Multi-site Benchmarking of Deep Learning Models for Intraparenchymal Hemorrhage Segmentation on NCCT

Kauê T N Duarte[*1]
Abhijot S Sidhu[1,2,3]
Murilo C Barros[4]
Taha Aslan[6]
Donghao Zhang[5]
Jianhai Zhang[1]
Devansh Bhatt[1]
Brij Karmur[1]
Mohamed AlShamrani[6]
Wu Qiu[5]
Aravind Ganesh[1]
Bijoy K Menon[1]

[1] *Cummings School of Medicine, University of Calgary, Calgary, AB, Canada.*
[2] *Graduate Program in Biomedical Engineering, University of Calgary, Calgary, Canada.*
[3] *Seaman Family MR Research Centre, Foothills Medical Centre, Calgary, Canada.*
[4] *School of Technology, University of Campinas, Limeira, Brazil.*
[5] *College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, China.*
[6] *Calgary Stroke Program, Department of Clinical Neurosciences, Foothills Medical Centre, University of Calgary, Calgary, Canada.*

Editors: Under Review for MIDL 2026

Abstract

Intraparenchymal hemorrhage (IPH) is a critical and often fatal subtype of hemorrhagic stroke, requiring rapid and accurate diagnosis on non-contrast computed tomography (NCCT) scans for effective treatment. While deep learning (DL) models, more specifically using convolutional neural networks (CNNs), offer potential for automating IPH segmentation, their real-world clinical utility is often limited by the lack of explicit data integration across diverse hospital sites with varying imaging protocols. This study conducted a multi-site benchmarking of four prominent CNN architectures: baseline U-Net, Attention U-Net, Feature Pyramid Network (FPN), and Trans U-Net, for IPH segmentation on a heterogeneous dataset from 17 clinical sites. Models were rigorously evaluated using F-measure (*a.k.a.*, Dice), Intersection over Union (IoU), and 95% Hausdorff Distance (d_{H95}). The advanced CNN variants (Attention U-Net, FPN, Trans U-Net) significantly outperformed the baseline U-Net in F-measure and IoU (*e.g.*, FPN F-measure: 0.868 vs. U-Net: 0.819, $p < 0.001$), with no significant difference among them. For boundary error, Attention U-Net and FPN demonstrated a substantial reduction in d_{H95} (42–50%) compared to the baseline, whereas Trans U-Net showed improvement, but it was not significant. These models exhibited robust cross-site generalization across hemorrhage volumes, with minimal site-specific effects

* Corresponding author

on performance. This study demonstrates that advanced CNN variants can be adopted for IPH segmentation to standardize and potentially accelerate IPH diagnosis.

Keywords: stroke, intraparenchymal hemorrhage, artificial intelligence, medicine, computed tomography.

1. Introduction

Stroke is a major cause of death and long-term disability globally. Each year, more than 12 million cases and over 7 million deaths are reported (Feigin et al., 2025). Among these cases, hemorrhagic stroke is one of the deadliest types, as it causes a rupture of cerebral blood vessels and subsequent intracranial bleeding. Although accounting for a smaller portion of the stroke cases, this type is associated with a high fatality rate.

Non-contrast computed tomography (NCCT) is a medical imaging modality commonly used to detect stroke. It not only provides rapid, accessible information on intracranial hemorrhage (ICH) but also plays a critical role in emergency diagnosis and treatment planning. Fast detection of hemorrhage can positively salvage brain function and increase the patient's survival rate (Ahmed and Prakasam, 2025). This urgency, in a clinical setting, can affect the time to diagnosis, potentially leading to delays or oversights.

Among the subtypes of ICH, intraparenchymal hemorrhage (IPH) represents a critical and challenging pathology. IPH is characterized by bleeding in the brain tissue ($\sim$15% of total stroke cases), which leads to a high mortality rate (with 30-day mortality rates of 40-50%) among the ICH subtypes (Roy et al., 2015). This mortality rate is nearly double that of the fatality caused by ischemic stroke (Rothwell et al., 2004; Woo et al., 2022). A baseline hematoma volume is one of the strongest independent predictors of mortality, with patients with volumes $> 30mL$ experiencing a mortality rate $> 50\%$ (Abulhasan et al., 2023). This volume often drives therapeutic choices, such as selecting candidates for minimally invasive evacuation or deciding on surgery after follow-up imaging (Polster et al., 2021). Unlike other subtypes of ICH, like subdural or subarachnoid hemorrhage, where surgical evacuation is primarily anatomically guided, IPH management often relies heavily on precise volumetric quantification. However, accurately measuring IPH is challenging due to factors such as irregular lesion boundaries, variable texture patterns, and proximity to complex anatomical structures, which can affect measurements. These typically require specialized, robust analytical tools to segment and measure these regions in different sites.

Artificial Intelligence (AI) techniques have accelerated stroke diagnosis by automating manual tasks, such as detection and segmentation, while maintaining high accuracy rates. Among AI types, convolutional neural networks (CNNs) automatically extract information from images and are mainly used for classification tasks, such as the ASPECTs score. For semantic segmentation, CNN-based U-Net variants and their numerous adaptations are often adopted, as they build an encoder-decoder architecture that not only extracts features but also reconstructs them in image space (Lin et al., 2025). These models have demonstrated high confidence in distinguishing pathological tissue from healthy tissue (Duarte et al., 2024). The advanced U-Net and its variants are continually improving, either by integrating attention mechanisms to delineate lesion boundaries more accurately or by using fractal pyramid networks to capture fine-grained and global contextual details.

One major challenge for clinical translation is domain shift across sites, which introduces variability across scanner vendors, acquisition protocols, and site practices. Variability in AI contexts can degrade model performance if not adequately tested in real-world clinical settings. The literature often refers to models trained on curated datasets, yet lacks systematic, comparative validation of these architectures for IPH segmentation across multiple clinical sites. Additionally, numerous studies have trained models on public data, addressing multiple hemorrhage types simultaneously, rather than optimizing for the complexities of IPH individually (Ahmed and Prakasam, 2025; Piao et al., 2023). The focus on architectural novelty can also overshadow deeper investigation of IPH's intrinsic features.

We propose a study focused on IPH segmentation across multiple sites. We benchmarked four CNN architectures (U-Net, Attention U-Net, FPN, Trans U-Net) on a heterogeneous, multi-site NCCT dataset and report F-measure, IoU and d_{H95}. We evaluate generalizability by assessing statistical values using several metrics.

The remainder of this paper is organized as follows. Section 2 reviews related work relevant to the study. Section 3 details the materials, methods, and statistical definitions employed. Section 4 presents the results, and Section 5 provides an analysis of these findings. Finally, Section 6 outlines the conclusions and suggests directions for future research.

2. Related Work

The accurate and timely segmentation of intracranial hemorrhage on NCCT is essential for acute stroke management, impacting diagnostic speed, treatment planning, and, ultimately, patient outcomes (Ahmed and Prakasam, 2025). However, manual interpretation by radiologists can be time-consuming and is often subject to inter-observer variability (Inkeaw et al., 2022). Thus, deep learning models can solve this by automating segmentation, thereby reducing diagnostic delays and standardizing analysis (Piao et al., 2023).

In response to pressing clinical needs, researchers have concentrated on developing advanced segmentation architectures. Models such as the U-Net framework and its variants, U-Net++, Attention U-Net, and ResU-Net, have demonstrated strong performance on curated public benchmarks (Lin et al., 2025). More recently, transformer-based models and hybrid architectures, such as TransHarDNet, have been explored to capture long-range dependencies (Piao et al., 2023) more effectively. These models consistently achieve high Dice Similarity Coefficients, sometimes reaching 0.85 on their respective test sets (Ahmed and Prakasam, 2025; Lin et al., 2025), highlighting the considerable potential of deep learning for this task. Zhang *et al.*(Zhang et al., 2025) proposed a multi-task study with the focus of understanding the use of DL for several hemorrhage applications.

However, this predictive performance is often obtained and validated using homogeneous or publicly available datasets collected with standardized imaging protocols (Roy et al., 2015). When implemented in real-world clinical settings, these models encounter a notable domain shift, resulting in lower performance (Inkeaw et al., 2022). This shift is driven by variations in scanner vendors, acquisition parameters (e.g., tube voltage and slice thickness), and reconstruction kernels across hospitals. Although some studies have begun to address this issue through approaches such as multi-window input optimization (Inkeaw et al., 2022), a gap remains in the validation of segmentation models on large, heterogeneous, multi-site datasets. The comparative analysis of how CNN variants can generalize across

multiple clinical sites remains underexplored, despite its vital importance for real-world applications.

This validation paper studies deep learning models for IPH segmentation using a multi-site dataset characterized by significant protocol heterogeneity. The model was designed explicitly for segmenting parenchymal hemorrhage. Unlike existing studies, our work uniquely quantifies the performance and generalization of this specialized model across a diverse, multi-hospital private dataset, rather than simply developing a new architecture based on public data or addressing a wide array of hemorrhage types. This approach provides a crucial real-world evaluation of the challenges of AI deployment in stroke care, particularly highlighting how different architectural strategies maintain performance across varied clinical settings.

3. Materials and Methods

3.1. Dataset and Participant Demographics

The study utilized a multi-site dataset with IPH segmentation. In total, $N = 239$ subjects were included from 17 clinical sites across Canada (labelled A-Q in accordance with our ethics board) participating in the ACT Trial imaging collection (Menon et al., 2022). All imaging data included manual ground-truth annotations for intraparenchymal hemorrhage (IPH), along with descriptive information such as age, sex, and other factors. Table 1 summarizes the demographic and clinical characteristics of the study population.

ACT Trial Imaging Dataset. We used CT imaging data and corresponding hemorrhage segmentation masks provided by the ACT Trial investigators (Menon et al., 2022). Site identifiers were anonymized in accordance with ethical and data-sharing agreements. Hemorrhagic stroke diagnoses were confirmed by board-certified neurologists using standardized clinical criteria. Ground-truth segmentation masks were generated using a semi-automated workflow that combines algorithmic lesion proposals with expert manual correction and review.

3.2. Data Preparation

To improve the quality of the NCCT scans, we applied skull stripping using SynthStrip, adjusted to CT (Hoopes et al., 2022). All 3D volumes were standardized to dimensions that are multiples of 64 through zero-padding. Each volume was split into $64 \times 64 \times 64$ patches to facilitate memory-efficient processing. To address class imbalance between hemorrhagic and non-hemorrhagic voxels, we employed a selective patching strategy that retained patches containing at least one hemorrhagic lesion voxel during training, validation, and testing. Intensity normalization was carried out in two steps: (1) we cropped the intensity from -30 to 100 Hounsfield units (HU); (2) we performed min-max normalization, mapping the image intensities to the range $[0.0, 1.0]$.

3.3. Deep Learning Architectures

We implemented and compared four state-of-the-art CNN variants for IPH segmentation:

Table 1: Demographic and clinical characteristics of the study population across the 17 anonymized clinical sites (A-Q). Data are presented as Mean ± Standard Deviation for Age (years) and IPH Volume (cm^3). Sex is reported as the count of male patients with the corresponding percentage in parentheses. The sample size (N) for each site is also provided.

	A	**B**	**C**	**D**	**E**
Age	71.37 ± 15.00	59.00 ± 18.38	72.97 ± 13.74	78.50 ± 12.07	78.78 ± 12.62
Sex	30 (55.6%)	2 (100.0%)	21 (52.5%)	4 (50.0%)	5 (55.6%)
Vol	12.69 ± 22.81	24.54 ± 32.42	25.29 ± 50.69	19.18 ± 28.48	3.32 ± 4.37
N	54	2	40	8	9

	F	**G**	**H**	**I**	**J**
Age	70.57 ± 12.14	75.75 ± 11.57	74.17 ± 9.39	72.60 ± 19.50	72.43 ± 12.99
Sex	3 (42.9%)	3 (37.5%)	2 (33.3%)	2 (40.0%)	18 (64.3%)
Vol	3.54 ± 3.09	17.05 ± 36.22	10.17 ± 14.15	7.66 ± 12.61	7.50 ± 17.22
N	7	8	6	5	28

	K	**L**	**M**	**N**	**O**
Age	76.64 ± 12.18	80.89 ± 11.89	84.50 ± 6.81	73.58 ± 11.68	77.33 ± 7.51
Sex	2 (18.2%)	7 (77.8%)	8 (50.0%)	12 (46.2%)	2 (66.7%)
Vol	4.87 ± 5.42	34.84 ± 40.85	18.71 ± 32.98	12.25 ± 28.40	6.88 ± 8.50
N	11	9	16	26	3

	P	**Q**	**Total**
Age	75.20 ± 15.32	87.50 ± 2.12	74.40 ± 13.29
Sex	3 (60.0%)	2 (100.0%)	126 (52.71%)
Vol	0.85 ± 1.42	0.44 ± 0.56	14.27 ± 30.36
N	5	2	239

1. *Baseline U-Net*: We implemented the original U-Net architecture (Ronneberger et al., 2015) as our baseline model. This encoder-decoder network with skip connections provides a robust foundation for medical image segmentation.

2. *Attention U-Net*: This architecture enhances the traditional U-Net by incorporating attention gates in the skip connections (Oktay et al., 2018). The attention mechanisms selectively emphasize relevant spatial features while suppressing irrelevant regions, particularly beneficial for detecting small hemorrhagic lesions and precise boundary delineation.

3. *Feature Pyramid Network (FPN)*: The FPN architecture (Lin et al., 2017) builds a multi-scale feature pyramid through top-down pathways and lateral connections. This design enables effective feature extraction across multiple scales, which is advantageous for detecting hemorrhagic lesions of varying sizes and shapes.

4. *Trans U-Net*: This architecture leverages a hybrid of a CNN+Transformer design (Chen et al., 2021), combining U-Net local feature extraction and the Vision Transformer (ViT). The integration of ViT and U-Net yields improved global context for IPH masks and is believed to enhance boundary delineation.

All architectures utilized a VGG16 backbone (Simonyan and Zisserman, 2014) for feature extraction in the encoder pathway, consistent with previous work demonstrating its effectiveness for medical image segmentation tasks (Duarte et al., 2024).

3.4. Model Training and Evaluation

Model training was conducted for a maximum of 300 epochs with an initial learning rate of 5×10^{-4}. We employed the Adam optimizer with standard parameters and reduced the learning rate when the loss plateaued. For each architecture, we trained separate 2D models using axial (2DAxi), coronal (2DCor), and sagittal (2DSag) projections, and obtained final predictions by averaging across these projections (2.5D model).

Loss Function. To address the class imbalance between *True* and *False* elements in the IPH masks, we used a composite loss function combining Dice Loss (DL, eq. 1) and Binary Focal Loss (FL, eq. 2).

$$DL = \frac{(1+\beta^2)\cdot TP}{(1+\beta^2)\cdot TP + \beta^2 \cdot FN + FP} \tag{1}$$

where β corresponds to a balance coefficient, TP, FP, and FN represent the true positive, false positive, and false negative voxels, respectively.

$$FL = -\,GT\alpha(1-PT)^{\gamma}\log(PT) \quad -(1-GT)\alpha PT^{\gamma}\log(1-PT) \tag{2}$$

where GT is the ground-truth values, PT represents the predicted truth, $\alpha = 0.25$ and $\gamma = 2.0$ are values that were fine-tuned through a calibration process.

Performance Metrics. The class imbalance between IPH and non-IPH voxels rendered accuracy an unsuitable performance metric, as the large number of true negatives (TN) would disproportionately influence the results. To better evaluate model performance, we

utilized the F-measure, intersection-over-union (IoU), and Hausdorff distance. The F-measure, *a.k.a.* dice coefficient for binary segmentation, is a commonly adopted metric in image segmentation:

$$\mathrm{F} = 2 \times \frac{P \times R}{P + R} \tag{3}$$

which represents the harmonic mean of precision (P) and recall (R):

$$P = \frac{TP}{TP + FP} \text{ and } R = \frac{TP}{TP + FN}.$$

IoU quantifies the overlap between predictions and ground truth:

$$\mathrm{IoU} = \frac{TP}{FP + TP + FN}. \tag{4}$$

Throughout training, the model achieving the highest IoU value was retained as optimal.

The Hausdorff distance measures the separation between the predicted and ground-truth IPH boundaries. For two point sets A and B representing these boundaries, the d_{H95} is defined as:

$$d_H(x,y) = \max\{d_{AB}, d_{BA}\} = \max\left\{\max_{a\in A}\left\{\min_{b\in B}\{d(a,b)\}\right\}, \max_{b\in B}\left\{\min_{a\in A}\{d(a,b)\}\right\}\right\} \tag{5}$$

where a and b represent elements of sets A and B, respectively, and $d(a,b)$ is the Euclidean distance between them. We used the 95^{th} percentile of the Hausdorff distance distribution (d_{H95}) to assess performance. Superior performance is indicated by higher F-measure and IoU values, and a lower d_{H95} value.

Implementation Details. We ran experiments on a four-node cluster (8× Tesla V100 16GB GPUs; 754 GB total system RAM. To enable a fair comparison with other U-Net variants, each brain projection was trained independently in parallel, substantially reducing the total training time. Model development was carried out using Python 3.6 in Jupyter Notebook, and the resulting code was subsequently converted to Python scripts to enable cluster execution. The full source code and Keras-based implementations are publicly available [1].

3.5. Statistical Analysis

Five-fold cross-validation was used to evaluate all four U-Net models, and results are reported as the mean ± standard deviation. Throughout the analysis, appropriate tests were performed to assess the validity of the model assumptions, including tests of residual normality. Significance level was set to $\alpha = 0.05$ for all statistical tests. The R statistical package (https://www.r-project.org/) was used.

To evaluate the effect of segmentation style on performance metrics PM (F-measure, IoU, and d_{H95}), separate linear regression models were fitted for each segmentation style to assess the relationship between performance metrics and covariates, including ground

1. https://github.com/KaueTND/ip-hemorrhagic-stroke-segmentation

truth hemorrhage volume (*IPH_vol*), *Age*, *Sex*, and site variability (*Sitename*). The model structure for each style was:

$$PM \sim IPH\_vol + Age + Sex + Sitename \tag{6}$$

Subsequently, a one-way analysis of variance (ANOVA) was conducted to test for overall differences in performance across the four segmentation styles (U-Net, Attention U-Net, FPN, and Trans U-Net). Post-hoc pairwise comparisons were performed using t-tests with Bonferroni correction to control for multiple comparisons.

4. Results

We defined four separate linear regression models for each segmentation style to evaluate performance across multiple sites, while controlling for age, sex, and ground-truth hemorrhage volume (*IPH_vol*), and site was incorporated as a fixed effect.

For F-measure, the linear regression models revealed significant positive associations between *IPH_vol* and performance across all styles (U-Net: 1.778×10^{-06}, $p < 0.001$; Attention U-Net: 1.402×10^{-06}, $p < 0.001$; FPN: 1.322×10^{-06}, $p < 0.001$; Trans U-Net: 1.416×10^{-06}, $p < 0.001$). One-way ANOVA demonstrated significant overall differences between styles ($F(3, 952) = 6.889$, $p < 0.001$), with post-hoc pairwise comparisons revealing that the advanced styles significantly outperformed the baseline U-Net ($p < 0.01$). In contrast, no significant differences were observed among the three advanced architectures ($p \geq 0.684$). Table 2 groups the F-Measure values for each orientation across the CNN variants. Figure 1 shows the IPH segmentation per orientation and CNN variant. In addition, Figure 2 presents the IPH volumes per F-measure in each CNN variant.

The IoU results followed similar patterns, with *IPH_vol* showing strong positive associations in all regression models (U-Net: 2.523×10^{-06}, $p < 0.001$; Attention U-Net: 2.116×10^{-06}, $p < 0.001$; FPN: 2.011×10^{-06}, $p < 0.001$; Trans U-Net: 2.126×10^{-06}, $p < 0.001$). ANOVA confirmed significant style differences ($F(3, 952) = 8.277$, $p < 0.001$), with the advanced styles demonstrating superior performance compared to baseline ($p < 0.001$) and no significant differences among advanced styles ($p \geq 0.684$). Table 3 shows the IoU values across orientation and CNN variants.

For d_{H95}, linear regression models showed significant negative associations between *IPH_vol* and boundary error across all styles (U-Net: -5.209×10^{-05}, $p = 0.030$; Attention U-Net: -4.642×10^{-05}, $p = 0.020$; FPN: -3.201×10^{-05}, $p = 0.052$; Trans U-Net: -6.274×10^{-05}, $p = 0.022$). ANOVA revealed significant overall differences ($F(3, 952) = 4.201$, $p = 0.006$), with pairwise comparisons confirming that the advanced styles achieved significantly lower boundary errors than the baseline (Attention U-Net vs U-Net: $p = 0.040$; FPN vs U-Net: $p = 0.009$; Trans U-Net vs U-Net: $p = 0.704$), while no significant difference was found among advanced styles ($p \geq 0.162$). Table 4 groups the d_{H95} values over orientation and CNN variants.

Site-specific effects showed minimal consistent impact on model performance across the regression models. Only isolated sites demonstrated marginal effects (*i.e.*, Site F for F-measure in U-Net: 0.118, $p = 0.038$ and Trans U-Net: 0.091, $p = 0.072$; Site N for d_{H95} in U-Net: 7.474, $p = 0.004$, FPN: 5.087, $p = 0.004$, and Trans U-Net: 9.511, $p = 0.001$),

indicating robust generalization of all models across different acquisition sites and protocols. Figure 3 measures the F-Measure grouped per clinical site and CNN variant.

5. Discussion

In this study, we compared and evaluated different CNN variants for IPH segmentation, with a particular focus on generalizability across data from other sites. We evaluated three metrics (F-Measure, IoU, d_{H95}) because they are the most relevant in the literature. We conducted four independent linear regression analyses, followed by ANOVA comparisons, and found that the advanced models (Attention U-Net, FPN, and Trans U-Net) significantly outperformed the baseline U-Net across F-Measure and IoU metrics. For boundary error (d_{H95}), Attention U-Net and FPN showed significant improvement over baseline, while Trans U-Net did not demonstrate a statistically significant reduction. This consistent performance across different metrics indicates that advanced models tend to estimate the IPH volume more accurately. This pattern is also well-identified in Tables 2-4. The 2DCor orientation achieved higher F-Measure and IoU scores, whereas 2.5D achieved the lowest d_{H95}, indicating better alignment of the IPH boundary with the ground truth. These patterns can also be highlighted in Figure 1, where there is an increasing number of FP and FN in 2DAxi, and fewer in 2DCor and 2.5D.

The Attention U-Net integrates attention blocks that adjust segmentations using high-resolution features from the encoding layers (via skip connections). On the other hand, FPN can capture multi-scale contextual information and identify texture patterns that the Attention U-Net sometimes misses. Trans U-Net, with its transformer-based architecture, is designed to capture long-range dependencies and may provide additional contextual information. In IPH segmentation, these types of models are suitable, as IPH often exhibits irregular borders and heterogeneous intensities (Lin et al., 2025).

In our multi-site comparison across CNN variants, the advanced models (Attention U-Net, FPN, and Trans U-Net) showed minimal significant differences in F-Measure and IoU, indicating generalizability even when protocols and scanner vendors differ. For boundary error (d_{H95}), Trans U-Net did not show a significant difference compared to baseline, while Attention U-Net and FPN did. This is of particular interest, as these models are often

Table 2: F-Measure for IPH Segmentation. Performance is compared across four CNN variants (Attention U-Net, baseline U-Net, FPN, and Trans U-Net), evaluated on axial (2DAxi), coronal (2DCor), sagittal (2DSag) projections, and their ensemble (2.5D). Results are reported as Mean ± Standard Deviation. The best model per orientation is highlighted in **bold**.

Style	Attention U-Net	FPN	Trans U-Net	U-Net
Orientation				
2DAxi	0.609±0.244	**0.635±0.225**	0.629±0.228	0.617±0.221
2DCor	0.865±0.129	**0.868±0.127**	0.863±0.132	0.819±0.153
2DSag	0.849±0.155	**0.854±0.14**	0.847±0.145	0.811±0.164
2.5D	0.851±0.165	**0.862±0.146**	0.855±0.151	0.823±0.164

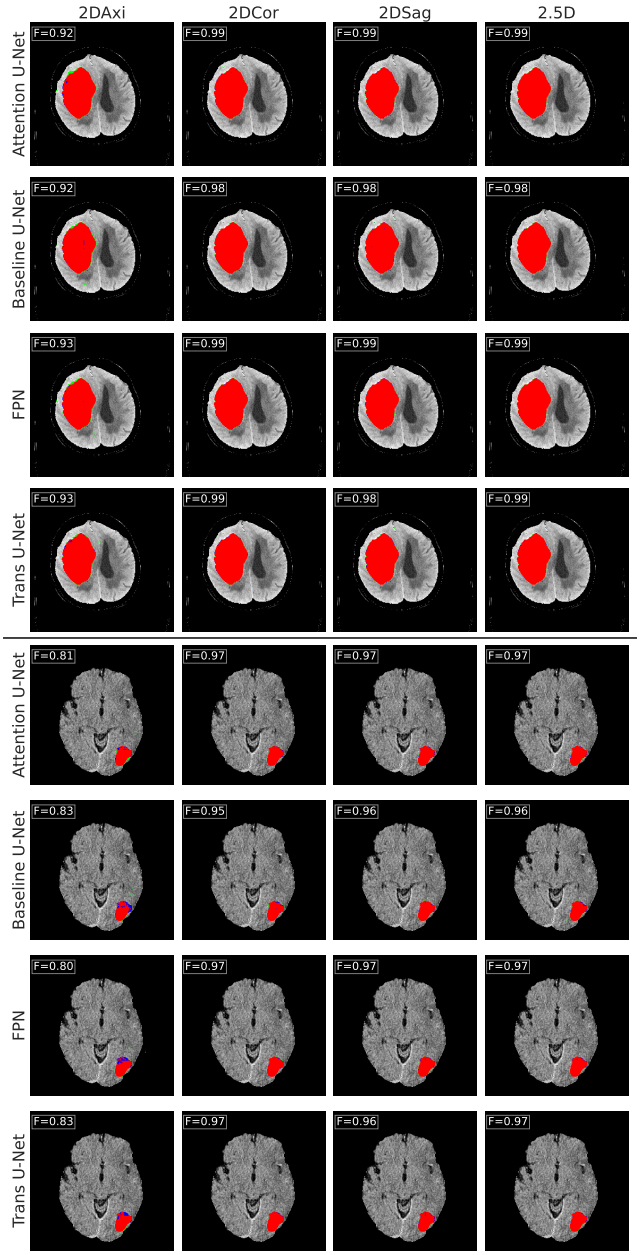

Figure 1: Qualitative comparison of IPH segmentation results across different CNN variants and imaging orientations. Two patients (A–B) are shown, representing large, small IPH volume, respectively. Rows 1–4 show Patient A (85-year-old male, IPH volume: 144 cm^3), rows 5–8 show Patient B (79-year-old female, IPH volume: 7.9 cm^3), Each row group illustrates segmentations from axial, coronal, sagittal, and 2.5D views using Attention U-Net, baseline U-Net, FPN, and Trans U-Net models. F-Measure value is reported in the top left corner.

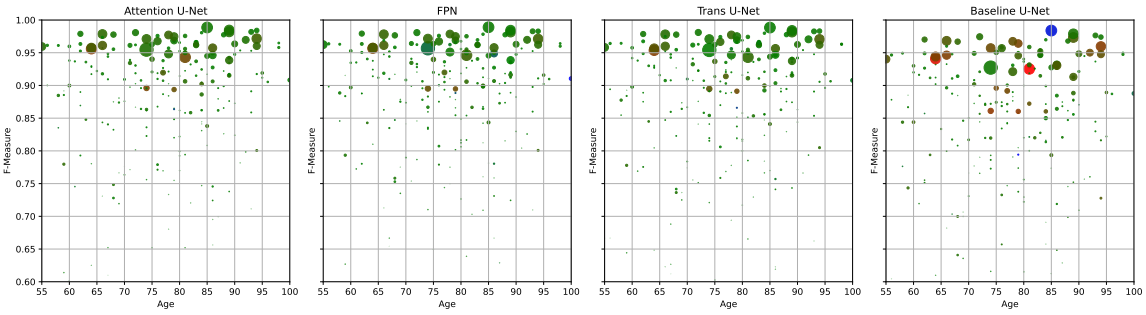

Figure 2: Scatter plot of subject age (x-axis) and F-Measure (y-axis) for IPH segmentation. Each point represents an individual subject, with the marker size corresponding to the IPH volume (in mm^3) computed from the model's predicted segmentation mask. Colours go from red (underestimated according to ground-truth) → green (correctly labelled) → blue (overestimation compared to ground-truth).

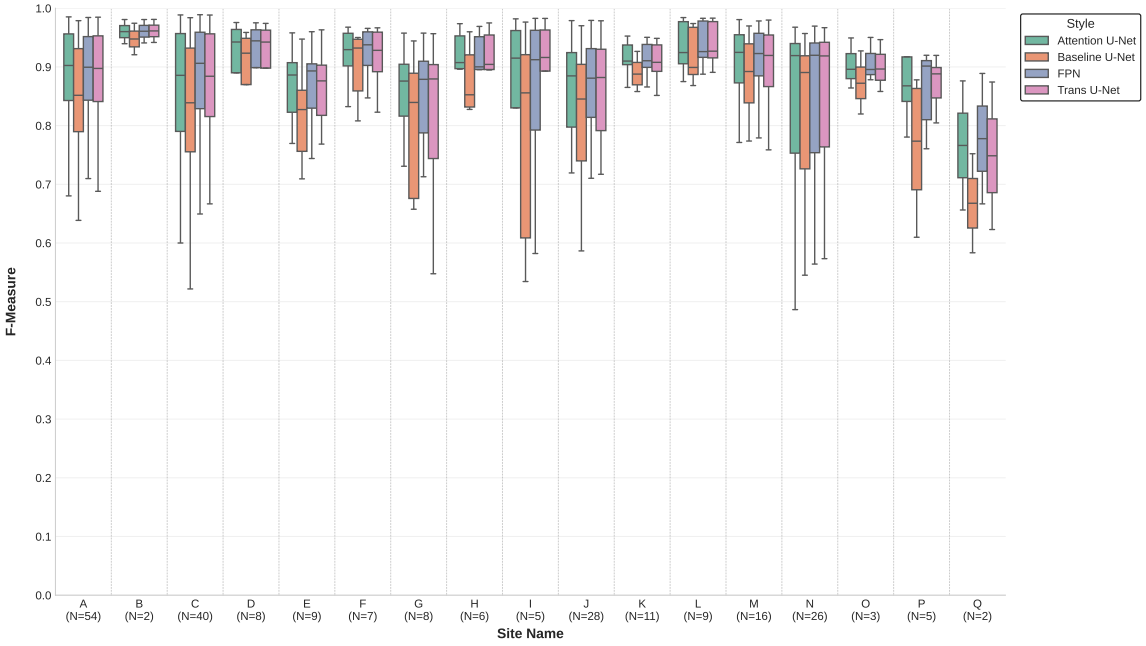

Figure 3: Boxplot comparison of F-Measure scores for IPH segmentation across the 17 clinical sites (A-Q), stratified by CNN variant. N indicates the sample size per site.

deployed in external sources where the protocol can vary drastically. On the other hand, the baseline U-Net showed minimal site preferences, which did not affect F-Measure scores.

Although the models showed better performance with larger, more confluent lesions, there was a drastic reduction in d_{H95} with Attention U-Net and FPN compared to baseline U-Net, along with better detection of smaller lesions. Although not statistically significant, Trans U-Net performed well in regards of d_{H95} compared to baseline. There is a consistent positive association between ground-truth IPH volume and segmentation performance across all models, as indicated by F-Measure ($p < 0.001$) and IoU ($p < 0.001$). While demonstrating significant negative associations with d_{H95} across all styles (U-Net: $p = 0.030$; Attention U-Net: $p = 0.020$; FPN: $p = 0.052$; Trans U-Net: $p = 0.022$). This consistent pattern indicates that all models perform better on larger IPH, which are generally easier to segment. However, the advanced models (Attention U-Net, FPN, and Trans U-Net) showed higher F-Measure and IoU means when compared across the same patients.

The low d_{H95} values in Attention U-Net and FPN demonstrated consistent boundary detection across IPH volumes, as evidenced by non-significant interaction terms ($p > 0.05$) and strong main effects. This indicates that the boundary-precision advantages of Attention U-Net, FPN, and Trans U-Net are maintained regardless of IPH burden, highlighting their robust generalization across varying pathology loads in multi-site applications.

Our post-hoc analyses confirmed our findings regarding multi-site consistency. The pairwise comparisons revealed no statistically significant differences among the advanced models (Attention U-Net, FPN, and Trans U-Net) for F-measure and IoU ($p \geq 0.684$). For d_{H95}, there were no significant differences among the advanced models ($p \geq 0.162$), but only Attention U-Net and FPN showed a significant improvement over the baseline U-Net ($p < 0.05$). This improvement was noticeable when evaluating d_{H95}, showing a reduction of 42% and 50% in IPH boundary error for Attention U-Net and FPN, respectively. Accurate boundary definition is key to differentiate healthy tissue from affected tissue, as well as being essential for volume calculation (Roy et al., 2015).

Although the findings for site-specific analysis were promising, minimal site-specific effects were observed in our linear regression models. We observed isolated sites showed marginal effects (*i.e.*, Site F for F-measure in baseline U-Net: $p = 0.0389$ and in Trans U-Net: $p = 0.0722$; Site N for d_{H95} in U-Net: $p = 0.004$, FPN: $p = 0.004$, and Trans U-

Table 3: IoU for IPH Segmentation. Performance is compared across four CNN variants (Attention U-Net, baseline U-Net, FPN and Trans U-Net), evaluated on axial (2DAxi), coronal (2DCor), sagittal (2DSag) projections, and their ensemble (2.5D). Results are reported as Mean ± Standard Deviation. The best model per orientation is highlighted in **bold**.

Style	Attention U-Net	FPN	Trans U-Net	U-Net
Orientation				
2DAxi	0.479±0.236	**0.502±0.225**	0.496±0.227	0.481±0.221
2DCor	0.78±0.164	**0.785±0.161**	0.777±0.167	0.718±0.187
2DSag	0.762±0.182	**0.767±0.173**	0.757±0.179	0.709±0.194
2.5D	0.767±0.192	**0.779±0.176**	0.77±0.182	0.726±0.194

Net: $p = 0.001$). This indicates a potential room for improvement. Our multi-site validation addresses a critical gap in previous segmentation studies, which were often limited to single-institution or publicly available datasets (Inkeaw et al., 2022). The consistent performance across sites suggests that the advanced architectures learn feature representations that are robust to site-specific variations in imaging protocols, making them suitable for broader clinical deployment.

When contextualized within the broader literature, our multi-site results demonstrate competitive, if not superior, performance and generalizability. While (Inkeaw et al., 2022) reported a median Dice coefficient of 0.37 for IPH segmentation (in multi ICH segmentation) and (Lin et al., 2025) achieved Dice scores around 0.91 for cerebral contusion segmentation, our FPN model achieved an F-measure of 0.868 (in Dice metric) while demonstrating robust multi-site performance. The CNN variant efficiencies of models like FPN, Attention U-Net, and Trans U-Net suggest a promising path toward developing solutions that are both highly accurate and computationally feasible for real-time use in emergency settings across multiple healthcare institutions (Piao et al., 2023).

6. Summary and Conclusions

In this work, we investigated the use of CNN variants for IPH segmentation, aligning with current findings on the best techniques in the literature. In essence, we tested statistical models to assess the best CNN variant in our experiments.

Our findings demonstrate that advanced deep learning architectures, specifically Attention U-Net, FPN, and Trans U-Net, significantly improve automated segmentation of intraparenchymal hemorrhage in NCCT scans, as measured by F-Measure and IoU. For boundary definition, Attention U-Net and FPN achieved substantially better performance than the baseline U-Net. In contrast, Trans U-Net yielded comparable volumetric overlap but did not yield a significant improvement in boundary precision. The most notable benefits of the advanced models were observed in minor, more challenging hemorrhages. By delivering more precise and reliable segmentation across a variety of sites in Canada, this work shows the potential to optimize labor-intensive manual IPH segmentation, thereby

Table 4: d_{H95} for IPH Segmentation. Performance is compared across four CNN variants (Attention U-Net, baseline U-Net, FPN, and Trans U-Net), evaluated on axial (2DAxi), coronal (2DCor), sagittal (2DSag) projections, and their ensemble (2.5D). Results are reported as Mean ± Standard Deviation. Values are in mm. The best model per orientation is highlighted in **bold**.

Style	Attention U-Net	FPN	Trans U-Net	U-Net
Orientation				
2DAxi	**10.066±9.643**	11.624±8.507	10.371±8.964	11.465±9.336
2DCor	3.376±9.066	**2.916±7.464**	4.855±12.558	5.862±11.051
2DSag	4.578±11.038	**4.203±10.113**	5.594±10.934	6.871±10.526
2.5D	1.6±4.974	**1.574±4.55**	1.615±4.837	2.542±6.356

reducing stroke care time. We plan to integrate these models into clinical workflows and extend this work to segment and classify other ICH subtypes.

Despite these promising results, some limitations should be acknowledged. While the advanced models excelled across most IPH volumes, segmenting very small or nascent bleeds remains challenging. Future work could explore progressive resolution training or optimized loss functions to improve sensitivity and reduce misclassifications of hemorrhage voxels. Additionally, although our models demonstrated robust performance across multiple sites, further validation on external datasets from diverse geographic and demographic populations will be essential to confirm true generalizability.

References

Yasser B Abulhasan, Jeanne Teitelbaum, Karim Al-Ramadhani, Kyle T Morrison, and Michele R Angle. Functional outcomes and mortality in patients with intracerebral hemorrhage after intensive medical and surgical support. *Neurology*, 100(19):e1985–e1995, 2023. doi: 10.1212/WNL.0000000000207132.

S. Nafees Ahmed and P. Prakasam. Intracranial hemorrhage segmentation and classification framework in computer tomography images using deep learning techniques. *Scientific Reports*, 15:17151, 2025.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *CoRR*, abs/2102.04306, 2021. URL https://arxiv.org/abs/2102.04306.

Kauê T. N. Duarte, Abhijot S. Sidhu, Murilo C. Barros, David G. Gobbi, Cheryl R. McCreary, Feryal Saad, Richard Camicioli, Eric E. Smith, Mariana P. Bento, and Richard Frayne. Multi-stage semi-supervised learning enhances white matter hyperintensity segmentation. *Frontiers in Computational Neuroscience*, 18, 2024. ISSN 1662-5188. doi: 10.3389/fncom.2024.1487877. URL https://www.frontiersin.org/journals/computational-neuroscience/articles/10.3389/fncom.2024.1487877.

Valery L Feigin, Michael Brainin, Bo Norrving, Sheila O Martins, Jeyaraj Pandian, Patrice Lindsay, Maria F Grupper, and Ilari Rautalin. World stroke organization: Global stroke fact sheet 2025. *International Journal of Stroke*, 20(2):132–144, 2025. doi: 10.1177/17474930241308142.

Andrew Hoopes, Jocelyn S. Mora, Adrian V. Dalca, Bruce Fischl, and Malte Hoffmann. Synthstrip: skull-stripping for any brain image. *NeuroImage*, 260:119474, 2022. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2022.119474.

Papangkorn Inkeaw, Salita Angkurawaranon, Piyapong Khumrin, Nakarin Inmutto, Patrinee Traisathit, Jeerayut Chaijaruwanich, Chaisiri Angkurawaranon, and Imjai Chitapanarux. Automatic hemorrhage segmentation on head ct scan for traumatic brain injury using 3d deep learning model. *Computers in Biology and Medicine*, 146:105530, 2022.

Tsung-Yi Lin, Piotr Dollár, and et al. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017. doi: 10.1109/CVPR.2017.106.

Xinxin Lin, Emniao Zou, Wenci Chen, Xinxin Chen, and Le Lin. Advanced multi-label brain hemorrhage segmentation using an attention-based residual u-net model. *BMC Medical Informatics and Decision Making*, 25:286, 2025.

Bijoy K Menon, Brian H Buck, Nishita Singh, Yan Deschaintre, Mohammed A Almekhlafi, Shelagh B Coutts, Sibi Thirunavukkarasu, and et al. Intravenous tenecteplase compared with alteplase for acute ischaemic stroke in canada (act): a pragmatic, multicentre, open-label, registry-linked, randomised, controlled, non-inferiority trial. *The Lancet*, 400 (10347):161–169, 2022. ISSN 0140-6736. doi: https://doi.org/10.1016/S0140-6736(22)01054-6.

Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention U-Net: Learning where to look for the pancreas, 2018. URL https://doi.org/10.48550/arXiv.1804.03999.

Zhegao Piao, Yeong Hyeon Gu, Haillin Jin, and Seong Joon Yoo. Intracerebral hemorrhage ct scan image segmentation with hardnet based transformer. *Scientific Reports*, 13:7208, 2023.

Sean P. Polster, Julián Carrión-Penagos, Seán B. Lyne, Barbara A. Gregson, Ying Cao, Richard E. Thompson, Agnieszka Stadnik, Romuald Girard, Patricia Lynn Money, Karen Lane, Nichol McBee, Wendy Ziai, W. Andrew Mould, Ahmed Iqbal, Stephen Metcalfe, Yi Hao, Robert Dodd, Andrew P. Carlson, Paul J. Camarata, Jean-Louis Caron, Mark R. Harrigan, Mario Zuccarello, A. David Mendelow, Daniel F. Hanley, Issam A. Awad, on behalf of the MISTIE III, STICH I, and II investigators. Intracerebral hemorrhage volume reduction and timing of intervention versus functional benefit and survival in the mistie III and STICH trials. *Neurosurgery*, 88(5), 2021. ISSN 0148-396X. doi: 10.1093/neuros/nyaa572. URL https://journals.lww.com/neurosurgery/fulltext/2021/05000/intracerebral_hemorrhage_volume_reduction_and.7.aspx.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4. URL https://doi.org/10.48550/arXiv.1505.04597.

Peter M Rothwell, A J Coull, M F Giles, S C Howard, L E Silver, L M Bull, S A Gutnikov, P Edwards, D Mant, C M Sackley, et al. Change in stroke incidence, mortality, case-fatality, severity, and risk factors in oxfordshire, uk from 1981 to 2004 (oxford vascular study). *The Lancet*, 363(9425):1925–1933, 2004. doi: 10.1016/S0140-6736(04)16405-2.

Snehashis Roy, Sean Wilkes, Ramon Diaz-Arrastia, John A. Butman, and Dzung L. Pham. Intraparenchymal hemorrhage segmentation from clinical head ct of patients with trau-

matic brain injury. In *Medical Imaging 2015: Image Processing*, volume 9413, page 94130I. SPIE, 2015.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556*, 09 2014.

Daniel Woo, Jane Khoury, Mary A Haverbusch, Padmini Sekar, Matthew L Flaherty, Charles J Moonaw, Pooja Khatri, Opeolu Adeoye, Dawn Kleindorfer, and Joseph P Broderick. Risk factors associated with mortality and neurologic disability after intracerebral hemorrhage in a racially and ethnically diverse cohort. *JAMA Network Open*, 5 (3):e221103–e221103, 2022. doi: 10.1001/jamanetworkopen.2022.1103.

Donghao Zhang, Yimin Chen, Kauê TN Duarte, Taha Aslan, Mohamed AlShamrani, Brij Karmur, Yan Wan, Shengcai Chen, Bo Hu, Bijoy K Menon, and Wu Qiu. Benchmarking dinov3 for multi-task stroke analysis on non-contrast ct, 2025. URL https://arxiv.org/abs/2509.23132.