Random Expert Sampling for Deep Learning Segmentation of Acute Ischemic Stroke on Non-contrast CT

Sophie Ostmeier, MD[1], Brian Axelrod, PhD[1], Benjamin Pulli, MD[1], Benjamin F.J. Verhaaren, MD, PhD[2], Abdelkader Mahammedi, MD[1], Yongkai Liu, PhD[1], Christian Federau, MD, MS[3], Greg Zaharchuk, MD, PhD[1], and Jeremy J. Heit, MD, PhD[1,*]

[*]*corresponding author's email: jheit@stanford.edu*
[1]*Stanford School of Medicine, 453 Quarry Rd, Palo Alto, CA 94304, USA*
[2]*KU Leuven, Medical Imaging Research Center, Leuven, Belgium*
[3]*AI Medical AG, Zurich, Switzerland*

Short Title

Random Expert Sampling in Deep Learning for Stroke

Word Count

3273 words in total

Abstract

Purpose

To automatically identify the presence and location of ischemic brain tissue on Non-Contrast CT

Materials and Methods

The data set consisted of 260 Non-Contrast CTs from 233 patients of acute ischemic stroke patients recruited in the DEFUSE 3 trial. A benchmark U-Net was trained on the reference annotations of three experienced neuroradiologists to segment ischemic brain tissue using majority vote and random expert sampling training schemes. We used a one-sided Wilcoxon signed-rank test on a set of segmentation metrics to compare bootstrapped point estimates of the training schemes with the inter-expert agreement and ratio of variance for consistency analysis. We further compare volumes with the 24h-follow-up DWI (final infarct core) in the patient subgroup with full reperfusion and we test volumes for correlation to the clinical outcome (mRS after 30 and 90 days) with the Spearman method.

Results

Random expert sampling leads to a model that shows better agreement with experts than experts agree among themselves and better agreement than the agreement between experts and a majority-vote model performance (Surface Dice at Tolerance 5mm improvement of 61% to 0.70±0.03 and Dice improvement of 25% to 0.50±0.04). The model-based predicted volume similarly estimated the final infarct volume and correlated better to the clinical outcome than CT perfusion.

Conclusion

A model trained on random expert sampling can identify the presence and location of acute ischemic brain tissue on Non-Contrast CT similar to CT perfusion and with better consistency than experts. This may further secure the selection of patients eligible for endovascular treatment in less specialized hospitals.

1   Introduction

Stroke is a leading cause of death and disability globally [1, 2]. Timely treatments, such as thrombolysis or thrombectomy, represent effective treatment options to improve clinical outcomes for patients suffering from acute ischemic stroke [3, 4, 5, 6]. Non-Contrast CT (NCCT) is widely used to differentiate ischemic from hemorrhagic stroke. It can be further used to estimate the extent of irreversibly damaged brain tissue as an imaging biomarker for the appropriate triage of patients [7, 8].

Advances have been made to facilitate the delineation of ischemic brain tissue with ruled-based algorithms for perfusion imaging [9]. But resources, specialized technicians, and physicians are needed for interpretations. The most utilized (96-100%) and cheapest imaging modality for acute stroke patients is the NCCT [7, 10].

The ASPECT score is an established semi-quantitative method to evaluate treatment eligibility in ischemic stroke patients based on NCCT by dividing the affected hemisphere into 10 structural regions [3, 4]. However, the extent and relative quantity of damaged brain tissue within a region to be rendered has not been defined [11, 12]. The individual regions cover different amounts of brain tissue which results in their unequal weighing into the score. The location of the stroke has a minor role in the ASPECT score value. Although it has been shown that the location of the stroke estimates the long-term effect on patients' outcome more precisely [13, 14, 15, 16].

A segmentation tool for acute ischemic brain tissue on NCCT may not only accurately quantify the volume but also allows the mapping of the stroke location to symptoms to better guide treatment decisions. However, a low signal-to-noise ratio limits rule-based algorithms and expert neuroradiologists to segment ischemic stroke on NCCT. An automated pipeline is therefore of high interest [17, 18, 19].

Supervised deep learning models have been widely applied for the segmentation of medical images and show promise for applications in stroke imaging, but rely on accurate expert reference annotations segmentation during training [20, 21]. To enhance the accuracy of ground truth segmentation, multiple expert annotations are commonly fused with a majority or with the probabilistic fusion algorithm STAPLE. The result is a binary ground truth (a voxel is part of the ischemic core or not). Note that STAPLE is not applicable in the context of this study, where sometimes only two, one, or no experts agree on ischemic stroke segmentation [22]. Instead, we use random expert sampling, meaning randomly choosing an expert segmentation as ground truth for each training example. We then compare the model's performance to the standard majority vote. We hypothesize that randomly sampling experts during training is better in approximating the ground truth of the ischemic core better than a majority vote due to its ability to encode the true probability when using only binary expert labels. For ischemic brain segmentation, the true probability of neurons being irreversibly damaged by ischemia maps to a continuous function of time, collateral blood flow, and clinical variables [23, 24, 25]. If we consider a Bernoulli random variable corresponding to a single pixel (a coin flip with probability p) is defined as 1 with probability p and 0 with probability $1 - p$. Consider this variable with $p = 50\%$ with three samples corresponding to the

annotations of three experts that happen to be $(0, 0, 1)$. If we used a majority vote pre-processing scheme, this would be summarized as a single zero, and any resulting model or estimator would attempt to predict a zero. However, maximizing the cross entropy loss function of the deep learning model (identical to taking the maximum likelihood estimate, (Supplemental Material, Section 12.1.1) would result in an estimate of $1/3$. In fact, we'd expect a better estimate in 75% of three-expert datasets. Finally, we correlate the prediction of the random expert model to the clinical outcome and compare it to CT perfusion and follow-up MR imaging.

2 Methods

2.1 Study Design and Data

In this post hoc analysis of the randomized DEFUSE 3 trial, 260 Head NCCT examinations from 233 patients with enrollment between May 2016 through May 2017 were included. The primary trial outcome determined thrombectomy eligibility for patients with acute ischemic stroke with an onset time within 16 hours. Detailed information of the inclusion and exclusion criteria can be found in `https://clinicaltrials.gov/ct2/show/NCT02586415` and [9]. Furthermore, 156 consecutive patients without an ischemic stroke from a different institutional database were included to evaluate the image-classification task (relevant stroke vs. no relevant stroke volume). These patients were screened for stroke (from 2011 to 2017) and confirmed to have no stroke with a normal DWI scan 24h after the initial NCCT, but were diagnosed with an alternative condition on the basis of all available clinical and radiologic data available to the treating physician [26]. We included patients from the University Hospital Lausanne (UHL) (1723 patients) as an external generalization cohort. We randomly chose 35 patients of this cohort where each patient was admitted between 2016 through 2019, had an onset time between 3h to 24h and an initial NCCT study. Institutional review board approval was obtained, and consent was waived.

2.2 Image Reference Annotations

Three experienced fellowship-trained neuroradiologists (experts A, B, and C) with 4, 4, and 5 years of post-fellowship experience (A.M., B.F.J.V., J.J.H.). Manual segmentations were done in Horos (Horosproject.org, version 4.0.0). The experts were given the information that each patient had a large vessel occlusion, but the side of the vessel occlusion was not provided. If no abnormal hypodensity was found, experts were offered the option of no segmentation. All manual segmentations were checked for correct coregistration to the NCCT. For the data set of the healthy patient, an empty segmentation mask was generated. The external generalization cohort was segmented by one fellowship-trained neuroradiologists with 2 years of post-fellowship training.

2.3 Data Preparation for Training

The cohort of 233 patients was randomly split into five subgroups for five-fold cross-validation with 208 patients for the training, and 52 for the test sets. If a patient had undergone multiple baseline NCCT (n=27 patients), the additional scans were only added to the corresponding patient NCCT in the training set to prevent information leakage (Figure 1). The 156 healthy cases were randomly and equally added to the test sets (n=33 per test set) with empty ground truth annotations.

2.4 Model Configuration and Training

A preprocessing pipeline and data loader were created to allow on the fly training with majority vote and random expert sampling (Figure 2). A majority vote means to sum the reference masks of all experts, set the voxels with a value greater than half the number of experts to one, and all other voxels to zero. Random expert sampling means choosing an expert for each training example with an equal probability.

A baseline nnUNet architecture and default hyperparameter include a patch size of 28x256x256 and spacing of (3.00, 0.45, 0.45), 7 stages with two 3D convolutions per stage, leaky ReLU as activation function, Dice + Cross-Entropy loss function, a batch size of 2, SGD optimizer with 0.99 Nestov momentum and He initialization (Section 12, Fig. S1). The model was trained with NCCT as input and manual annotations of experts as ground truth to output a predicted segmentation mask of acute ischemic stroke on NCCT [27]. As all models share the same core nnUNet component and for fairness and ease of comparability, we let all models undergo the same training schedule with default hyperparameters to prevent information leakage.

2.5 Model Testing

All analyses have been performed on the aggregated test sets of the five folds. This procedure is explored and validated in previous literature [28]. The results of the aggregated test sets (n=389) were tested for normal distribution with the Shapiro and shown as median and 95% CI estimated by the bootstrapping (R=1000)[29] We tested the best-performing model (overall highest metrics value) on the external generalization cohort.

2.6 Outcome Measures and Statistical Analysis

R (Version 2022.02.3) was used for statistical analysis.

All three expert reference annotations per NCCT were compared to the inter-expert agreement, to the prediction of the majority vote model, and to the prediction of the random expert sampling model (Figure 3). The per-patient median was used for further analysis.

The model-based predictions were evaluated for segmentation error, image classification, volume classification, correlation to CT ischemic core, 24h-follow-up volume, and clinical outcome prediction.

To access the segmentation and image-classification task a clinically motivated threshold of 1ml was chosen to differentiate between relevant stroke and no relevant stroke volume. For example, after determining no relevant stroke volume on the NCCT, the location, and volume of ischemic brain tissue of 0.5 ml are unlikely to influence the triage of patients and excluded from the segmentation evaluation. Expert A, B, and C had 155, 226, and 206 manual segmentations > 1ml and 78, 7, and 27 manual segmentations < 1ml, respectively.

All NCCTs with a median reference annotation above 1ml were evaluated for segmentation error with the following metrics.

- Volume-based metrics (Volumetric similarity and absolute volume difference [ml])

- Overlap metrics (Dice, Precision and Recall)

- Distance metrics Hausdorff distance with the 95 percentile [mm], and the surface dice at tolerance 5mm

The definition can be found in the supplemental material 12, Table S1 [30].

The Surface Dice at Tolerance measures the distances between each surface voxel of the reference and predicted mask. The chosen tolerance allows a maximum distance between surface voxels on the reference and predicted masks to be considered true positive voxels. We use this metric to account for possible pathophysiological and modality-related lower signals that cause more variability in the outer compared to the inner region of the ischemic stroke [31, 32].

For the image-classification task, the predictions were categorized in stroke volume above or below 1ml including the healthy cases. Sensitivity, specificity, F-score, Correct Classification Ratio, and Area under the curve served as metrics.

We used each metric to evaluate how close the model-based predictions were to experts (accuracy) and how consistent the model-based predictions were across the patient population (precision).

Accuracy was tested for statistical superiority with a one-sided Wilcoxon sign rank test and a lower boundary of 0.0 ($p < 0.05$, Figure 3). We evaluate for precision with the ratio of variances and standard error (bootstrapped, R=1000). We then tested for statistical superiority with a delta of 0.0.

To evaluate for ischemic core volume size that alters clinical treatment decisions based on ANGEL-ASPECT and SELECT-2 [4, 3], the predicted segmentations were classified in <1ml, <50ml, <100ml and >100ml and compared to the median expert volume with Cohen's and Fleiss' kappa.

To put the results into perspective, we computed the Spearman correlation coefficient of the ASPECT and ischemic core volumes estimated by experts, deep learning models, and CT perfusion to the clinical outcomes. We compared the ischemic core volumes of the random expert sampling model and CT perfusion (30% CBF) to the 24h-follow-up DWI (final infarct core) with Bland-Altman plots in patients with full reperfusion (TICI ≥2B). We analyzed differences in correlation coefficients with Fisher's z-test.

All p-values were adjusted for multiple comparisons with the Holm-Bonferroni method.

2.7   Model and Data Set Availability

The model and data set are available upon request to the corresponding author.

3   Results

3.1   Patient Characteristics

We analyzed 233 randomized and non-randomized patients in the DEFUSE 3 trial (121 women (52%), median (IQR) age, 69 (59-78) years). The expert ischemic core volume and ischemic core (CBF<=30%) and penumbra (Tmax . =6) volume with perfusion imaging were 8[3-26]ml, 11[2-28] and 104[62-157] ml, respectively (median [IQR] ml). The median onset to image time of the 50 patients witnessed was 10[8-12]h. Further patient characteristics are summarized in Table 1.

3.2   Model Performance

The model-based segmentation with the random expert sampling showed statistically significant better agreement with experts than inter-expert agreement and majority vote training scheme (median $\pm$ 95% CI (bootstrapped), Surface Dice at Tolerance 0.7$\pm$0.03 vs. 0.56$\pm$0.03, 0.68$\pm$0.05, Dice 0.50$\pm$0.04 vs. 0.31$\pm$0.04, 0.29.$\pm$0.01, Absolute Volume Difference of 5.36$\pm$1.32 ml vs. 10.20$\pm$2.09 ml, 7.91$\pm$1.77 ml) (Table 2).

The consistency of the model-based segmentation with the random expert sampling across the patient population was significantly better for VS, AVD, and Precision but insignificant elsewhere. However, most non-significant comparisons have a standard deviation that includes variance ratios of 1. For these metrics, no conclusion about the segmentation precision superiority or inferiority can be made and the variance is most likely similar.

The model-based segmentation with the random expert sampling showed a smaller average volume difference and 95% CI when compared to the median expert volume with Bland-Altman plots, while the majority vote tends to underestimate the median expert volume (Figure 4) (mean and 95% CI volume difference of random expert sampling vs. majority vote, 0.8(-25.8-27.5) ml vs. 7.2(-26-42.4) ml).

Qualitatively, the probability predictions of random expert sampling correlate better to the fields segmented by the experts than to the predictions for the majority vote which may show overfitting to the single hypodense voxels and image noise (Figure 6).

The random expert sampling model indicated similar performance to classifying cases into relevant and not relevant stroke volume vs. inter-expert agreement and the majority vote model, respectively (AUC 0.92$\pm$0.02 vs. 0.93$\pm$0.02, 0.90$\pm$0.02) (Table 3)

3.3 Evaluation of NCCT ischemic Core Volume as Imaging Biomarker in Clinical Practice

Cohen's kappa for volume classification (<1ml, <50ml, <100ml and <100ml) of the random expert sampling model is higher than Fleiss' kappa for inter-expert agreement and Cohen's kappa for the majority vote model (0.52 ±0.06 vs. 0.32 ±0.04 and 0.24 ±0.05).

When comparing the initial ischemic core volume to the clinical outcome (mRS after 30 and 90 days), random experts sampling model volume and ASPECTS correlate significantly to the mRS scores after 30 and 90 days and the CTP volume does not (Table 4).

Subgroup analysis of patients with full reperfusion (TICI ≥2B, n=51) shows no significant difference when comparing the average volume differences for the random expert sampling volume prediction and CT perfusion volume to the 24h-follow-up DWI volume (Fisher's z-test p-value = 0.355, Figure 5). Analysis of the whole patient population shows similar results (Supplemental Material, Figure S2).

4 Discussion

In this study, we found that random expert sampling training of a benchmark deep learning model leads to better agreement with experts than experts among themselves for the segmentation of the ischemic core on NCCT. The model-based ischemic core shows similar volume agreement with the final infarct volume and better correlation to the clinical outcomes as CT perfusion.

The significance of these findings is threefold. First, the model outperforms experts of the highest human expertise when jointly trained by the experts themselves. Second, the model's volume demonstrates similar clinical predictive value as the ischemic core volume estimates of current rule-based algorithms. Third, we found that contrary to prior deep learning applications, fusion techniques such as a majority vote cause the model to overfit to single voxel values and segment less meaningful clinical information [33, 34, 35, 22].

ECASS I, ESCAPE, DAWN, DEFUSE 3, SELECT-2 and ANGEL-ASPECT are clinical trials that have included the measures of ischemic core on NCCT as inclusion and treatment criteria for patients with acute ischemic stroke [6, 36, 9, 5, 4, 3].

The ECASS I and TWIST trial specified hypodensity consistent with ischemic brain tissue of more than 1/3 of the medial cerebral artery territory as exclusion criteria[37, 38]. Equivalently, ESCAPE and (IMS)-III Trial and retrospective studies suggest Alberta Stroke Program Early CT Score (ASPECTS) of < 5, < 6, < 7 or < 8 for estimating a large ischemic infarct and to guide endovascular treatment decisions [36, 39, 38, 40]. However, the benefits of thrombectomy have recently also been demonstrated for patients with ASPECTS 3 − 5 in SELECT-2 and ANGEL-ASPECT [3, 4]. Although ASPECTS is widely used as a predictor for patients to benefit from endovascular treatment using different thresholds, it is limited by inter-rater variability, a modest correlation to

ischemic core volumes and location [41, 42, 43, 12].

Besides the ASPECT score, the ASPECT-ANGEL and SELECT-2 trials have also included patients with an ischemic core volume of $70 - 100$ml and > 50ml that was quantified with CT perfusion [3, 4].

We proposed a deep learning segmentation method of the ischemic core on NCCT that may alleviate limitations of perfusion imaging and the ASPECTS score.

Different reference standard definitions of the ischemic core exist across various stroke imaging modalities (MR and CT Perfusion, DWI). Sarraj et. al and SELECT investigators have shown that the ischemic core volume on CT perfusion is biased and poorly overlaps with the final infarct volume (median (IOR) absolute volume difference 17.9 (5.6–43.9)ml and Dice score 0.1 (0.0-0.4)) [44].

Diffusion-weighted imaging (DWI) within a short time after the NCCT is used in prior works as ground truth for the segmentation or detection of the ischemic core on NCCT [45, 46, 47, 48]. However, especially in earlier time windows (< 1h) cytotoxic edema is the predominant image abnormality depicted as diffusion restriction on DWI [49]. Vasogenic edema depicted as hypodensity on non-contrast Computed Tomography (NCCT) develops $> 1 - 4$h after stroke onset and suggests irreversibly damaged brain tissue (ischemic core) [50]. The image correlation of the underlying ischemic core pathology of a DWI lesion on NCCT especially in the earlier time frames seem unclear and voxel-wise comparisons of NCCT and DWI has been scarcely studied [51].

Deep learning medical image segmentation is more often supervised by human experts' annotations and fusion methods (e.g. majority vote) may approximate an error-free ground truth when collecting annotations from multiple experts. In a segmentation task with reference annotation of uncertainty, small target lesions, or empty segmentations reference annotations require at least highly skilled experts to minimize the errors, and multiple experts' segmentations are needed to approximate the distribution of interpretations, which is resource-intensive and timely [30]. For acute ischemic stroke segmentation on NCCT, fusion methods for categorical voxel classes may limit the ability of the model to learn segmentation tasks where experts inherently disagree. We found that advanced fusion methods, such as STAPLE or SIMPLE are not applicable and tested majority voting with modest performance [22, 34].

We propose a training requiem for multi-expert training that maximizes the encoded information of the NCCT interpretations from three expert neuroradiologists. We used three experts so that there is a "tie-breaker" between two raters. A larger number of experts is not feasible. The model trained with this combined encoded information (random expert sampling methods) agreed more with experts when compared to the inter-human-reader agreement (Dice score 0.51±0.04 vs. 0.31±0.04 (Table 2)). When comparing the model-based volume and CT Perfusion volume to the final infarct volume we could find no significant difference in final infarct estimation.

Neuroimaging is a significant cost driver in acute ischemic stroke patients [52]. In addition, reduced imaging modalities and scan time may increase the net monetary benefit of earlier endovascular treatment ($10,593 for every 10 minutes) [53, 54]. However, additional and prospective studies regarding clinical outcomes are needed

before the long-term cost-effectiveness of using NCCT and deep learning tools can be concluded [55].

This study has limitations. First, we only use a patient cohort of one multicenter clinical trial with the primary outcome measure of selecting patients that present within 16h. For the image classification, we do not include stroke mimics or hemorrhagic stroke patients. Further research may test the proposed method in a broader and prospective patient population. Second, we include three experts. The distribution of image interpretation may require a larger set of neuroradiologists which would be resource-intensive but may improve performance. Third, the volume of the stroke is only one imaging feature that correlates with the clinical outcome and may influence decision-making. In future work, deep learning methods may offer voxel-based lesion-symptom mapping for better estimation of clinical outcome prediction.

In clinical practice, our model may identify, quantify, and localize acute ischemic stroke on NCCT comparable to CT perfusion. An accurate and precise deep learning segmentation methodology allows less specialized on-call clinicians to utilize the information for treatment decisions given by NCCT. This possibly enhances the impact of NCCT in endovascular treatment decisions for ischemic strokes as a basic, cheap, and widely available imaging modality.

5 Acknowledgments

-

7 Disclosures

Sophie Ostmeier: none

Brian Axelrod: none

Benjamin F.J. Verhaaren: none

Abdelkader Mahammedi: none

Benjamin Pulli: none

Yongkai Liu: none

Christian Federau: Founder and CEO of AI Medical AG

Greg Zaharchuk: co-founder, equity of Subtle Medical, funding support GE Healthcare, consultant Biogen Jeremy

J. Heit: Consultant for Medtronic and MicroVention, Member of the medical and scientific advisory board for iSchemaView

8 Supplemental Material

Supplemental Methods

Tables S1–S2

Figure S1

9 References

[1] V. L. Feigin, B. A. Stark, C. O. Johnson, G. A. Roth, C. Bisignano, G. G. Abady, M. Abbasifard, M. Abbasi-Kangevari, F. Abd-Allah, V. Abedi *et al.*, "Global, regional, and national burden of stroke and its risk factors, 1990–2019: a systematic analysis for the global burden of disease study 2019," *The Lancet Neurology*, vol. 20, no. 10, pp. 795–820, 2021.

[2] N. Lakomkin, M. Dhamoon, K. Carroll, I. P. Singh, S. Tuhrim, J. Lee, J. T. Fifi, and J. Mocco, "Prevalence of large vessel occlusion in patients presenting with acute ischemic stroke: a 10-year systematic review of the literature," *Journal of NeuroInterventional Surgery*, vol. 11, no. 3, pp. 241–245, 2019. [Online]. Available: https://jnis.bmj.com/content/neurintsurg/11/3/241.full.pdf

[3] X. Huo, G. Ma, X. Tong, X. Zhang, Y. Pan, T. N. Nguyen, G. Yuan, H. Han, W. Chen, M. Wei, J. Zhang, Z. Zhou, X. Yao, G. Wang, W. Song, X. Cai, G. Nan, D. Li, A. Y.-C. Wang, W. Ling, C. Cai, C. Wen, E. Wang, L. Zhang, C. Jiang, Y. Liu, G. Liao, X. Chen, T. Li, S. Liu, J. Li, F. Gao, N. Ma, D. Mo, L. Song, X. Sun, X. Li, Y. Deng, G. Luo, M. Lv, H. He, A. Liu, J. Zhang, S. Mu, L. Liu, J. Jing, X. Nie, Z. Ding, W. Du, X. Zhao, P. Yang, L. Liu, Y. Wang, D. S. Liebeskind, V. M. Pereira, Z. Ren, Y. Wang, and Z. Miao, "Trial of endovascular therapy for acute ischemic stroke with large infarct," *New England Journal of Medicine*, p. null, 2023.

[4] A. Sarraj, A. E. Hassan, M. G. Abraham, S. Ortega-Gutierrez, S. E. Kasner, M. S. Hussain, M. Chen, S. Blackburn, C. W. Sitton, L. Churilov, S. Sundararajan, Y. C. Hu, N. A. Herial, P. Jabbour, D. Gibson, A. N. Wallace, J. F. Arenillas, J. P. Tsai, R. F. Budzik, W. J. Hicks, O. Kozak, B. Yan, D. J. Cordato, N. W. Manning, M. W. Parsons, R. A. Hanel, A. N. Aghaebrahim, T. Y. Wu, P. Cardona-Portela, N. Pérez de la Ossa, J. D. Schaafsma, J. Blasco, N. Sangha, S. Warach, C. D. Gandhi, T. J. Kleinig, D. Sahlein, L. Elijovich, W. Tekle, E. A. Samaniego, L. Maali, M. A. Abdulrazzak, M. N. Psychogios, A. Shuaib, D. K. Pujara, F. Shaker, H. Johns, G. Sharma, V. Yogendrakumar, F. C. Ng, M. H. Rahbar, C. Cai, P. Lavori, S. Hamilton, T. Nguyen, J. T. Fifi, S. Davis, L. Wechsler, V. M. Pereira, M. G. Lansberg, M. D. Hill, J. C. Grotta, M. Ribo, B. C. Campbell, and G. W. Albers, "Trial of endovascular thrombectomy for large ischemic strokes," *New England Journal of Medicine*. [Online]. Available: https://doi.org/10.1056/NEJMoa2214403

[5] R. G. Nogueira, A. P. Jadhav, D. C. Haussen, A. Bonafe, R. F. Budzik, P. Bhuva, D. R. Yavagal, M. Ribo, C. Cognard, R. A. Hanel, C. A. Sila, A. E. Hassan, M. Millan, E. I. Levy, P. Mitchell, M. Chen, J. D. English, Q. A. Shah, F. L. Silver, V. M. Pereira, B. P. Mehta, B. W. Baxter, M. G. Abraham, P. Cardona, E. Veznedaroglu, F. R. Hellinger, L. Feng, J. F. Kirmani, D. K. Lopes, B. T. Jankowitz, M. R. Frankel,

V. Costalat, N. A. Vora, A. J. Yoo, A. M. Malik, A. J. Furlan, M. Rubiera, A. Aghaebrahim, J.-M. Olivot, W. G. Tekle, R. Shields, T. Graves, R. J. Lewis, W. S. Smith, D. S. Liebeskind, J. L. Saver, and T. G. Jovin, "Thrombectomy 6 to 24 hours after stroke with a mismatch between deficit and infarct," *New England Journal of Medicine*, vol. 378, no. 1, pp. 11–21, 2017. [Online]. Available: https://www.nejm.org/doi/full/10.1056/NEJMoa1706442

[6] W. Hacke, M. Kaste, E. Bluhmki, M. Brozman, A. Dávalos, D. Guidetti, V. Larrue, K. R. Lees, Z. Medeghri, T. Machnig *et al.*, "Thrombolysis with alteplase 3 to 4.5 hours after acute ischemic stroke," *New England journal of medicine*, vol. 359, no. 13, pp. 1317–1329, 2008.

[7] Y. Kim, S. Lee, R. Abdelkhaleq, V. Lopez-Rivera, B. Navi, H. Kamel, S. I. Savitz, A. L. Czap, J. C. Grotta, L. D. McCullough, T. M. Krause, L. Giancardo, F. S. Vahidy, and S. A. Sheth, "Utilization and availability of advanced imaging in patients with acute ischemic stroke," *Circ Cardiovasc Qual Outcomes*, vol. 14, no. 4, p. e006989, 2021, kim, Youngran Lee, Songmi Abdelkhaleq, Rania Lopez-Rivera, Victor Navi, Babak Kamel, Hooman Savitz, Sean I Czap, Alexandra L Grotta, James C McCullough, Louise D Krause, Trudy Millard Giancardo, Luca Vahidy, Farhaan S Sheth, Sunil A eng Research Support, Non-U.S. Gov't 2021/03/25 Circ Cardiovasc Qual Outcomes. 2021 Apr;14(4):e006989. doi: 10.1161/CIRCOUTCOMES.120.006989. Epub 2021 Mar 24. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/33757311https: //www.ahajournals.org/doi/pdf/10.1161/CIRCOUTCOMES.120.006989?download=true

[8] R. McDonough, J. Ospel, and M. Goyal, "State of the art stroke imaging: A current perspective," *Can Assoc Radiol J*, p. 8465371211028823, 2021, mcDonough, Rosalie Ospel, Johanna Goyal, Mayank eng 2021/09/28 Can Assoc Radiol J. 2021 Sep 27:8465371211028823. doi: 10.1177/08465371211028823. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/34569306https: //journals.sagepub.com/doi/pdf/10.1177/08465371211028823

[9] G. W. Albers, M. P. Marks, S. Kemp, S. Christensen, J. P. Tsai, S. Ortega-Gutierrez, R. A. McTaggart, M. T. Torbey, M. Kim-Tenser, T. Leslie-Mazwi, A. Sarraj, S. E. Kasner, S. A. Ansari, S. D. Yeatts, S. Hamilton, M. Mlynash, J. J. Heit, G. Zaharchuk, S. Kim, J. Carrozzella, Y. Y. Palesch, A. M. Demchuk, R. Bammer, P. W. Lavori, J. P. Broderick, and M. G. Lansberg, "Thrombectomy for stroke at 6 to 16 hours with selection by perfusion imaging," *New England Journal of Medicine*, vol. 378, no. 8, pp. 708–718, 2018. [Online]. Available: https://www.nejm.org/doi/full/10.1056/NEJMoa1713973https: //www.nejm.org/doi/pdf/10.1056/NEJMoa1713973?articleTools=true

[10] J. J. Wang, A. Boltyenkov, J. M. Katz, J. O'Hara, M. Gribko, and P. C. Sanelli, "Striving for socioeconomic equity in ischemic stroke care: Imaging and acute treatment utilization from a comprehensive stroke center,"

Journal of the American College of Radiology, vol. 19, no. 2, Part B, pp. 348–358, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1546144021007341

[11] T. G. Phan, G. A. Donnan, M. Koga, L. A. Mitchell, M. Molan, G. Fitt, W. Chong, M. Holt, and D. C. Reutens, "The aspects template is weighted in favor of the striatocapsular region," *Neuroimage*, vol. 31, no. 2, pp. 477–481, 2006.

[12] J. Schroeder and G. Thomalla, "A critical review of alberta stroke program early ct score for evaluation of acute stroke imaging," *Frontiers in Neurology*, vol. 7, 2017. [Online]. Available: https://www.frontiersin.org/article/10.3389/fneur.2016.00245

[13] N. A. Weaver, H. J. Kuijf, H. P. Aben, J. Abrigo, H.-J. Bae, M. Barbay, J. G. Best, R. Bordet, F. M. Chappell, C. P. Chen *et al.*, "Strategic infarct locations for post-stroke cognitive impairment: a pooled analysis of individual patient data from 12 acute ischaemic stroke cohorts," *The Lancet Neurology*, vol. 20, no. 6, pp. 448–459, 2021.

[14] Y. Wang, J. M. Juliano, S.-L. Liew, A. M. McKinney, and S. Payabvash, "Stroke atlas of the brain: Voxel-wise density-based clustering of infarct lesions topographic distribution," *NeuroImage: Clinical*, vol. 24, p. 101981, 2019.

[15] S. Meyer, S. S. Kessner, B. Cheng, M. Bönstrup, R. Schulz, F. C. Hummel, N. De Bruyn, A. Peeters, V. Van Pesch, T. Duprez *et al.*, "Voxel-based lesion-symptom mapping of stroke lesions underlying somatosensory deficits," *NeuroImage: Clinical*, vol. 10, pp. 257–266, 2016.

[16] G. Broocks, F. Flottmann, M. Ernst, T. D. Faizy, J. Minnerup, S. Siemonsen, J. Fiehler, and A. Kemmling, "Computed tomography–based imaging of voxel-wise lesion water uptake in ischemic brain: relationship between density and direct volumetry," *Investigative radiology*, vol. 53, no. 4, pp. 207–213, 2018.

[17] W. L. Nowinski, J. Walecki, G. Półtorak-Szymczak, K. Sklinda, and B. Mruk, "Ischemic infarct detection, localization, and segmentation in noncontrast ct human brain scans: review of automated methods," *PeerJ*, vol. 8, p. e10444, 2020.

[18] X. Chen, S. Lin, X. Zhang, S. Hu, and X. Wang, "Prognosis with non-contrast ct and ct perfusion imaging in thrombolysis-treated acute ischemic stroke," *European Journal of Radiology*, vol. 149, p. 110217, 2022.

[19] H. El-Hariri, L. A. S. M. Neto, P. Cimflova, F. Bala, R. Golan, A. Sojoudi, C. Duszynski, I. Elebute, S. H. Mousavi, W. Qiu *et al.*, "Evaluating nnu-net for early ischemic change segmentation on non-contrast computed tomography in patients with acute ischemic stroke," *Computers in biology and medicine*, vol. 141, p. 105033, 2022.

[20] C.-F. Liu, J. Hsu, X. Xu, S. Ramachandran, V. Wang, M. I. Miller, A. E. Hillis, A. V. Faria, M. Wintermark, S. J. Warach, G. W. Albers, S. M. Davis, J. C. Grotta, W. Hacke, D.-W. Kang, C. Kidwell, W. J. Koroshetz, K. R. Lees, M. H. Lev, D. S. Liebeskind, A. G. Sorensen, V. N. Thijs, G. Thomalla, J. M. Wardlaw, M. Luby, S. The, and V. I. investigators, "Deep learning-based detection and segmentation of diffusion abnormalities in acute ischemic stroke," *Communications Medicine*, vol. 1, no. 1, p. 61, 2021. [Online]. Available: https://doi.org/10.1038/s43856-021-00062-8https://www.nature.com/articles/s43856-021-00062-8.pdf

[21] Y. Yu, Y. Xie, T. Thamm, E. Gong, J. Ouyang, C. Huang, S. Christensen, M. P. Marks, M. G. Lansberg, G. W. Albers *et al.*, "Use of deep learning to predict final ischemic stroke lesions from initial magnetic resonance imaging," *JAMA network open*, vol. 3, no. 3, pp. e200 772–e200 772, 2020.

[22] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation," *IEEE transactions on medical imaging*, vol. 23, no. 7, pp. 903–921, 2004.

[23] J. L. Saver, "Time is brain—quantified," *Stroke*, vol. 37, no. 1, pp. 263–266, 2006.

[24] M. P. Marks, M. G. Lansberg, M. Mlynash, J.-M. Olivot, M. Straka, S. Kemp, R. McTaggart, M. Inoue, G. Zaharchuk, R. Bammer *et al.*, "Effect of collateral blood flow on patients undergoing endovascular therapy for acute ischemic stroke," *Stroke*, vol. 45, no. 4, pp. 1035–1039, 2014.

[25] K. Johnston, A. Connors Jr, D. Wagner, W. Knaus, X.-Q. Wang, and E. C. Haley Jr, "A predictive risk model for outcomes of ischemic stroke," *Stroke*, vol. 31, no. 2, pp. 448–455, 2000.

[26] S. Christensen, M. Mlynash, J. MacLaren, C. Federau, G. W. Albers, and M. G. Lansberg, "Optimizing deep learning algorithms for segmentation of acute infarcts on non-contrast material-enhanced ct scans of the brain using simulated lesions," *Radiol Artif Intell*, vol. 3, no. 4, p. e200127, 2021, christensen, Soren Mlynash, Michael MacLaren, Julian Federau, Christian Albers, Gregory W Lansberg, Maarten G eng 2021/08/06 Radiol Artif Intell. 2021 May 12;3(4):e200127. doi: 10.1148/ryai.2021200127. eCollection 2021 Jul. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/34350404https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8328101/pdf/ryai.2021200127.pdf

[27] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nat Methods*, vol. 18, no. 2, pp. 203–211, 2021, isensee, Fabian Jaeger, Paul F Kohl, Simon A A Petersen, Jens Maier-Hein, Klaus H eng Research Support, Non-U.S. Gov't 2020/12/09 Nat Methods. 2021 Feb;18(2):203-211. doi: 10.1038/s41592-020-01008-z. Epub 2020 Dec 7. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/33288961https://www.nature.com/articles/s41592-020-01008-z.pdf

[28] G. C. Cawley and N. L. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *The Journal of Machine Learning Research*, vol. 11, pp. 2079–2107, 2010.

[29] D. Lakens, "Equivalence tests: A practical primer for t tests, correlations, and meta-analyses," *Social psychological and personality science*, vol. 8, no. 4, pp. 355–362, 2017.

[30] S. Ostmeier, B. Axelrod, J. Bertels, F. Isensee, M. G. Lansberg, S. Christensen, G. W. Albers, L.-J. Li, and J. J. Heit, "Evaluation of medical image segmentation models for uncertain, small or empty reference annotations," *arXiv preprint arXiv:2209.13008*, 2022.

[31] S. Nikolov, S. Blackwell, A. Zverovitch, R. Mendes, M. Livne, J. De Fauw, Y. Patel, C. Meyer, H. Askham, B. Romera-Paredes *et al.*, "Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy," *arXiv preprint arXiv:1809.04430*, 2018.

[32] L. Maier-Hein, A. Reinke, E. Christodoulou, B. Glocker, P. Godau, F. Isensee, J. Kleesiek, M. Kozubek, M. Reyes, and M. A. Riegler, "Metrics reloaded: Pitfalls and recommendations for image analysis validation," *arXiv preprint arXiv:2206.01653*, 2022.

[33] O. Vincent, C. Gros, and J. Cohen-Adad, "Impact of individual rater style on deep learning uncertainty in medical imaging segmentation," *arXiv preprint arXiv:2105.02197*, 2021.

[34] T. R. Langerak, U. A. van der Heide, A. N. T. J. Kotte, M. A. Viergever, M. van Vulpen, and J. P. W. Pluim, "Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (simple)," *IEEE Transactions on Medical Imaging*, vol. 29, no. 12, pp. 2000–2008, 2010.

[35] B. A. Landman, A. J. Asman, A. G. Scoggins, J. A. Bogovic, F. Xing, and J. L. Prince, "Robust statistical fusion of image labels," *IEEE Transactions on Medical Imaging*, vol. 31, no. 2, pp. 512–522, 2012.

[36] M. Goyal, A. M. Demchuk, B. K. Menon, M. Eesa, J. L. Rempel, J. Thornton, D. Roy, T. G. Jovin, R. A. Willinsky, B. L. Sapkota *et al.*, "Randomized assessment of rapid endovascular treatment of ischemic stroke," *New England Journal of Medicine*, vol. 372, no. 11, pp. 1019–1030, 2015.

[37] W. Hacke, M. Kaste, C. Fieschi, D. Toni, E. Lesaffre, R. Von Kummer, G. Boysen, E. Bluhmki, G. Höxter, M.-H. Mahagne *et al.*, "Intravenous thrombolysis with recombinant tissue plasminogen activator for acute hemispheric stroke: the european cooperative acute stroke study (ecass)," *Jama*, vol. 274, no. 13, pp. 1017–1025, 1995.

[38] A. M. Zha, H. Kamal, J. A. Jeevarajan, O. Arevalo, L. Zhu, C. M. Ankrom, E. E. Bonfante-Mejia, T. D. Cossey, T. C. Wu, A. D. Barreto *et al.*, "Non-contrast head ct-based thrombolysis for wake-up/unknown onset stroke is safe: A single-center study and meta-analysis," *International Journal of Stroke*, vol. 17, no. 3, pp. 354–361, 2022.

[39] M. D. Hill, A. M. Demchuk, M. Goyal, T. G. Jovin, L. D. Foster, T. A. Tomsick, R. von Kummer, S. D. Yeatts, Y. Y. Palesch, and J. P. Broderick, "Alberta stroke program early computed tomography score to select patients for endovascular treatment: Interventional management of stroke (ims)-iii trial," *Stroke*, vol. 45, no. 2, pp. 444–449, 2014.

[40] A. J. Yoo, O. O. Zaidat, Z. A. Chaudhry, O. A. Berkhemer, R. G. González, M. Goyal, A. M. Demchuk, B. K. Menon, E. Mualem, D. Ueda *et al.*, "Impact of pretreatment noncontrast ct alberta stroke program early ct score on clinical outcome after intra-arterial stroke therapy," *Stroke*, vol. 45, no. 3, pp. 746–751, 2014.

[41] P. A. Barber, A. M. Demchuk, J. Zhang, and A. M. Buchan, "Validity and reliability of a quantitative computed tomography score in predicting outcome of hyperacute stroke before thrombolytic therapy," *The Lancet*, vol. 355, no. 9216, pp. 1670–1674, 2000.

[42] M. Goyal, B. K. Menon, W. H. van Zwam, D. W. Dippel, P. J. Mitchell, A. M. Demchuk, A. Dávalos, C. B. Majoie, A. van der Lugt, and M. A. De Miquel, "Endovascular thrombectomy after large-vessel ischaemic stroke: a meta-analysis of individual patient data from five randomised trials," *The Lancet*, vol. 387, no. 10029, pp. 1723–1731, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S014067361600163X?via%3Dihub

[43] P. A. Barber, A. M. Demchuk, J. Zhang, and A. M. Buchan, "Validity and reliability of a quantitative computed tomography score in predicting outcome of hyperacute stroke before thrombolytic therapy," *The Lancet*, vol. 355, no. 9216, pp. 1670–1674, 2000, doi: 10.1016/S0140-6736(00)02237-6. [Online]. Available: https://doi.org/10.1016/S0140-6736(00)02237-6https: //www.sciencedirect.com/science/article/pii/S0140673600022376?via%3Dihub

[44] A. Sarraj, B. C. Campbell, S. Christensen, C. W. Sitton, S. Khanpara, R. F. Riascos, D. Pujara, F. Shaker, G. Sharma, M. G. Lansberg *et al.*, "Accuracy of ct perfusion–based core estimation of follow-up infarction: effects of time since last known well," *Neurology*, vol. 98, no. 21, pp. e2084–e2096, 2022.

[45] H. Kuang, B. K. Menon, S. I. L. Sohn, and W. Qiu, "Eis-net: Segmenting early infarct and scoring aspects simultaneously on non-contrast ct of patients with acute ischemic stroke," *Medical Image Analysis*, vol. 70, p. 101984, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S136184152100030X

[46] W. Qiu, H. Kuang, E. Teleg, J. M. Ospel, S. I. Sohn, M. Almekhlafi, M. Goyal, M. D. Hill, A. M. Demchuk, and B. K. Menon, "Machine learning for detecting early infarction in acute stroke with non-contrast-enhanced ct," *Radiology*, vol. 294, no. 3, pp. 638–644, 2020, qiu, Wu Kuang, Hulin Teleg, Ericka Ospel, Johanna M Sohn, Sung Il Almekhlafi, Mohammed Goyal, Mayank Hill, Michael D Demchuk, Andrew M Menon, Bijoy K eng CIHR/Canada Research Support, Non-U.S. Gov't 2020/01/29 Radiology. 2020

Mar;294(3):638-644. doi: 10.1148/radiol.2020191193. Epub 2020 Jan 28. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/31990267https://pubs.rsna.org/doi/pdf/10.1148/radiol.2020191193

[47] R. Gauriau, B. C. Bizzo, D. S. Comeau, J. M. Hillis, C. P. Bridge, J. K. Chin, J. Pawar, A. Pourvaziri, I. Sesic, E. Sharaf *et al.*, "Head ct deep learning model is highly accurate for early infarct estimation," *Scientific Reports*, vol. 13, no. 1, p. 189, 2023.

[48] J. Lu, Y. Zhou, W. Lv, H. Zhu, T. Tian, S. Yan, Y. Xie, D. Wu, Y. Li, Y. Liu *et al.*, "Identification of early invisible acute ischemic stroke in non-contrast computed tomography using two-stage deep-learning model," *Theranostics*, vol. 12, no. 12, p. 5564, 2022.

[49] G. W. Albers, "Diffusion-weighted mri for evaluation of acute stroke," *Neurology*, vol. 51, no. 3 Suppl 3, pp. S47–S49, 1998.

[50] J. E. D. Almandoz, S. R. Pomerantz, R. G. González, and M. H. Lev, "Imaging of acute ischemic stroke: Unenhanced computed tomography," *Acute Ischemic Stroke*, pp. 43–56, 2011.

[51] M. G. Lansberg, G. W. Albers, C. Beaulieu, and M. P. Marks, "Comparison of diffusion-weighted mri and ct in acute stroke," *Neurology*, vol. 54, no. 8, pp. 1557–1561, 2000.

[52] E. W. Christensen, C. E. Pelzl, J. Hemingway, J. J. Wang, M. X. Sanmartin, J. J. Naidich, E. Y. Rula, and P. C. Sanelli, "Drivers of ischemic stroke hospital cost trends among older adults in the united states," *Journal of the American College of Radiology*, 2022.

[53] M. Straka, G. W. Albers, and R. Bammer, "Real-time diffusion-perfusion mismatch analysis in acute stroke," *Journal of Magnetic Resonance Imaging*, vol. 32, no. 5, pp. 1024–1037, 2010. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/jmri.22338

[54] W. G. Kunz, M. G. Hunink, M. A. Almekhlafi, B. K. Menon, J. L. Saver, D. W. Dippel, C. B. Majoie, T. G. Jovin, A. Davalos, S. Bracard, and et al., "Public health and cost consequences of time delays to thrombectomy for acute ischemic stroke," *Neurology*, vol. 95, no. 18, 2020.

[55] T. N. Nguyen, M. Abdalkader, S. Nagel, M. M. Qureshi, M. Ribo, F. Caparros, D. C. Haussen, M. H. Mohammaden, S. A. Sheth, S. Ortega-Gutierrez, and et al., "Noncontrast computed tomography vs computed tomography perfusion or magnetic resonance imaging selection in late presentation of stroke with large-vessel occlusion," *JAMA Neurology*, vol. 79, no. 1, p. 22, 2022.

10 Tables

Table 1: Patient characteristics

Categories	Characteristic	Randomized-Train[1]	Non-randomized-Train[1]	Total-Train[1]	Total-Test	p-value[7]
General	Total Number	146	87	233	35	
	Age	70 (59-79)	67 (58-76)	69 (59-78)	71 (64-78)	0.4
	Female %	52	52	52	54	
Imaging Characteristics	Expert A Volume [ml]	9 (4-21)	22 (6-69)	12 (5-32)		
	Expert B Volume [ml]	14 (5-27)	14 (0-60)	14 (4-37)		
	Expert C Volume [ml]	3 (1-7)	7 (0-43)	4 (1-12)		
	Ischemic Core Volume [ml]	9 (2-27)	18 (0-77)	11 (2-38)		
	Tmax6 Volume [ml][2]	117 (78-158)	69 (3-150)	104 (62-157)		
	ASPECTS on Baseline CT[3]	8 (7-9)	8 (5-10)	8 (7-9)		
Process	Witnessed Number	50		50		
	Wake-Up Number	77		77		
	Unwitnessed Number	19		19		
	Onset to Image Time [h]	10 (8-12)		10 (8-12)	10 (5-13)	0.43
Follow-Up	24h DWI Number	146		146		
	24h DWI Volume [ml]	39 (24-110)		39 (24-110)		
Clinical Outcome	mRS[5] at Baseline	0 (0-0)		0 (0-0)	0 (0-0)	
	mRS[5] at 90 days	4 (2-5)		4 (2-5)	2 (1-3)	0.002

[1] Median (1st - 3rd quantile), if not otherwise indicated, [2] Time-to-Maximum after 6 seconds, [3] Alberta Stroke Program Early CT Score, [5] modified Ranking Scale, [6] data for non-randomized patients not available, [7] double sided Wilcoxon test

Table 2: Segmentation, Internal Evaluation

| Categories | Metric | Internal Evaluation | | | | | | External Evaluation | |
		Random Sampling[5]		Interexpert[5]		p-value random-inter-expert[6]	Majority Vote[5]		p-value random-majority[6]	Random Sampling[7]	
Volume	VS	0.71	± 0.04	0.55	± 0.06	<0.0001	0.45	± 0.09	<0.0001	0.49	±0.08
	AVD [ml]	5.36	± 0.87	8.85	± 1.46	<0.0001	7.91	± 1.74	<0.0001	5.47	±4.12
Overlap	Dice	0.51	± 0.04	0.36	± 0.05	<0.0001	0.45	± 0.05	<0.0001	0.37	±0.07
	Precision	0.61	± 0.07	0.53	± 0.03	<0.05	0.79	± 0.04	non-sig	0.90	±0.06
	Recall	0.60	± 0.05	0.47	± 0.03	<0.0001	0.38	± 0.06	<0.0001	0.24	±0.06
Distance	HD 95 [mm]	13.60	± 2.01	17.54	± 2.01	<0.0001	10.92	± 1.72	non-sig	18.11	±7.15
	SDT 5mm	0.71	± 0.03	0.60	± 0.05	<0.0001	0.75	± 0.03	non-sig	0.59	±0.07

[1] VS = Volumetric Similarity,
[2] AVD = Absolute Volume Difference,
[3] HD 95 = Hausdorff Distance 95th percentile,
[4] SDT = Surface Dice at Tolerance,
[5] Median ± 95% CI (bootstrapped) compared to Expert A, B, C,
[6] p-values of one-sided Wilcoxon sign rank test,
[7] Median ± 95% CI (bootstrapped) compared to Expert D

Table 3: Image Classification with 1ml threshold

| Categories | Metric | Internal Evaluation | | | | | | External Evaluation | |
		Random Sampling[1]		Interexpert[1]		Majority Vote[1]		Random Sampling [2]	
Image-level	Sensitivity	0.94	± 0.02	0.91	± 0.02	0.65	± 0.03	0.95	±0.05
classification	Specificity	0.70	± 0.03	0.99	± 0.03	0.97	± 0.01	0.36	±0.07
	F-score	0.85	± 0.02	0.85	± 0.02	0.77	± 0.02	0.78	± 0.05
	AUC	0.92	± 0.02	0.93	± 0.02	0.90	± 0.03	0.74	±0.09

[1] Median ± 95% CI (bootstrapped) compared to Expert A, B, C
[2] Median ± 95% CI (bootstrapped) compared to Expert D

Table 4: Correlation to the clinical outcome for ASPECTS and volume estimates

Predictor	Rho of mRS 30 days[1]	p-value	Rho of mRS 90 days [1]	p-value
ASPECT	-0.18	<0.05	-0.19	<0.05
Ischemic core volume of CTP	0.16	non-sig	0.15	non-sig
Median expert volume	0.17	<0.05	0.16	non-sig
Volume of majority vote model	0.13	non-sig	0.15	non-sig
Volume of random expert sampling model	0.21	<0.01	0.19	<0.05

[1] Spearman's Rho and p-value for correlation
[2] Alberta Stroke Program Early CT Score
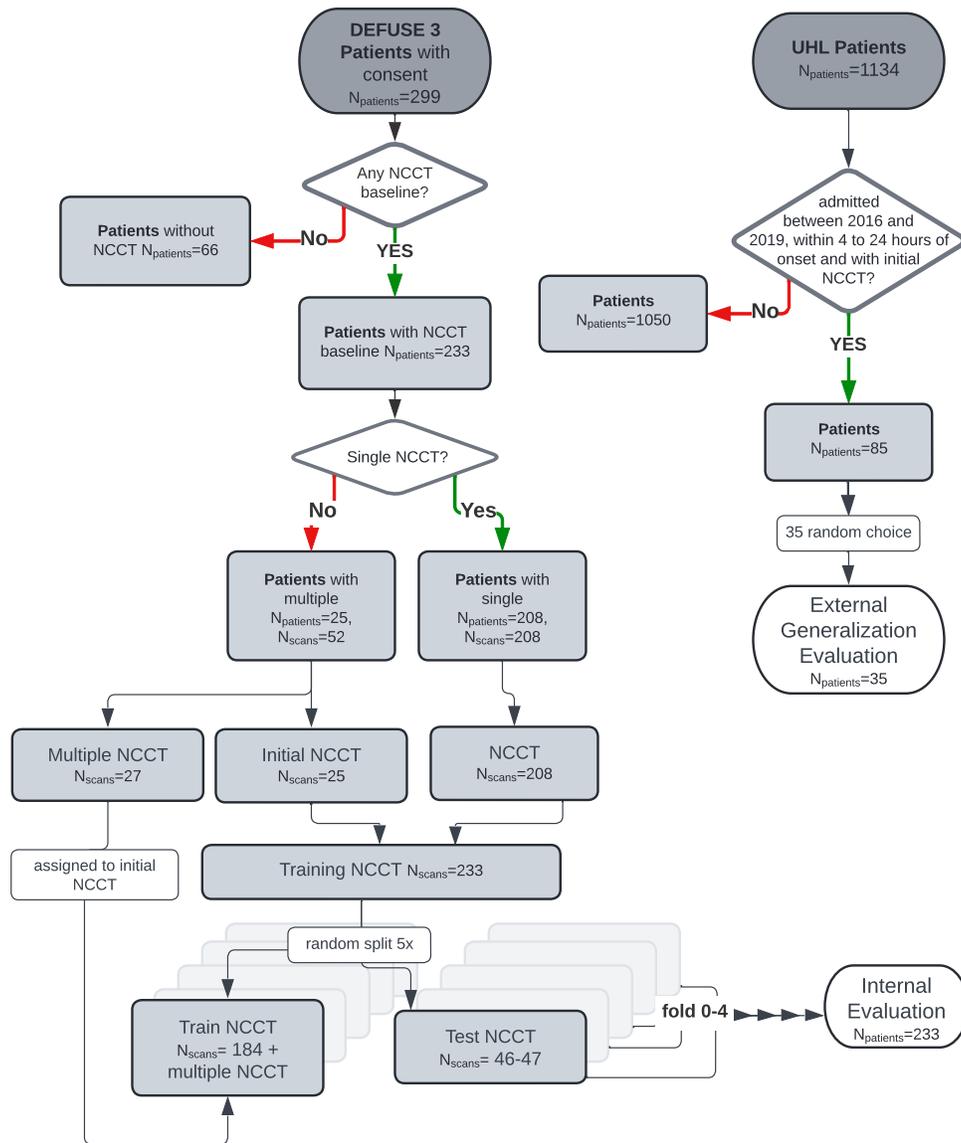[3] Computer Tomography Perfusion, Cerebral blood flow (CBF) <30%

11   Figures

Figure 1: Flowchart of the data partition. 233 patients with their initial NCCT were randomly split into 5 folds of training and test sets. 25 patients had multiple NCCT. Those were only assign to the initial NCCT when in the training set. The external generalization cohort included 35 patients.
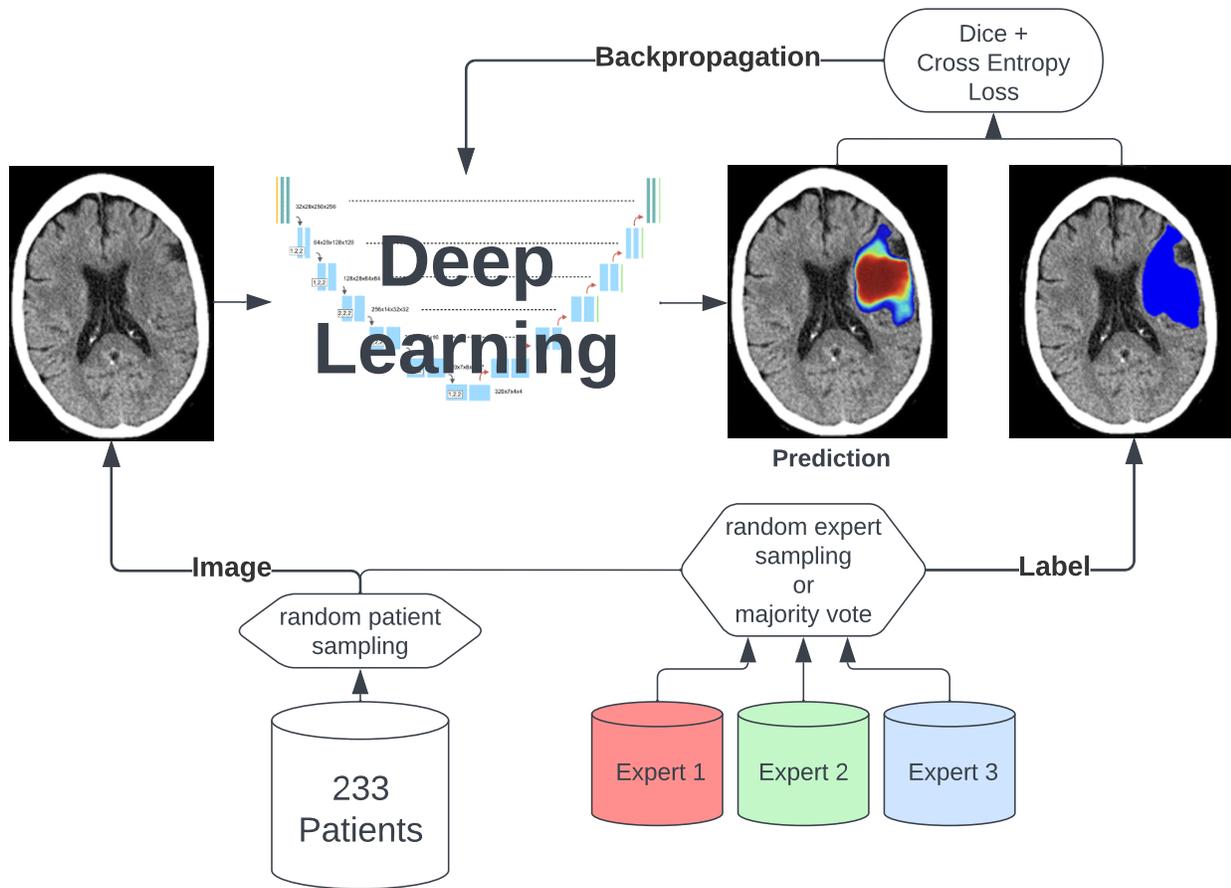
Figure 2: Training scheme pipeline with sampling strategy for random expert sampling and majority vote
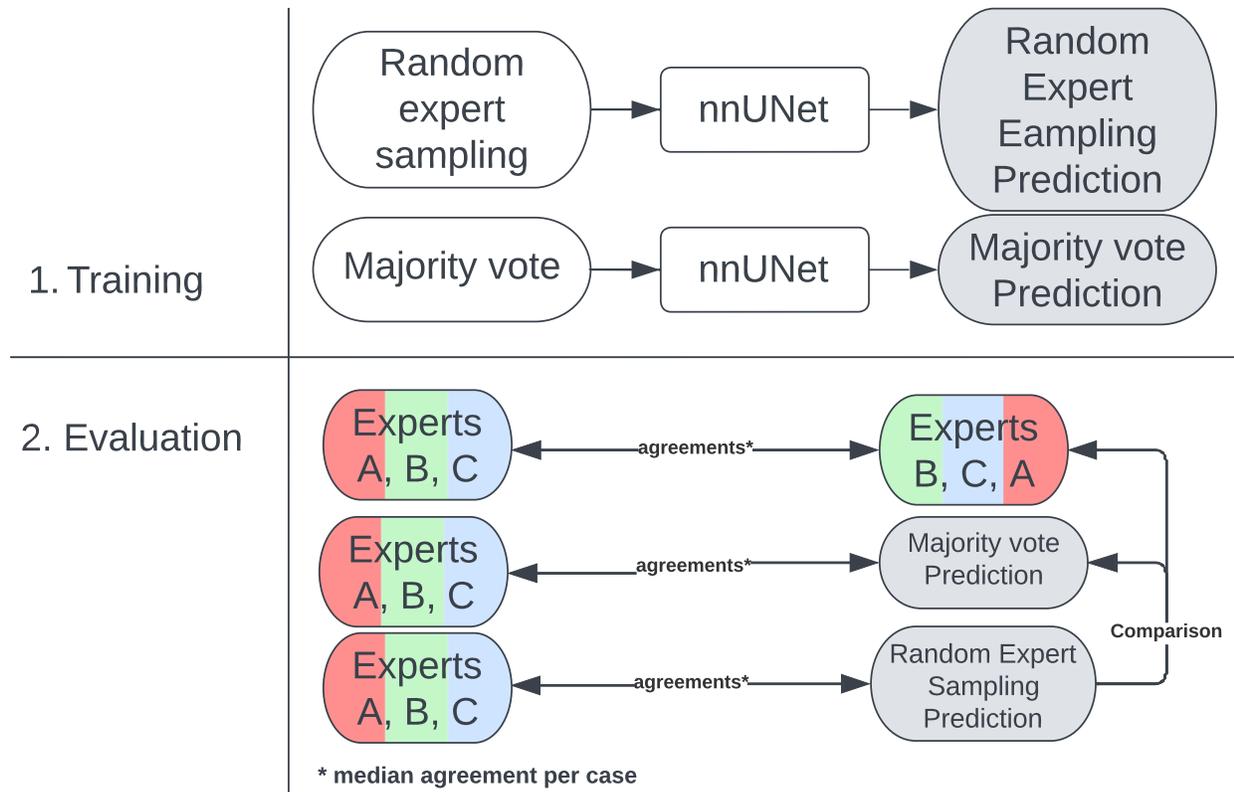
Figure 3: First two model were trained on majority vote and random expert sampling. Second, the median agreement per case for inter-expert agreement, model-expert agreement for the prediction of majority vote and random expert sampling was the basis to compare random expert sampling to the majority vote and inter-expert agreement.
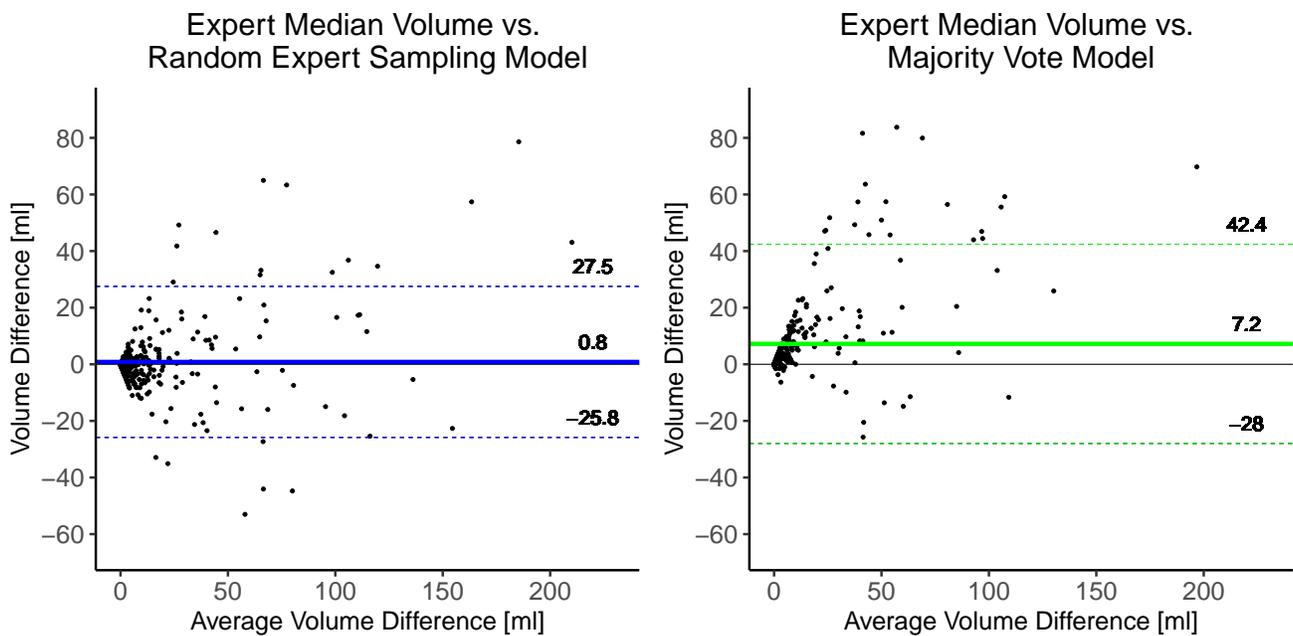

Figure 4: Bland-Altman for Random Expert Sampling (blue) and Majority Vote Model Volume (green) estimates compare to Median Expert Volume.
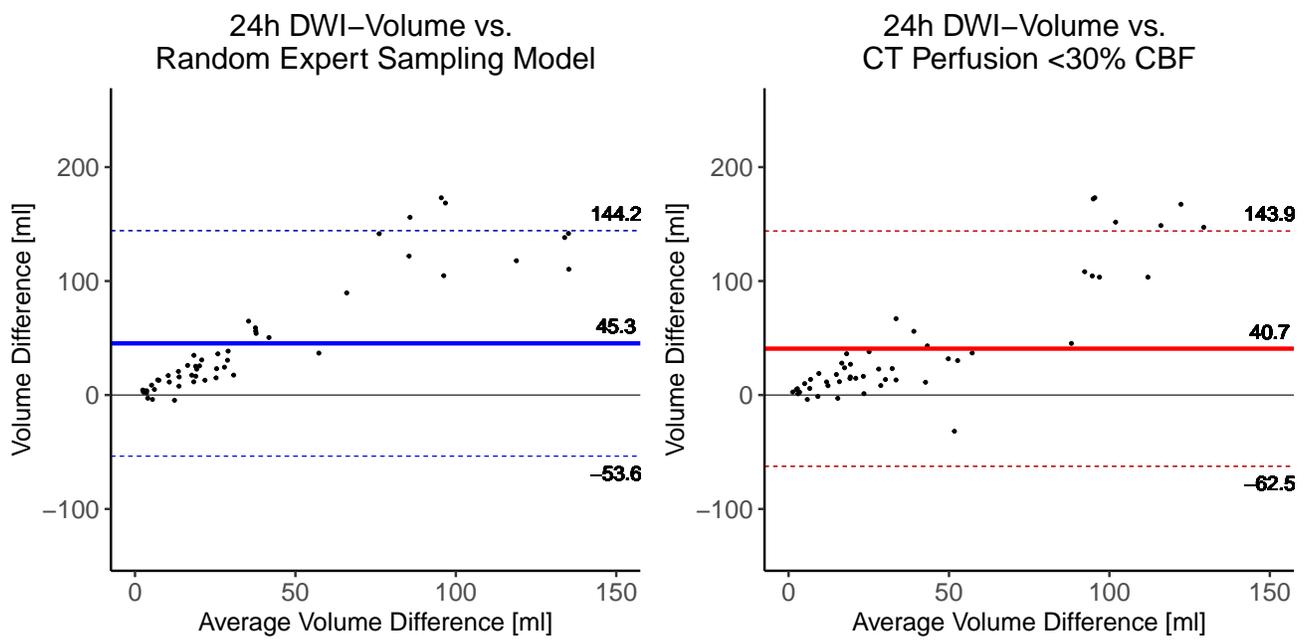
Figure 5: Bland-Altman for Random Expert Sampling Model Volume (blue) and CTP Ischemic Core Volume ¡30% (red) compared to 24h DWI-Volume for all reperfusors (TICI≥2B).
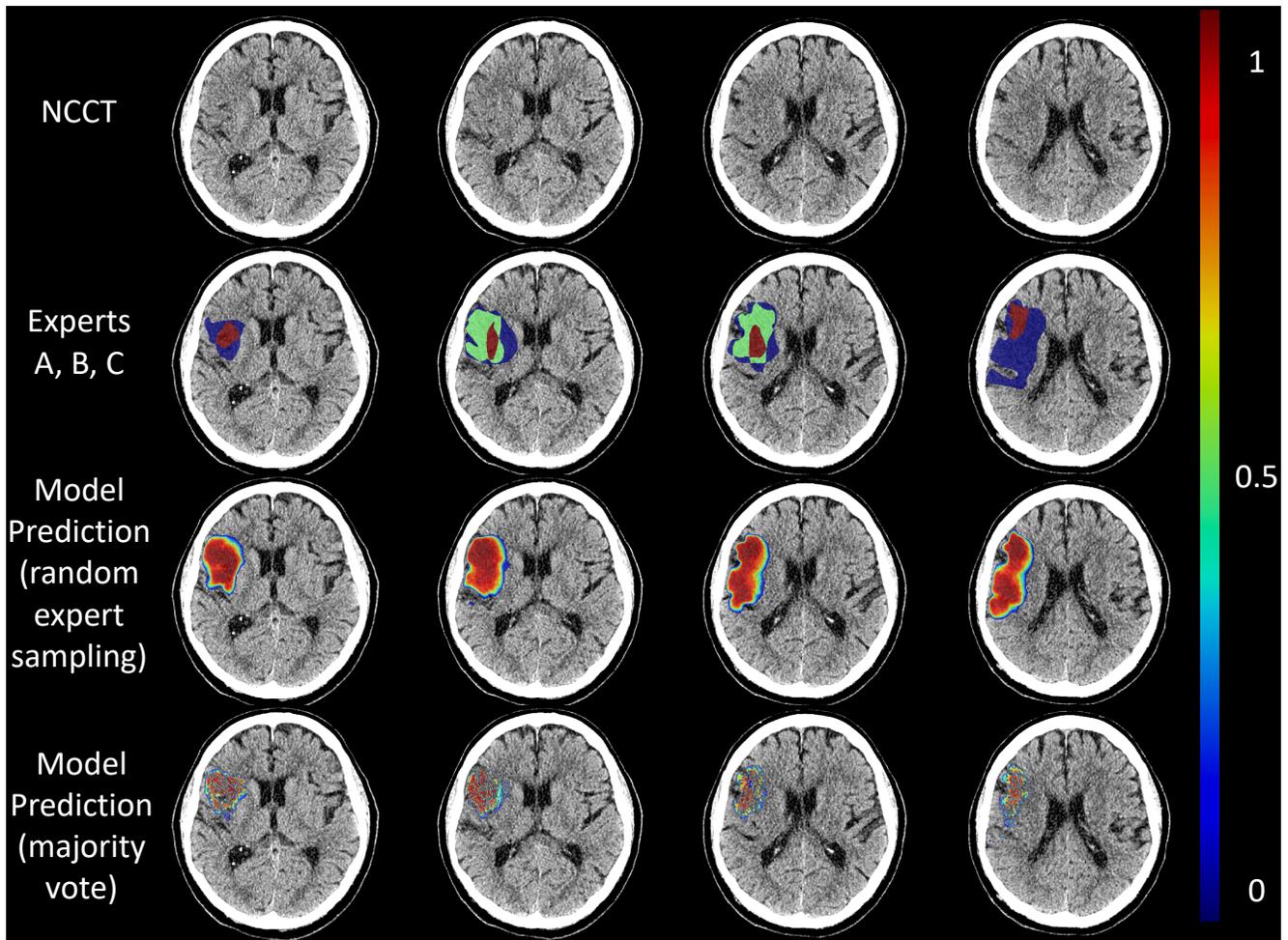
Figure 6: Example of a heatmap of expert compared to models trained on random expert sampling and majority vote. Low values are represented by blue to green colors, and high values by yellow to red colors. The prediction of a majority vote underestimates acute ischemic brain tissue shown.

12 Supplementary Material

12.1 Supplemental Methods

12.1.1 Mathematical derivation: Cross Entropy Loss as Product of Bernoulli Maximum Likelihood estimation

While using majority vote preprocessing of data results in a regression problem, optimizing the cross entropy loss corresponds to a maximum likelihood estimation by recovering the probability distribution that is most likely to have produced the data.

In fact, maximizing the cross entropy loss corresponds to maximizing the likelihood of producing the data when assuming each voxel labeling is independent.

Under that assumption, the log-likelihood of getting a labeling $\{k_i\}$, i indexing voxels can be transformed into the cross entropy loss with the following standard transformation:

$$
\begin{aligned}
\log\_\text{likelihood} &= \log\left(\prod_{i|k_i=1} p_i \prod_{i|k_i=0} (1-p_i)\right) \\
&= \log\left(\prod_{i|k_i=1} p_i \prod_{i|k_i=0} (1-p_i)\right) \\
&= \sum_{i|k_i=1} \log(p_i) + \sum_{i|k_i=0} \log(1-p_i) \\
&= \sum_{i|k_i=1} k_i \log(p_i) + \sum_{i|k_i=0} (1-k_i)\log(1-p_i) \\
&= \sum_i k_i \log(p_i) + (1-k_i)\log(1-p_i)
\end{aligned}
$$

12.2 Tabels

Table S1: **Definitions of Performance Metrics for Medical Image Segmentation**

Category	Metric	Abbreviation	Definition
Volume	**Volumetric Similarity**	VS	$1 - \dfrac{\left\|\left\|V_p^1\right\| - \left\|V_{ra}^1\right\|\right\|}{\left\|V_p^1\right\| + \left\|V_{ra}^1\right\| + \epsilon}$
	Absolute Volume Difference	AVD	$\frac{1}{m}\sum\limits_{i=1}^{m}\left\|V_{ra}^i - V_p^i\right\|$
Overlap	**Dice Similarity Coefficient**	Dice	$\frac{2 \times TP}{2 \times TP + FN + FP}$
	Recall = Sensitivity	Recall	$\frac{TP}{TP + FN}$
	Precision	Precision	$\frac{TP}{TP + FP}$
Distance	**Hausdorff Distance, q = 95th percentile**	HD 95	$\max\left(h(A,B), h(B,A)\right)$ with $h(A,B) = \max\limits_{a \in A}\min\limits_{b \in B}\|b - a\|$
	Surface Dice at Tolerance	SDT	$\frac{\|S_p \cap B_{ra}^t\| + \|S_{ra} \cap B_p^t\|}{\|S_p\| + \|S_{ra}\|}$
Image-level classification	**Correct Classification Rate**	CCR	$\frac{\text{number of correctly detected subjects}}{\text{number of all subjects}}$
	Sensitivity	Sensitivity	$\frac{TP_i}{TP_i + FN_i}$
	Specificity	Specificity	$\frac{TN_i}{TN_i + FP_i}$
	Area Under the Curve	AUC	

Table S2: **Segmentations precision**

Categories	Metric[1]	Random expert sampling variance	Inter-expert variance	Ratio[2]		p-value	Majority vote variance	Ratio[3]		p-value
Volume	VS	0.05	0.06	0.91	0.23	non-sig	0.09	0.56	0.12	<0.0001
	AVD [ml]	263.38	418.18	0.63	0.25	<0.05	487.67	0.54	0.19	<0.0001
Overlap	Dice	0.06	0.05	1.08	0.18	non-sig	0.06	0.90	0.13	non-sig
	Precision	0.07	0.07	1.12	0.19	non-sig	0.10	0.72	0.09	<0.0001
	Recall	0.08	0.07	1.14	0.17	non-sig	0.06	1.37	0.23	non-sig
Distance	HD 95 [mm]	306.87	159.76	1.92	0.86	non-sig	338.46	0.91	0.31	non-sig
	SDT 5mm	0.06	0.05	1.22	0.29	non-sig	0.07	0.86	0.16	non-sig

[1] VS = Volumetric Similarity, AVD = Absolute Volume Difference, HD 95 = Hausdorff Distance 95th percentile, SDT = Surface Dice at Tolerance
[2] Ratio of inter-expert variance over random expert sampling variance ± Standard deviation,
[3] Ratio of majority vote over variance over random expert sampling variance ± Standard deviation

12.3   Figures
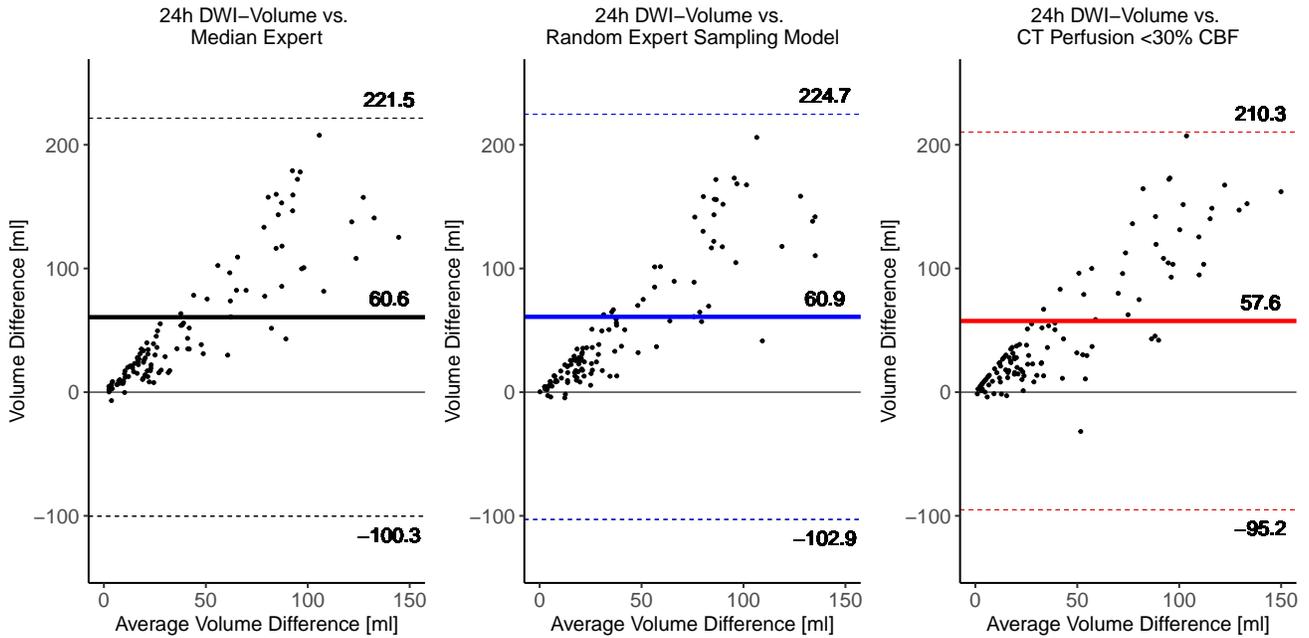
Figure S1: **Model Architecture**

Figure S2: Bland-Altman for Median Expert Volume (black), Random Expert Sampling Model Volume (blue) and CTP Ischemic Core Volume <30% (red) compared to 24h DWI-Volume for over entire patient population).