SUPPLEMENTAL MATERIAL

Suitability of Deep Weakly Supervised Learning to detect Acute Ischemic Stroke and Hemorrhagic Infarction Lesions Using Diffusion-weighted Imaging

Image acquisition, patient characteristics and consistency test, CNN architecture, evaluation metrics

Image Acquisition

MRI measurements were acquired from three MR scanners, with two 3T MR scanners (Skyra, Siemens and Trio, Siemens) and one 1.5TMRscanner (Avanto, Siemens). DWI images were acquired using a spin-echo type echo-planar (SE-EPI) sequence with *b* values of 0 and 1000 s/mm2. Following acquisition, ADC maps were calculated from the diffusion scan raw data in a pixel-by-pixel manner. The parameters are summarized in Supplementary Tab. 1.

Supplementary Table1 Scan Parameters

	Skyra		Trio		Avanto	
	DWI	GRE	DWI	GRE	DWI	GRE
Repetition (ms)	5200	220	3100	566	3800	576
Echo time (ms)	80	2.46	99	20	102	20.4
Number of excitations	1	1	3	1	3	1
Field of view (mm^2)	240×240	240×240	200×200	230×230	240×240	240×240
Matrix size	130×130	180×288	132×132	166×256	192×192	173×256
Slice thickness（mm）	5	5	6	6	5	5
Slice spacing（mm）	1.5	1.5	1.8	1.8	1.5	1.5
Number of slices	21	21	17	17	21	21

DWI, diffusion-weighted imaging; GRE, gradient recalled echo

Supplementary Table 2 Patient Characteristics and Consistency Test

Training Set	AIS Data (n=417)		HI Data (n=240)
Age, mean [min-max]	62 [31-86]		68 [24-93]
Male sex, number (%)	168 (40%)		94 (39%)
Classification, number (%)			
Normal	7036 (82%)		4101 (76%)
AIS	1597 (18%)		426 (8%)
HI	0		840 (16%)
Consistency test (manual 1-2)			
Kappa coefficient	0.97		0.98
Testing Set	**AIS Data (n=319)**		**HI Data (n=65)**
Age, mean [min-max]	66 [23-88]		64 [33-89]
Male sex, number (%)	124 (39%)		19 (29%)
Classification, number			
AIS, mean [min-max]	TI	5.1 [1-26]	5.9 [1-27]
	LI	1 [1-1]	
HI, mean [min-max]	0		4.1 [1-26]
Lesion volume, mm^3			
AIS, mean [min-max]	TI	3487.8 [102-28701]	8050.1 [215-46951]
	LI	41.6 [5-313]	
HI, mean [min-max]	0		925.4 [26-12420]
Consistency test (manual 1-2)			
ICC [95% CI]	TI	0.98 [0.97-0.99]	AIS 0.97 [0.95-0.98]
	LI	0.96 [0.95-0.97]	HI 0.99 [0.99-1]

AIS, acute ischemic stroke; HI, hemorrhagic infarction; LI, lacunar infarction; TI, territorial infarction; ICC, intraclass correlation coefficient

CNN Architecture

Different from the classical networks such as AlexNet and visual geometry group network (VGG), we used a global average pooling layer followed by a dense layer, which indicated the probability that the current slice contained a lesion, instead of using several fully connected layers at the top of the convolution layer. Each image slice was resampled to a voxel size of 0.87 mm×0.87 mm and then cropped to a matrix size of 256×256. All of the images were then normalized to images with zero mean and unit variance. In the training stage, the feature maps in the last convolution layer were processed by a global average pooling (GAP) layer, which outputs the mean value of each feature map. The mean values were further processed by a dense layer for classification. In the testing stage, we directly output the feature maps of the last convolutional layer, and used the weighted sum as the localization results to generate a class activation map (CAM). The weights were obtained by copying the weights of the last dense layer. A probability map can then be obtained by normalizing the pixel intensities as

$$x_{\mathrm{i}} = \frac{x_i}{\max\limits_{i \in \mathrm{CAM}} x_i} \times \hat{y}_{cls},$$

where x_i is the intensity of pixel i on the CAM, and $\hat{y}_{cls}$ is the output value of the classifier, which indicates the probability that any lesion is found in the slice.

CNNs, such as VGG and residual neural network (ResNet), were initially designed for classification. In the classification task, determining the kind of object presented in the image is the goal; therefore, it is not necessary to preserve the spatial location information of an object. Such CNNs were thus designed with very small sized feature maps in the last several convolution layers. In our task, we needed to answer two questions: whether a lesion appears and the location of the lesion. Therefore, we had to extract the sematic information and preserve the spatial information at the same time. To this end, we used a truncated version of the well-applied CNN by only using the output of the convolution layer, which provided feature maps with heights and widths that were at most 8 times smaller than the original input.

Transfer learning techniques in which the network weights were initialized by using the results of the pretrained ImageNet were used to improve the performance of the network on small datasets. The network is then fine-tuned by using the stochastic gradient descent (SGD) method with the Nesterov momentum as the optimizer, an initial learning rate of 0.001 and a momentum of 0.9. During training, 300 image slices were randomly chosen from the training set for validation. A dynamic training policy was adopted, in which we monitor the loss value for the validation samples at the end of each training epoch, and the learning rate is

reduced by a factor of $\sqrt{0.1}$ if the validation loss does not improve for 10 epochs. An early-stopping method, in which the training stops if no progress is made in 30 epochs, was adopted to avoid overfitting.

Evaluation Metrics

To evaluate the performance of the CAM-based methods, we proposed several lesion-wise metrics using 3D connected component analysis. In particular, for a single subject, probability map was first generated for each individual slice, and the probability maps were stacked on the z-axis to generate the predicted probability map of the subject. We thenconverted the predicted probability map to a binary segmentation map by thresholding and then measured the per-subject mean numbers of false positive lesions (mFP-L), false negative lesions (mFN-L) and true positive lesions. A false negative lesion (FN-L) was defined as a connected volume on the ground truth label that had no overlapping volume with any connected volumes on the predicted segmentation. A false positive lesion (FP-L) was defined as a connected volume on the predicted segmentation that had no overlapping volume with that on the ground truth. If a lesion appears on both the ground truth and predicted segmentations, we defined it as a true positive lesion (TP-L). The mFP-L and the mFN-L were then calculated by respectively averaging the FN-Ls and FN-Ls for all tested subjects. We further defined the lesion-wise sensitivityand precision as

$$\text{Sensitivity} = \text{Recall} = \frac{\text{TPL}}{\text{TPL} + \text{FNL}}$$

and

$$\text{Precision} = \text{Positive predictive value} = \frac{\text{TPL}}{\text{TPL} + \text{FPL}}$$

respectively, to evaluate the lesion-wise performance.

In addition, the subject-wise detection rate also matters in clinical diagnosis. We used the number of failed detected subjects (FD-S) to evaluate the subject-level performance.