Supplementary

Table S1 Sample size of artificial intelligence studies related to stroke

Title	Journal	Train data	Test data
Evaluation of Diffusion Lesion Volume Measurements in Acute Ischemic Stroke Using Encoder-Decoder Convolutional Network	Stroke	296	134
Machine Learning for Detecting Early Infarction in Acute Stroke with Non-Contrast-enhanced CT	Radiology	157	100
Deep Learning-Derived High-Level Neuroimaging Features Predict Clinical Outcomes for Large Vessel Occlusion	Stroke	250	74
Machine Learning Approach to Identify Stroke Within 4.5 Hours	Stroke	299	56

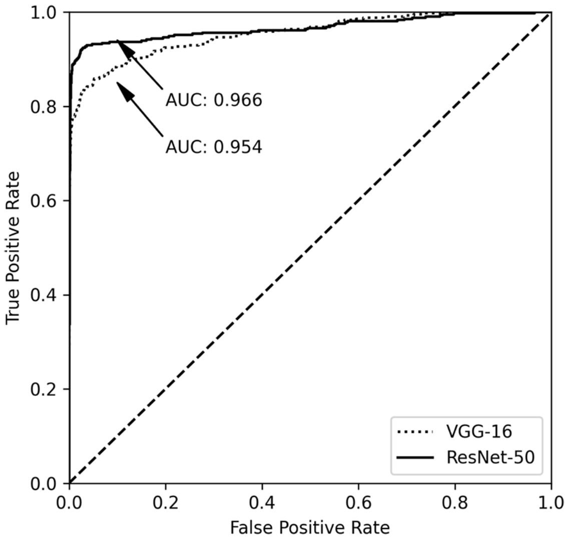

Figure S1 The receiver operating characteristic (ROC) curve of the residual neural (ResNet-50) and visual geometry group (VGG-16) network classifiers shows the false positive rate (x-axis) *vs.* the true positive rate (y-axis). The areas under the ROC curve (AUCs) for the ResNet-50 and VGG-16 networks were both superior in being able to identify lesions in acute ischemic stroke (AIS) image slices.

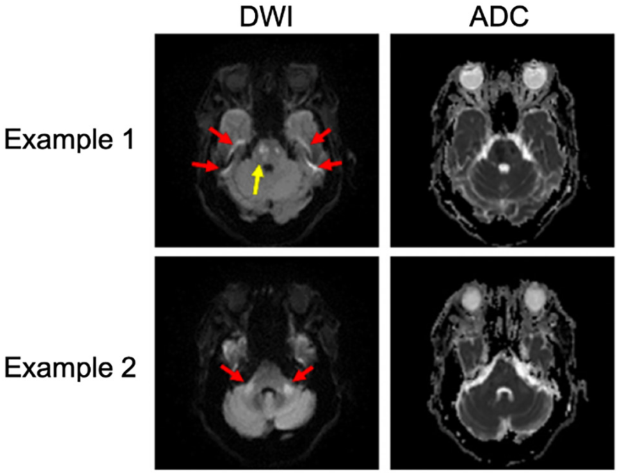

Figure S2 Challenge examples in ischemic stroke segmentation. In example 1, the yellow arrow identifies the hyperintensity that is a true acute ischemic stroke lesion, and the red arrows identify hyperintensity due to magnetic susceptibility artifacts. In example 2, the red arrows identify hyperintensity due to the T2 shine-through effect.

https://dx.doi.org/10.21037/qims-21-324

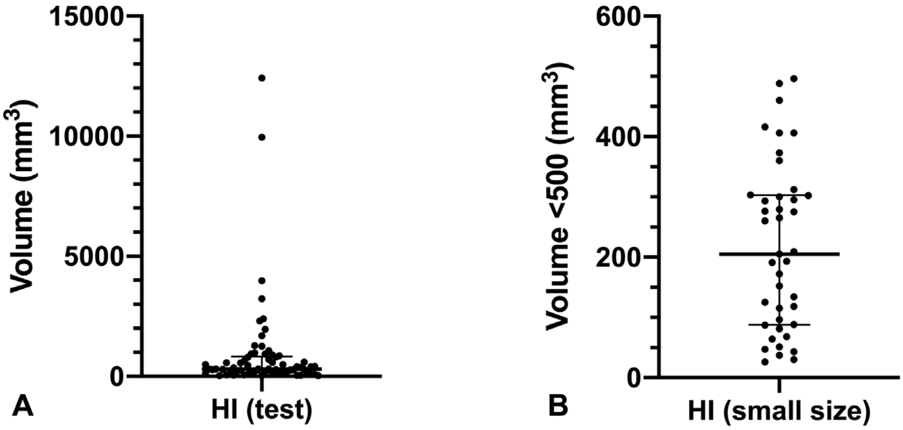

Figure S3 Scatter plots of lesions volume in the hemorrhagic infarction (HI) test set. (A) The volume (median and interquartile range) of HI in the test set was measured by the ground truth (n=65). (B) The volume in small HI lesion volume cases (n=41).

Convolutional neural network (CNN) architecture

Unlike the classical networks, such as AlexNet and visual geometry group network (VGG-16), 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 0.87×0.87 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 output 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 CAM. The weights were obtained by copying the weights of the last dense layer. A probability map could then be obtained by normalizing the pixel intensities as follows:

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

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-16 and residual neural network (ResNet-50), 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. These CNNs were thus designed with very small-sized feature maps in the last several convolution layers. In our task, we aimed to determine two issues: whether a lesion can be detected and the location of the lesion. Therefore, we needed to extract the sematic information and simultaneously preserve the spatial information. 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 through use of the ImageNet pretrained weights were used to improve the performance of the network on small data sets. The whole network was 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 monitored the loss value for the validation samples at the end of each training epoch, and the learning rate was reduced by a factor of $\sqrt{0.1}$ if the validation loss did not improve for 10 epochs. Data augmentation methods, including random flipping along 2 axes and random rotation, were adopted to prevent overfitting, where the rotation were restricted within a range of [–30°,30°]. An early-stopping method, in which the

https://dx.doi.org/10.21037/qims-21-324

training is stopped if no progress is made in 30 epochs, was also adopted to avoid overfitting.

Statistical analysis

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, a 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 then converted the predicted probability map to a binary segmentation map by thresholding and subsequently 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 region on both the ground truth and predicted segmentation overlapped with each other, 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 sensitivity and precision as follows:

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

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

to evaluate the lesion-wise performance. In addition, the subject-wise detection rate is important in clinical diagnosis. We used the number of failure-to-detect subjects (FD-S) to evaluate the subject-level performance.

To verify the consistency of the labels that were twice given by the experts, the intraclass correlation coefficient (ICC) and κ coefficient were computed between the 2 lesion measurements. Two-paired-sample Wilcoxon and Kruskal-Wallis tests were performed to determine whether the VGG-16 and ResNet-50 were significantly different in terms of parameters. The full and weak labeling time, as well as the human and machine reading time, was compared using the 2-paired sample Wilcoxon test.