Dual-Family Ensembling and Failure-Mode Analysis for Ischemic Stroke Lesion Segmentation in Single-Sequence T1w MRI

Wei Lu[0000−0001−6510−2532]

ArteryFlow Technology Co., Ltd., Hangzhou, China
wei.lu@arteryflow.com

Abstract. We evaluate an ensemble of two nnU-Net ResEnc-L families trained with different objectives: the default Dice–cross-entropy loss and a modified Focal-Tversky–cross-entropy loss that favours recall. In a 1452-case out-of-fold evaluation using the organisers' metric code, exploratory paired tests favoured the ensemble on four of the five official metrics and detected no significant regression on the fifth. Official Dice is reported only after instance matching at an intersection-over-union of 0.25; consequently, 177 cases scored zero despite standard Dice reaching 0.365 in some. Per-case oracle sweeps could recover 45 of them by thresholding or size filtering, and the minimum-lesion-size filter itself caused 18 failures, one third of the reference lesions being smaller than its threshold. A largest-component fallback produced seven Dice and lesion-F1 wins, no losses and 1445 ties. Lesion core thickness was the strongest post-hoc correlate of performance, and the contrast-associated Dice gap was twice as large among thin lesions. Re-selecting post-processing on disjoint halves of the sites retained two thirds of its apparent gain. All code is released under Apache 2.0.

Keywords: Stroke lesion segmentation · T1-weighted MRI · nnU-Net · Model ensembling · Failure analysis · Instance-wise evaluation

1 Introduction

ISLES'26 uses a *single* native-space T1-weighted volume as input [5]. Earlier editions supplied diffusion-weighted imaging, where acute infarcts are bright and top methods reached Dice around 0.82 [2]; on T1w, lesions are hypointense and often indistinguishable from enlarged perivascular spaces, leukoaraiosis or partial-volume effects, and native space removes the spatial prior that template registration would provide [7].

This paper describes our ISLES'26 submission: an ensemble of two nnU-Net ResEnc-L [3, 4] families with different objectives, evaluated on all 1452 training cases out-of-fold under the official metrics. We additionally attribute residual errors to absent lesion signal, post-processing and the instance-matching rule. Challenge reports rarely separate these three, and all are present here in measurable proportions.

2 Methods

2.1 Data and cross-validation

We use the ISLES'26 training release derived from ATLAS R3.0 [7]: skull-stripped T1w volumes with expert annotations at native resolution, five of them with empty reference masks. Lesion burden is heavy-tailed, the median of the 4905 reference components being 0.16 mL. One case was excluded after a mask audit and later confirmed faulty by the organisers, leaving $N = 1452$ of 1453 released scans. One centre (R032, 36 cases) reports an impossible through-plane spacing, which the organisers verified and kept; those cases fail more often (17.1% zero-Dice vs. 12.1%) but are at most 6 of the 177 failures below.

Cases carry a site identifier, and we build **centre-aware** five-fold splits in which a site never straddles two folds,[1] so that every out-of-fold prediction is also an out-of-distribution prediction with respect to acquisition. This avoids centre leakage and is intended to approximate the cross-centre generalisation the hidden test set will demand.

2.2 Model families and complementary losses

Both families are nnU-Net v2.8.0 with the ResEnc-L preset (`nnUNetResEncUNetLPlans`, `3d_fullres`), patch $160 \times 192 \times 160$, batch 3, 102.3 M parameters, default schedule and augmentation. **Family A** uses the stock Dice + cross-entropy objective; **Family B** replaces the Dice term with a Focal-Tversky loss [9, 1],

$$\mathrm{TI} = \frac{\mathrm{TP}}{\mathrm{TP} + \alpha\,\mathrm{FP} + \beta\,\mathrm{FN}}, \qquad \mathcal{L}_{\mathrm{FT}} = (1 - \mathrm{TI})^{\gamma}, \tag{1}$$

with $\alpha = 0.3$, $\beta = 0.7$ and $\gamma = 4/3$, applied at all deep-supervision scales and summed with cross-entropy. Our modified variant uses γ as the exponent directly, whereas the original formulation writes $1/\gamma$; an exponent above one down-weights already-accurate predictions. Penalising false negatives more heavily makes Family B over-segment, so the two families have complementary operating characteristics and reach their optima at different thresholds (0.35 for A, 0.60 for B).

2.3 Ensembling, post-processing and outputs

The ensemble averages the per-voxel foreground softmax of both families with equal weight over all five folds of each (ten models at inference), thresholds at 0.45, and deletes 26-connected components below 0.05 mL. If that would empty the mask, the single largest component is restored. This is the **largest-component fallback**, validated in Sect. 3.2. We emit both the binary mask

[1] Two cases carry an *empty* site field in the released metadata and were grouped as a centre of their own, so they sit in a different fold from the other 167 cases of their site. The other 54 sites are unaffected; excluding the two changes the out-of-fold Dice below by +0.0003.

Table 1. Out-of-fold performance ($N = 1452$), each method at its own optimal threshold with the same 0.05 mL size filter and without the largest-component fallback (Sect. 3.2). Mean rank is over the three methods shown, lower being better.

Method	rank	Dice	Lesion-F1	PR-AUC	AVD	$\lvert\Delta n\rvert$
Ensemble (thr 0.45)	**1.870**	**0.6259**	**0.5829**	**0.7237**	6.184	**1.900**
Family A only (thr 0.35)	2.054	0.6171	0.5754	0.7078	**6.094**	1.962
Family B only (thr 0.60)	2.075	0.6196	0.5760	0.7125	6.460	1.906

and the *unthresholded* averaged probability map, since the official PR-AUC is computed on the soft map.

2.4 Evaluation protocol

We import the organisers' `eval_utils.py` at commit `c430f15ed80c` rather than reimplementing the metrics, because the official definition changed twice during the challenge. The five metrics are Dice, absolute volume difference (AVD), absolute lesion-count difference, lesion-wise F1 and PR-AUC; instance matching uses `panoptica` [6] at IoU $\geq$ **0.25**, and Dice is its `global_bin_dsc`. Methods are compared by **rank-then-aggregate** over the five metrics [8], with paired Wilcoxon signed-rank tests on per-case values.

The evaluation has two limitations. Out-of-fold, each case is scored by the two models, one per family, whose folds excluded its site, whereas the deployed system averages all ten; the comparison below therefore isolates the ensembling effect under a strict no-leakage protocol rather than measuring the deployed configuration, which cannot be evaluated without leakage. Thresholds and post-processing rules were selected on the same 1452 cases used for the paired tests, so all p-values are exploratory and uncorrected for multiplicity. We therefore report per-case win, loss and tie counts alongside them.

3 Results

3.1 Out-of-fold ensemble comparison

Exploratory paired tests agreed with Table 1. The ensemble improved Dice, lesion-F1, count difference and PR-AUC over Family A, and Dice, lesion-F1, AVD and PR-AUC over Family B, with no significant regression on any metric in either comparison; the PR-AUC gains were +0.0159 and +0.0113 at $p < 10^{-52}$, small p-values reflecting that softmax averaging improves the voxel *ranking* on nearly every case even where the thresholded mask is unchanged. The apparent detection advantage of Family B depended on the matching rule: significant under a laxer rule (overlap ≥ 1 voxel; +0.0122, $p < 0.01$) but absent under the official IoU ≥ 0.25 rule (+0.0006, $p = 0.96$).

3.2 Attribution of zero-Dice cases

Because `global_bin_dsc` returns 0 when no instance clears IoU 0.25, 177 of 1452 ensemble cases (12.2%) receive exactly zero, including cases with standard Dice as high as 0.365. We attributed each of those 177 by sweeping thresholds from 0.05 to 0.80 with the size filter enabled and disabled, recording the best achievable instance-pair IoU. The causes split four ways: (A) the size filter deleted the component that would have matched, 18 cases (10.2%); (B) a different threshold suffices, 27 (15.3%); (C) signal is present but shape or location never reaches IoU 0.25 at any setting, 46 (26.0%); and (D) essentially no probability mass falls inside the reference lesion, 86 (48.6%). Under this per-case oracle, alternative post-processing recovers groups A and B, **45 cases or 25.4% of the failure tail**, though no single deployable rule recovers all 45 without affecting other cases.

The minimum-size filter caused group A. Its 0.05 mL threshold lies within the reference size distribution: **33.1% of all 4905 reference lesions are smaller than it**, and 17 cases have no reference lesion above it at all, so they can never score above zero under our own pipeline. Disabling it raises the best achievable IoU on those 18 cases from 0.000 to between 0.26 and 0.78, but costs lesion-F1 heavily ($0.583 \rightarrow 0.538$) by admitting spurious specks.

Across 39 post-processing configurations evaluated on all 1452 cases, the largest-component fallback (keep the size filter, but restore the largest component if filtering would empty the mask) improved Dice, lesion-F1, count difference and AVD while leaving PR-AUC unchanged. Relative to Table 1 it lifts Dice to **0.6270**, lesion-F1 to **0.5856**, count difference to **1.870** and AVD to **6.183** mL, and cuts zero-Dice cases from 177 to 170; PR-AUC is unchanged at 0.7237 by construction, since the fallback edits only the binary mask. Because the rule fires only when the output would otherwise be empty, every case with a non-empty post-filter mask is bit-identical, and the paired record agrees: Dice and lesion-F1 each show **7 wins, 0 losses, 1445 ties** ($p = 0.018$), and count difference is better on 44 cases and worse on none. A confidence-based exemption (retain an undersized component whose peak probability exceeds 0.95) reaches Dice 0.6284 and 162 zero-Dice cases, but perturbs 185 cases in both directions for 92 wins against 93 losses; we kept the fallback because it modified fewer cases with no observed OOF regression. One risk remains, since the organisers state the test set may contain lesion-free scans, on which forcing a prediction turns a perfect score into zero. Across all five such training cases the triggering condition never occurs: three produce no component and two a single false positive of 0.48 and 0.77 mL that already scores zero. This model fires in blobs rather than specks, where the fallback is inert.

How much survives out-of-sample selection? The rules above were selected and reported on the same cases. Separating the two roles by site (1000 random partitions of the 55 sites into balanced halves; select the best of 21 configurations on one half, score it on the other), the mean-rank advantage over an untuned reference (argmax thresholding, no size filter) falls from +0.95 to +0.65, staying positive in 93.9% of partitions. About **two thirds of the**

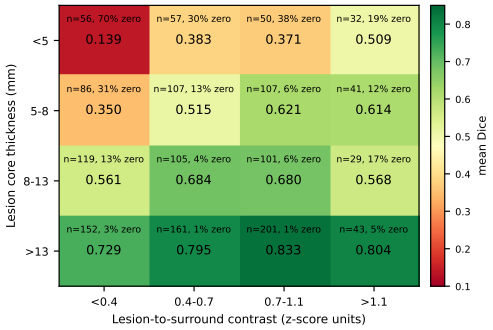

Fig. 1. Mean official Dice by core thickness and lesion-to-surround contrast ($n = 1447$ out-of-fold; cells with $n \geq 8$). Labels give mean Dice, case count and the zero-score fraction.

apparent gain from post-processing search transfers across sites, while the configuration chosen on half the sites ranks only 6th of 21 on the other half (median). The thresholds we report are one defensible setting, not an optimum.

3.3 Associations with residual performance

Among sites with at least ten cases, the zero-Dice rate ranges from 0% (five sites) to 53%. Lesion volume explains much of this variation (zero-Dice falls from 43.8% below 0.5 mL to 1.1% above 50 mL) but not all: within the 0.5–2 mL stratum, mean Dice per site still ranges from 0.31 to 0.77 across the eighteen sites contributing at least seven such cases.

We therefore measured, per case, the lesion-to-surround contrast (mean z-score difference between the lesion and a spacing-aware 1–5 mm shell, under nnU-Net's own normalisation), the contrast-to-noise ratio, and shape descriptors including *core thickness* (twice the maximum inscribed-sphere radius). Rank correlations with per-case Dice ($n = 1447$): core thickness $\rho = +0.616$ ($p = 3.5 \times 10^{-152}$), surface-to-volume ratio -0.599, volume $+0.562$, contrast $+0.231$, and contrast-to-noise ratio only $+0.072$. All are computed from the reference mask, so they explain performance only after the fact. Thickness correlates with Dice more strongly than volume does and is not merely a proxy for it: within 2–10 mL, lesions thinner than 7 mm average Dice 0.426 against 0.663 above 11 mm. Contrast alone is a weak correlate.

The association between contrast and Dice differs by thickness (Fig. 1). Splitting the cohort at the median contrast (0.62 in z-score units), the Dice gap between high- and low-contrast cases is $+0.172$ for lesions thinner than 8 mm ($n = 536$) but only $+0.085$ for thicker ones ($n = 911$); a case-level bootstrap puts the difference between these gaps at $+0.087$, 95% interval $[+0.028, +0.147]$. The contrast-associated gap is therefore about twice as large among thin lesions. Consistent with this interaction, the lowest-contrast site has a median core thickness of 11.7 mm and no zero-Dice cases, whereas the thin, low-contrast cell of Fig. 1 has a mean Dice of 0.139 and a 70% zero rate. A preprocessing artefact (two

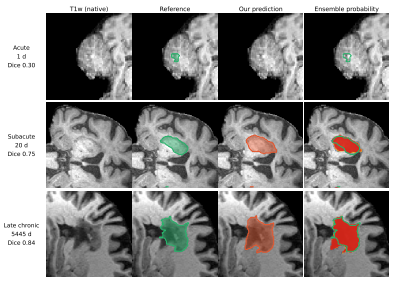

Fig. 2. One case near the median Dice of each chronicity stratum, at the slice with the largest reference cross-section; 95 mm square crops centred on the reference centroid, aspect-corrected for voxel spacing. Green: reference; orange: our output; right: probability.

sites with non-zero image background, which perturbs intensity normalisation) showed no evidence of explaining their performance ($p = 0.74$).

3.4 Negative results

Ninety post-processing configurations and one fine-tuning run were evaluated on all 1452 cases; none outperformed the fallback across the official metrics. Hysteresis thresholding (nine lo/hi pairs) lifts lesion-F1 by 0.008 ($p < 0.01$) and brings zero-Dice cases to 165, but enlarges *every* lesion, so AVD degrades from 6.18 to 6.56–7.16 mL; a peak-relative threshold reduces Dice and lesion-F1 while increasing AVD.

Chronicity-conditioned thresholds. Days-post-stroke (metadata we do not use for prediction) correlates with per-case Dice at $\rho = +0.273$ ($p = 2 \times 10^{-20}$), and acute cases (<14 days, $n = 124$) average Dice 0.349 with a 37.1% zero rate against 6–12% elsewhere. Lowering the threshold for acute cases only, leaving 89% of the cohort bit-identical, changed nothing: every paired test on the acute subgroup was non-significant from 0.40 down to 0.25, and the acute zero-Dice count rose from 45 to 46. Knowing which cases are hard told us nothing about whether post-processing could rescue them.

Instance-balanced patch sampling. nnU-Net draws foreground patch centres uniformly over foreground *voxels*, so the chance of centring on a lesion is proportional to its volume: lesions below 0.05 mL are 33.1% of all lesions but only **0.1995%** of foreground volume. We rebuilt the stored sampling table to be uniform over connected components (1.84% → 16.64% selection probability for small lesions) and fine-tuned the Dice+CE family for 150 epochs. Fine-tuning reduced PR-AUC ($\Delta = -0.0044$, $p = 8 \times 10^{-5}$) and the resulting ensemble differs from the current one on no metric. A pre-specified check found the effect confined to the smallest stratum as intended (<0.5 mL: ΔDice +0.0049), but two recovered cases in 192 cannot offset the calibration loss. With per-sample rather than batch Dice, a patch holding ∼20 foreground voxels gives an unstable loss, and this change raised the share of such patches from ∼0.2% to ∼17%.

4 Discussion and Conclusion

Gating Dice at IoU ≥ 0.25 shifts the value of a change from improving cases that already match to moving near-misses above the threshold, and can favour changes that redistribute accuracy rather than provide a consistent gain, as the confidence exemption did, which per-case win and loss counts expose but aggregate means hide. The attribution further showed that our own minimum-lesion-size threshold lay above the 33rd percentile of the reference size distribution and guaranteed a zero score on 17 cases, which aggregate metrics did not reveal.

Site-level differences were not explained by contrast alone, because contrast-to-noise ratio was nearly uncorrelated with Dice. Core thickness showed the strongest association, with contrast modifying the outcome mainly among thin lesions. This pattern may reflect limited preservation of thin structures during encoding and decoding, but external validation is needed before using it to prioritise architectural changes over intensity augmentation. Post-processing cannot reach the largest group of failures: 86 of the 177 zero-Dice cases carried essentially no probability mass inside the reference lesion. All code is released under Apache 2.0 at `github.com/luwill/isles26-dual-family-ensemble`.

Disclosure of Interests. The author has no competing interests to declare.

References

1. Abraham, N., Khan, N.M.: A novel focal Tversky loss function with improved attention U-Net for lesion segmentation. In: IEEE ISBI 2019, pp. 683–687 (2019)
2. Hernandez Petzsche, M.R., et al.: ISLES 2022: a multi-center MRI stroke lesion segmentation dataset. Sci. Data **9**, 762 (2022)
3. Isensee, F., et al.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
4. Isensee, F., et al.: nnU-Net revisited: a call for rigorous validation in 3D medical image segmentation. In: MICCAI 2024. LNCS, Springer (2024)
5. ISLES'26 challenge (ATLAS reloaded): task description, data release and evaluation code. `isles-26.grand-challenge.org` and `github.com/ezequieldlrosa/isles26` (2026)
6. Kofler, F., et al.: Panoptica – instance-wise evaluation of 3D semantic and instance segmentation maps. arXiv:2312.02608 (2023)
7. Liew, S.L., et al.: A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. Sci. Data **9**, 320 (2022). Dataset: `fcon_1000.projects.nitrc.org/indi/retro/atlas.html`
8. Maier-Hein, L., et al.: Why rankings of biomedical image analysis competitions should be interpreted with care. Nat. Commun. **9**, 5217 (2018)
9. Salehi, S.S.M., Erdogmus, D., Gholipour, A.: Tversky loss function for image segmentation using 3D fully convolutional deep networks. In: MLMI 2017, pp. 379–387 (2017)