From Baseline to Bilateral Asymmetry: A Systematic nnU-Net Comparison for Ischemic Stroke Lesion Segmentation in ISLES'26

Zayim, Beyza[1 [0009−0002−0054−8466]⋆] ; Aissiou, Ikram[2 [0009−0008−2845−3606]⋆]; Collins, Stephan[3 [0000−0002−1533−3380]] ; Lalande, Alain[4 [0000−0002−7970−366X]] ; Meriaudeau, Fabrice[1 [0000−0002−8656−9913]]

[1] Université Bourgogne Europe, CNRS, ICMUB UMR 6302, 21000 Dijon, France
Beyza.Zayim@u-bourgogne.fr
[2] University of Algiers 1 Benyoucef Benkhedda Algiers, Algeria
aissiouikram47@gmail.com
[3] Inserm UMR1231, Center of Translational Molecular Medicine, Mouse NeuroGenetics Laboratory, University of Bourgogne Europe, Dijon, France
[4] Université Bourgogne Europe, CHU Dijon Bourgogne, Pôle Imagerie, CNRS, ICMUB UMR 6302, 21000 Dijon, France

Abstract. Automated segmentation of ischemic stroke lesions from structural MRI is challenging due to lesion variability and pronounced class imbalance. We present our contribution to the ISLES'26 challenge, systematically evaluating four nnU-Net-based configurations on 1453 T1-weighted cases. To isolate architectural and data-level effects, all models shared identical preprocessing (reorientation and N4 bias correction). The evaluated configurations, all implemented within the nnU-Net framework, include: (1) the standard baseline, (2) a residual encoder variant, (3) the residual encoder variant trained with extended data augmentation, and (4) the residual encoder variant trained with an explicit bilateral (left-right) asymmetry input channel. On our internal hold-out test split, the residual encoder provided the strongest foundational improvement in median overlap and boundary accuracy. Building on this, the extended augmentation configuration achieved the best overall boundary precision (lowest HD95) and reduced missed lesion voxels. Conversely, while the bilateral asymmetry channel yielded mixed results on the internal test set, it demonstrated superior generalization on external data, outperforming the standalone residual encoder on the preliminary ISLES'26 leaderboard.

Keywords: Ischemic stroke · Lesion segmentation · nnU-Net · Residual encoder · Data augmentation · Bilateral asymmetry · ISLES 2026.

1 Introduction

Accurate delineation of ischemic stroke lesions on MRI is a prerequisite for quantifying infarct volume, guiding treatment decisions, and enabling downstream

⋆ Equal contribution.

prognostic modeling. Manual segmentation is time-consuming and subject to inter-rater variability, which motivates the development of automated tools such as those benchmarked in the ISLES challenge series. The ISLES'26 edition poses this problem on a large, single-channel T1-weighted cohort, framed as a binary voxel-wise segmentation task (lesion vs. background).

The self-adapting nnU-Net and its residual-encoder variants consistently achieve state-of-the-art performance in biomedical segmentation using core convolutional designs [1, 3]. While robust data augmentation simulating intensity and contrast variations enhances their generalization on brain MRI [2], plain intensity-based models still overlook an important anatomical cue: since stroke lesions are typically unilateral [4], the healthy contralateral hemisphere could—yet does not currently—serve as a natural, patient-specific reference.

Motivated by these two directions, we structure our ISLES'26 submission as a controlled ablation. Starting from a common preprocessing pipeline, we compare different configurations all implemented within the nnU-Net framework, including: (i) the standard baseline, (ii) a residual-encoder variant, (iii) the residual-encoder variant trained with additional augmentation informed by prior data-augmentation literature for brain MRI segmentation [2], and (iv) the residual-encoder variant given an extra input channel that encodes bilateral asymmetry through a flip-register-subtract procedure.

2 Methods

2.1 Dataset and task

We use the ISLES'26 training release, comprising 1453 single-channel T1-weighted MRI volumes with corresponding binary lesion masks (background vs. foreground) [5, 6]. As the official challenge test set was not available prior to manuscript submission, we first set aside 243 cases as a held-out internal test set, common across all four configurations. The remaining 1210 cases were used for 5-fold cross-validation; and all models reported here were trained on fold 0 of this split (967 training / 243 validation cases), ensuring identical training and validation data across configurations. All results are computed on this held-out test subset.

2.2 Common preprocessing

Every configuration in this study is trained on data that has undergone the same two preprocessing steps prior to nnU-Net's own resampling and normalization: (1) reorientation of each volume to a canonical (RAS) orientation, so that anatomical axes are consistent across cases regardless of how the original NIfTI files were stored, and (2) N4 bias-field correction [7] to remove low-frequency intensity inhomogeneities caused by scanner coil sensitivity. Beyond this shared step, nnU-Net-internal preprocessing was left at its automatically configured defaults: resampling all images to the median voxel spacing of the dataset, and applying intensity normalization based on the data modality and background heuristics (e.g., Z-score normalization using foreground statistics) [1].

2.3 Configuration 1: Baseline

Our baseline network is the default nnU-Net 3D full-resolution configuration (plain convolutional encoder–decoder, nnUNetPlans), trained directly on the reoriented, bias-corrected data described above, following the standard nnU-Net training protocol (deep supervision, combined Dice and cross-entropy loss, SGD optimization, polynomial learning-rate decay, 1000 epochs) on a single Tesla V100 (32GB) GPU without distributed training or cascaded architectures.

2.4 Configuration 2: Residual encoder

The second configuration replaces the plain encoder with a 3D Residual Encoder U-Net, using the nnUNetResEncUNetL plans produced by the nnUNetPlannerResEncL experiment planner on the same preprocessed data as the baseline. Training followed the same protocol as Configuration 1 on the same GPU.

2.5 Configuration 3: Residual encoder with extended Gaussian Mixture Model-based data augmentation

The third configuration preserves the residual-encoder architecture and training schedule from Configuration 2. However, it extends nnU-Net's default augmentation pipeline by incorporating additional intensity and contrast transformations inspired by the Gaussian Mixture Model-based data augmentation approach described by Meyer et al. [2]. Specifically, three augmentations are applied recursively to the entire image, augmenting the input volumes on the label part, while an additional three augmentations are selectively applied only to the lesion regions. These transformations are applied on top of the reoriented, bias-corrected inputs, aiming to further increase variability in tissue contrasts and intensities while maintaining anatomical and lesion consistency. This setup evaluates whether stronger and more diverse data augmentation, focusing on realistic intensity variations in both global and lesion-specific regions, can more effectively improve the model's generalization to unseen multi-scanner data, without altering network capacity.

2.6 Configuration 4: Residual encoder with bilateral asymmetry channel

The fourth configuration retains the same residual-encoder architecture and training protocol, but adds a second input channel that encodes hemispheric intensity asymmetry, computed as follows. For each case: (1) the original T1-weighted image is loaded; (2) a working copy is reoriented to canonical RAS orientation so that a left–right flip is well defined irrespective of the image's original storage orientation; (3) this canonical copy is mirror-flipped along the left–right axis; (4) the flipped image is rigidly registered back onto the original (unflipped) image in physical space, which corrects for the reorientation step as well as any head tilt or positioning asymmetry; (5) the registered, flipped image

is resampled onto the original image's voxel grid; and (6) the absolute voxel-wise difference between the original image and the registered, resampled flipped image is computed. The absolute-difference map is saved with the same affine transformation matrix as the original image to ensure exact voxel-wise alignment. It is then supplied to nnU-Net as a second channel alongside the original intensity image. The original image itself is never modified. The underlying assumption is that stroke lesions, which are typically unilateral, should produce localized differences in this asymmetry map, giving the network an additional, anatomically motivated cue on top of the raw intensity.

2.7 Evaluation

All four configurations were evaluated on the common held-out test split described in Section 2.1. For Configuration 3, the fold 0 training cases were augmented with additional transformed copies derived from those same images, with augmentation confined strictly to the training split to avoid leakage into validation data. We report median Dice similarity coefficient, median Intersection-over-Union (IoU), median 95th-percentile Hausdorff Distance (HD95), and median false-positive (FP) and false-negative (FN) voxel counts per case, computed over the test cases. Notably, healthy cases were excluded from this test set evaluation. Because these cases completely lack a ground-truth target mask, standard overlap metrics like Dice and IoU are mathematically ill-defined; even a single false-positive voxel forces the score to zero, which would artificially skew the aggregated median performance.

3 Results

Table 1 summarizes the performance of the four completed configurations (trained on fold 0) on our internal held-out test split.

Table 1. Median results (Dice, IoU, FP, FN voxel counts, HD95) on our internal held-out test split, for models trained on fold 0 of the internal cross-validation. Unhealthy cases only (zero-Dice entries excluded), across four configurations. The best results are in bold.

Configuration	Dice [Q25–Q75]	IoU	FP	FN	HD95 (mm) [Q25–Q75]
1: Baseline	0.763 [0.491–0.852]	0.616	1047.5	**1364.5**	8.88 [3.74–39.30]
2: Residual encoder	**0.783** [0.501–0.862]	**0.644**	**965.0**	1592.0	8.31 **[3.32–37.21]**
3: Residual encoder + augmented	0.773 [0.491–0.864]	0.630	1255.0	1502.5	**8.06** [3.44–37.55]
4: Residual encoder + new channel	0.778 **[0.509–0.867]**	0.637	1051.5	1671.5	8.63 [3.50–35.35]

Switching from the plain encoder (Configuration 1) to the residual encoder (Configuration 2) yields the largest improvement in both median Dice (+0.020)

and median IoU (+0.028) among all comparisons, and produces the lowest false-positive count of the four configurations (965.0 voxels), as shown in Table 1. It also achieves the lower median HD95 of the two (8.31 mm vs. 8.88 mm; Table 1), indicating that among successfully detected lesions, boundary delineation is both more accurate and more precise than the baseline. However, this improvement coincides with a higher false-negative voxel count (1592.0 vs. 1364.5) and a higher total-miss count (8 vs. 5 cases where the lesion was entirely undetected), suggesting that the residual encoder's overall gains stem from tighter, more accurate boundaries on the lesions it does detect, while occasionally failing to detect the lesion at all slightly more often than the baseline.

Adding extended augmentation on top of the residual encoder (Configuration 3) slightly reduces Dice and IoU relative to Configuration 2 (0.773 vs. 0.783; 0.630 vs. 0.644), while increasing false positives (1255.0 vs. 965.0) and reducing false negatives (1502.5 vs. 1592.0); it nonetheless still outperforms the baseline on both metrics (Table 1). Notably, Configuration 3 achieves the lowest median HD95 of the four configurations (8.06 mm vs. 8.31 mm for Configuration 2; Table 1), indicating more accurate boundary delineation despite the small drop in overlap-based metrics — the surface distance to the reference contour is smaller on average even though voxel-level overlap and FP count both worsen slightly.

The bilateral-asymmetry channel (Configuration 4) behaves differently: it improves over the baseline in Dice (0.778 vs. 0.763) but falls marginally short of the residual-encoder-only configuration on our internal test set, and, unlike Configuration 3, does not reduce false negatives — its FN count (1671.5) is the highest of the four configurations, while its FP count (1051.5) sits between Configurations 2 and 3. Configuration 4 does achieve the highest upper-quartile Dice of any configuration (IQR upper bound 0.867), suggesting the asymmetry feature may benefit already well-segmented cases even though it does not improve the median.

Notably, this internal ranking did not hold externally: in the Preliminary Docker Evaluation (Sanity Check) leaderboard of the ISLES'26 Challenge, Configuration 4 outperformed Configuration 2 following submission of both. This discrepancy suggests that the asymmetry-channel modification may provide a generalization benefit not fully captured by our internal test-set metrics, and that its interaction with the residual-encoder architecture may be additive on out-of-distribution data even where it appears neutral or slightly negative internally.

4 Discussion and Conclusion

Our ablation study highlights the differing impacts of architectural choices, augmentation strategies, and domain-specific priors on ischemic lesion segmentation. First, evaluating performance on our internal hold-out test split demonstrates that adopting a residual encoder within the nnU-Net framework provides the strongest foundational performance. This observation is consistent with recent

literature indicating that residual architectures reliably improve upon standard plain encoders in large-scale medical segmentation tasks [3].

Second, while the standard residual encoder excels at general detection and overlap, coupling it with our extended data augmentation scheme specifically optimizes spatial precision. This configuration achieved the best boundary-based accuracy of our experiments, yielding the lowest HD95 scores. This suggests that aggressive augmentation acts as a complementary regularizer, encouraging highly precise and conservative lesion delineations.

Finally, explicitly encoding a clinically meaningful prior—the bilateral asymmetry channel—proved crucial for external generalization. Although this configuration's performance appeared modest on our internal test split, it demonstrated clear superiority externally. On the ISLES'26 Preliminary Docker Evaluation leaderboard, the configuration incorporating bilateral asymmetry outperformed the standalone residual encoder. This indicates that guiding the network with explicit hemispheric differences effectively buffers against the distribution shifts inherent to multi-center challenge data.

5 Acknowledgements

We would like to thank all organizers of the ISLES'26 Challenge, especially Dr. Ezequiel de la Rosa for all his efforts.

References

1. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
2. Meyer, M. I., de la Rosa, E., Pedrosa de Barros, N., Paolella, R., Van Leemput, K., Sima, D. M.: A Contrast Augmentation Approach to Improve Multi-Scanner Generalization in MRI. Frontiers in Neuroscience **15**, 708196 (2021). https://doi.org/10.3389/fnins.2021.708196
3. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., Jaeger, P.F.: nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation. arXiv preprint arXiv:2404.09556 (2024)
4. Esposito, E., Shekhtman, G., Chen, P.: Prevalence of spatial neglect post-stroke: A systematic review. Annals of Physical and Rehabilitation Medicine **64**(5), 101459 (2021)
5. Absher, J., et al.: The stroke outcome optimization project: Acute ischemic strokes from a comprehensive stroke center. Scientific Data **11**(1), 839 (2024)
6. Liew, S.L., et al.: A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. Scientific Data **9**(1), 320 (2022), `https://fcon_1000.projects.nitrc.org/indi/retro/atlas.html`
7. Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C.: N4ITK: improved N3 bias correction. IEEE Transactions on Medical Imaging **29**(6), 1310–1320 (2010)