ABBREVIATED TITLE PAGE

Title: Multiclass Segmentation of Intracerebral Hemorrhage, Intraventricular Hemorrhage, and Perihematomal Edema: Public Dataset and Benchmark Model

Keywords

Intracerebral hemorrhage, Intraventricular hemorrhage, Perihematomal edema, Non-contrast CT, Deep learning, Multiclass segmentation, U-Net.

ABBREVIATIONS

CNN	Convolutional neural network
DSC	Dice Similarity Coefficient
DL	Deep learning
ICC	Intraclass Correlation Coefficient
ICH	Intracerebral hemorrhage
IVH	Intraventricular hemorrhage
mRS	Modified Rankin scale
NR1-RE	Neuroradiologist 1-repeat annotation
NR1-RS	Neuroradiologist 1-reference standard
NR2	Neuroradiologist 2
NCCT	Non-contrast computed tomography
OR	Odds ratio
PHE	Perihematomal edema
SN	Stroke neurologist

KEY POINTS

1. On external testing, the model achieved median Dice similarity coefficients of 0.90 (intracerebral hemorrhage, ICH), 0.67 (intraventricular hemorrhage, IVH), and 0.53 (perihematomal edema, PHE)
2. Performance was within inter-annotator variability for ICH and IVH, with median volume differences below 2.5 mL, with a 20-72-fold reduction in annotation time.
3. PHE showed the lowest segmentation performance among both model and human annotators, reflecting inherently ambiguous boundaries.

SUMMARY STATEMENT

Our externally tested deep learning model accurately segmented intracerebral hemorrhage, intraventricular hemorrhage, and perihematomal edema on CT scans, enabling rapid automated volumetric quantification for hemorrhagic stroke research and clinical workflows.

ABSTRACT

Background: Accurate mapping and volume assessment of spontaneous intracerebral hemorrhage (ICH) and its common complications, intraventricular hemorrhage (IVH) and perihematomal edema (PHE), on non-contrast computed tomography (NCCT) help identify patients at risk of poor outcomes or guide treatments but remain challenging. Conventional methods (e.g., ABC/2 formula) and manual segmentation are imprecise and labor-intensive. While deep learning (DL) models show promise, few simultaneously segment all three components, and no publicly available multiclass model or dataset exists.

Purpose: We developed and externally tested a DL model for multiclass segmentation of ICH, IVH, and PHE on NCCT.

Materials and Methods: The 2019 Intracranial Hemorrhage Detection Challenge dataset was repurposed for segmentation, with expert neuroradiologist annotations for ICH, IVH, and PHE serving as the reference standard on 359 selected NCCTs. We trained a DL segmentation model using the nnU-Net framework with five-fold cross-validation (n = 341) with internal testing on 18 cases. External testing was conducted on 183 independent baseline NCCTs of patients with spontaneous ICH at our institution. Performance was evaluated using the Dice similarity coefficient (DSC) and volume agreement. Intra- and inter-annotator variability on 30 exams from the test set served as a benchmark.

Results: On external (local) testing, the model achieved the highest median DSC for ICH (0.90), with lower performance for IVH (0.67) and PHE (0.53). It did not differ from inter-annotator variability for ICH and inter-specialty variability for IVH, with small median volume differences (<2.5 mL for ICH, <2.0 mL for IVH). Intraclass correlation coefficients (ICCs) for volumetric agreement were 0.98 (ICH), 0.84 (IVH), and 0.80 (PHE).

Conclusion: We present an externally tested DL model for multiclass segmentation of ICH, IVH and PHE on NCCT. We aim to publicly release our dataset and model to improve reproducibility, support clinical and research applications, and provide a benchmark for performance comparison.

INTRODUCTION

Spontaneous intracerebral hemorrhage (ICH), a devastating type of stroke, accounts for approximately 15% of strokes and carries high morbidity and mortality.[1–3] Larger ICH volumes worsen outcomes and increase the risk of hematoma expansion, an independent predictor of poor outcomes.[4,5] A 2022 meta-analysis including 30 studies found just 1 mL increase was associated with an odds ratio (OR) of 1.12 for poor functional outcomes (modified Rankin scale (mRS) 4−6 at ≥6 weeks post-ICH).[4] Additional complications, including intraventricular hemorrhage (IVH) and perihematomal edema (PHE), can further worsen outcomes.[6,7] Even small volumes of IVH may cause hydrocephalus, hemosiderin deposition and hippocampal injury.[6,8] Precise identification and quantification of these lesions would enable multiple downstream applications to support diagnosis, clinical interventions and prognostication during the acute care pathway.

Due to its widespread availability and rapid image acquisition, non-contrast computed tomography (NCCT) is the most common modality for the initial diagnosis of ICH. ICH volume is commonly estimated using the ABC/2 method, a convenient but inaccurate approach, especially with larger or irregular hematomas, where it can overestimate volume.[9,10] Estimation of IVH and PHE relies on even less precise assessments.[11] Voxel-level manual annotation remains the reference standard, but it is too time-consuming and impractical in busy clinical settings, making these important metrics unavailable for current clinical workflows.

To address these limitations, recent efforts have turned toward automated segmentation using deep learning (DL).[12–14] DL models have achieved near-human performance in segmenting ICH

(representative studies are provided in **Supplementary Table S1**). These models promise scalable, reproducible volumetric analysis and the potential to streamline both clinical and research workflows. In parallel, the field of ICH has emerged from decades of therapeutic nihilism, with a shift towards proactive interventions including intensive care bundles,[15] novel minimally invasive surgical approaches,[16] and ongoing anti-expansion trials (e.g. FASTEST Part 2).[17] Yet even recent landmark trials have relied on subjective CT interpretation, including the ABC/2 formula, to assess hematoma volume for eligibility and stratification. Automated segmentation could standardize patient selection and streamline this time-sensitive care pathway by supporting prognostication, extracting ICH regions of interest for downstream prediction models, detecting complications on longitudinal follow-up, and guiding surgical planning – all through interpretable outputs suited to a human-in-the-loop workflow.

However, early segmentation models were limited to binary ICH segmentation.[18,19] Despite the prognostic importance of IVH and PHE,[20,21] their automated segmentation has received attention only recently.[12,13,22] In addition, many studies lack external testing and/or use homogeneous populations such as clinical trial datasets. Finally, most studies do not publish annotation protocols, and many datasets and models remain inaccessible, limiting reproducibility and benchmarking.

This study aimed to address these gaps by (1) curating a publicly available multiclass segmentation dataset for ICH, IVH, and PHE on NCCT focused on the downstream assessment of spontaneous ICH, and (2) developing and evaluating a DL model capable of segmenting all three lesion types.

MATERIALS AND METHODS

Study design

This retrospective study was conducted with data obtained under approval from our institutional research ethics board. Informed consent from patients was waived due to the retrospective nature of the analysis and the use of de-identified data. We obtained written consent from all annotators. All data processing, including model training and inference, was performed on secure local computers within the institution's research infrastructure.

Dataset

Training and internal testing (public): We selected relevant examples of intracerebral hemorrhage from the 2019 Intracranial Hemorrhage Detection Challenge dataset, an artificial intelligence model development challenge organized by the RSNA to detect and classify intracranial hemorrhage on NCCT scans.[23] The exam selection process is described in the **Supplementary Material**.

External testing (local): Data from our institution were retrieved from our PACS (March 2016–May 2023) through a retrospective spontaneous ICH study in patients aged 18–100, including only one baseline exam per patient. Patient selection and data processing for all datasets are further detailed in the **Supplementary Method**.

Annotations

In the first phase, trainees generated preliminary segmentations de novo, which were then reviewed and refined by a neuroradiologist (L.L.G., 12 years of experience) (*n*=157). In the second phase, we trained an nnU-Net model on these initial annotations to generate preliminary segmentations, with a similar review and refinement process by trainees followed by the same neuroradiologist (*n*=202), mirroring the methodology used in other studies.[24,25] The external test dataset was constructed separately using a method similar to that of the first phase detailed above. Further details are provided in the **Supplementary Method**. These reference standard annotations were used to train and evaluate the model (Neuroradiologist 1-reference standard (NR1-RS)).

Deep learning network and training

The data flow for model training and evaluation is shown in **Figure 1**. Based on preliminary evaluation of different architectures and prior performance for the same task,[12,13,22] we opted for the nnU-Net framework (v2.5.1) for final model development.[26] Implementation details are available in the **Supplementary Method**.

Intra- and inter-annotator reliability study

Since no similar model is currently publicly available, we compared the model's performance in terms of intra- and inter-annotator reliability to that of expert annotators, using the same reference annotations (NR1-RS) for all comparisons. To this end, a random subset of 30 NCCTs from the

external dataset, all exhibiting ICH, IVH, and PHE, was constructed. This sample size was chosen to achieve sufficient statistical power while considering the feasibility of a multi-expert study. First, the same neuroradiologist who performed the reference standard annotations repeated the annotations after a 6-month washout period, this time using a de novo annotation workflow (Neuroradiologist 1-repeat annotation (NR1-RE)). Second, we assessed inter-annotator agreement using independent de novo segmentations performed by a second neuroradiologist (NR2, 3 years of experience) and a stroke neurologist (SN, 10 years of experience) (details in the **Supplementary Method**).

Performance evaluation and statistical analysis

Segmentation performance was primarily measured using the Dice Similarity Coefficient (DSC) (see **Supplemental Method** for implementation).[27] On the external test dataset, automated and reference volume agreement was assessed using the single fixed raters' ICC and Bland–Altman analysis.[28] Finally, a qualitative evaluation of the segmentation masks was conducted.

Subanalyses stratified by demographic characteristics (age and sex) and anticoagulation status were performed. We also evaluated radiological features that could impact segmentation performance as assessed by DSC. We first assessed how lesion volumes affect DSC. Second, the impact of IVH presence on ICH segmentation was assessed between cases with and without IVH. Third, we studied the impact of Barras density and shape expansion markers, radiological signs indicating content heterogeneity and/or margin irregularity, as potential indicators of segmentation task difficulty.[29] These markers were assessed using a consensus, as previously reported.[30] Fourth,

we compared the segmentation performance for ICH and PHE according to the anatomical location of ICH. Since IVH may be multifocal or extend distally from the primary ICH site via CSF pathways, this label was not evaluated by ICH anatomical location. Finally, we evaluated the impact of removing trace volume (<1 mL) specifically for IVH.[13]

Shapiro-Wilk tests showed non-normal distribution for DSCs; thus, our analyses used non-parametric tests. The median (interquartile range [IQR]) and mean ± standard deviation (SD) were used as summary statistics. Statistical significance was set at $p < 0.05$, with correction applied for multiple comparisons within each group. Statistical tests are specified where relevant in the results section. Ninety-five percent (95%) confidence intervals (CIs) were calculated through bias-corrected and accelerated bootstrap resampling (1,000 iterations). Analyses were conducted using Python and R, as detailed in the **Supplementary Method**.

Data Availability

Selected NIfTI volumes from the 2019 RSNA Intracranial Hemorrhage dataset, along with corresponding segmentation masks, will be made publicly available through the RSNA Medical Imaging and Resource Archive (MIRA), pending RSNA approval and in accordance with the original challenge data-sharing restrictions. The trained model is publicly available and can be accessed at: https://avnir-models-documentation.readthedocs.io/models/iips.html. Access to the external test dataset from our institution can be requested from the corresponding author. The request will be reviewed by the institutional data security board, and access may be granted depending on their decision.

RESULTS

Patient Characteristics

Of the 359 annotated RSNA NCCTs, 341 were used for model training with five-fold cross-validation. Given the availability of a large independent external test set, we prioritized maximizing training data and reserved only 18 randomly held-out cases for internal testing using the same model ensembling approach as the external evaluation. One hundred eighty-three (183) NCCTs from our institution were used for external testing. **Table 1** summarizes the characteristics of the training and internal test datasets, while **Table 2** summarizes the external test set.

Performance analysis on the internal test set

On the internal test set, median DSC for ICH was 0.90 (IQR: [0.84–0.94]), higher than that of IVH (0.83 [0.47–0.87]) and PHE (0.71 [0.55–0.80]). Illustrative examples are shown in **Figure 2**. Five-fold cross-validation performance across all 341 training subjects is detailed in **Table 3**.

Performance analysis on the external test set

On the external test set, the median DSC for ICH was 0.90 (IQR: [0.86–0.92]), higher than that of IVH (0.67 [0.20–0.79]) and PHE (0.53 [0.38–0.67]). Filtering for IVH >1 mL (n=70) increased the median DSC to 0.75 [0.65–0.81]. When compared to the internal test set (n=18), no significant differences were observed for ICH (p = .42) or IVH (p = .11) after Bonferroni correction, though

PHE showed a significant decline (p = .003). However, cross-validated internal estimates (n=341, single-model inference) showed significantly higher performance than the external test set (model ensembling) across all three labels (all p < .001, **Table 3**). **Supplementary Figure S1** shows DSC boxplots according to volume quartiles. **Figure 3** presents representative examples.

Examples from our qualitative review, including failure analysis and improved segmentation performance compared with reference standards, are illustrated in **Supplementary Figures S2 and S3**, respectively.

Intra- and inter-annotator reliability study

Results of the inter-annotator study are summarized in **Figure 4, Supplementary Tables S2 and S3**). Illustrative examples are shown in **Figure 5**. Median total annotation time (all labels combined) ranged between 7.8 to 25.0 minutes depending on the annotator (**Supplementary Figure S4**). Mean inference time per prediction was 20.7 seconds for the nnU-Net model.

For ICH segmentation, the model did not differ significantly from inter-annotator comparisons with NR2 and SN. However, the median DSC was significantly lower (p < .001) than the intra-annotator agreement (NR1-RE), but the difference was relatively small (0.90 and 0.92). Among human annotators, a statistically significant (p = .001) DSC difference was observed only between SN and NR1-RE (0.91 and 0.92). For volumetric measurements, the absolute median volume difference between the model and human annotators remained below 2.5 mL, comparable in magnitude to differences observed between pairs of human annotators.

For IVH segmentation, the model achieved a median DSC of 0.70, which was significantly lower than both neuroradiologists (NR1-RE, 0.80; NR2, 0.79) but not the stroke neurologist (SN, 0.78). A significant DSC difference was also observed between the neuroradiologists (NR1-RE and NR2) and SN. For volumetric measurements, the model produced slightly smaller volumes than human annotators, although differences remained modest (<2 mL).

For PHE segmentation, no significant DSC differences were observed in any pairwise comparisons, including between the model and all human annotators. However, the model consistently produced larger volumes than human annotators, with median differences ranging from 6.75 to 14.05 mL.

Agreement for volumetric calculation on the external test set

Automated volume measurements showed excellent agreement with manual measurements for ICH (ICC = 0.98; 95% CI: 0.97–0.98), and good agreement for IVH (0.84; 0.80–0.88) and PHE (0.80; 0.74–0.85), all statistically significant ($p < .001$). Bland–Altman plots are displayed in **Figure 6**.

Segmentation performance in subgroups on the external test set

Subgroup analyses based on clinical and imaging variables are summarized in **Table 4** and **Supplementary Table S4**. Sex and age showed no consistent influence on volume or segmentation performance. By anatomical location, median DSC was lower for infratentorial than

supratentorial ICH (values in **Supplementary Table S4**). Median ICH volume was also smaller in infratentorial ICH (2.43 mL) compared to supratentorial deep (13.11 mL) and lobar (39.91 mL) ICH. For PHE, median DSC was highest in supratentorial lobar (0.57), followed by deep (0.52), and lowest in infratentorial ICH (0.08). Median DSC was comparable in patients with and without anticoagulation for ICH (0.90 and 0.89) and IVH (0.69 and 0.66) (similar volumes). ICH segmentation performance was comparable between patients with and without IVH (median DSC 0.89 and 0.90), despite the larger median ICH volumes observed in examples with IVH (20.95 and 12.97 mL, $p = .04$).

Additional subgroup analysis for Barras expansion markers is shown in **Supplementary Figure S5**. The presence of expansion markers was associated with larger volumes and higher DSC for ICH and PHE. However, after stratifying for volume, no differences in DSC were observed.

DISCUSSION

In this study, we curated a public dataset and developed an nnU-Net–based DL model for multiclass segmentation of ICH, IVH, and PHE on NCCT. The model showed the strongest performance in ICH segmentation, with lower DSCs for IVH and PHE, a finding consistent with previous studies.[12,13,22]

ICH has high contrast compared to the background brain parenchyma and PHE, an attribute that may explain its relative robustness to dataset shift and high agreement with human annotators.

While IVH also has high density, ambiguity can arise at the ICH-IVH boundary and the choroid plexus, as well as from frequent low-volume (**Supplementary Figure S1)**.[13] PHE demonstrated the greatest performance degradation on external testing, but even experts had difficulty agreeing on its extent and margins, likely because PHE may exhibit minimal contrast relative to the surrounding parenchyma, particularly in regions with significant microangiopathic white matter changes. This is an important topic to address, as emerging literature now emphasizes quantifying PHE to aid outcome prediction and, in the future, to guide neuroprotective interventions.[7,20,31]

Overall, our findings indicate that task-dependent revisions of automated masks may still be warranted, depending on the specific downstream use. Interestingly, our qualitative review revealed that in some cases the model's inferences were better than the reference-standard manual annotations (**Supplementary Figure S3**), yet in other instances errors occurred **(Supplementary Figure S2)**. Nevertheless, automated inference substantially accelerates the annotation workflow with an estimated 20-72-fold reduction in annotation time.

Our study prioritized maximizing the use of training data to optimize model performance, reserving only 18 cases for internal testing. While this limited the statistical power of direct comparisons between internal and external performance, cross-validation estimates across all 341 training subjects showed significant performance degradation across all lesion types, despite this comparison being conservative, as the internal nn-UNet cross-validation predictions used a single model while external predictions utilized a five-model ensemble. The public release of both the annotated dataset and model weights will enable the research community to further refine

segmentation labels, expand the dataset, and/or fine-tune the model on target-domain data to improve generalizability.

Our ICH and IVH segmentation performances are comparable to a recent multi-institutional study by Nawabi et al.[13], with our model additionally showing near-zero ICH bias and smaller IVH bias on Bland-Altman analysis compared to the systematic underestimation they reported on external validation. Conversely, PHE segmentation performances were lower than those reported previously.[12,13,18] When compared specifically to Nawabi et al,[13] PHE remained the most challenging label for both models, demonstrating wider limits of agreement and proportional bias at larger volumes. Nonetheless, comparisons with previous studies should be interpreted with caution. While our training dataset was smaller than those used in recent multi-institutional efforts, some prior studies used homogeneous clinical trial datasets and lacked external testing, which may artificially inflate performance estimates.

We examined various factors that could influence segmentation performance. First, the model showed similar performance across age, sex, and anticoagulation subgroups on the external test set, an important finding given the lack of exam-level demographic and clinical data in the RSNA training set. Second, the volume-dependence of DSC reflects a known property of the metric.[27] Third, the presence of IVH did not affect ICH segmentation accuracy, consistent with Zhao et al.'s findings.[22] However, boundary ambiguity at the ICH–IVH interface likely attenuates DSC gains otherwise expected with larger volumes. Fourth, although expansion markers might indicate greater difficulty in segmentation, we did not find any effect on performance. Fifth, infratentorial ICH underperformance likely reflects training data underrepresentation compounded by smaller

lesion sizes and posterior fossa beam hardening artifacts. Finally, despite a higher proportion of deep ICH in the external test set relative to the training data, supratentorial performance remained comparable across locations.

Beyond technical performance, automated segmentation of ICH, IVH, and PHE could potentially impact ICH management in several ways. In the hyperacute setting, such a model enables standardized CT analysis at presentation, supporting outcome prediction, including identification of patients most likely to benefit from anti-expansion therapy, either by calculating lesional volumes or as a preprocessing step for radiomics or DL models. Moreover, as dynamic complications emerge over the first hours to days, serial segmentation-derived imaging biomarkers can be integrated with clinical predictors to continuously refine outcome prediction — with the same pipeline enabling automated detection of interval worsening across all three classes. For the ICH component specifically, additional clinical applications include refined patient selection for treatment (e.g., surgical trials relying on fixed ICH volume thresholds),[32] guidance of surgical trajectories in minimally invasive procedures,[33] and improved prognostication, as the impact of ICH volume on outcomes varies by anatomical location.[34] Reliable IVH quantification could further support functional prognostication when combined with ICH[6] and help predict the need for external ventricular drainage.[35] Monitoring PHE progression over the initial days can further support prognostication[7] and serve as a tool in future trials aimed at reducing secondary brain injury.[15]

Several limitations should be noted. The reference annotations used to train and evaluate the model are inherently biased toward a single annotator, as obtaining consensus from multiple annotators

across the entire dataset was deemed prohibitive, a factor that also explains the limited sample size in the inter-annotator study. The test set was retrospectively collected from a single center, which was biased toward one CT vendor. Model performance in non-acute settings, other intracranial hemorrhage etiologies, or the pediatric population remains to be established.

In conclusion, we developed and externally tested a DL model for multiclass segmentation of ICH, IVH, and PHE on NCCT that can support downstream clinical applications. The model demonstrated relative robustness for ICH, intermediate performance for IVH, and greater susceptibility to dataset shift for PHE. Future investigations should focus on improving the segmentation performance on unseen data and assessing integration into real-world practice settings, including whether automated quantification improves risk stratification, reduces CT interpretation variability, or meaningfully influences management strategies in patients with ICH.

REFERENCES

1. An SJ, Kim TJ, Yoon B-W. Epidemiology, Risk Factors, and Clinical Features of Intracerebral Hemorrhage: An Update. J Stroke 2017;19(1):3–10.

2. Feigin VL, Abate MD, Abate YH, et al. Global, regional, and national burden of stroke and its risk factors, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021. Lancet Neurol 2024;23(10):973–1003.

3. Sheth KN. Spontaneous Intracerebral Hemorrhage. N Engl J Med 2022;387(17):1589–96.

4. Hammerbeck U, Abdulle A, Heal C, Parry-Jones AR. Hyperacute prediction of functional outcome in spontaneous intracerebral haemorrhage: systematic review and meta-analysis. Eur Stroke J 2022;7(1):6–14.

5. Al-Shahi Salman R, Frantzias J, Lee RJ, et al. Absolute risk and predictors of the growth of acute spontaneous intracerebral haemorrhage: a systematic review and meta-analysis of individual patient data. Lancet Neurol 2018;17(10):885–94.

6. Yogendrakumar V, Ramsay T, Fergusson D, et al. New and expanding ventricular hemorrhage predicts poor outcome in acute intracerebral hemorrhage. Neurology 2019;93(9):e879–88.

7. Samarasekera N, Tuck S, Wang X, et al. Perihematomal Edema and Functional Outcome After Intracerebral Hemorrhage: A Meta-Analysis of Individual Participant Data. Stroke 2026;STROKEAHA.125.053991.

8. Chen Q, Tang J, Tan L, et al. Intracerebral Hematoma Contributes to Hydrocephalus After Intraventricular Hemorrhage via Aggravating Iron Accumulation. Stroke 2015;46(10):2902–8.

9. Huttner HB, Steiner T, Hartmann M, et al. Comparison of ABC/2 Estimation Technique to Computer-Assisted Planimetric Analysis in Warfarin-Related Intracerebral Parenchymal Hemorrhage. Stroke 2006;37(2):404–8.

10. Webb AJS, Ullman NL, Morgan TC, et al. Accuracy of the ABC/2 Score for Intracerebral Hemorrhage. Stroke 2015;46(9):2470–6.

11. Graeb DA, Robertson WD, Lapointe JS, Nugent RA, Harrison PB. Computed tomographic diagnosis of intraventricular hemorrhage. Etiology and prognosis. Radiology 1982;143(1):91–6.

12. Kok YE, Pszczolkowski S, Law ZK, et al. Semantic Segmentation of Spontaneous Intracerebral Hemorrhage, Intraventricular Hemorrhage, and Associated Edema on CT Images Using Deep Learning. Radiol Artif Intell 2022;4(6):e220096.

13. Nawabi J, Baumgaertner GL, Schulze-Weddige S, et al. Cross-institutional automated multilabel segmentation for acute intracerebral hemorrhage, intraventricular hemorrhage,

and perihematomal edema on CT. Radiol Adv [Internet] 2025 [cited 2025 July 12];2(2). Available from: https://academic.oup.com/radadv/article/doi/10.1093/radadv/umaf012/8090148

14. Weissflog JS, Keller EJ, Neymeyer ML, Morotti A, Dowlatshahi D, Nawabi J. Systematic review of commercial artificial intelligence tools for the detection and volume quantification in intracerebral hemorrhage. Eur Radiol [Internet] 2025;Available from: https://doi.org/10.1007/s00330-025-11834-4

15. Hsieh P-F, Liu Y-B, Cordonnier C, Tsai H-H. Therapeutic strategies in intracerebral hemorrhage: from hematoma stabilization to alleviating secondary brain injury. Res Pract Thromb Haemost 2026;10(3):103423.

16. Alkhiri A, Alamri AF, Almaghrabi AA, et al. Minimally Invasive Surgery for Spontaneous Intracerebral Hemorrhage: Meta-Analysis of High-Quality Randomized Clinical Trials. Ann Neurol 2025;97(1):185–94.

17. Broderick JP, Naidech AM, Elm JJ, et al. Recombinant factor VIIa versus placebo for spontaneous intracerebral haemorrhage within 2 h of symptom onset (FASTEST): a multicentre, double-blind, randomised, placebo-controlled, phase 3 trial.

18. Ironside N, Chen C-J, Mutasa S, et al. Fully Automated Segmentation Algorithm for Hematoma Volumetric Analysis in Spontaneous Intracerebral Hemorrhage. Stroke 2019;50(12):3416–23.

19. Sharrock MF, Mould WA, Ali H, et al. 3D Deep Neural Network Segmentation of Intracerebral Hemorrhage: Development and Validation for Clinical Trials. Neuroinformatics 2021;19(3):403–15.

20. Selim M, Norton C. Perihematomal edema: Implications for intracerebral hemorrhage research and therapeutic advances. J Neurosci Res 2020;98(1):212–8.

21. Yogendrakumar V, Ramsay T, Fergusson DA, et al. Redefining Hematoma Expansion With the Inclusion of Intraventricular H emorrhage Growth. Stroke 2020;51(4):1120–7.

22. Zhao X, Chen K, Wu G, et al. Deep learning shows good reliability for automatic segmentation and volume measurement of brain hemorrhage, intraventricular extension, and peripheral edema. Eur Radiol 2021;31(7):5012–20.

23. Flanders AE, Prevedello LM, Shih G, et al. Construction of a Machine Learning Dataset through Collaboration: The RSNA 2019 Brain CT Hemorrhage Challenge. Radiol Artif Intell 2020;2(3):e190211.

24. Monteiro M, Newcombe VFJ, Mathieu F, et al. Multiclass semantic segmentation and quantification of traumatic brain injury lesions on head CT using deep learning: an algorithm development and multicentre validation study. Lancet Digit Health 2020;2(6):e314–22.

25. Wasserthal J, Breit H-C, Meyer MT, et al. TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. Radiol Artif Intell 2023;5(5):e230024.

26. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021;18(2):203–11.

27. Reinke A, Tizabi MD, Baumgartner M, et al. Understanding metric-related pitfalls in image analysis validation. Nat Methods 2024;21(2):182–94.

28. Martin Bland J, Altman DouglasG. STATISTICAL METHODS FOR ASSESSING AGREEMENT BETWEEN TWO METHODS OF CLINICAL MEASUREMENT. The Lancet 1986;327(8476):307–10.

29. Barras CD, Tress BM, Christensen S, et al. Density and Shape as CT Predictors of Intracerebral Hemorrhage Growth. Stroke 2009;40(4):1325–31.

30. Ducroux C, Nehme A, Rioux B, et al. NCCT Markers of Intracerebral Hemorrhage Expansion Using Revised Crite ria: An External Validation of Their Predictive Accuracy. Am J Neuroradiol 2023;44(6):658–64.

31. Shi N, Zhang Y, Ma Y, Zhao H. Small Intracerebral Hemorrhage: Baseline CT Edema-to-Hematoma Ratio as an Independent Predictor of Early Expansion. Am J Neuroradiol 2026;ajnr.A9261.

32. De Oliveira Manoel AL, Msheik A, Schuind S, Al Rumaihi GIK, Taccone F. Surgery for intracerebral hemorrhage: new evidence and changing perspectives. Crit Care 2025;30(1):7.

33. Gan Z, Xu X, Li F, Kikinis R, Zhang J, Chen X. Deep learning–based clinical decision support system for intracerebral hemorrhage: an imaging-based AI-driven framework for automated hematoma segmentation and trajectory planning. Neurosurg Focus 2025;59(1):E5.

34. Morotti A, Li Q, Nawabi J, et al. Volume Tolerance and Prognostic Impact of Hematoma Expansion in Deep and Lobar Intracerebral Hemorrhage. Stroke 2025;56(5):1224–31.

35. Chen A, R X, E Z, J C, R Z, J L. Development and validation of a nomogram for predicting early acute hydrocephalus after spontaneous intracerebral hemorrhage: a single-center retrospective study. Sci Rep [Internet] 2024 [cited 2025 Mar 1];14(1). Available from: https://pubmed.ncbi.nlm.nih.gov/39548258/

TABLES

Table 1: Imaging characteristics in the training and internal testing set	
Characteristics	**Training and internal testing (*n*=359)**
Slice thickness [mm], median (range)	5.000 (2.995–6.675)
Simple	166 (46.2)
Complex	193 (53.8)
Motion artifacts	93 (25.9)
Metal artifacts	83 (23.1)
Ventriculostomy drain	90 (25.1)
Surgical changes (other than ventriculostomy drain)	36 (10.0)
Trauma	17 (4.7)
Infarct	8 (2.2)
Spontaneous	292 (81.3)
Location*	
Supratentorial	455 (90.6)
Lobar	312 (62.2)
Frontal	120 (23.9)
Parietal	70 (13.9)
Occipital	43 (8.6)
Temporal	79 (15.7)
Deep	143 (28.5)
Basal ganglia	66 (13.1)
External capsule	2 (0.4)
Internal capsule	8 (1.6)
Thalamus	67 (13.3)
Infratentorial	47 (9.4)
Cerebellum	34 (6.8)
Lobar	17 (3.4)
Deep	17(3.4)
Pons	2 (0.4)
Other brainstem	11 (2.2)

Unless otherwise stated, data are number of cases with percentages in parentheses. Characteristics are based on a neuroradiologist interpretation without additional clinical context as this was not provided in the RSNA dataset.
*The location does not add up to the total number of cases because one case could have hemorrhages in multiple locations.

Table 2: Patient and imaging characteristics in the external testing set	
Characteristics	**Test (*n* = 183)**
Age [years], median [IQR] (range)	77 [67–93] (39–99)
Sex	
Male	111 (60.7)
Female	72 (39.3)
CT manufacturer	
Siemens	173 (94.5)
Philips	6 (3.3)
GE HealthCare	3 (1.6)
Toshiba	1 (0.5)
Slice thickness [mm], median (range)	4.000 (3.000–5.000)
Location	
Supratentorial	162 (88.5)
Lobar	67 (36.6)
Frontal	36 (19.7)
Parietal	9 (4.9)
Occipital	13 (7.1)
Temporal	9 (4.9)
Deep	95 (51.9)
Basal ganglia	36 (19.7)
External capsule	9 (4.9)
Internal capsule	8 (4.4)
Thalamus	42 (23.0)
Infratentorial	21 (11.5)
Cerebellum	12 (6.6)
Deep	9 (4.9)
Lobar	3 (1.6)
Pons	8 (4.4)
Midbrain	1 (0.5)
ICH	
Count	183 (100.0)
Volume [ml], median [IQR]	16.6 [6.3–40.6]
IVH	
Count	88 (48.1)
Volume [ml], median [IQR]	7.2 [1.8–22.2]
PHE	
Count	172 (94.0)
Volume [ml], median [IQR]	8.0 [3.5–17.3]
Anticoagulation status	
Yes	78 (42.6)
No	98 (53.6)
Unknown	7 (3.8)
Expansion markers	
Barras density	93 (50.8)
Barras margins	103 (56.3)

ICH = intracerebral hemorrhage, IVH = intraventricular hemorrhage, PHE = perihematomal edema. Unless otherwise stated, data are number of cases with percentages in parentheses.

Table 3: nnU-Net segmentation performance by label. Comparison of internal five-fold cross validation (CV, *n*=341) and internal held-out test set (*n*=18) against external test dataset (*n*=183)

Label	DSC (95% CI)			*p*-value	
	Internal 5-fold CV	Internal test	External test	5-fold CV and external test	Internal and external test
ICH	0.93 (0.92–0.94)	0.90 (0.86–0.94)	0.90 (0.89–0.91)	<.001*	.42
IVH	0.79 (0.75–0.82)	0.83 (0.31–0.88)	0.67 (0.56–0.72)	<.001*	.11
PHE	0.75 (0.72–0.77)	0.71 (0.55–0.80)	0.53 (0.48–0.56)	<.001*	.003*

DSC = Dice similarity coefficient, ICH = intracerebral hemorrhage, IVH = intraventricular hemorrhage, PHE = perihematomal edema. All metrics correspond to the median DSC. Comparison performed using the Mann-Whitney U test with Bonferroni correction for multiple comparisons. After Bonferroni correction applied separately within each comparison pair (per-test $\alpha^* = 0.05/3 = 0.017$ per comparison set), comparisons that reached significance ($p < \alpha^*$) are marked with an asterisk (*).

Table 4: Stratified nnU-Net segmentation performance based on demographic characteristics (test dataset)						
	ICH		**IVH**		**PHE**	
	Median DSC [IQR]	**Mean DSC (± SD)**	**Median DSC [IQR]**	**Mean DSC (± SD)**	**Median DSC [IQR]**	**Mean DSC (± SD)**
Sex						
Female (*n* = 74)	0.90 [0.86–0.92]	0.86 (± 0.12)	0.63 [0.02–0.78]	0.48 (± 0.33)	0.51 [0.36–0.66]	0.50 (± 0.21)
Male (*n* = 112)	0.90 [0.85–0.92]	0.86 (± 0.15)	0.70 [0.35–0.79]	0.55 (± 0.32)	0.53 [0.39–0.67]	0.50 (± 0.23)
Difference in Median DSC (95% CI, *p*-value)	0.00 (-0.02–0.02, *p*=.75)		0.07 (-0.07–0.35, *p*=.25)		-0.03 (-0.09–0.05, *p*=.72)	
Age						
<50 (*n* = 15)	0.92 [0.86–0.93]	0.88 (± 0.11)	0.70 [0.00–0.79]	0.47 (± 0.38)	0.47 [0.31–0.61]	0.42 (± 0.25)
50–70 (*n* = 41)	0.89 [0.86–0.92]	0.89 (± 0.04)	0.66 [0.49–0.79]	0.56 (± 0.32)	0.48 [0.38–0.62]	0.49 (± 0.19)
>70 (*n* = 130)	0.90 [0.86–0.92]	0.85 (± 0.16)	0.67 [0.27–0.78]	0.52 (± 0.33)	0.54 [0.38–0.68]	0.52 (± 0.23)
Overall Difference in Median DSC (Kruskal-Wallis *p*-value)	*p*=.40		*p*=.82		*p*=.27	
Difference in Median DSC (95% CI, *p*-value pairwise comparison)	<50 and 50-70: -0.03 (-0.06–0.03, *p*=.20)		<50 and 50-70: -0.04 (-0.26–0.74, *p*=.52)		<50 and 50-70: 0.01 (-0.15–0.19, *p*=.59)	
	<50 and >70: -0.02 (-0.04–0.04, *p*=.19)		<50 and >70: -0.03 (-0.16–0.70, *p*=.65)		<50 and >70: 0.07 (-0.07–0.24, *p*=0.19)	
	50-70 and >70: 0.01 (-0.01–0.04, *p*=.87)		50-70 and >70: 0.00 (-0.19–0.16, *p*=.72)		50-70 and >70: 0.06 (-0.01–0.15, *p*=.26)	

DSC = Dice similarity coefficient, ICH = intracerebral hemorrhage, IVH = intraventricular hemorrhage, PHE = perihematomal edema. *n* is the number of cases with ICH. *P*-values were calculated using the Mann–Whitney U test for two-group comparisons and the Kruskal–Wallis test for comparisons involving three or more groups with pairwise comparisons performed using the Dunn test.

FIGURES

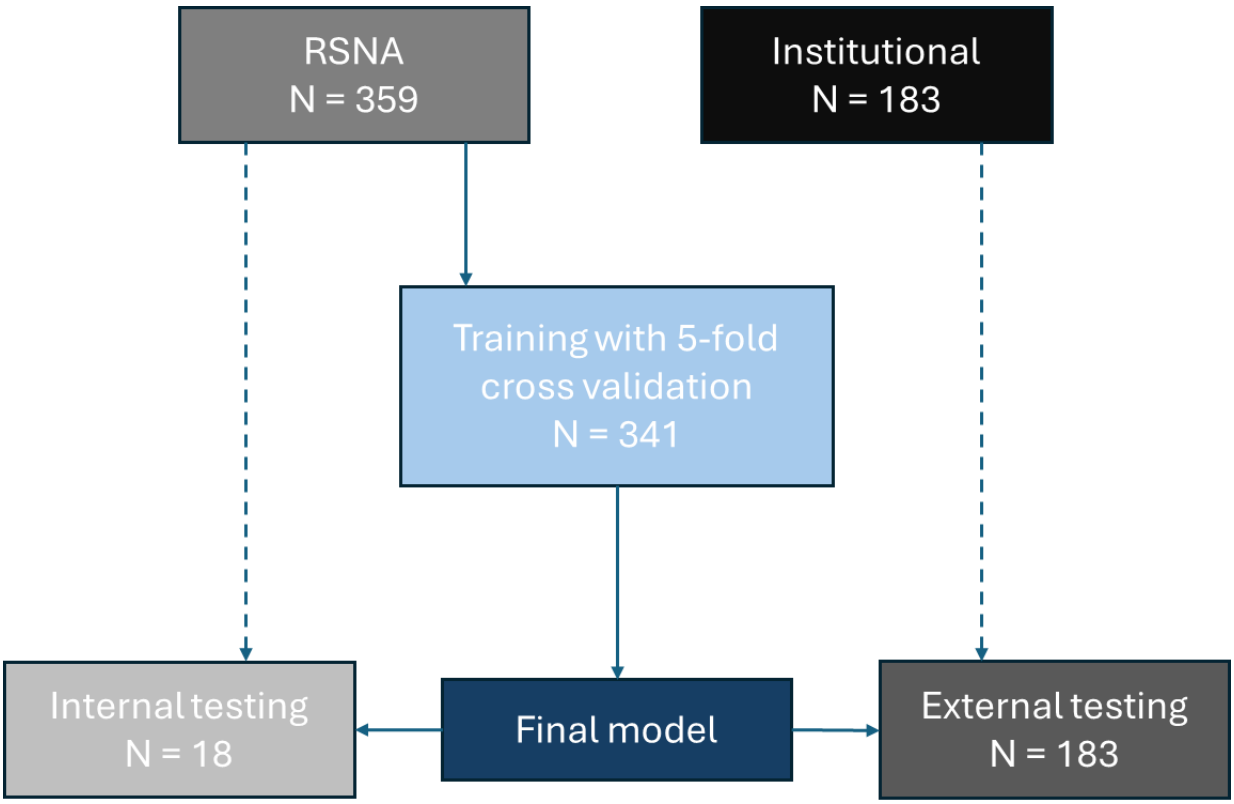


Figure 1. Study design for model training, cross-validation and testing. A total of 359 non-contrast head CT (NCCT) from the RSNA ICH dataset and 183 institutional examples were included. For model training, 341 NCCTs from the RSNA dataset were used for training with 5-fold cross-validation. The final model (ensemble) was evaluated on two cohorts: 18 hold-out RSNA NCCTs (internal testing) and 183 institutional NCCTs (external testing). Dashed lines indicate the origin of the test sets, which were not included in training or validation.

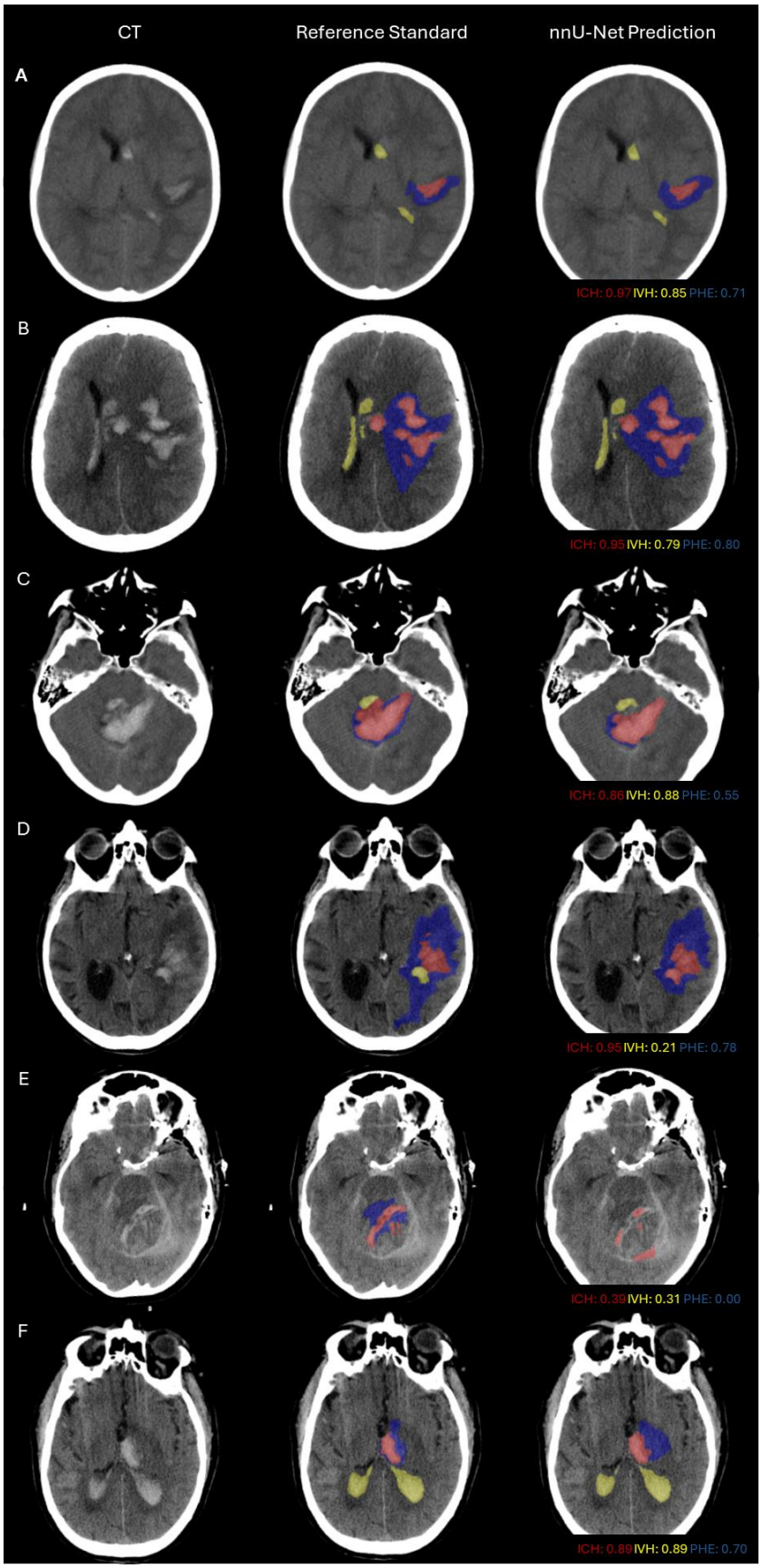

CT
Reference Standard
nnU-Net Prediction
A
ICH: 0.97 IVH: 0.85 PHE: 0.71
B
ICH: 0.95 IVH: 0.79 PHE: 0.80
C
ICH: 0.86 IVH: 0.88 PHE: 0.55
D
ICH: 0.95 IVH: 0.21 PHE: 0.78
E
ICH: 0.39 IVH: 0.31 PHE: 0.00
F
ICH: 0.89 IVH: 0.89 PHE: 0.70

Figure 2. Representative examples from the internal test set showing nnU-Net segmentation predictions compared with reference standard annotations. Each row displays the same axial slice across three columns: the original non-contrast CT (left), reference standard segmentation (middle), and nnU-Net prediction (right). Hemorrhage types are color-coded: intracerebral hemorrhage (ICH) in red, intraventricular hemorrhage (IVH) in yellow, and perihematomal edema (PHE) in blue. The Dice Similarity Coefficient (DSC) for each lesion is reported in the lower right corner of the nnU-Net prediction panels. Note: the slices shown are representative axial slices; some lesion voxels, including IVH, may be present in other slices not displayed.
(A) Well-delineated ICH, IVH, and PHE with clearly defined margins, facilitating accurate segmentation.
(B) Multifocal hemorrhages pose a challenge due to non-contiguous foci and ill-defined PHE margins.
(**C–D**) Ambiguous boundaries between ICH and IVH make precise delineation difficult.
(**E**) Postoperative example with pneumocephalus. The complex morphology of the ICH makes it challenging to define the extent of hemorrhage. There is ambiguity regarding the presence and extent of PHE.
(**F**) Artifacts degrade image quality with variable impact on segmentation accuracy.

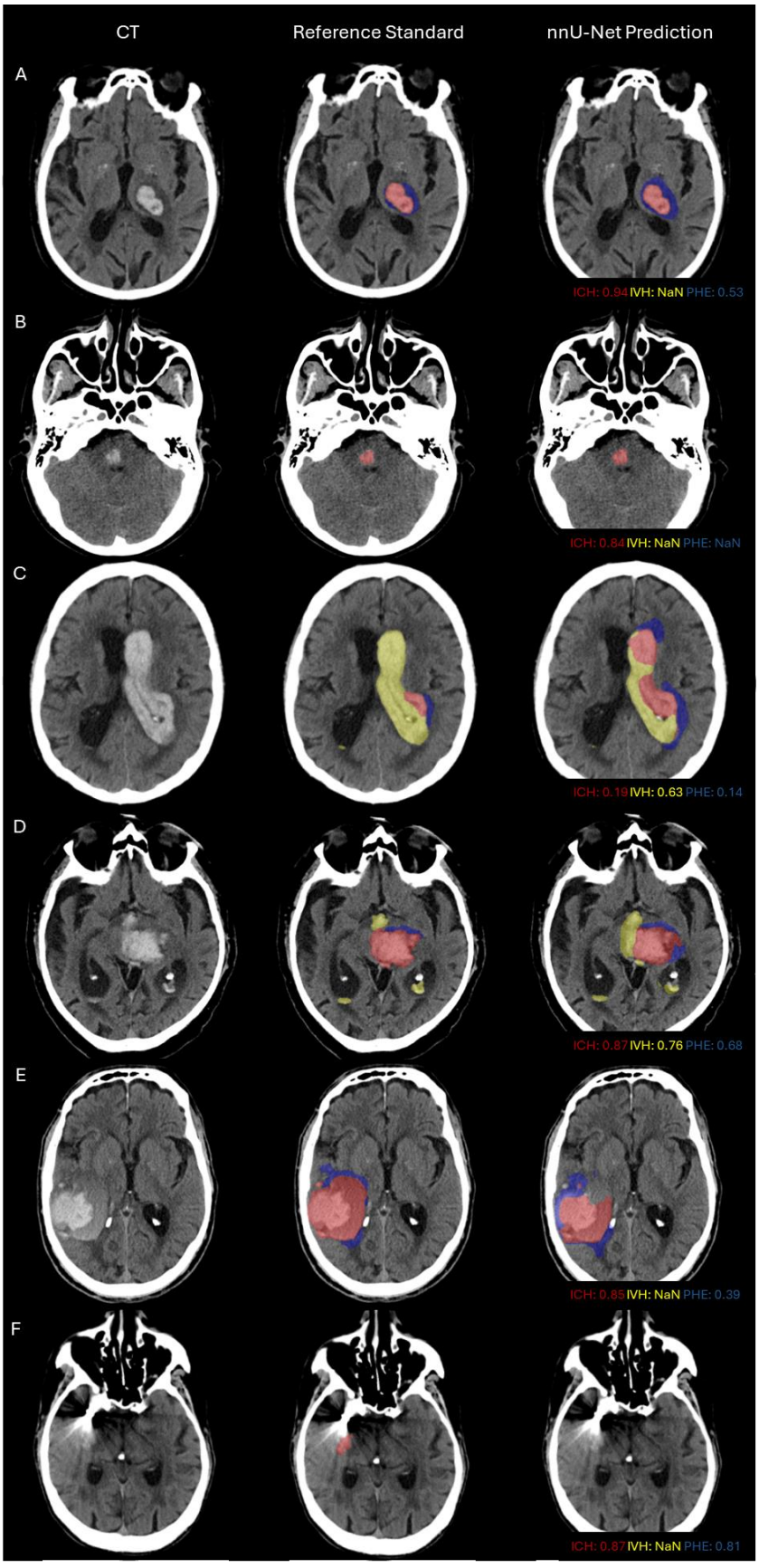

CT
Reference Standard
nnU-Net Prediction
A
ICH: 0.94 IVH: NaN PHE: 0.53
B
ICH: 0.84 IVH: NaN PHE: NaN
C
ICH: 0.19 IVH: 0.63 PHE: 0.14
D
ICH: 0.87 IVH: 0.76 PHE: 0.68
E
ICH: 0.85 IVH: NaN PHE: 0.39
F
ICH: 0.87 IVH: NaN PHE: 0.81

Figure 3. Representative cases from the external test set showing nnU-Net predictions compared with reference standard annotations for ICH, IVH, and PHE. Each row displays the same axial slice across three columns: the original non-contrast CT (left), reference standard segmentation (middle), and nnU-Net prediction (right). Hemorrhage types are color-coded: intracerebral hemorrhage (ICH) in red, intraventricular hemorrhage (IVH) in yellow, and perihematomal edema (PHE) in blue. The Dice Similarity Coefficient (DSC) for each lesion is reported in the lower right corner of the nnU-Net prediction panels. NaN (Not a Number) indicates cases where both the predicted mask and the reference standard mask were empty for a given lesion, making the DSC undefined. Details on how these cases were handled in the analysis are provided in the Supplemental Material.
(**A**) Well-delineated spontaneous lobar hematoma with excellent accuracy of the model, but over-segmentation of PHE.
(**B**) Small pontine hematoma without definite PHE.
(**C–D**) Confluent ICH and IVH with ambiguous boundaries and ICH location.
(**E**) Heterogeneous ICH with an isodense component (Blend sign), which makes it challenging to differentiate from brain parenchyma. This also impacts PHE accuracy.
(**F**) Metal artifacts impair ICH visibility (ICH segmentation was less impacted on other slices without artifacts).

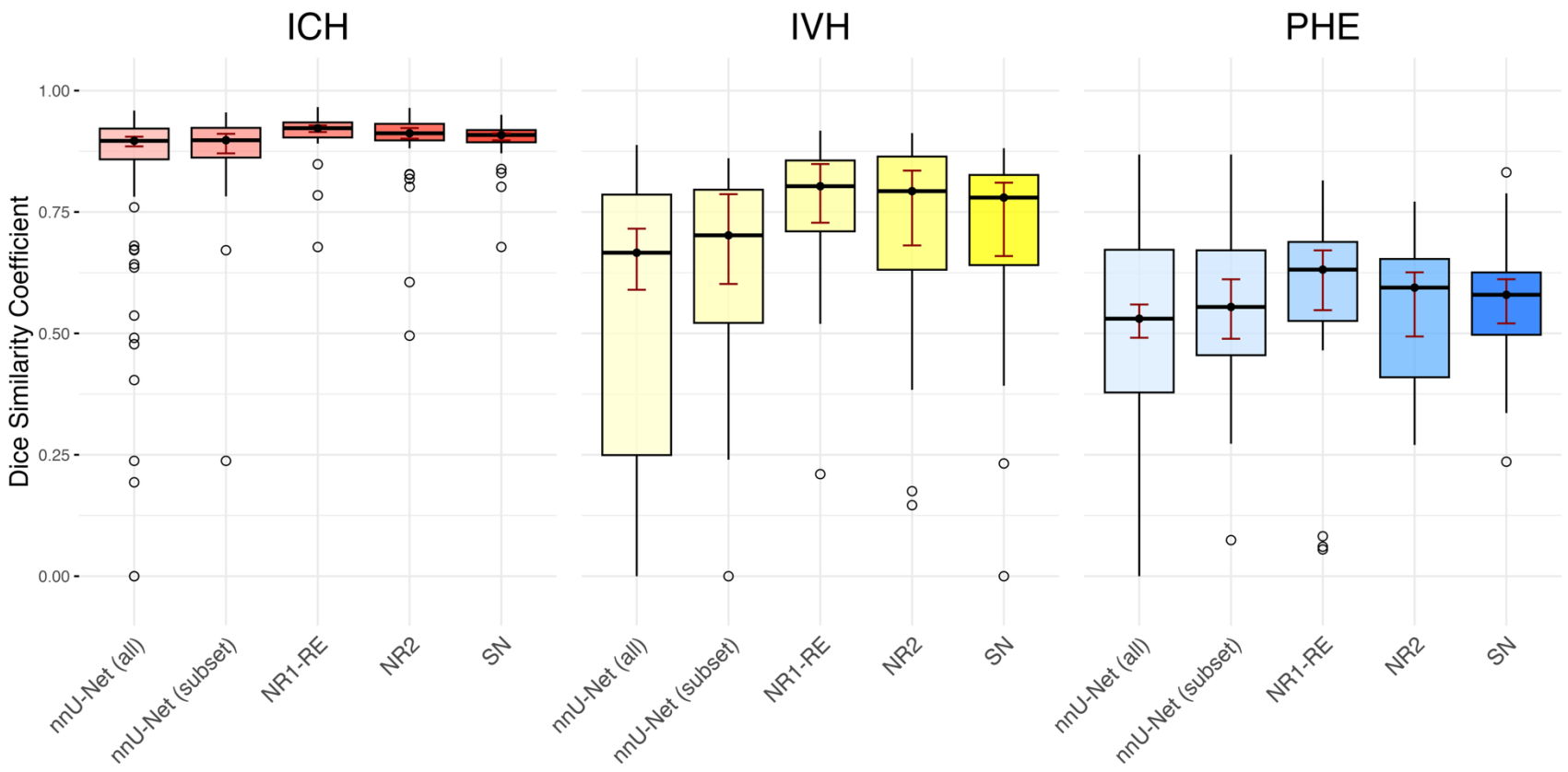


Figure 4. Intra- and inter-annotator agreement compared with model performance on the external test set. Boxplots show Dice Similarity Coefficients (DSC) for intracerebral hemorrhage (ICH, red), intraventricular hemorrhage (IVH, yellow), and perihematomal edema (PHE, blue). For each lesion type, five groups are displayed (left to right): nnU-Net (all), model performance on the full external test cohort (n = 183) compared with NR1-RS; nn-UNet (subset), model performance restricted to the 30-case subset used for the annotator study; NR1-RE, intra-annotator agreement between NR1-RS and repeat de novo annotations by the same neuroradiologist after a 6-month washout period; NR2, inter-annotator agreement between NR1-RS and a second neuroradiologist; and SN, inter-annotator agreement between a stroke neurologist and NR1-RS. Hollow circles represent outliers, and thin red vertical lines represent the 95% confidence intervals for the median.

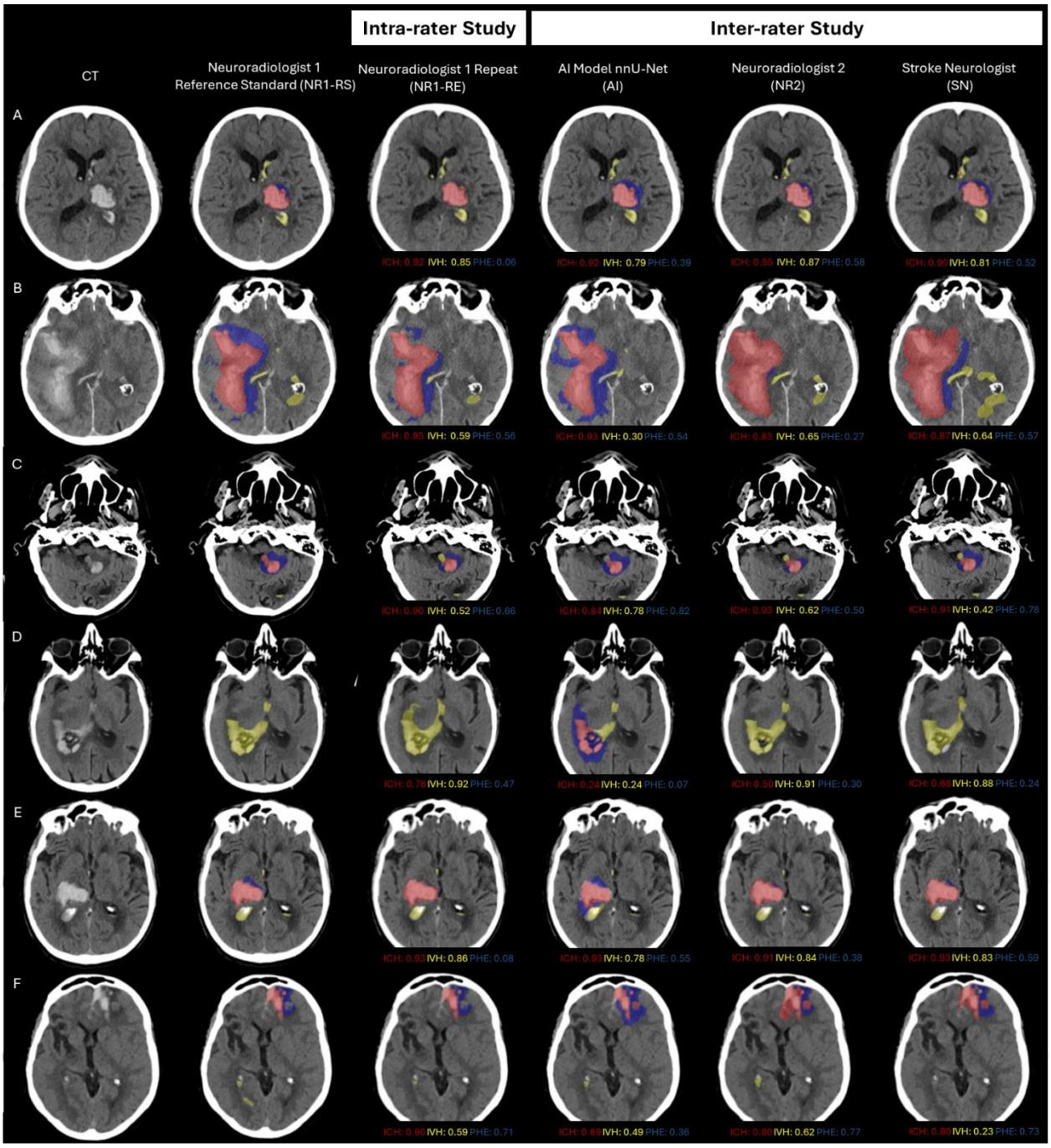

Intra-rater Study
Inter-rater Study
CT
Neuroradiologist 1 Reference Standard (NR1-RS)
Neuroradiologist 1 Repeat (NR1-RE)
AI Model nnU-Net (AI)
Neuroradiologist 2 (NR2)
Stroke Neurologist (SN)
A
ICH: 0.92 IVH: 0.85 PHE: 0.06
ICH: 0.92 IVH: 0.79 PHE: 0.39
ICH: 0.95 IVH: 0.87 PHE: 0.58
ICH: 0.90 IVH: 0.81 PHE: 0.52
B
ICH: 0.95 IVH: 0.59 PHE: 0.56
ICH: 0.93 IVH: 0.30 PHE: 0.54
ICH: 0.83 IVH: 0.65 PHE: 0.27
ICH: 0.87 IVH: 0.64 PHE: 0.57
C
ICH: 0.90 IVH: 0.52 PHE: 0.66
ICH: 0.84 IVH: 0.78 PHE: 0.82
ICH: 0.93 IVH: 0.62 PHE: 0.50
ICH: 0.91 IVH: 0.42 PHE: 0.78
D
ICH: 0.78 IVH: 0.92 PHE: 0.47
ICH: 0.24 IVH: 0.24 PHE: 0.07
ICH: 0.58 IVH: 0.91 PHE: 0.30
ICH: 0.68 IVH: 0.88 PHE: 0.24
E
ICH: 0.93 IVH: 0.86 PHE: 0.08
ICH: 0.93 IVH: 0.78 PHE: 0.55
ICH: 0.91 IVH: 0.84 PHE: 0.38
ICH: 0.93 IVH: 0.83 PHE: 0.59
F
ICH: 0.90 IVH: 0.59 PHE: 0.71
ICH: 0.89 IVH: 0.49 PHE: 0.36
ICH: 0.88 IVH: 0.62 PHE: 0.77
ICH: 0.88 IVH: 0.23 PHE: 0.73

Figure 5. Illustrative cases from the intra-annotator and inter-annotator agreement study. Each row (A–F) shows the same axial slice from left to right: original non-contrast CT, reference standard annotation of the first neuroradiologist (NR1-RS), repeat annotation by the same neuroradiologist (NR1-RE) for intra-annotator assessment, nnU-Net automated segmentation (AI), annotation by the second neuroradiologist (NR2), and annotation by the stroke neurologist (SN) for inter-annotator comparison. Hemorrhage types are color-coded: intracerebral hemorrhage (ICH) in red, intraventricular hemorrhage (IVH) in yellow, and perihematomal edema (PHE) in blue. Dice similarity coefficients (DSC) for each lesion class are displayed in the lower right corner of the segmentation panels.
(**A**) Well-defined ICH and IVH with strong agreement across all labels. PHE remains challenging.
(**B**) Large, heterogeneous ICH with complex morphology and isodense components, with differences in segmentation masks across sources. Isodense components make differentiation between ICH and background brain parenchyma challenging. IVH and PHE masks were also variable.
(**C-D**) Challenging cases with unclear distinction between ICH and IVH components, resulting in segmentation variability across annotators and the automated model. Ambiguity between ICH and IVH directly impacts PHE.
(**E**) Microangiopathic changes are visible, complicating the delineation of PHE.
(**F**) Trace IVH, which is difficult to identify and segment accurately on non-contrast CT.

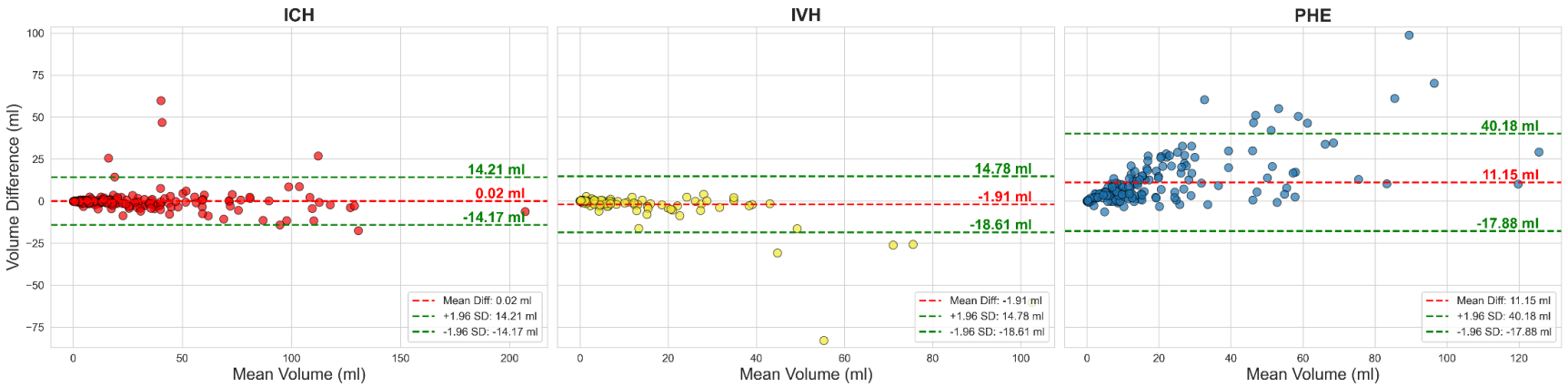


Figure 6. Bland–Altman plots comparing predicted and reference standard volumes for intracerebral hemorrhage (ICH, in red), intraventricular hemorrhage (IVH, in yellow), and perihematomal edema (PHE, in blue) on the external test set. Each plot shows the volume difference (model prediction minus reference standard) on the y-axis against the mean volume between the predicted and reference standard segmentations on the x-axis. The red dashed line indicates the mean difference (bias), while the green dashed lines represent the 95% limits of agreement (±1.96 standard deviations from the mean difference).