This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Diagnosing the Recall Tail: A Class-Imbalance Study of nnU-Net for Detection and Segmentation of Heterogeneous Moderate–Severe TBI Lesions

Geofray Paul J

Independent Researcher
geofraypaul1223@gmail.com

Abstract. Lesions in moderate-to-severe traumatic brain injury (msTBI) vary enormously in size, number and tissue involvement, which breaks standard neuroimaging pipelines. It also creates a class-imbalance problem that most challenge reports treat as background detail: in the AIMS-TBI 2026 training set, 41.7% of subjects have no lesion at all. We examine how this imbalance shapes a self-configuring nnU-Net trained on a single 16 GB GPU, and serve both challenge tasks from it. The surface-distance metric separating mid-field entries from leading ones tracks a recall tail rather than boundary quality: roughly 7% of scored cases, the outright misses and gross mislocalisations, carry more than half the mean average symmetric surface distance (ASSD), and removing them drops the cross-validation mean from 8.55 mm to 6.19 mm while the median holds near 2.4 mm. The tail is almost entirely a small-lesion effect: the failure rate falls monotonically with lesion volume and reaches zero above the cohort median, confining the deficit to a characterisable subgroup. Three cheap interventions fail to touch it. Lowering the softmax threshold and batch-pooling the Dice loss are reported as negative results with a mechanism; test-time mirroring, which we scored against the organisers' hidden labels, improves Dice and HD95 slightly while ASSD moves the wrong way. Because surviving lesion volume separates the classes cleanly, a subject-level volume rule reaches a balanced accuracy of 0.846 on the official test set and 0.924 held-out; segmentation reaches a lesion-containing Dice of 0.505. On imbalanced benchmarks, the shape of the failure distribution deserves reporting alongside the aggregate.

Keywords: Traumatic brain injury · Lesion segmentation · Lesion detection · Class imbalance · nnU-Net · Surface distance metrics.

1 Introduction

Lesions are the hallmark of moderate-to-severe traumatic brain injury, and they are difficult precisely because they are so irregular. They may be focal or diffuse, vary by orders of magnitude in size and number, often appear bilaterally, and cross grey matter, white matter and cerebrospinal fluid without respecting

anatomical boundaries, which confounds registration, normalisation and parcellation and makes msTBI lesion analysis harder than the more regular profiles seen in stroke, multiple sclerosis or tumour. Every response is unsatisfactory in some direction: ignoring lesions corrupts downstream measurements, excluding large-lesion patients discards the cases that matter most, and manual segmentation does not scale, while existing automated tools either demand several sequences that multi-site consortia cannot supply uniformly or work only on CT. The AIMS-TBI challenge therefore asks for a pipeline running on T1-weighted (T1w) MRI alone, and its 2026 edition splits the problem across two leaderboards: *detection*, scored over every subject, and *segmentation*, scored only on lesion-bearing subjects.

We treat the class imbalance as the object of study rather than a nuisance. The paper contributes **a diagnosis**, that the surface-distance metric most separating entries is dominated by a thin tail of catastrophic cases rather than by generally poor boundaries, which we quantify, connect to the 41.7% empty-mask imbalance and resolve against lesion volume; **three negative results**, since the cheapest levers available, a lower foreground threshold, a batch-pooled Dice loss and test-time mirroring, each fail to recover the misses for a identifiable reason, together moving the problem from calibration to recall; and **a dual-task system under a hard compute limit**, one nnU-Net on one 16 GB GPU serving a geometry-safe segmentation container and a volume-threshold detection rule, with competitive detection and a segmentation result whose failure modes we characterise rather than hide.

2 Related Work

Self-configuring segmentation and prior editions. nnU-Net [1] remains the sensible starting point: it fixes architecture, patch and batch size, normalisation and schedule from a dataset fingerprint, and offers residual-encoder (ResEnc) presets that buy accuracy at higher compute cost. We use ResEnc-M, the largest preset fitting in 16 GB. The challenge ran at MICCAI 2024 and 2025, where the best model reported an overall Dice of 0.637 and a lesion-only Dice of 0.515 [6] and a top-six ensemble inside nnU-Net reached 0.597 [5]. Rather than build a stronger ensemble, we ask why surface-distance rank behaves as it does when the data is this imbalanced.

Imbalance, loss design, and what the metrics reward. The soft Dice loss is standard for imbalanced segmentation, but its gradients destabilise when foreground is sparse or absent; generalised Dice weighting [2] and the Tversky losses [3] address this by reweighting and by trading precision for recall, while nnU-Net's `batch_dice` switch pools the loss across the batch on the argument that this stabilises gradients when many images have no foreground (Section 4.4). The scoring conventions matter as much as the loss. Dice is bounded below at zero and HD95 tolerates a thin tail, but ASSD averages the whole boundary and is unbounded, so a handful of bad cases dominate it; detection uses balanced accuracy, punishing both the always-present and always-absent predictor; and final

ranking follows BraTS case-wise rank aggregation with significance testing [4]. Under these conventions the shape of the failure distribution matters more than its mean.

3 Data and Methods

3.1 Dataset

The challenge supplies T1w MRI with expert lesion masks from the ENIGMA msTBI cohort [6]. We kept the 552 subjects with both a T1w image and a lesion mask, dropping 7 that carried only diffusion data since the evaluation is T1w-only. Of the 552, 230 (41.7%) are lesion-free and 322 lesion-bearing. That near-even split is convenient for calibrating a detection threshold and brutal at the voxel level. The cohort is heterogeneous in almost every other respect: ages 6 to 85 (265 paediatric, 287 adult), three scanner vendors (Siemens 272, Philips 167, GE 113), time since injury 0 to 5801 days, lesion volume 14 to 214,208 voxels, and 83% of lesion-bearing subjects multifocal (median 5 components).

3.2 Preprocessing

Masks were binarised with (mask > 0), since 143 encoded subtypes as values 2 to 6. We did not resample; 549 of 552 volumes are 1 mm isotropic at 256^3. All 230 lesion-free subjects were kept as all-background training cases: dropping them would discard 42% of the data and remove exactly the negatives detection depends on.

3.3 Model and compute

We used nnU-Net v2 with the ResEnc-M plan in `3d_fullres`; the framework chose a patch size of $128 \times 160 \times 112$, batch size 2, and the default Dice and cross-entropy loss. The 16 GB memory budget set the ceiling on architecture, patch and batch size, and fixed ResEnc-M as the largest usable preset. A separate constraint was wall-clock time: the default 1000-epoch schedule would have taken roughly 150 hours across five folds, so we trained each fold for 250 epochs instead, about 39 hours in total. This is a compressed complete schedule rather than a truncated one: the trainer variant sets `num_epochs` to 250 and the same value is passed to nnU-Net's polynomial learning-rate scheduler as its step ceiling, so the rate follows $\eta_0(1 - t/250)^{0.9}$ and anneals to zero at epoch 250 rather than being interrupted part-way through a 1000-epoch decay. The reduced budget remains a real limitation (Section 5), but the model is fully annealed. For the challenge's T4 target we repinned the runtime to PyTorch 2.5.1 with CUDA 12.4 and checked it reproduced training numerically (Section 4.6).

3.4 Serving both tasks from one model

Both containers run the same five-fold ensemble with test-time augmentation off by default for timing headroom; Section 4.7 reports a second submission with mirroring enabled. For segmentation we label components with 26-connectivity, drop anything below 13 voxels, one below the smallest true lesion, and copy the input geometry onto the `int8` output. We never keep only the largest component, since 83% of lesion-bearing subjects are multifocal. Detection sums surviving volume above the 10-voxel scoring floor and calls a lesion present if the total reaches τ. Both containers fail safely, emitting an empty mask or `false` rather than crashing, which matters because the ranking assigns worst-case scores to missing outputs. Median inference was about 35 s per case on a T4 without augmentation and 90 s with it, against a 7 min limit.

3.5 Choosing τ

The threshold τ is the only free parameter in the detection rule. We fixed it on the five-fold out-of-fold (OOF) predictions, choosing the value maximising balanced accuracy, with leave-one-fold-out selection giving an optimism estimate. This produced $\tau = 100$ voxels, an OOF balanced accuracy of 0.859 and a negligible in-sample gap. We did not tune τ on any later evaluation set.

3.6 Evaluation

Three levels, increasingly severe: five-fold cross-validation over the 552 training subjects, used only for diagnosis and ablation; an independent held-out set of 102 labelled subjects with zero ID overlap; and the official test phase, whose ground truth the organisers hold. Boundary metrics use per-case header spacing. Segmentation is scored on subjects whose ground-truth lesion exceeds 50 voxels, per the challenge convention, with failures assigned worst-case values.

4 Experiments and Results

4.1 Where the surface-distance penalty comes from

Our development submission landed mid-field on Dice with a comparable HD95, but its ASSD was several times worse than anything else on the board. Since ASSD averages the entire boundary and is unbounded, it is the one metric a few catastrophic cases can dominate, which pointed us at the failure distribution rather than the boundaries. We recomputed per-case Dice, HD95 and ASSD over the five-fold OOF predictions, handling empty predictions and ground truths explicitly (distances flagged, never substituted). On the scored subset (ground-truth volume above 50 voxels, $n = 305$), 22 cases (7.2%) fail outright: 9 predict nothing and 13 predict in the wrong place with zero overlap. The other 283 sit at a median ASSD of 2.3 mm.

Table 1: Complete failures (Dice = 0) by ground-truth lesion volume, scored subset ($n = 305$, quintiles of ground-truth volume).

Quintile	Volume (voxels)	n	Failures	Rate	Median Dice
Q1	61–377	61	17	27.9%	0.31
Q2	380–1118	61	4	6.6%	0.52
Q3	1128–3709	61	1	1.6%	0.62
Q4	3726–14,325	61	0	0.0%	0.73
Q5	15,324–214,208	61	0	0.0%	0.80

The failure mass is concentrated: the ten worst cases by ASSD carry 27.6% of the total and the twenty worst 42.9%, every one a miss or a stray speck, the worst pitting a 61-voxel lesion against a 2-voxel prediction at 105 mm. Removing the 22 Dice-zero failures drops the mean ASSD from 8.55 to 6.19 while the median barely moves, so the mean describes a tail rather than a central tendency. On the cases the model finds, boundary accuracy is adequate; the rank penalty comes from recall.

4.2 What falls into the tail

The tail is not drawn at random. The 22 failures carry a median ground-truth volume of 165 voxels (IQR 108 to 330) against 2544 (IQR 648 to 11,084) for the 283 remaining cases, significant at $p < 0.001$ (Mann–Whitney U). Table 1 resolves this by volume quintile: the failure rate falls monotonically from 27.9% in the smallest quintile to zero in the largest two, and not one failure occurs above the cohort median volume of 2034 voxels. Seventeen of the 22 sit in the smallest quintile, and 21 of 22 below 1000 voxels. They are not concentrated in any one fold (χ^2, $p = 0.37$), so this is a property of the lesions rather than of an unlucky split.

The recall deficit is therefore not diffuse but a small-lesion detection floor: above roughly 1000 voxels the model essentially does not miss. Recall-specific interventions should accordingly be evaluated on small lesions rather than on a cohort mean that would hide a real gain on 7% of cases. Crossing the tail against the scanner, age and chronicity covariates of Section 3.1 remains open.

4.3 Lowering the threshold does not help

If the misses were merely under-called, dropping the foreground softmax threshold below 0.5 ought to rescue them. Sweeping it over the saved five-fold probabilities, the best value near 0.25 gained +0.004 mean lesion Dice over the argmax baseline (0.570 against 0.566; leave-one-fold-out 0.571), which is noise. Per case the result is sharper than the aggregate suggests. Of the 22 failures, 20 remain at exactly Dice 0 at the optimal threshold. The two that move are both mislocalisations, reaching Dice 0.007 and 0.021, which is not a recovery in any useful sense; none of the 9 empty predictions moves at all. That split is the mechanism: lowering the threshold can nudge a prediction that already sits in the wrong place,

but an empty prediction has no sub-threshold mass to recover, so no threshold can turn a miss into a hit. These misses are not a calibration artefact.

4.4 Nor does batch-pooled Dice

With 42% of masks empty, the pooled form of the Dice loss should give steadier gradients. We trained the fold with the most failures under an isolated plan with `batch_dice`=True at the same 250-epoch budget, so the loss is the only change. The aggregates improve sharply but describe the wrong thing. Over the 111 cases mean Dice rises from 0.512 to 0.576, median from 0.679 to 0.732, and mean ASSD falls from 6.01 mm to 4.25 mm, yet not one lesion-bearing miss is recovered. Twelve cases move from Dice 0 to > 0 and all twelve are lesion-free scans where the baseline hallucinated and `batch_dice` correctly stayed silent, while two lesion-bearing cases regress to Dice 0, one of them a new miss, so true misses rise from 1 to 3. Zero-overlap predictions across all 111 cases fall from 35 to 26, but that count is dominated by lesion-free scans; restricted to the scored lesion-bearing subset it moves only from 3 to 2. The direction of the aggregate reverses on that subset too: over the 53 scored cases mean Dice falls from 0.597 to 0.571. The whole of the headline gain therefore comes from empty scans. What `batch_dice` buys is specificity on lesion-free images, which helps detection and leaves recall untouched.

This result is preliminary. It rests on a single fold under one plan rather than the complete cross-validation, so it should be read as suggestive: it shows that batch pooling did not recover misses here, not that batch pooling fails to help recall in general. A full five-fold ablation would be needed to settle the point.

4.5 Detection: the classes separate on volume

Since lesion volume is the discriminative feature, we swept τ on the held-out set. The separation is stark: lesion-free subjects have a 90th percentile of only 24 surviving voxels against a lesion-bearing median of 3069. At the deployed $\tau = 100$, held-out balanced accuracy is 0.924 (sensitivity 0.889, specificity 0.958), comfortably above the 0.859 OOF estimate. The held-out optimum is $\tau^\star = 50$ at 0.950, and we did not take it: the advantage rests entirely on four small lesions of 37 to 89 voxels clearing the lower bar on this one sample, and $\tau = 100$ sits on a downslope rather than a plateau (balanced accuracy 0.866, 0.939, 0.950, 0.942, 0.924, 0.914, 0.887 at $\tau = 10$ to 200), the profile of a threshold fitted to noise. We kept the pre-registered $\tau = 100$.

4.6 Held-out performance and container parity

On the 54 lesion-bearing held-out subjects, segmentation reaches a mean lesion Dice of 0.567, median 0.681, median HD95 22.1 mm, median ASSD 2.42 mm, with 1 miss and 2 mislocalisations: the same tail structure the cross-validation showed. The native evaluation reproduces the deployed containers on smoke cases (detection Booleans identical, masks at Dice 0.9998 to 1.0), so these are container numbers, not an optimistic proxy.

4.7 The official test set

Detection reached a balanced accuracy of **0.846** (sensitivity 0.863, specificity 0.829) at the deployed $\tau = 100$, placing among the leading entries; we report the standing as provisional, since final ranking applies case-wise rank aggregation with significance testing. Specificity is the informative number, 0.958 held-out against 0.829 on test: the operating point transferred imperfectly, and on the true distribution specificity rather than sensitivity binds, so chasing $\tau^\star = 50$ would have traded away the weaker of the two for sensitivity the test set did not reward.

Segmentation reached a lesion-containing Dice of **0.5048**, HD95 30.76 mm and ASSD 11.15 mm. Against the held-out estimate the test Dice is lower and the ASSD far higher, which we read as the recall tail of Section 4.1 reasserting itself on a harder distribution rather than a collapse in boundary quality, since an unbounded surface average above 10 mm is what a handful of misses produces.

A second submission let us score one further intervention against ground truth we never saw. Enabling test-time mirroring, every other setting fixed, gained +0.014 mean lesion Dice in cross-validation (bootstrap 95% CI $[+0.009, +0.019]$; $n = 302$ cases with a ground-truth component above 50 voxels; 210 improved and 71 worsened, of which 207 and 67 moved by more than 0.0005 Dice). The gain sat entirely in cases already located: one failure recovered, to Dice 0.085, two further cases fell to zero, and the Dice-zero count rose rather than fell. The test set reproduced the pattern, Dice improving by 0.0010 and HD95 by 0.55 mm while ASSD worsened by 0.34 mm. Averaging over mirrored predictions refines boundaries the model has found and cannot recover those it misses, so bounded metrics improve while the unbounded surface average moves the wrong way. This is the third cheap intervention to leave the recall tail intact, and the only one confirmed against hidden labels.

5 Discussion and Limitations

The practical lesson is that on an imbalanced benchmark the failure distribution deserves reporting next to the aggregate. The same 7% recall tail is nearly invisible in Dice and HD95 rank yet dominates ASSD rank. Anyone reading only the mean would conclude their boundaries are poor and spend compute on a bigger model or a longer schedule, neither of which addresses the problem. We tested the three cheapest interventions available and none recovered a miss. Section 4.2 locates the deficit precisely: it is a small-lesion detection floor rather than a diffuse weakness, which makes recall-specific work both the right target and a measurable one. A lesion-focused loss, aggressive foreground oversampling, or a recall-tuned model paired with a separate precision filter should all be assessed on the small-volume strata of Table 1, where the deficit lives, rather than on a cohort mean that would hide a real gain on 7% of cases.

One may ask whether the tail reflects the reduced schedule rather than a structural property. Three observations argue against it, none decisive alone. An

undertrained network produces diffuse, low-confidence foreground whose misses are recoverable by lowering the decision threshold; the sweep in Section 4.3 recovers none of the nine empty predictions and leaves 20 of 22 failures at exactly zero, so the model is not hesitating but confidently predicting nothing. The failure distribution is bimodal rather than shifted, which is not the signature of a uniform convergence deficit. And the batch-pooled fold of Section 4.4, trained at the identical budget, raised mean fold Dice from 0.512 to 0.576 without recovering a single miss. A full-schedule replication remains the proper way to close the question, and we did not have the compute for one. The scope of the interventions was likewise fixed deliberately: each costs nothing beyond a configuration change or an inference flag, which is what makes these negative results portable. Tversky and focal Tversky [3], generalised Dice [2] and aggressive foreground oversampling all require retraining and fall outside that scope. They are the designated next step rather than an open question.

Several caveats bound these results. The 16 GB limit fixed ResEnc-M as the largest usable preset and a 150-hour full schedule was infeasible, so our segmentation result is a characterised baseline and not a ceiling; neither constraint reflects a limit of the method. The `batch_dice` result is a single fold, although the mirroring ablation and the failure concentration behind the central diagnosis both span the full cross-validation set. Diffusion data went unused, since the test set is T1w-only. The held-out threshold optimum proved sensitive to a handful of small lesions; we fixed τ in advance, and the test set suggests that was right.

6 Conclusion

We studied class imbalance in a single self-configured nnU-Net serving both msTBI tasks on one consumer GPU. The surface-distance penalty separating leaderboard entries is a recall tail rather than a boundary deficit; it is concentrated almost entirely in small lesions, vanishing above the cohort median volume, and it resists the three cheapest fixes available. The system reached a detection balanced accuracy of 0.846 on test and 0.924 held-out, and a segmentation Dice of 0.505, whose inflated surface distance is itself further evidence for the diagnosis. On imbalanced benchmarks, the location of a model's failures is as informative as its aggregate scores.

Acknowledgments. Large language models were used in the preparation of this work: to assist in generating and refining the analysis and experiment code, and in drafting and editing the manuscript text. The author designed and ran all experiments, verified every reported result against the underlying data, and takes full responsibility for the content and accuracy of this article.

The data were provided through the AIMS-TBI 2026 Challenge and shared via the ENIGMA Brain Injury working group. The author thanks the challenge organisers at the University of Utah, the ENIGMA Brain Injury consortium and the contributing sites, and confirms the data were used solely for the challenge in accordance with the signed data use agreement.

Disclosure of Interests. The author has no competing interests to declare that are relevant to the content of this article.

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. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M.J.: Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. In: DLMIA/ML-CDS, pp. 240–248 (2017)
3. Salehi, S.S.M., Erdogmus, D., Gholipour, A.: Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks. In: MLMI, pp. 379–387 (2017)
4. Baid, U., et al.: The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification. arXiv:2107.02314 (2021)
5. Dhamat, G., Sundaresan, V.: Traumatic Brain Injury Segmentation using an Ensemble of Encoder-decoder Models. arXiv:2509.24684 (2025)
6. Deutscher, E., et al.: The AIMS-TBI 2025 MICCAI Challenge. In: Bakas, S., et al. (eds.) MICCAI 2025 Challenges, LNCS, vol. 16377. Springer (2026). `https://doi.org/10.1007/978-3-032-16370-7_24`