselin S, Cardoso MJ, Vercauteren T. 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 IPMI 2017 Lecture Notes in Computer Science, vol 10265*: Springer, Cham; 2017;348-360

6. Casamitjana A, Puch S, Aduriz A, Vilaplana V. 3D Convolutional Neural Networks for Brain Tumor Segmentation: A Comparison of Multi-resolution Architectures In: Crimi A. MB, Maier O., Reyes M., Winzeck S., Handels H. , ed. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries BrainLes 2016 Lecture Notes in Computer Science, vol 10154* Springer, Cham; 2016;150-161

7. Xue Y, Xu T, Zhang H, Long LR, Huang X. SegAN: Adversarial Network with Multi-scale L1 Loss for Medical Image Segmentation. *Neuroinformatics* 2018;16:383-392

8. Nie D, Wang L, Adeli E, Lao C, Lin W, Shen D. 3-D Fully Convolutional Networks for Multimodal Isointense Infant Brain Image Segmentation. *IEEE Transactions on Cybernetics* 2018:1-14

9. Brosch T, Saalbach A. Foveal fully convolutional nets for multi-organ segmentation. *SPIE Medical Imaging*: SPIE; 2018;9

10. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)2016; p. 770-8.

11. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J Mach Learn Res. 2014;15:1929-58.

Table A1: Diffusion-weighted MRI Acquisition parameters. Shown are median [IQR] values and statistical significance of differences between the Training and Evaluation Cohorts.

Characteristic	**Training (N=116)**	**Evaluation (N=151)**	**P-value**
TR (ms)	5000 [5000-5000]	5000 [5000-6000]	<0.0001
TE (ms)	88.9 [85.9-94.5]	89.7 [85.3-99.2]	0.35
FOV (mm)	220 [220-220]	220 [220-220]	0.28
Reconstructed Matrix	256x256	128x128 (N=34) 256x256 (N=117)	<0.0001
Slices	24 [23-26]	24 [23-26]	0.70
Slice Spacing (mm)	6 [6-6]	6 [6-6]	0.60
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)	<0.0001

Data are reported as Median [Interquartile Range].

Table A2: Five different CNNs trained on two diffusion maps (DWI+ADC) and their ensemble (E2). Performance was robust across all single CNNs (CNN #1 - #5). The ensemble had significantly better Dice performance than all single models ($p<0.0001$). †Excludes one subject with automatically segmented lesion volume of zero since precision is undefined in this circumstance.

	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 † [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]

All performance metrics in Median [Interquartile Range] %. CNN=Convolutional Neural Network, E2=Ensemble of 5 CNNs trained on DWI+ADC.

Table A3: Five different CNNs trained on three diffusion maps (DWI+ADC+LOWB) and their ensemble (E3). Performance was robust across all single CNNs (CNN 1 to 5). The ensemble outperformed all single models ($p<0.0001$) in terms of Dice.

	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]

All metrics denoted in % as median [Interquartile Range]. CNN=Convolutional Neural Network, E3=Ensemble of 5 CNNs trained on DWI+ADC+LOWB, LOWB=Low b-value diffusion-weighted image.

Table A4: Ensemble results as a function of lesion location for the evaluation cohort. †Excludes one subject with automatically segmented lesion volume of zero since precision is undefined in this circumstance.

	Cortical (N=104)	**Subcortical** (N=30)	**Multiple** (N=4)	**Cerebellum** (N=8)	**Brainstem** (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]	0.0003
PLV (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]	<0.0001
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]	0.0019
Precision	85.6 [68.1–93.9]	66.3† [38.6–82.3]	93.3 [76.1–96.9]	67.9 [12.2–86.0]	75.9† [18.5–94.5]	0.013
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]	<0.0001

Supplemental Figures

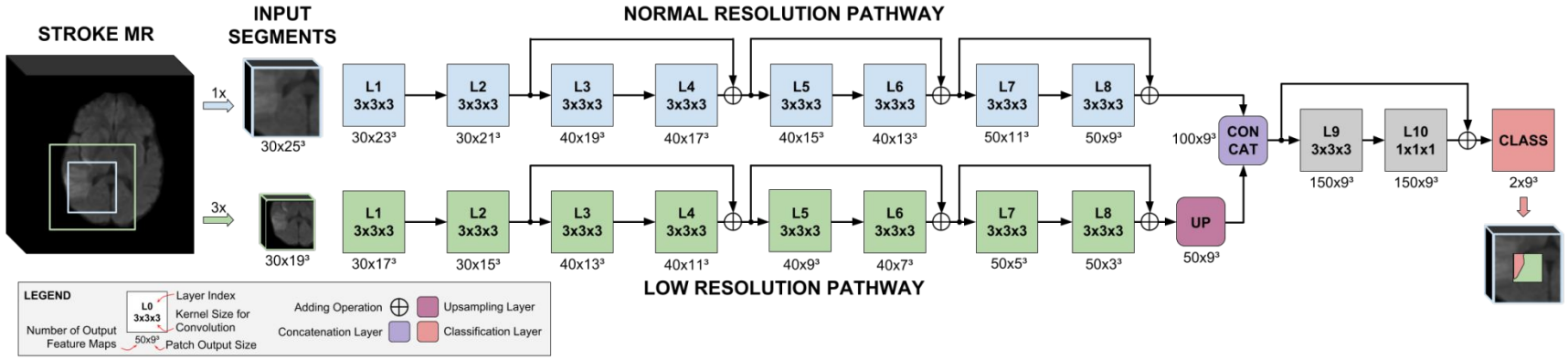
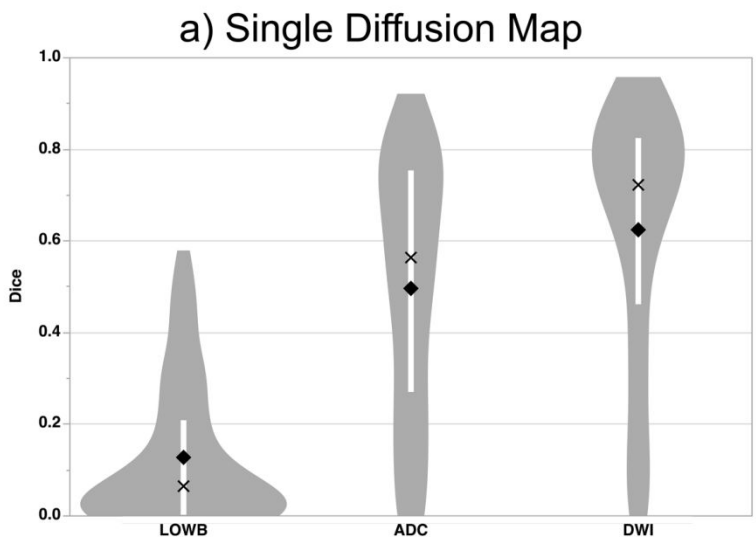
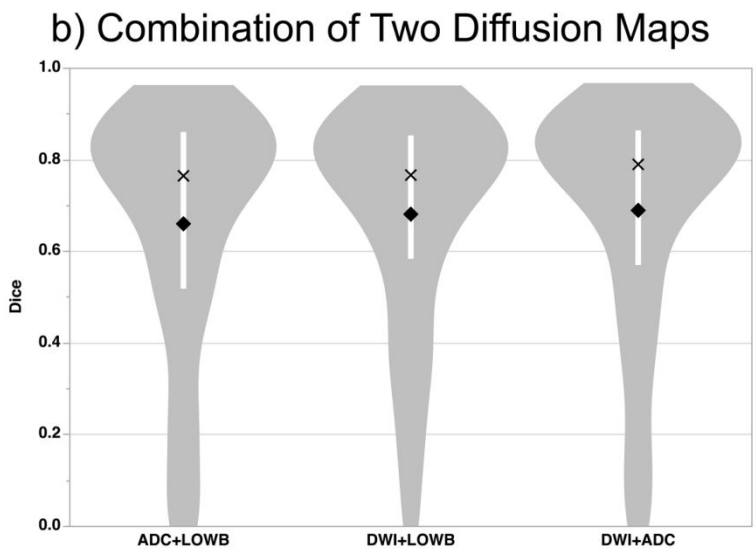


Figure A1: The employed DeepMedic architecture operates on two different receptive fields. One with original resolution and one isotropically downsampled by a factor of 3. Each receptive field was processed by an individual but equally constructed pathways. Each included 8 convolutional layers (L1-L8) with 3x3x3 kernel size and 3 residual connections between outputs of layer 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 normal resolution pathway's output (i.e. 9x9x9). Both outputs were then concatenated (CONCAT), processed by two further convolutional (L9 & L10) layers with 3x3x3 and 1x1x1 kernel size, respectively, and one residual connection. The final classification layer (CLASS) provided the lesion prediction. Although the figure shows only the DWI channel, multiple channels can be easily utilized.


a) Single Diffusion Map
Dice
1.0
0.8
0.6
0.4
0.2
0.0
LOWB
ADC
DWI


b) Combination of Two Diffusion Maps
Dice
1.0
0.8
0.6
0.4
0.2
0.0
ADC+LOWB
DWI+LOWB
DWI+ADC

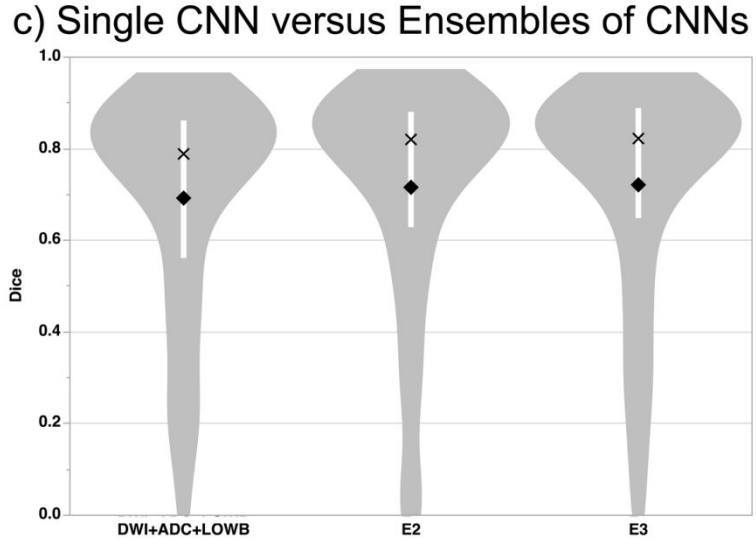

c) Single CNN versus Ensembles of CNNs
Dice
1.0
0.8
0.6
0.4
0.2
0.0
DWI+ADC+LOWB
E2
E3

Figure A2: Distribution of Dice Scores of All Models. The models trained on single diffusion parametric maps are shown in the first row (a), two parametric maps in the second row (b), and all 3 parametric maps along with ensemble results from 5 separate convolutional neural networks (CNNs). Of the individual models, the one based on DWI did best while the individual ADC model performed moderately well while the LOWB model performed worst. Dice scores using two parametric maps did better with DWI+ADC yielding the best performance ($p<0.05$). Adding LOWB maps to the CNN (DWI+ADC+LOWB) did not improve Dice scores over the DWI+ADC model ($p=0.49$). Both ensembles (bottom), each consisting of 5 CNNs trained either on DWI+ADC (E2) or DWI+ADC+LOWB (E3), outperformed all other models ($p<0.0001$), but offered a similar performance when compared to each other ($p=0.66$). The white bar within the violin plot shows the IQR, mean is represented as diamond, median is marked as cross. ADC=apparent diffusion coefficient, CNN=convolutional Neural Network, DWI=Isotropic Diffusion-Weighted Image, LOWB=Low b-value diffusion-weighted image.

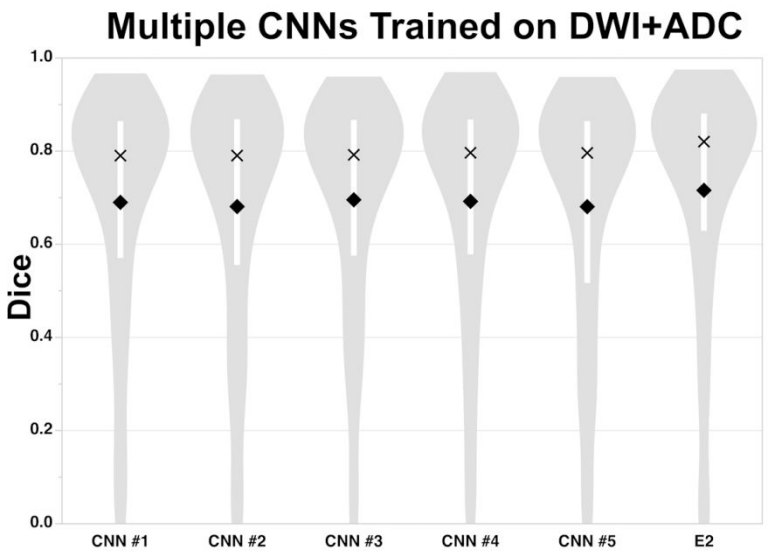


Figure A3: Distribution of Dice Scores for all Five CNNs Trained on DWI+ADC Maps and Their Ensemble (E2). The white bar within the violin plot shows the IQR, mean is represented as diamond, median is marked as cross. ADC=apparent diffusion coefficient, CNN=Convolutional Neural Network, DWI=Isotropic Diffusion-Weighted Image, E2=Ensemble of 5 CNNs trained on DWI+ADC.

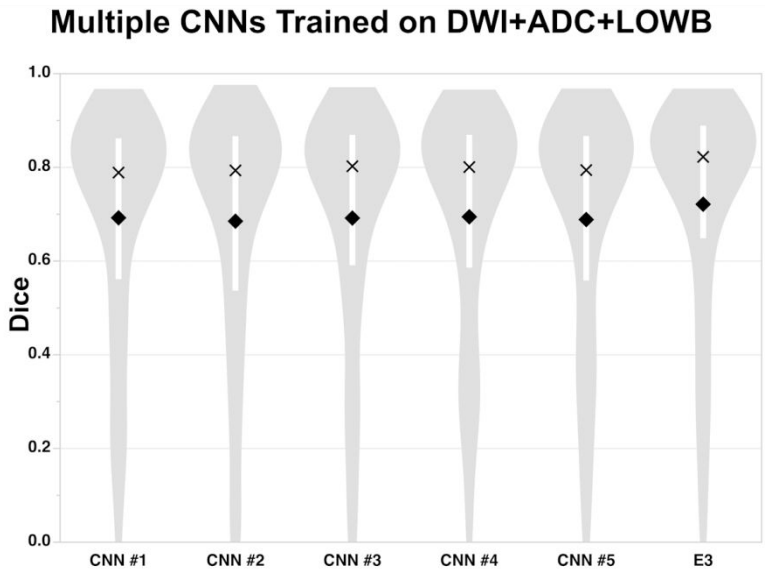
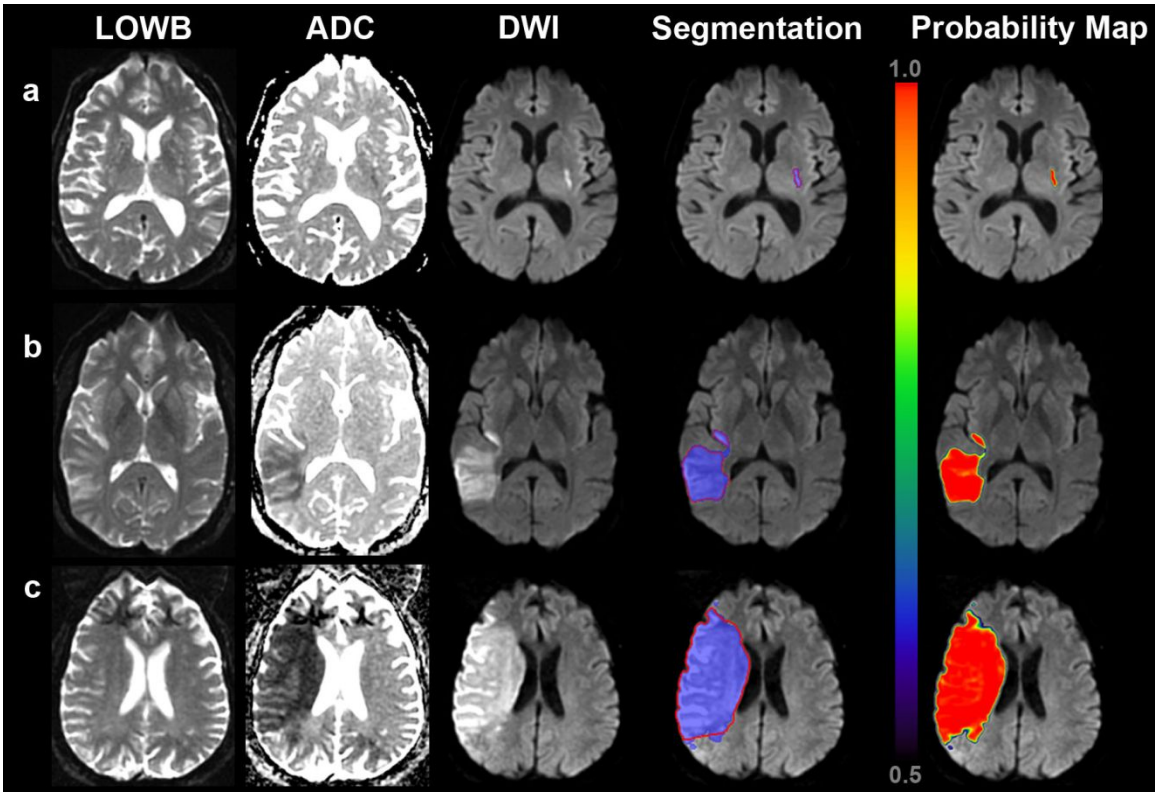
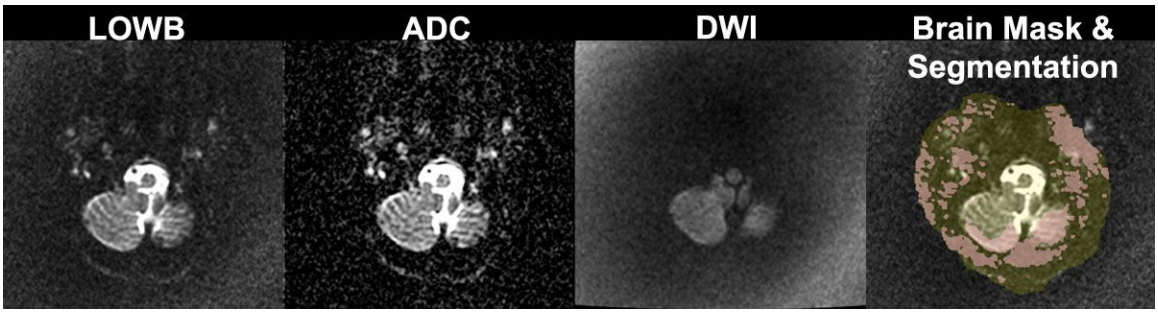


Figure A4: Distribution of Dice Scores for all Five CNNs Trained on DWI+ADC+LOWB Maps and Their Ensemble (E3). The white bar within the violin plot shows the IQR, mean is represented as diamond, median is marked as cross. ADC=apparent diffusion coefficient, CNN=Convolutional Neural Network, DWI=Isotropic Diffusion-Weighted Image, E3=Ensemble of 5 CNNs trained on DWI+ADC+LOWB, LOWB=Low b-value diffusion-weighted image.



Supplemental Figure A5: Example Segmentation Results of 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 Figure 2.



Supplemental Figure A6: Example of poor segmentation results. LOWB, ADC, DWI maps are shown along with automatically extracted brain mask (yellow) overlaid on LOWB image from an 86-year-old woman, presenting with admission NIHSS Score of 12, and imaged approximately 6 h 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) grossly encompasses normal tissue and background. This is due to the poor automated brain extraction as a result of scanner inhomogeneity artifacts. The measured lesion volume was 14.1 cm^3 while the automated lesion volume was 133.7 cm^3. ADC=apparent diffusion coefficient, DWI=Isotropic Diffusion-Weighted Image, LOWB=Low b-value diffusion-weighted image.