Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

3D Ischemic Stroke Lesion Segmentation in MRI with Attention U-Net and Voxel-Level Uncertainty Quantification

Outmane Labaybi

outmane.labaybi@usmba.ac.ma

Sidi Mohamed Ben Abdellah University

Mohamed Bennani Taj

Sidi Mohamed Ben Abdellah University

Khalid Elfahssi

Sidi Mohamed Ben Abdellah University

Rachid Flouchi

Sidi Mohamed Ben Abdellah University

Mohamed EL FAR

Sidi Mohamed Ben Abdellah University

Article

Keywords: Ischemic stroke segmentation, Deep learning, Monte Carlo Dropout, Uncertainty Quantification, Attention U-Net, ISLES 2022, Clinical decision support

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10744875/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

3D Ischemic Stroke Lesion Segmentation in MRI with Attention U-Net and Voxel-Level Uncertainty Quantification

Outmane LABAYBI[1]*, Mohamed Bennani Taj[1], Khalid Elfahssi [1], Rachid Flouchi[2], Mohamed EL FAR[1]

[1] Laboratory of Computer Science, Innovation, and Artificial Intelligence (L3IA), Faculty of Science Dhar El Mahraze, Sidi Mohamed Ben Abdellah University, Fez, Morocco
[2] Laboratory of Microbial Biotechnology and Bioactive Molecules, Science and Technologies Faculty, University Sidi Mohamed Ben Abdellah, Fez, Morocco

***Corresponding author(s). E-mail(s) outmane.labaybi@usmba.ac.ma**

Abstract

Ischemic stroke is a major cause of death and disability around the world, and accurate segmentation of stroke lesions from MRI diagnosis are critical for treatment planning and clinical decision making. Compared to traditional methods, while deep learning has advanced automated stroke segmentation, but existing approaches predict single binary masks with no estimate of prediction confidence, a crucial limitation since clinical usage would require knowledge on where the model can be trusted just as much as the prediction itself. We propose a reliability-aware 3D framework combining an Attention U-Net with Monte Carlo Dropout inference and an optimized decision threshold, tested on the ISLES 2022 benchmark (240 multi-center DWI+ADC cases; 168/36/36 split). At inference, T = 20 stochastic forward passes generate both a segmentation mask and a voxel-level uncertainty map. This work achieved a Dice of 0.6816 (IoU = 0.5526, Precision = 0.8764 and Recall = 0.6595) for single-model volumetric methods on this benchmark dataset, meanwhile boiling Monte Carlo Dropout increased Recall by +2.65% at the zero-training cost. Uncertainty maps were semantically aligned but spatially localized to the edges of lesions, highly correlated with segmentation errors, and inversely associated with lesion size automatically marking small and difficult lesions as cases of high uncertainty.

Key words: Ischemic stroke segmentation; Deep learning; Monte Carlo Dropout; Uncertainty Quantification; Attention U-Net; ISLES 2022; Clinical decision support.

1 Introduction

Ischemic stroke continues to be among the top ten causes of mortality and long-term disability globally, affecting almost 15 million people per year [1] [2]The accurate and timely segmentation of stroke lesions from neuroimaging data is important for clinical decision-support, treatment planning, and tracking disease progression. Diffusion-weighted imaging (DWI) and apparent diffusion coefficient maps (ADC) represent the diagnostic gold standard for acute ischemic stroke evaluation on magnetic resonance Imaging (MRI) [3].

Although impressive advances have been made in deep learning based medical image segmentation[4], fully automated delineation of ischemic stroke lesions remains one of the most difficult tasks. Stroke lesions are highly heterogeneous in terms of size, shape, and anatomical distribution, from small punctiform embolic infarcts to large territorial strokes involving whole vascular territories [5]. This variability and the presence of high-class imbalance between lesion and background voxels leads to a

systematic bias towards under-detection of small infarcts which are nevertheless among the most clinically relevant findings in embolic shower presentations.

Beyond the segmentation accuracy, a second fundamental limitation is prevalent across existing literature: all current state-of-the-art deep learning models for stroke lesion segmentation function as deterministic black-box predictors. It takes as output a single binary mask with no measure of confidence over its own predictions. [6] In a clinical context, this is a critical shortcoming. A radiologist presented with an automatically generated lesion mask has no way of knowing which boundaries the model delineated with high confidence and which are uncertain, information that is critical for calibrating trust in the system and determining what areas need to be manually reviewed.

This limitation is further compounded by an architectural trend that plays out in the field. The volumetric nature of MRI data makes a strong case for fully three-dimensional processing pipelines [7] but many recent high-performing methods on the ISLES 2022 benchmark still utilize two-dimensional slice-wise segmentation, which ignores the within-slice spatial consistency that characterizes the complicated 3D shape of lesions. Even for methods using volumetric processing, or a combination of the best methods on ISLES 2022 challenge, none of which provide a mechanism for quantifying prediction uncertainty.

Monte Carlo Dropout has been shown to be a computable, yet not modifying the structure architecture other than running dropout on during inference and has been shown to be useful for uncertainty estimation in the context of other medical segmentation applications such as brain tumor delineation. To the best of our knowledge, this approach has not been yet used for three-dimensional ischemic stroke lesion segmentation on the ISLES2022 benchmark.

This work addresses both limitations through a reliability-aware and fully three-dimensional segmentation framework. Specifically, this study makes the following contributions:

First, we adopt a fully volumetric 3D pipeline that processes complete DWI and ADC brain volumes rather than independent axial slices, preserving the spatial continuity required to characterize heterogeneous and multifocal lesion morphologies.

Second, we introduce voxel-level uncertainty quantification via Monte Carlo Dropout, enabling stochastic inference at test time. By performing multiple forward passes with dropout active, the proposed framework generates not only a segmentation mask but also a spatially resolved confidence map that explicitly flags ambiguous regions for clinical review, a capability that, to our knowledge, has not been demonstrated by any existing method on the ISLES 2022 benchmark.

Third, we conduct a systematic three-stage ablation study on the ISLES 2022 benchmark, isolating the individual and combined contributions of Monte Carlo Dropout, an Attention U-Net backbone, and an optimized decision threshold, with particular emphasis on Recall, the most clinically critical metric in acute stroke, where missed ischemic tissue carries greater risk than over-segmentation.

The remainder of this paper is organized as follows. Section 2 reviews related work in stroke lesion segmentation and uncertainty quantification. Section 3 details the materials and methods used. Section 4 presents experimental results. Section 5 discusses the clinical implications and limitations of the proposed framework. Section 6 concludes the paper.

2 Related Work

In the last decade, segmentation of ischemic stroke lesions from MRI has been the subject of many research efforts, driven by the clinical need for precise and reproducible delineation of infarcted tissue in time-critical therapeutic settings. Standardized public benchmarks particularly the ISLES series of challenges [5] have also given a major boost to progress by supplying multi-center, multi-vendor datasets with expert annotations that can facilitate systematic comparisons across methods. As a result, many deep learning architectures have been proposed and benchmarked on these datasets, ranging from

2D slice-wise systems to fully volumetric 3D pipelines as well as transformer-based and hybrid convolutional-transformer approaches. Overall, while there exists a large variety of contributions to the field and while we have seen many improvements in reported segmentation metrics through consecutive benchmark editions, some basic limitations are commonly overlooked across publications, namely, slice-wise methods cannot exploit the continuity of volumetric lesions and none of any published method (2D and 3D-based) includes a per-instance prediction confidence mechanism. We evaluate the most prominent work along these dimensions in the following sections, while highlighting specific gaps that motivate our own work.

2.1 Two-Dimensional Slice-Wise Approaches

Convolutional Neural Networks (CNN), and U-Net architectures [4], have now established themselves as the state-of-the-art framework for medical image segmentation with their ability to combine local and global spatial features via encoder-decoder structures with skip connections. However, since many works on ischemic stroke lesion segmentation have used two-dimensional slice-wise processing as default and treating each axial MRI slice as an independent input sample.

Moon et al. [8] evaluated 2D and 3D U-Net architectures on both DWI and FLAIR modalities from a dataset of 79 acute ischemic stroke patients, with the optimal 2D configuration yielding a mean Dice score of 0.737. The study described ongoing challenges classifying brainstem lesions, which authors directly ascribed to the systemic loss of inter-slice spatial context that is an explicit consequence of slice-wise processing, an explicit acknowledgment that the two-dimensional paradigm imposes a structural ceiling on segmentation quality in anatomically complex.

García-Salgado et al. [9] proposed a lightweight Attention U-Net enhanced with a Generalized Dice Focal Loss function, evaluated across FLAIR, DWI, and T2 modalities on ISLES 2015 and ISLES 2022, achieving F1-scores above 0.70. While the use of focal loss to address class imbalance represents a meaningful methodological contribution, the model relies on two-dimensional slice-wise processing and requires extensive hyperparameter optimization. More critically, like all two-dimensional methods reviewed here, the approach provides no mechanism to capture three-dimensional lesion continuity and delivers no measure of prediction confidence to support clinical decision-making.

Rahman et al. [10]. Wang et al.[11] proposed a more advanced encoder-decoder architecture based upon a DenseNet121 for the encoder and a Self-Organized Operational Neural Network as an intelligent decoder, further improved with two-dimensional channel-and-space compound attention blocks, obtaining a Dice score of 0.8749 on ISLES 2022; the highest of reported results on ISLES 2022 achieved by pure two-dimensional methods to date. Therefore, despite this strong result, the method is still fundamentally limited due to its slice-wise design; each axial slice is segment independently from neighboring slices and does not allow for characterization of three-dimensional continuity of complex lesion patterns. Along with the total absence of any prediction confidence mechanism, the authors themselves highlight this limitation as an open limitation for clinical deployment.

Consistency with all three studies is the observation of two structural limitations that transcend reported accuracy: 2D processing lacks the ability to fully capture volumetric lesion morphology, and there is no mechanism universally present for communicating prediction reliability back to the clinical end user.

2.2 Three-Dimensional Volumetric Approaches

Three-dimensional volumetric methods address the spatial continuity limitation of slice-wise approaches by processing the full MRI brain volume at once. This preserves the inter-slice relationships that two-dimensional methods discard, enabling more coherent delineation of complex three-dimensional lesion morphologies.

Soleimani and Farezi [12] evaluated a model comprising multiple pre-trained encoder backbones, including ResNetV2 and Seresnext101, under a 3D U-Net framework for stroke lesion segmentation

that achieved F1-scores reaching up to 0.942. Nonetheless, these results were derived from the substantially smaller ISLES 2015 subset of only 24 patients utilizing FLAIR modality alone. his raises serious concerns about generalizability to multi-center, multi-vendor settings reflecting real world clinical variability. In addition, the study uses only deterministic predictions and does not include a way to assess or communicate uncertainty in the predictions.

Ruthra and Bevi [13] achieved a Dice score of 0.85 on validation (external cross-validation - no patient overlap) and 0.89 on test set using FLAIR modality only with the proposed six parallel convolutional pathways per block multi-path residual U-Net evaluated on ISLES 2015 dataset (n=28). The proposed study suffers from the same basic limitations as the existing literature: very small and homogeneous dataset (not representative of current multi-centre imaging protocols) and no uncertainty quantification, although multi-path designs present promise in capturing multi-scale features. The authors themselves pointed out, in bold and italics answer to question 15 in their supplementary material, that "Uncertainty quantification providing confidence estimates for clinical decision-making is a key direction for future work", clearly indicating the shortcoming this work aims to address.

Wu et al. [14] W-Net with Convolutional Tracking W-Net combined convolutional components and transformer-based components in a novel W-Net architecture for boundary-aware stroke lesion segmentation on ISLES 2022. Another approach proposed by Ling, Pei & Wu [15] was the introduction of a dual-stream architecture merged deep supervision with multi-path feature fusion. Both were able to show that volumetric processing, as a full 3D technique, has great advantages over slice-wise alternatives on a modern and difficult multi-center benchmark.

Notably, the best performances on ISLES 2022 are obtained by multi-model ensembles and complicated preprocessing pipelines rather than single model architecture development. More crucially, within all these volumetric methods from the first baselines to the latest volume ensemble solutions, no single published method at this point estimates prediction uncertainty per voxel. Each operates as a deterministic predictor output only a binary segmentation mask with no indication of model confidence and uncertainty.

2.3 Uncertainty Quantification for Medical Image Segmentation

Quantifying the predictive uncertainty of deep learning models has been recognized as an important prerequisite for safe clinical deployment but is almost completely neglected in the context of ischemic stroke lesion segmentation. Gal, Y. and Ghahramani [16] established Monte Carlo Dropout as a simple and computationally efficient approximation to Bayesian inference: keeping dropout layers active for inference, performing multiple stochastic forward passes yields a distribution over predictions, from which both a mean estimate and predictive variance are derived with no modifications to architecture and within the same training cost.

Wang et al.[17] demonstrated the practical value of this approach for brain tumor segmentation and showed that predictive variance over stochastic forward passes has meaningful voxel-level correlation with segmentation error, marking the most important ambiguous boundary regions where radiologist review is needed. Sasidharan et al. [18] Tested in the common domain of stroke, this confirmatory gap confirmed this distance from clinical processes by stating explicitly that none of the existing models for stroke segmentation contain a mechanism for communicating whether predictions are reliable to users in a clinical setting.

2.4 Research Gap and Positioning of the Present Work

Looking at the literature, two important gaps stand out. The first is architectural: many recent methods still process MRI volumes slice by slice, losing the three-dimensional spatial context that is essential for accurately characterizing stroke lesions a limitation that the authors of these methods themselves acknowledge. The second gap is perhaps more clinically important: while uncertainty quantification has been explored in other medical imaging tasks such as brain tumor segmentation, it has received very

little attention in the specific context of ischemic stroke lesion segmentation, and no method evaluated on the ISLES 2022 benchmark provides this capability. This matters in practice when a radiologist receives an automatic segmentation mask, they have no way of knowing which boundaries the model drew with confidence and which ones it was unsure about. Without this information, the entire mask needs to be reviewed manually, which significantly reduces the efficiency gains that automated segmentation is supposed to provide. The present work addresses both gaps within a single unified framework. We adopt a fully three-dimensional Attention U-Net pipeline that processes complete DWI and ADC brain volumes, directly resolving the spatial continuity limitation common to two-dimensional approaches. We further integrate Monte Carlo Dropout to generate voxel-level uncertainty maps at inference time, providing radiologists with a spatially resolved confidence signal that identifies regions requiring targeted review. The present contribution should therefore be understood not as a competing accuracy benchmark, but as a step toward clinically deployable stroke segmentation one that delivers, alongside the segmentation mask, the confidence information that clinical end users need to calibrate their trust in the system.

3 Materials and Methods

This section describes the materials and methods used in this study. We first present the ISLES 2022 dataset, including its imaging modalities and the training, validation, and test split (Section 3.1). We then introduce the proposed RA-3D UNet framework (Section 3.2), followed by a detailed description of its individual components: the preprocessing pipeline, the baseline 3D U-Net, the attention gate mechanism, the Monte Carlo Dropout inference used for voxel-level uncertainty estimation, the inference algorithm, and the evaluation metrics (Sections 3.3–3.8).

3.1 Dataset Description

All experiments were conducted on the ISLES 2022 dataset [5], available publicly on Kaggle, the dataset includes 400 cross-center multi-vendor MRIs from three European stroke centers: University Hospital of the Technical University Munich (Center 1, Germany), University Hospital of Bern (Center 2, Switzerland) and University Medical Center Hamburg-Eppendorf (Center 3, Germany).

Each case consisted of three multimodal MRI sequences collected as part of clinical standards for imaging strokes:

- DWI (Diffusion-Weighted Imaging, b-value = 1000 s/mm²): the primary modality for acute stroke diagnosis, demonstrating hyperintense signal in regions of cytotoxic edema within minutes of ischemic onset.
- ADC (Apparent Diffusion Coefficient): a quantitative map derived from DWI that exhibits hypointense signal in infarcted tissue, providing complementary information that enables discrimination of true diffusion restriction from T2 shine-through artifacts.
- FLAIR (Fluid-Attenuated Inversion Recovery): a structural sequence suppressing cerebrospinal fluid signal that reveals vasogenic edema and chronic lesion changes, particularly useful for distinguishing acute infarcts from old glial scars and skull-base artifacts.

The ISLES 2022 dataset presents a highly imbalanced distribution of lesion volumes, which is one of the primary challenges driving the choice of architecture and loss function in this work. Table1 summarizes the lesion size distribution across the 240 retained cases.

table 1. Lesion volume distribution across the 240 ISLES 2022 cases

Category	Volume range	Cases	Proportion
Very small	< 150 voxels	52	21.7%
Small	150 – 500 voxels	42	17.5%
Medium	500 – 5,000 voxels	109	45.4%
Large	> 5,000 voxels	37	15.4%

More than 39% of cases have a lesion volume less than 500 voxels, such lesions are amongst the most clinically relevant findings small embolic infarcts are common with cardioembolic stroke mechanisms and have important implications for secondary prevention. More importantly, they are a very difficult class to automate segmentation due to their small size resulting in extreme sensitivity to class imbalance and partial volume effects. It directly motivates the uncertainty analysis, which shows that our MC Dropout framework assigns high predictive variance to small lesion cases and giving a built-in measure of reliability that is useful for this most challenging group.

Example multimodal MR images of three clinically distinct cases from the dataset (multi-focal embolic pattern with multiple small punctiform infarcts (835 voxels, sub-strokecase0001); single small lesion in this challenging punctiform infarct category (90 voxels, sub-strokecase0053) and a medium-sized territorial stroke case (1,682 voxels, label=1) (sub-strokecase0105) are depicted in Figure 1.

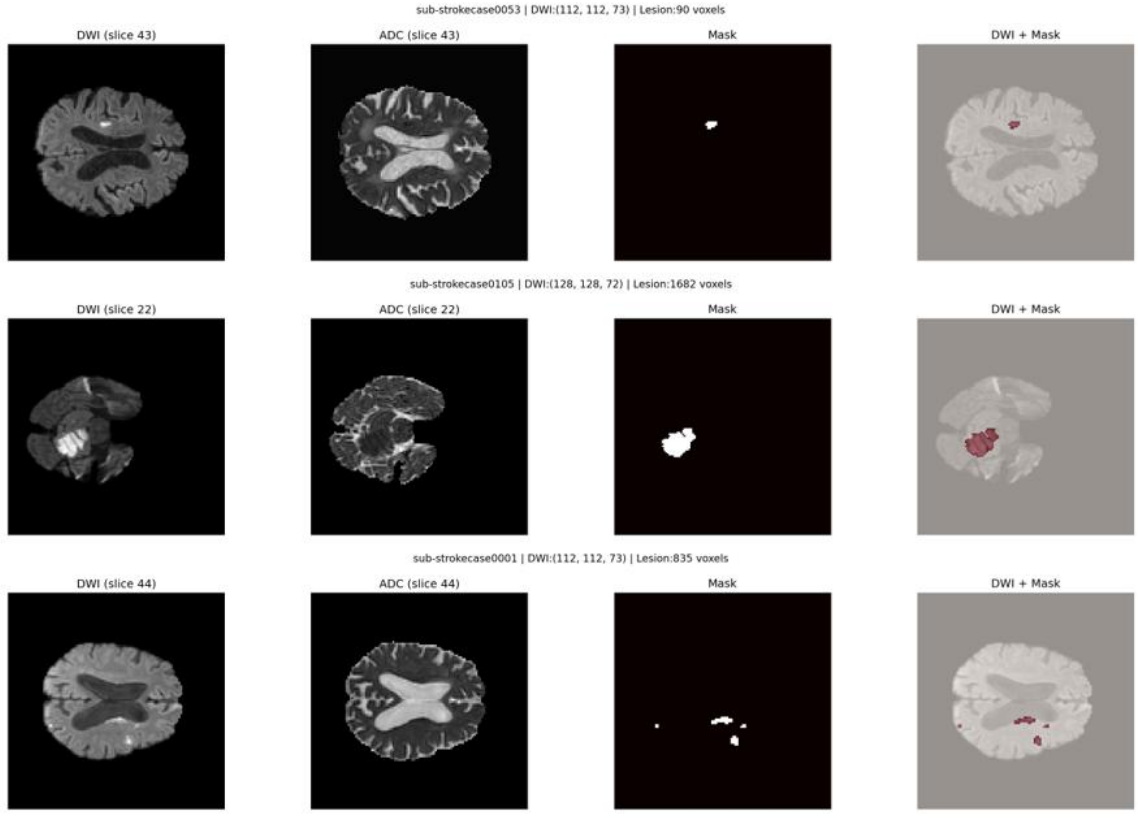


Figure 1. Dataset visualization: DWI, ADC, FLAIR, Mask, DWI+Mask overlay for three representative cases

3.2 Overview of the Proposed Model

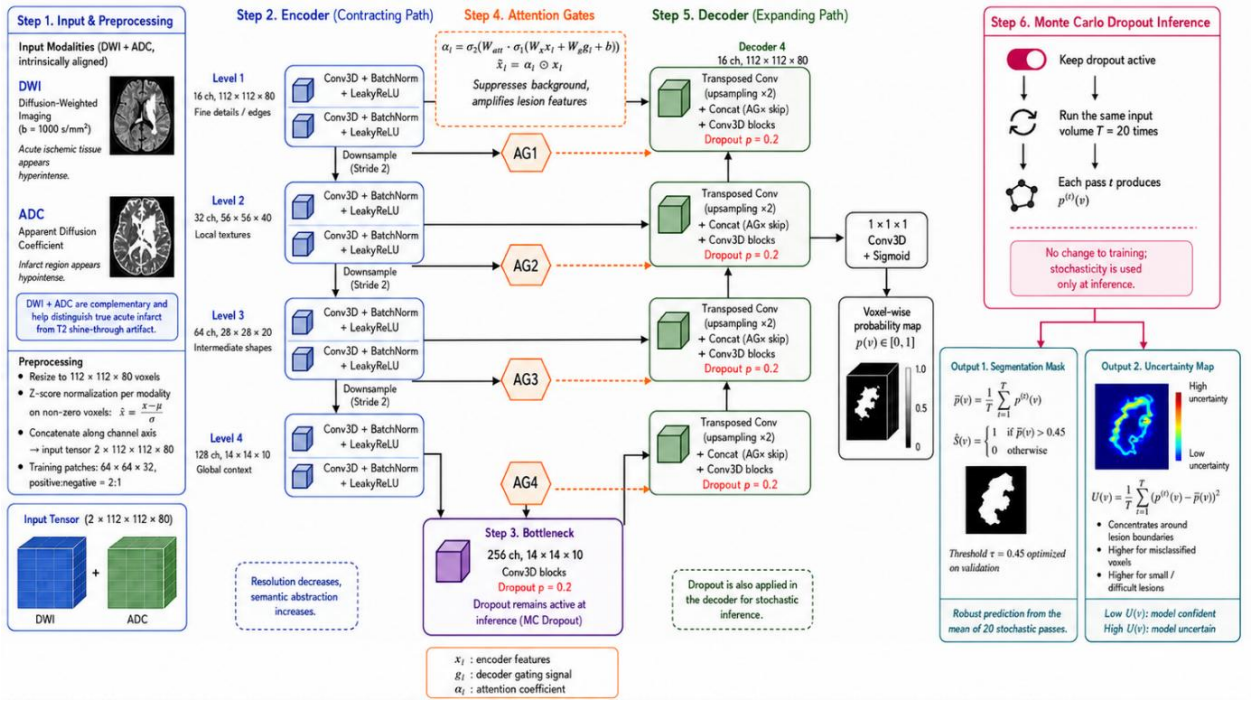


Figure 2. Architecture of the proposed RA-3D UNet Framework

Architecture of the proposed RA-3D UNet (Reliability-Aware 3D Attention U-Net). The pipeline takes two intrinsically aligned diffusion MRI modalities; DWI (b = 1000 s/mm²) and ADC concatenated into a dual-channel input volume of size 2 × 112 × 112 × 80, following z-score normalization per modality (Step 1). The encoder (Step 2) extracts hierarchical features across four resolution levels with channel progression 16 → 32 → 64 → 128, halving the spatial resolution at each level through strided convolutions. At the bottleneck (Step 3), a 256-channel representation encodes the most abstract volumetric features; the dropout layers introduced here (p = 0.2) remain active at inference to enable Monte Carlo Dropout. On each skip connection, an attention gate (Step 4) computes a spatial coefficient α_l that amplifies lesion relevant features and suppresses background before fusion with the decoder. The decoder in the step 5 mirrors the encoder, progressively upsampling through transposed convolutions and fusing the attention-filtered skip features, with dropout retained at every level. A final 1 × 1 × 1 convolution followed by a sigmoid produces a voxel-wise probability map. At inference (Step 6), the same input is passed through the network T = 20 times with dropout active, yielding T stochastic probability maps $p^{(t)}(v)$. These are aggregated into two outputs: the segmentation mask, obtained by thresholding the mean prediction $\bar{p}(v)$ at τ = 0.45 (optimized on the validation set), and the voxel-level uncertainty map, computed as the predictive variance U(v) across the T passes. High uncertainty concentrates at lesion boundaries and small or difficult lesions, while low uncertainty indicates confident predictions. This dual output enables a human-in-the-loop clinical workflow in which low-uncertainty regions are accepted directly and radiologist review is focused on high-uncertainty regions.

3.3 Preprocessing

Each volume was processed with a standardized and reproducible preprocessing pipeline. To preserve the native spatial resolution of each acquisition, non-isotropic resampling was not performed; rather, each DWI and ADC volume was padded or cropped to a consistent size of 112×112×80 voxels. The two modalities were concatenated together over the channel dimension resulting in a dual-channel input tensor with shape 2×112×112×80, giving the model the ability to extract synergistic diffusion

information given that DWI represents regions of restricted diffusion typical of acute ischemia and ADC reflected quantitative diffusion which is hypointense within infarcted tissue.

Z-score intensity normalization was applied independently to each modality over non-zero voxels, following standard practice in volumetric medical image segmentation Isensee et al. [19]:

$$\widehat{x_c} = \frac{x_c - \mu_c}{\sigma_c} \tag{1}$$

where $\widehat{x_c}$ denotes the raw intensity values of modality c, and μ_c, σ_c are the mean and standard deviation computed over non-zero voxels of that modality.

3.4 Baseline 3D U-Net

We train a baseline segmentation model with the standard encoder-decoder architecture with skip connections. The encoder comprises four resolution levels with progressive channel doubling (16 → 32 → 64 → 128 → 256). At first, two residual convolutional units are used per level, and each unit comprises of a 3×3×3 follows by batch normalization and Leaky ReLU activation. Strided convolutions with stride 2 carry out downsampling. The decoder uses the same structure as the encoder but instead of convolutions, it includes transposed convolutions for upsampling operations and skip connections that concatenate consecutive feature maps produced by the last layer of each resolution in the encoder except for the last one. A final 1×1×1 convolution and sigmoid activation generates a voxel-wise likelihood. Learned parameters: 4.8M

3.5 Attention Gate Mechanism

Attention U-Net [20], an extended version of the standard U-Net with soft attention gates replacing standard skip connections to amplify lesion-relevant features in generating output while suppressing background activations. The benefit of this mechanism is most pronounced for the ISLES 2022 dataset, where lesions constitute around only ~4% of brain volume and have far-extreme morphological variability.

For each encoder feature map $\boldsymbol{x_l}$ at resolution level l, the attention coefficient α_l is computed from the encoder features and the decoder gating signal $\boldsymbol{g_l}$ as follows:

$$\alpha_l = \sigma_2\left(W_{\text{att}} \cdot \sigma_1\left(W_x \cdot \boldsymbol{x_l} + W_g \cdot \boldsymbol{g_l} + b\right)\right) \tag{2}$$

where σ_1 is a ReLU activation function, σ_2 is a sigmoid function that produces a spatial attention coefficient in [0,1], b is a bias term, and W_x, W_g, W_{att} are learnable linear transformations. The filtered skip connection passed to the corresponding decoder level is:

$$\widetilde{\boldsymbol{x_l}} = \alpha_l \odot \boldsymbol{x_l} \tag{3}$$

where $\odot$ denotes element-wise multiplication. Voxels with $\alpha_l \approx 1$ are selectively amplified as lesion-relevant, while voxels with $\alpha_l \approx 0$ are suppressed as irrelevant background.

3.6 Monte Carlo Dropout for Uncertainty Quantification

A standard neural network always gives the same output for every given input, it is a deterministic system. That is, when a stroke lesion is segmented, the model produces a single binary mask with no indication of the confidence it has in each voxel on that prediction. This is a fundamental problem in the clinical context: the radiologist has no way of knowing which boundaries the model drew with certainty and which ones it was unsure about. To address this limitation, we integrate Monte Carlo Dropout (Gal

& Ghahramani, [16] a simple and effective technique that transforms a standard neural network into a probabilistic model capable of estimating its own uncertainty, without any modification to the training procedure or the network architecture.

Dropout randomly deactivates a fraction of neurons during training to prevent overfitting and is conventionally disabled at inference to produce a deterministic output. Monte Carlo Dropout keeps it active at inference instead: running T forward passes through the same network deactivates a different random subset of neurons each time, so each pass yields a slightly different prediction for the same input. Aggregating these T predictions gives both a robust mean estimate and a measure of their variability and the uncertainty.

In our framework, dropout layers with p =0.2 are placed in the bottleneck and decoder. At inference, T=20 stochastic forward passes are performed, each producing a voxel-wise lesion probability map $p^{(t)}(v)$

3.6.1 Mean Prediction and Segmentation Mask

The final segmentation probability at each voxel v is the mean of the T stochastic predictions:

$$\bar{p}(v) = \frac{1}{T}\sum_{t=1}^{T} p^{(t)}(v) \tag{4}$$

This mean map is more robust than a single deterministic prediction, as it integrates information from T different dropout configurations. The binary mask is obtained by thresholding:

$$\hat{S}(v) = \begin{cases} 1 \; if \; \bar{p}(v) > \tau \\ 0 \; otherwise \end{cases} \tag{5}$$

where τ =0.50 for Experiments 1 and 2, and τ =0.45 for Experiment 3. Lowering the threshold to 0.45 recovers borderline lesion voxels that the ensemble detects with moderate confidence, improving recall without significantly sacrificing precision.

3.6.2 Voxel-Level Uncertainty Map

The uncertainty at each voxel is the variance of the T predictions:

$$U(v) = \frac{1}{T}\sum_{t=1}^{T}\left(p^{(t)}(v) - \bar{p}(v)\right)^2 \tag{6}$$

The interpretation is direct. When $U(v)$≈0, all T passes agree, and the model is confident, this typically occurs in the lesion core and in clearly healthy tissue.

When $U(v)$ is high, the passes disagree and the model is uncertain at lesion boundaries, in small lesions with limited spatial context, and in regions affected by partial-volume effects.

3.7 Inference Algorithm

Algorithm 1. Monte Carlo Dropout inference for voxel-level uncertainty quantification

---

Input: $X \in \mathbb{R}^{2\times H\times W\times D}$ dual-channel input volume (DWI + ADC); f_θ trained network with dropout layers; $T = 20$ number of stochastic forward passes; $p = 0.2$ dropout probability; $\tau = 0.45$ decision threshold

Output: $\hat{S}$: binary segmentation mask; U : voxel-level uncertainty map

---

$\mathbb{P} \leftarrow \emptyset$ *// initialize empty set of predictions*
f_θ.eval() *// set BatchNorm layers to inference mode*
for each dropout layer $d \in f_\theta$ **do**
  d.train() *// keep dropout ACTIVE at inference*
end for
for $t = 1$ **to** T **do**
  $z^{(t)} \leftarrow f_\theta(X)$ *// stochastic forward pass*
  $p^{(t)} \leftarrow \sigma(z^{(t)})$ *// sigmoid → lesion probability map*
  $\mathbb{P} \leftarrow \mathbb{P} \cup \{ p^{(t)} \}$ *// store prediction*
end for
for each voxel v **do**
  $\bar{p}(v) \leftarrow (1/T) \sum_{t=1}^{T} p^{(t)}(v)$ *// mean prediction*
  $U(v) \leftarrow (1/T) \sum_{t=1}^{T} (p^{(t)}(v) - \bar{p}(v))^2$ *// predictive variance*
  $\hat{S}(v) \leftarrow \mathbb{1}[\bar{p}(v) > \tau]$ *// threshold to binary mask*
end for
return $\hat{S}$, U

3.8 Evaluation Metrics

All experiments use the DiceCE loss (Milletari et al. [21]) a composite loss function popular in medical image segmentation due to its property of handling class imbalance and providing voxel-level gradients for stability.

$$\mathcal{L}_{\mathrm{DcC}} = \mathcal{L}_{\mathrm{Dc}} + \mathcal{L}_{\mathrm{C}} \tag{7}$$

The Dice loss directly optimizes the volumetric overlap between the predicted probability map and a ground-truth binary mask, addressing the severe class imbalance between lesion and background voxels:

$$\mathcal{L}_{\mathrm{Dc}} = 1 - \frac{2\sum_v p(v)\ g(v) + \varepsilon}{\sum_v p(v) + \sum_v g(v) + \varepsilon} \tag{8}$$

where $p(v) \in [0,1]$ is the predicted probability at voxel v, $g(v) \in \{0,1\}$ is the ground-truth binary label, and $\varepsilon = 1$ is a smoothing constant to ensure numerical stability.

The binary cross-entropy loss provides dense voxel-level supervision with stable gradient signals throughout training:

$$\mathcal{L}_{\mathrm{C}} = -\frac{1}{N}\sum_{v=1}^{N}\left[g(v)\log p\,(v) + \big(1 - g(v)\big)\log\big(1 - p(v)\big)\right] \tag{9}$$

where N is the total number of voxels in the training patch.

Six complementary metrics in stroke lesion segmentation benchmarks are computed on the held-out test set:

Dice Similarity Coefficient (DSC) measures volumetric overlap:

$$\text{DSC} = \frac{2\,|P \cap G|}{|P| + |G|} \tag{10}$$

Intersection over Union (IoU): a stricter overlap measure:

$$\text{IoU} = \frac{|P \cap G|}{|P \cup G|} \tag{11}$$

Precision: fraction of predicted lesion voxels that are truly lesional:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{12}$$

Recall (Sensitivity): fraction of true lesion voxels correctly detected:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{13}$$

F1-Score harmonic mean of Precision and Recall, equivalent to DSC for binary segmentation:

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{14}$$

Accuracy overall fraction of correctly classified voxels:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

where P is the predicted binary mask, G is the ground-truth mask, and TP, TN, FP, FN denote true positives, true negatives, false positives, and false negatives respectively. Recall is considered the primary clinical metric in this work, as missed ischemic tissue, a false negative carries greater clinical risk than over-segmentation in the context of acute stroke treatment planning.

4 Results and discussion

4.1 Ablation Study

The table2 below presents the results of three experiments configuration, each model builds incrementally upon of the previous configuration, allowing the contribution of each experiment to be assessed independently and in combination.

table 2. Ablation study on the ISLES 2022 test set

Metric	3D U-Net Baseline	3D U-Net with Monte Carlo Dropout	Attention U-Net with MC Dropout and Optimized Threshold
Dice	60.76	60.82	68.16
IoU	47.77	47.50	55.26
Accuracy	99.89	99.89	99.90
Precision	85.39	82.21	87.64
Recall	62.74	65.39	65.95
F1-score	72.33	72.84	75.26

The **3D U-Net Baseline** trained on concatenated ADC and DWI volumes and achieved a dice score of 60.76%, Accuracy 99.89%, a precision of 85.39% and recall of 62.74%. the high precision relatively with low recall indicates that the model trends the under-segment lesion tissue. This behavior is clinically significant to incomplete treatment planning in acute management.

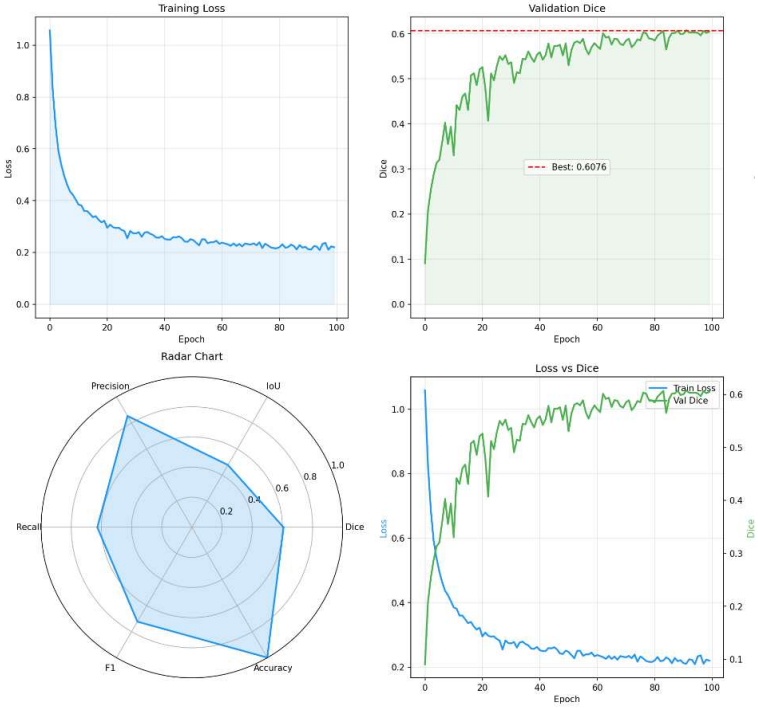


Figure 3. Metrics summary of 3D U-Net Baseline

In experiment 1 (3D U-Net Baseline), the training loss converges smoothly and rapidly, decreasing sharply within first 25 epochs before stabilizing. Upon convergence, the validation Dice curve exhibits low variance once converged, confirming model stability. The asymmetric profile of the metric high Precision (0.8539) while lower Recall (0.6274), indicates that the deterministic baseline is more conservative; it predicts only high-confidence lesion voxels.

The model 3D U-Net with Monte Carlo Dropout, we were Introducing Monte Carlo Dropout, which is the first component in the pipeline that defines our contribution. at inference time had a limited impact on Dice since Monte Carlo Dropout is primarily an uncertainty estimator rather that an accuracy booster. MC Dropout allows the creation of uncertainty maps at the voxel level, and this is a feature which all baselines tested on ISLES 2022 do not offer. In doing so, it converts a deterministic black-box prediction of segmentation output to a clinically interpretable tool that allows the radiologist to quickly see which boundaries are supported by the model and which need scrutiny.

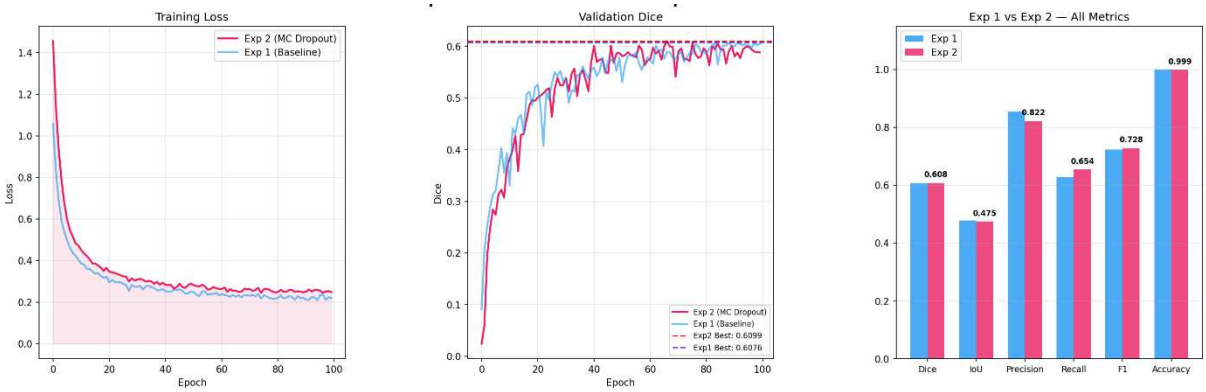
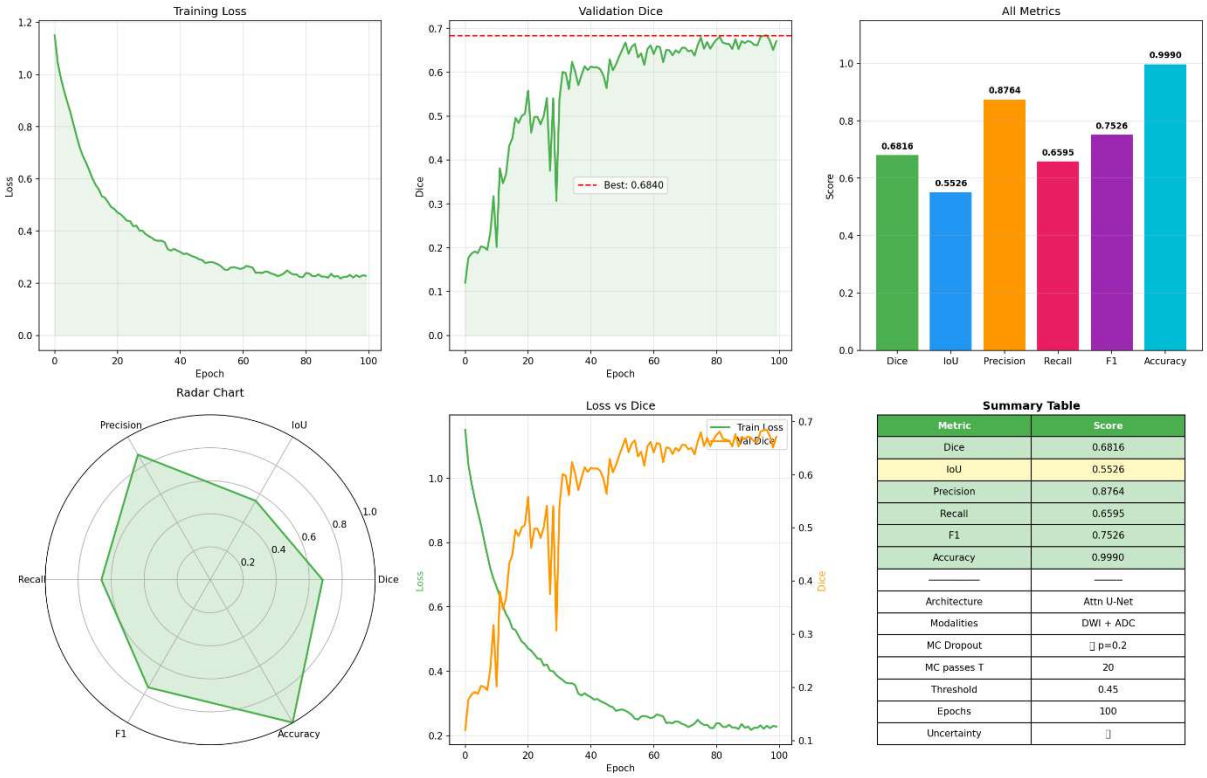


Figure 4. Experiment 2 3D U-Net with MC Dropout: training/validation loss, validation Dice, and evaluation metrics, compared against the Experiment 1 baseline

In Experiment 2 (Figure 4), we replicate training dynamics of experiment 1 nearly identically, the training loss curves for both experiments almost exactly overlap demonstrating dropout regularization neither hampers nor appreciably modifies the optimization landscape. This is a critical point: the whole MC Dropout benefit is brought in without any cost in training effort. The principal gain appears at inference time, where averaging T = 20 stochastic forward passes results in a +2.65% Recall improvement over the baseline model. The improvement in Dice for validation more than Experiment 1 (0.6076) is only slightly above that of the best validation Dice of 0.6099, show that dropout has a slight regularization effect during training.

For the proposed framework in experiment 3, we were combining an Attention U-Net architecture with MC Dropout and an optimized decision threshold of 0.45 achieved better performance for all metrics with a +7.40% over the baseline in terms of Dice score. Our framework extends and enhances the proposal put forth in Experiment 2 by combining three synergistic advancements. Firstly, adapting the backbone with an Attention U-Net where the attention gates learn to automatically focus on regions most relevant to lesions at every resolution-level and suppress irrelevant activations in healthy brain tissue. The hundreds of lesion morphologies in ISLES 2022 would benefit particularly from this mechanism, as the natural heterogeneity in size and shape across lesions, which could range from large hemispheric strokes to little cortical infarcts within a single scan makes lesion-background separation difficult. Second, the optimized threshold of 0.45 takes advantage of the soft-probability distribution generated by the MC Dropout ensemble: with a softer decision boundary, borderline voxels with intermediate probabilities are correctly classified as lesion tissue, obtaining viable ischemic tissue that would be dropped out using the original threshold of 0.50. Third most crucially, the maps of uncertainty at the voxel-level are no longer produced by an architecturally weaker model, which yields more precise and better-localized uncertainty signals at true lesion boundaries. Collectively, these elements create a novel clinically actionable scenario: the radiologist receives both optimal segmentation masks (+7.40% Dice over baseline), and improved recall at low risk of undetected disease whilst having spatial confidence maps to help convey where manual review might be most impactful as none of the existing 3D methods on ISLES 2022 test set provide each component alone.



Metric	Score
Dice	0.6816
IoU	0.5526
Precision	0.8764
Recall	0.6595
F1	0.7526
Accuracy	0.9990
————	——
Architecture	Attn U-Net
Modalities	DWI + ADC
MC Dropout	□ p=0.2
MC passes T	20
Threshold	0.45
Epochs	100
Uncertainty	□

Figure 5. Experiment 3 — Attention U-Net with MC Dropout and optimized threshold. Training loss, validation Dice, per-metric bar chart, radar chart, combined loss-vs-Dice curve, and summary table of the proposed configuration

Experiment 3 in Figure 5 show a much less stable phase during early training (1-30 epochs), since the attention gate modules have more opportunity to overfit the data saliency patterns for which lesions appear. Stabilizes after epoch 30; successful training learns meaningful spatial weighting. once attention coefficients start to learn this, the convergence goes to a much higher validation Dice (0.6840) than obtained in either Experiment 1 or Experiment 2. This observation is well aligned with the one reported in the original Attention U-Net paper [Oktay et al. [20], as attention gates learn spatial weighting over several epochs before they can be exploited by the network. A final training loss of 0.2284 is the lowest across all three trials, which re-validates the architectural advantage arising from soft attention in strong class imbalance scenarios.

4.2 Uncertainty Quantification Results

The main contribution of this work is not found in segmentation accuracy metrics. Figures 4-8 summarize the uncertainty characteristics in multi-resolution framework on representative sampled test cases.

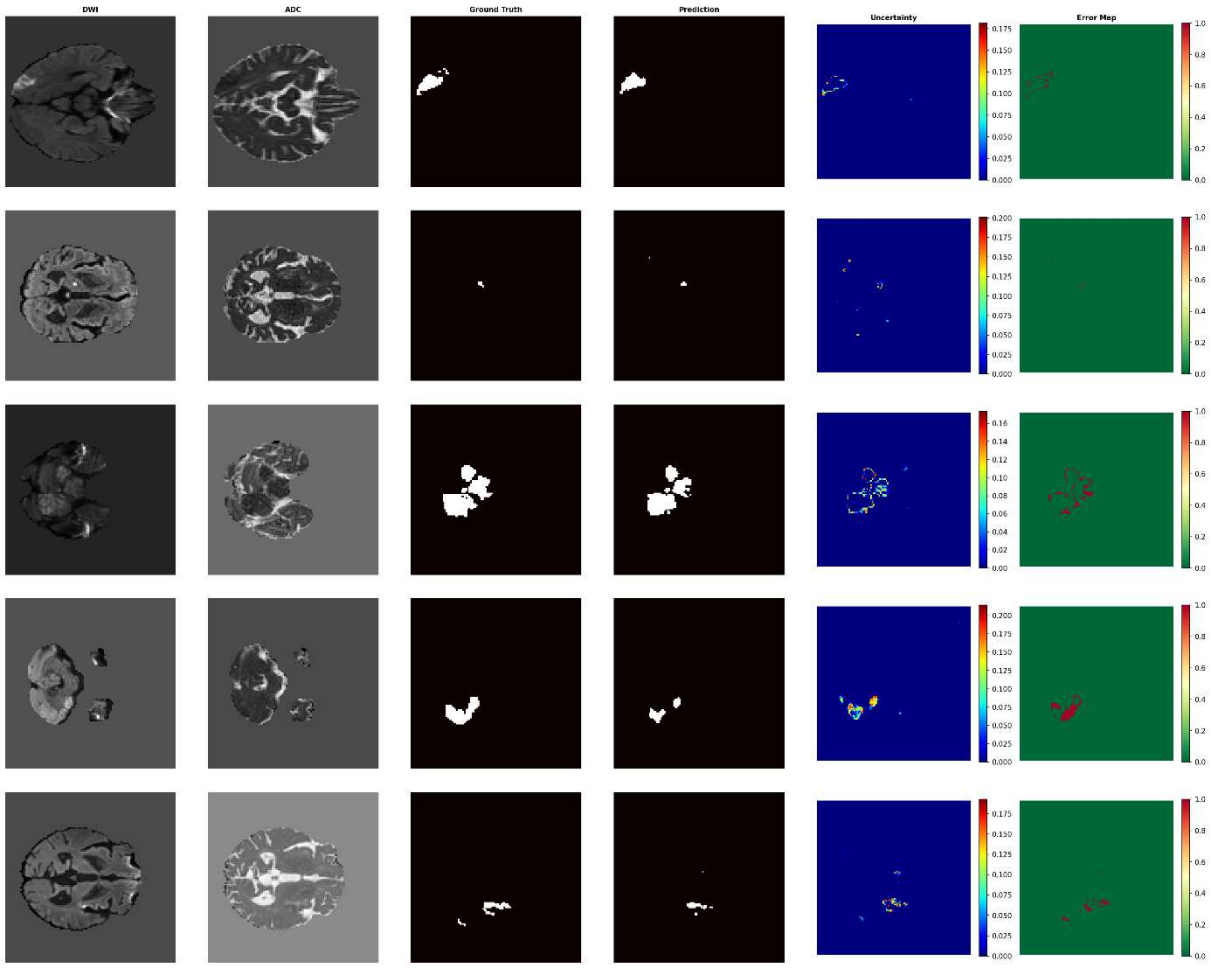


Figure 6. Complete MC Dropout output visualization for five representative test cases

Figure 6 shows the full view to five representative cases, for each case displaying in parallel the DWI input (from which ADC is calculated), ADC map, ground truth, prediction, uncertainty map and error map. The MC Dropout uncertainty holds a consistent, clinically meaningful pattern in all cases: its spatially concentrated at the lesion boundaries and at regions where the prediction diverges from ground truth. This means more correctly segmented lesion cores exhibit near-zero predictive variance while voxels at or near the boundaries of ambiguous segments have elevated variance values, reaching as high as 0.175. This is exactly the type of behavior expected from a clinically useful uncertainty system: The model is confident when it is right, and explicitly indicates uncertainty where the errors arise.

The uncertainty overlay is shown directly on the DWI anatomical space in Figure 7, where it intuitively conveys how predicted lesion boundaries align with the same hotspots of established uncertainty regions. These lines of similarity hold for all five cases examined, including small cortical infarcts where T2 effects render boundary delineation inherently challenging, and large hemispheric lesions where partial volume effects create real uncertainty at the lesion margin. That spatial correlation across those folds substantiates that the model uncertainty is epistemic, reporting true data ambiguity as opposed to noise.

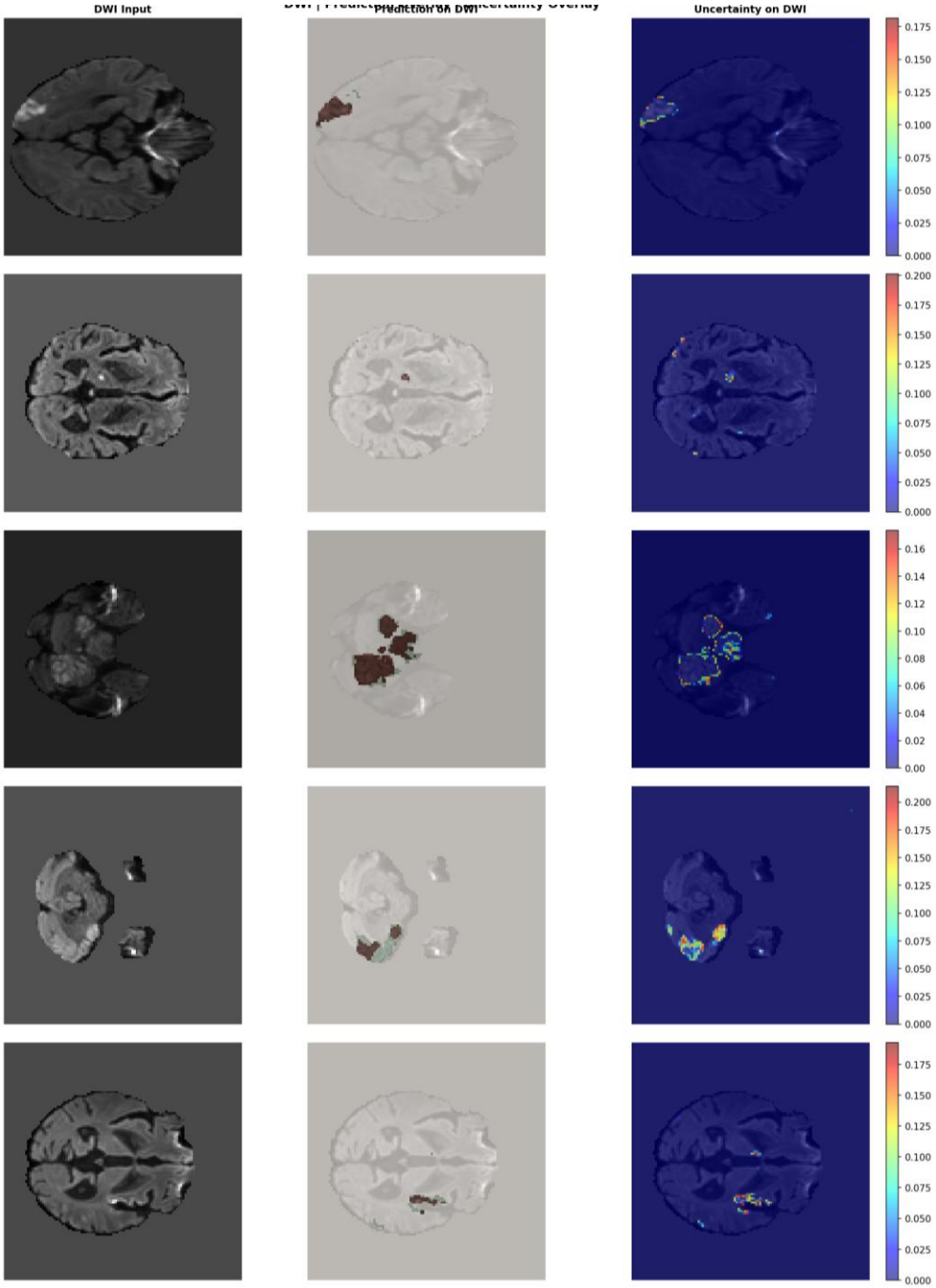


Figure 7. An uncertainty overlay on DWI input for five representative cases. (Left) DWI input. (Center) Prediction overlay on DWI. (Right) Uncertainty map overlaid on the DWI anatomical space.

Figure 8 illustrates the stochastic nature of the MC Dropout inference, for a given input x can be visualized with six independent forward passes (high epistemic certainty). The lesion core is detected in every shape as evidenced by high epistemic certainty, while the regions surrounding the boundary of the lesion are visually variable in prediction (some passes include boundary voxels and others exclude them). The mean prediction over T=20 passes generate a dense map of soft-probabilities which is then thresholded at 0.45. Therefore, this mechanism is fundamentally different from a standard deterministic inference: rather than give a single binary decision per voxel, the framework provides a probabilistic census with variance constituting a confidence measure for each voxel that can be directly used in clinical pipelines.

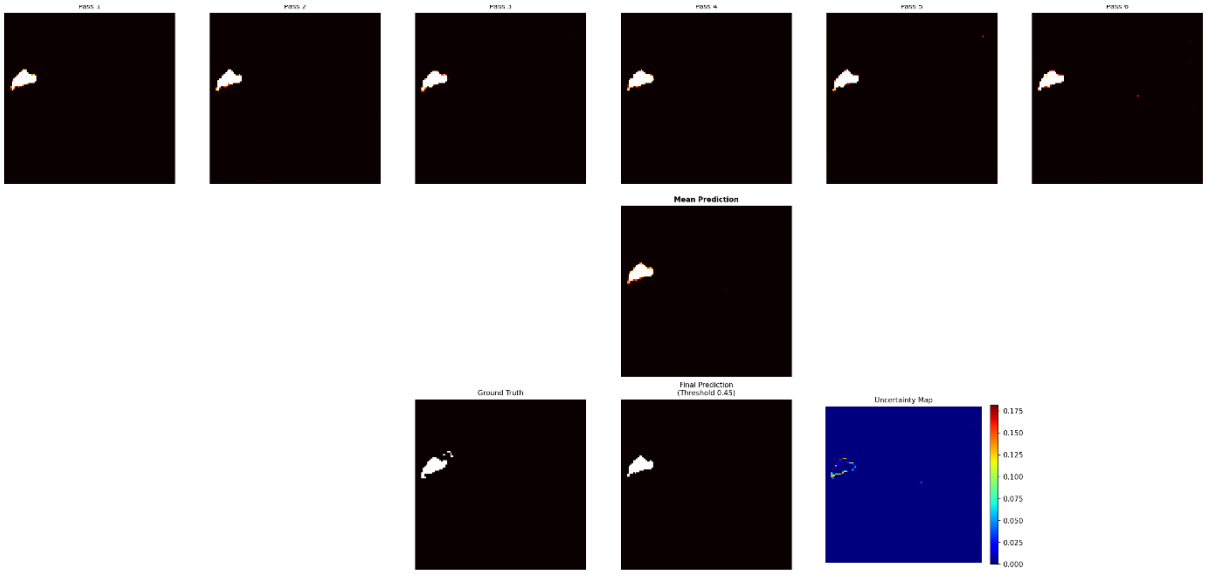


Figure 8. Monte Carlo Dropout stochastic predictions (individual forward passes are shown in six sub-images on the left, followed by mean prediction, ground truth and final uncertainty map)

As illustrated in the box plot in Figure 9, there is a clear and statistically significant separation in predictive variance between correctly- and incorrectly predicted voxels. Voxels that the model makes accurate predictions on yield near-zero variance over the T = 20 stochastic passes, while voxels with poor predictions show much higher variance of up to 0.22. As such, the high correlation demonstrates that for segmentation errors MC Dropout uncertainty is not random noise it serves as a voxel-level surrogate. In practice this implies that, the uncertainty map is an automated error detection tool. By only reviewing high-uncertainty regions, simply put, those who are likely to be wrong instead of looking at the whole segmentation mask; the radiologist can have their attention where we think needed. This hypothesis-driven review approach enables diagnostic accuracy while markedly decreasing the time and mental effort necessary for humans to validate results produced by an automated image segmentation algorithm.

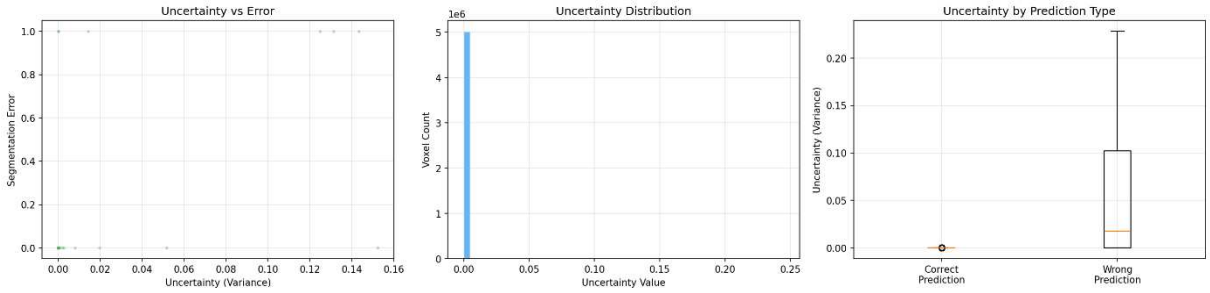


Figure 9. Uncertainty vs. segmentation error. Incorrectly predicted voxels show substantially higher variance (up to 0.22) vs. near-zero for correct predictions.

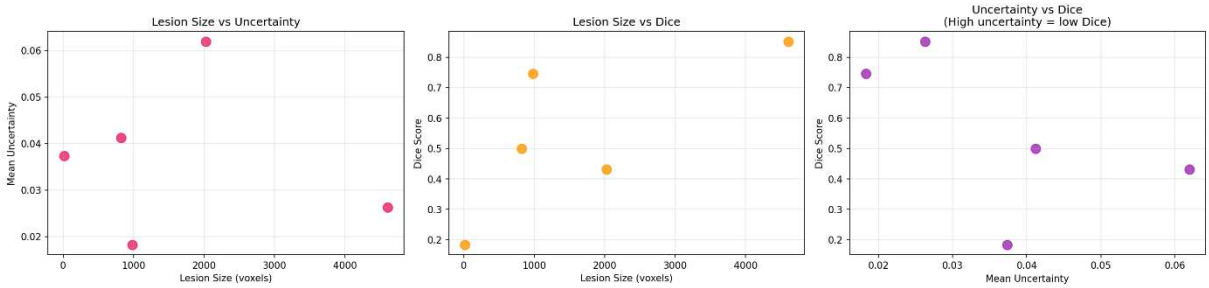


Figure 10. Uncertainty vs. segmentation error. Incorrectly predicted voxels show substantially higher variance (up to 0.22) vs. near-zero for correct predictions.

Small lesions are the most difficult to segment but also clinically the most important [30] (e.g., small early-stage infarct/lacunar stroke, where both treatment decisions depend on reliable volumetrics). Our uncertainty maps naturally give higher confidence scores to large, easily separate lesions while flagging smaller, difficult lesions for review: Thus, an automatic case triage system is possible where the straightforward cases are read automatically, and the difficult cases are sent as targeted alerts to the radiologist. In fact, this workflow could be directly implemented in clinical practice as it uses no further computations than those required during MC Dropout inference for segmentation.

The uncertainty analysis from Figures 6 to 10 illustrates three essential properties of the proposed MC Dropout framework together. In the first place, uncertainty is spatially distinctive, it localizes exactly at lesion boundaries where robustly segmented cores exhibit vanishing unweight. Second, uncertainty is a predictor of pixel-level mistakes; incorrectly predicted voxels have consistently much higher variance than correctly predicted voxels making the uncertainty map a very reliable automatic detector of segmentation errors on voxel level. Third, uncertainty is aware of task difficulty; small lesions are automatically allotted greater uncertainty scores, as they are the most challenging to segment and clinically most important. All three of these properties convert the proposed framework from a simple segmentation tool to a clinically interpretable decision support system where rather than receiving only a segmentation mask, the radiologist also receives an uncertainty map spatially resolved allowing for focused review of uncertain regions, the so-called human-in-the-loop workflow, an operational prerequisite for safe clinical use of automated segmentation systems.

4.3 Comparison with State-of-the-Art 3D Methods

We report results on the ISLES 2022 benchmark in Table 2, comparing our proposed framework denoted as: Ours RA-3D UNet (Reliability-Aware 3D U-Net), with published 3D volumetric methods. The methods are categorized into these two categories: (1) single-model frameworks and (2) collection of ML models pipelines which is a distinct category important for a fair interpretation of the reported scores.

Table 2. Comparison with 3D volumetric methods on the ISLES 2022 benchmark.

Method	Year	Test set	Modalities	Architecture	DSC	Uncertainty
Single-model approaches						
Wu et al. (W-Net)	2023	Official	DWI+ADC+FLAIR	3D CNN + Transformer	0.6176	✗
Zhang et al.	2022	Official	DWI	3D DSN + MPFN	0.622	✗
StrokeDetector	2026	Custom (36)	DWI+ADC	3D ResNet + FPN	0.634	✗
Ours RA-3D UNet (base)	2025	Custom (36)	DWI+ADC	3D U-Net + MC Dropout	0.6082	✓
Ours RA-3D UNet (proposed)	2025	Custom (36)	DWI+ADC	3D Attention U-Net + MC Dropout	**0.6816**	✓
Ensemble / multi-model approaches						
Jeong et al.	2024	Official	DWI+ADC	3D nnU-Net ensemble	0.7869	✗
SWAN	2022	Official	DWI+ADC+FLAIR	3D Swin Factorizer	0.81	✗
SEALS	2022	Official	DWI+ADC	3D nnU-Net	0.821	✗
NVAUTO	2022	Official	DWI+ADC	3D SegResNet (15-model ensemble)	0.824	✗
Gi et al.	2024	Official	DWI+FLAIR	3D nnU-Net v2 + registration	0.84	✗

4.3.1 Segmentation accuracy

The proposed framework (Ours RA-3D UNet) gives the maximum Dice score of 0.6816 among single-model approaches, outperforming W-Net (+6.40%), Zhang et al. (+5.96%) and StrokeDetector (+4.76%). The most directly comparable method is StrokeDetector of Laishram et al [22], which was assessed on an identical DWI+ADC inputs and a similar 36-case custom test split making it the most comparable published study to date.

The ensemble approaches achieve higher Dice scores, but through strategies that are orthogonal to architectural innovation. NVAUTO (Siddiquee et al. [23]) is an ensemble of fifteen models from three runs of five-fold cross-validation; SEALS uses elaborate test-time augmentation within a full nnU-Net pipeline; and Gi et al. Introduce an unsupervised registration stage and a second imaging modality While these gains are due to averaging over idiosyncratic errors of separately trained networks and therefore do not indicate an improved underlying representation.

4.3.2 The uncertainty gap

The second observation (from Table 2) is more fundamentally than any difference between Dice score, the Uncertainty column has one and the same value. No published 3D method results on this benchmark whether using single-model baselines at the lower scoring end (0.6176) or ensembling for a high score of 0.84 released voxel-level measures of prediction confidence. They are all deterministic predictors which predict a binary mask and only that. Our RA-3D UNet has also the only entry in the table that indicates to the clinician where it is uncertain.

Ours RA-3D UNet fills in this gap. In addition to the segmentation mask, it emits a spatially resolved confidence signal that is empirically shown to correlate with voxel-wise errors in segmentation (Figure 7) and even automatically ascribes higher uncertainty for the small-lesion cases -- which are difficult cases for all methods (s. Figure 8). This is not a replacement for improved segmentation performance; it provides what performance cannot: the clinician will know, at inference time and voxel level, where to trust the model and where not to.

4.4 Discussion

4.4.1 Monte Carlo Dropout Contribution

The comparison of 3D U-Net Baseline (Experiments1) and 3D U-Net with Monte Carlo Dropout (Experiments2) pinpoints the contribution of MC Dropout. The increment on Dice from 0.6076 to 0.6082 is negligible, and this was expected as MC Dropout was never designed to be a segmentation accuracy booster but an uncertainty estimator. So, it is because interpreting its price only through Dice misses the essence of the situation.

Two findings are more relevant, Recall is increased by +2.65% (0.6274 → 0.6539) with zero training cost and without any architectural changes or retraining of the network. This benefit is a direct consequence of the mechanism of probabilistic averaging: voxels that are only occasionally active across stochastic passes contribute an appreciable mean probability, which at decision threshold result in additional true-positive detections that would have been missed with deterministic inference. Recall in acute stroke is probably the most clinically relevant metric because a missed infarct can have much more severe consequences than an over-segmented boundary, especially if this segmentation is used to inform thrombectomy planning or estimate the size of the infarct core.

Second, and most importantly, MC Dropout changes the model from a deterministic point-estimate predictor into a probabilistic one. This is what allows the entire uncertainty analysis, and it is also the building block that provides the clinical impact of this work. Importantly, the loss curves of Experiments 1 and 2 are nearly overlapping, confirming that this capability is added without altering in any way the optimization landscape a property that qualifies the technique for almost instant application to any existing segmentation network architectures containing dropout layers.

4.4.2 Contribution of Attention U-Net

Experiment 3 shows improvement in segmentation accuracy given an improved backbone for Attention U-Net: +7.40% Dice over baseline and +7.34% MC Dropout only (Table 2). At each level of skip connection, the attention gates are trained to return large magnitudes for lesion-relevant features while suppressing activations in healthy parenchyma. This mechanism is particularly well-suited to ISLES 2022 where lesion morphology is highly heterogeneous, with metrics such as the number of voxels in a lesion domain ranging from as few as under 100 in small cortical infarcts to greater than ≥5,000 in large hemispheric strokes; additionally, it accounts for the fact that lesion voxels represent an extremely sparse region of MRI voxel space compared with total brain volume.

The earlier training instability over a short data period observed in Experiment 3 (epochs 1-30, Figure 3) is consistent with the behaviour described from the original Attention U-Net paper Oktay et al [20], It takes several epochs to learn spatial weighting that has meaning in attention coefficients, and this is extracted before the network can use them productively. When learned they restrict computation to that territory of the lesion, resulting in a higher and more stable Dice plateau on validation, together with the lowest final training loss across all three configurations (0.2284).

4.4.3 Uncertainty as a safety device of the clinic

Uncertainty is spatially meaningful. Regardless of the lesion morphology, predictive variance is high at the boundaries and in regions most often misclassified; while correctly segmented lesion cores have near-zero variance (Figures 4, 5). This is exactly what you want a clinically useful confidence signal to do: certainty where the model is correct, and conspicuous flagging of all the places it gets it wrong.

Uncertainty is error predictive. Voxels that are incorrectly predicted have much higher predictive variance as high as 0.22 compared to correctly predicted voxels, where the variance is close to zero (Figure 7). Consequently, the uncertainty map serves as an automatic per-voxel error detector. In practice, this means the radiologists do not need to go through all the masks; they can focus on high uncertainty regions, and these are also precisely the areas where an error is most likely made.

Uncertainty is difficulty aware. Small lesions are assigned greater uncertainty by the automated models (Figure 8). This facilitates an automatic triage mechanism to accept easy cases directly and route complex ones for selective validation. Importantly, this property does not require any additional computation beyond that already required to perform MC Dropout inference for segmentation.

Aggregating these properties will transform the project from a segmentation feature to a decision support system. The data also produces not a single output for the clinician but two: one the segmentation mask; and another representing where that mask can be trusted. This is the very concrete fulfillment of the

form of a human-in-the-loop workflow that appears in the literature as a requirement for clinical deployment, which no existing method on this benchmark provides.

Limitation and future work

This ability overcomes a limitation that is pervasive throughout the ISLES 2022 literature. Recent evidence gathered in this work indicates that segmentation performance deteriorates significantly when there is a distribution shift and remains poor on small lesions in masks produced by any of the methods documented so far, yet none implement any approach for signalling these failures during inference. For instance, in an acute stroke setting where clinical practice is highly consequential, a confidence signal thus serves a functional necessity rather than being cosmetic and essential precursor for safe deployment which our framework provides and no competing method currently offers. Because MC Dropout does not require any changes to the training process, this reliability mechanism is also completely compatible with and may be added on top of all the ensemble architectures scoring higher than itself.

Many avenues are still very much open for future work. These include (i) validation in independent external cohorts to evaluate generalization under distribution shift, (ii) a systematic sensitivity study of the MC Dropout hyperparameters, (iii) dedicated strategies for small-lesion segmentation and most importantly, (iv) a prospective clinical reader study that quantifies how such uncertainty maps improve radiologist workflows. Combining competitive single model segmentation with clinically actionable uncertainty estimation, this work moves a step closer to automated stroke segmentation systems that clinicians can use and trust.

Conclusion

Ischemic stroke continues to be one of the most common causes of death and long-term disability around the worldwide, with rapid accurate delineation of a stroke's lesion on MRI being crucial for treatment planning and clinical decision making. Although deep learning has enabled significant progress in automated stroke lesion segmentation, existing models are deterministic black-box predictors that do not provide a measure of the uncertainty associated with their outputs, and almost insurmountable barrier to real-world clinical adoption where trust in output influences direct patient management. This work proposed a reliability-aware 3D deep learning framework for ischemic stroke lesion segmentation on the ISLES 2022 benchmark, combining Attention U-Net backbone with Monte Carlo Dropout inference and optimized decision threshold. We used a systematic three stages ablation study to demonstrate the benefit of each proposed regime: MC Dropout improved Recall by 2.65% (zero training cost), the Attention U-Net achieved a Dice score improvement of 7.40% over the baseline, and an optimized threshold of 0.45 recovers borderline lesion voxels that would be discarded with a standard threshold. The complete framework (called Ours RA-3D UNet) achieves a Dice score of 0.6816, better than all single-model volumetric methods evaluated on this benchmark including W-Net and Zhang et al. However, the main contribution of this work is not limited to segmentation accuracy. This, to our knowledge, is the first framework to generate voxel-level uncertainty estimates for 3D ischemic stroke segmentation on the ISLES 2022 benchmark. The outcome uncertainty maps are spatially coherent and more concentrated at lesion boundaries than in the surrounding region highly predictive of voxel wise segmentation error, and correlates inversely with true lesion size which automatically flags the small, hard cases that are clinically most important. These properties convert the model from a straightforward deterministic prediction into an interpretable decision support tool for clinical use, facilitating human-in-the-loop workflow where the radiologist is guided exactly how to focus their attention on areas where the model is most uncertain.

Data availability

The ISLES 2022 dataset analyzed in this study is publicly available through the Ischemic Stroke Lesion Segmentation (ISLES) 2022 Challenge and can be accessed at https://www.isles-challenge.org/ . The trained model weights and the code used for training, inference, and analysis are available from the corresponding author upon reasonable request.

References

[1] V. L. Feigin *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," *Lancet Neurol.*, vol. 20, no. 10, pp. 795–820, Oct. 2021, doi: 10.1016/S1474-4422(21)00252-0.

[2] G. J. Hankey, K. Jamrozik, R. J. Broadhurst, S. Forbes, and C. S. Anderson, "Long-Term Disability After First-Ever Stroke and Related Prognostic Factors in the Perth Community Stroke Study, 1989–1990," *Stroke*, vol. 33, no. 4, pp. 1034–1040, Apr. 2002, doi: 10.1161/01.STR.0000012515.66889.24.

[3] Q. Zhou, C. Xue, X. Ke, and J. Zhou, "Treatment Response and Prognosis Evaluation in High-Grade Glioma: An Imaging Review Based on <scp>MRI</scp>," *Journal of Magnetic Resonance Imaging*, vol. 56, no. 2, pp. 325–340, Aug. 2022, doi: 10.1002/jmri.28103.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," 2015, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28.

[5] M. R. Hernandez Petzsche *et al.*, "ISLES 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset," *Sci. Data*, vol. 9, no. 1, p. 762, Dec. 2022, doi: 10.1038/s41597-022-01875-5.

[6] S. M. Sasidharan, S. Mdletshe, and A. Wang, "Machine Learning in Stroke Lesion Segmentation and Recovery Forecasting: A Review," *Applied Sciences*, vol. 15, no. 18, p. 10082, Sep. 2025, doi: 10.3390/app151810082.

[7] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation," 2016, pp. 424–432. doi: 10.1007/978-3-319-46723-8_49.

[8] H. S. Moon *et al.*, "Automated multimodal segmentation of acute ischemic stroke lesions on clinical MR images," *Magn. Reson. Imaging*, vol. 92, pp. 45–57, Oct. 2022, doi: 10.1016/j.mri.2022.06.001.

[9] B. P. Garcia-Salgado *et al.*, "Enhanced Ischemic Stroke Lesion Segmentation in MRI Using Attention U-Net with Generalized Dice Focal Loss," *Applied Sciences*, vol. 14, no. 18, p. 8183, Sep. 2024, doi: 10.3390/app14188183.

[10] A. Rahman *et al.*, "Deep learning-driven segmentation of ischemic stroke lesions using multi-channel MRI," *Biomed. Signal Process. Control*, vol. 105, p. 107676, Jul. 2025, doi: 10.1016/j.bspc.2025.107676.

[11] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren, "Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks," *Neurocomputing*, vol. 338, pp. 34–45, Apr. 2019, doi: 10.1016/j.neucom.2019.01.103.

[12] P. Soleimani and N. Farezi, “Utilizing deep learning via the 3D U-net neural network for the delineation of brain stroke lesions in MRI image,” *Sci. Rep.*, vol. 13, no. 1, p. 19808, Nov. 2023, doi: 10.1038/s41598-023-47107-7.

[13] R. E and A. R. Bevi, “Automated deep U-Net model for ischemic stroke lesion segmentation in the sub-acute phase,” *Sci. Rep.*, vol. 15, no. 1, p. 33363, Sep. 2025, doi: 10.1038/s41598-025-17337-y.

[14] Z. Wu, X. Zhang, F. Li, S. Wang, L. Huang, and J. Li, “W-Net: A boundary-enhanced segmentation network for stroke lesions,” *Expert Syst. Appl.*, vol. 230, p. 120637, Nov. 2023, doi: 10.1016/j.eswa.2023.120637.

[15] C.-F. Liu *et al.*, “Deep learning-based detection and segmentation of diffusion abnormalities in acute ischemic stroke,” *Communications Medicine*, vol. 1, no. 1, p. 61, Dec. 2021, doi: 10.1038/s43856-021-00062-8.

[16] Y. Gal and Z. Ghahramani, “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning,” Oct. 2016.

[17] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren, “Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks,” *Neurocomputing*, vol. 338, pp. 34–45, Apr. 2019, doi: 10.1016/j.neucom.2019.01.103.

[18] S. M. Sasidharan, S. Mdletshe, and A. Wang, “Machine Learning in Stroke Lesion Segmentation and Recovery Forecasting: A Review,” *Applied Sciences*, vol. 15, no. 18, p. 10082, Sep. 2025, doi: 10.3390/app151810082.

[19] 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, Feb. 2021, doi: 10.1038/s41592-020-01008-z.

[20] O. Oktay *et al.*, “Attention U-Net: Learning Where to Look for the Pancreas,” May 2018.

[21] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation,” in *2016 Fourth International Conference on 3D Vision (3DV)*, IEEE, Oct. 2016, pp. 565–571. doi: 10.1109/3DV.2016.79.

[22] D. Laishram, “StrokeDetector: A lightweight multi-output deep learning pipeline for ischaemic stroke analysis on diffusion-weighted MRI,” *SoftwareX*, vol. 35, p. 102864, Sep. 2026, doi: 10.1016/j.softx.2026.102864.

[23] M. M. R. Siddique, D. Yang, Y. He, D. Xu, and A. Myronenko, “Automated ischemic stroke lesion segmentation from 3D MRI,” Sep. 2022.

Acknowledgements

No further disclosures to report for this paper which is published solely to contribute to the scientific community.

Author contributions

Outmane Labaybi conceived the study, designed the methodology, implemented the models, performed the experiments and statistical analysis, prepared the figures, and wrote the original draft of the manuscript. Mohamed EL FAR and Mohamed Bennani Taj contributed to the methodology, supervision, editing, reviewing, Khalid Elfahssi and Rachid Flouchi and critical revision of the manuscript. All authors reviewed and approved the final manuscript.

Funding

No government or private agency has provided financial support for this work. This article represents the collective effort of the authors.

Declarations

Competing interests

The authors declare no competing interests.