Dual-Model nnU-Net Strategy for TBI Segmentation and Detection in T1-weighted MRI

Sakshi Rathi[1], Yuk Sham[1,2], and Uzma Samadani[1,3]

[1] Department of Bioinformatics and Computational Biology, University of Minnesota, Minneapolis, MN, USA
rathi036@umn.edu
[2] Department of Integrative Biology and Physiology, University of Minnesota, Minneapolis, MN, USA
[3] Surgical Services, Minneapolis VA Medical Center, Minneapolis, MN, USA

Abstract. We present our approach to the AIMS-TBI 2026 Grand Challenge for automated traumatic brain injury (TBI) lesion segmentation and detection from T1-weighted MRI. Starting from baseline nnU-Net v2 models, we identified lesion contrast heterogeneity as the primary failure mode and developed a custom trainer with domain-specific contrast augmentations. Our final detection submission uses a 5-fold baseline ensemble with conservative post-processing, while segmentation uses a single-fold model fine-tuned with contrast augmentations on combined training and validation data (652 cases). Note that V2 differs from V1 in multiple factors (additional data, augmentations, fine-tuning), so individual contributions are not isolated. Our approach achieved a detection score of 0.8612 (Sensitivity 0.775, Specificity 0.947) and segmentation Dice of 0.4957. We analyze the progression from baseline to final submission, the impact of post-processing strategies, and lessons learned regarding the detection–segmentation trade-off.

Keywords: Traumatic brain injury · Medical image segmentation · nnU-Net · Data augmentation · Lesion detection

1 Introduction

Traumatic brain injury affects millions worldwide, and accurate identification and delineation of lesions from neuroimaging is critical for treatment planning and prognosis. The AIMS-TBI 2026 Challenge evaluates algorithms on two complementary tasks: (1) binary segmentation of TBI lesions, and (2) detection of lesion presence, using T1-weighted MRI scans.

Key challenges include: **extreme size variability** with volumes ranging from <1 mL to >200 mL (median 0.74 mL); **heterogeneous appearance** varying with lesion age, type, and location (Fig. 1); **class imbalance** with ∼42% negative cases; and **small lesion detection** where many lesions occupy fewer than 100 voxels.

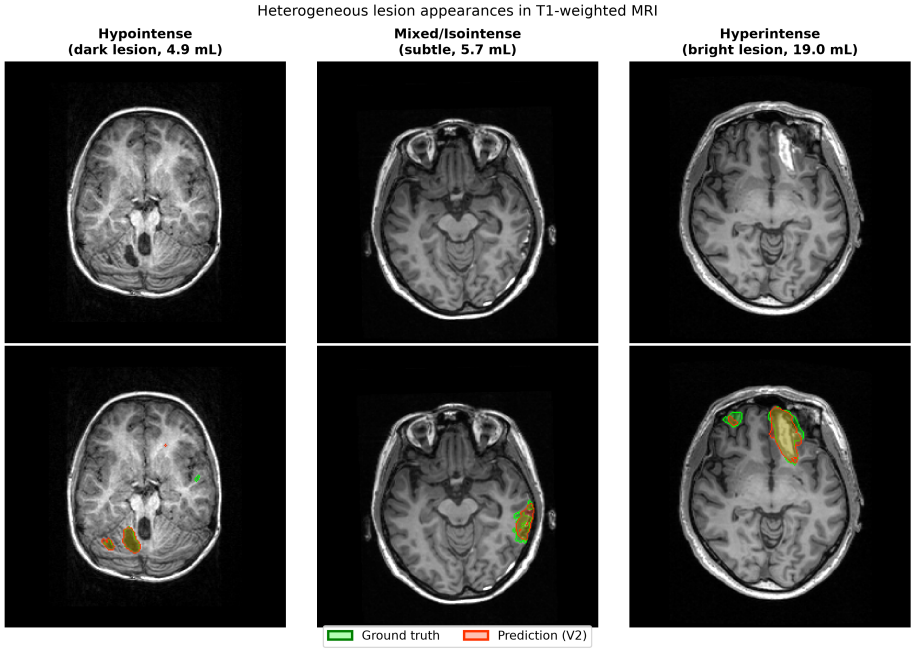

Fig. 1. Heterogeneous lesion appearances in T1-weighted MRI. Three cases illustrating hypointense (dark), mixed-intensity (subtle), and hyperintense (bright) lesions. Top: raw MRI. Bottom: ground truth (green) and V2 model predictions (red).

2 Methods

2.1 Dataset

The challenge provides 552 T1-weighted MRI scans with binary lesion annotations (321 positive, 231 negative) and 100 validation cases. All images were resampled to 1.0 mm isotropic resolution. We define two nnU-Net datasets: Dataset001 contains the original 552 training cases, and Dataset002 contains the combined 552 training and 100 validation cases (652 total). The lesion volume distribution (Fig. 3) is heavily right-skewed: median 0.74 mL (IQR: 0.02–4.30 mL), spanning four orders of magnitude.

2.2 Stage 1: Baseline Models

We trained three nnU-Net v2 (v2.8.1) configurations with 5-fold cross-validation on Dataset001 using a custom trainer (AIMSTBITrainer) that added CSV/text logging and 50% foreground oversampling (vs. default 33%):

1. **3d_fullres (nnUNetPlans)**: 3D PlainConvUNet, 6 stages, patch size 128× 160 × 112, instance normalization, LeakyReLU.

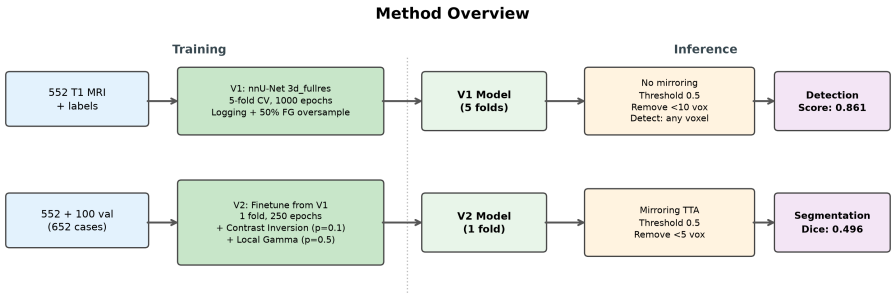

Fig. 2. Method overview. V1 (baseline 5-fold) serves detection via ensemble-driven specificity; V2 (augmented, fine-tuned on 652 cases) serves segmentation via sensitivity to diverse lesion appearances.

2. **ResEncM (nnUNetResEncUNetMPlans)**: Residual encoder variant with larger capacity.
3. **2D (nnUNetPlans)**: 2D baseline for comparison.

All used Dice + CE loss, SGD with polynomial LR decay (initial 0.01), trained for 1000 epochs on 2× NVIDIA H100 80GB GPUs with DDP.

2.3 Stage 2: Ensemble

We constructed a weighted softmax ensemble of 3d_fullres and ResEncM (weights 1.0/1.2) by averaging softmax probability maps across all 10 models (5 folds × 2 configs). The 2D model was excluded because of consistently inferior performance (∼0.54 vs 0.63+ Dice).

2.4 Stage 3: Failure Analysis and Custom Augmentation

After initial Docker submission showed strong detection but weak segmentation, we ran a failure analysis on validation results and identified **lesion contrast heterogeneity** as the primary failure mode (Fig. 1). The baseline models struggled with lesions whose appearance deviated from the typical hypointense pattern.

We extended AIMSTBITrainer with domain-specific augmentations:

Contrast Inversion ($p = 0.10$): Globally inverts image contrast, forcing the model to learn shape and boundary features rather than intensity polarity.

Local Gamma Transform ($p = 0.50$): Applies lesion-specific contrast modification with target ratio $\in [0.5, 1.6]$, simulating hypointense, mixed-intensity, and hyperintense lesion appearances.

Figure 2: Distribution of lesion volumes in training set (N=321 lesion-positive cases)

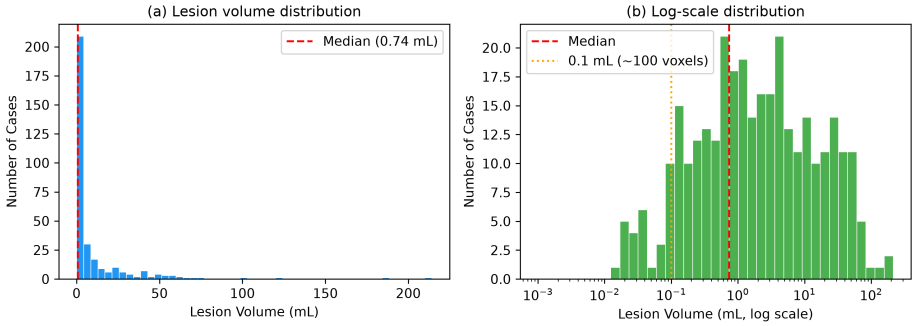

Fig. 3. Lesion volume distribution (N=321 positive cases). Most lesions are sub-milliliter; many fall below 0.1 mL (∼100 voxels) where segmentation becomes unreliable.

2.5 Stage 4: Fine-tuning

The augmented model (V2) was trained on Dataset002 (652 cases) as a single fold, initialized from V1 fold_0 weights (trained for 1000 epochs), and fine-tuned for 250 additional epochs—yielding 615 total logged epochs in the training curve (Fig. 4), as the epoch counter continued from the checkpoint. This improved sensitivity substantially but introduced more false positives, motivating post-processing optimization.

2.6 Stage 5: Post-Processing Optimization

The fine-tuned model's increased false positive rate required careful post-processing. We explored:

- Probability threshold at 0.5
- Connected component filtering (5–100 voxel minimum)
- Gaussian smoothing ($\sigma = 1.0$) and morphological closing ($r = 1$)
- Volume floors (50–100 voxels)

Ultimately, aggressive filtering improved validation Dice but removed small real lesions on the hidden test set. We adopted minimal post-processing (threshold + remove <10 voxels) for the final segmentation submission.

2.7 Final Submissions

Segmentation: V2 model (AIMSTBITrainer with contrast augmentations, single fold, Dataset002), mirroring TTA enabled, relaxed post-processing (threshold 0.5, remove <5 voxels). The augmented model's ability to handle diverse lesion intensities was critical for maximizing Dice.

Figure 3: Training curves for V2 model (Dataset002, fold_0)

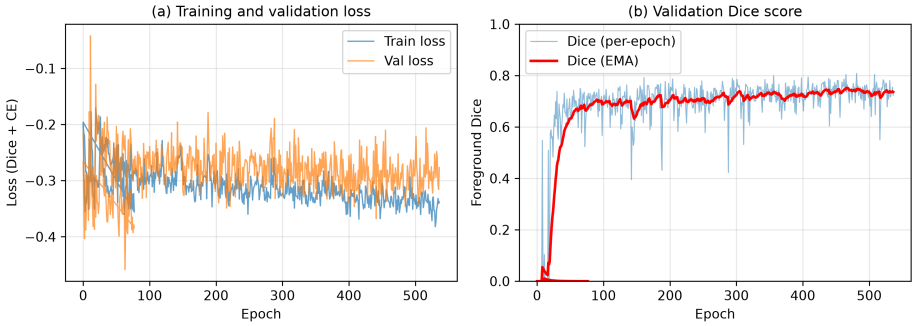

Fig. 4. Training curves for fine-tuned V2 model (615 logged epochs). (a) Loss convergence. (b) Exponential moving average Dice plateaus at approximately 0.77.

Detection: V1 model (AIMSTBITrainer with logging only, 5-fold, Dataset001), no mirroring, conservative post-processing (threshold 0.5, remove <10 voxels). The 5-fold ensemble's probability averaging naturally suppresses false positives, yielding high specificity for binary detection.

3 Results

3.1 Method Progression

Table 1 summarizes results across development stages, from baseline through final hidden test set submission.

The ensemble substantially improved specificity (0.687 → 0.970) and balanced accuracy (0.917) at the cost of slightly lower Dice. The augmented V2 model substantially improved validation Dice (0.554 → 0.637) and sensitivity (0.85 → 0.944).

Hidden test set results. Detection: Sensitivity 0.775, Specificity 0.947, Score 0.8612. Segmentation: Dice 0.4957, HD95 32.26 mm, ASSD 11.38 mm.

3.2 Post-Processing Ablation

Table 2 compares two post-processing strategies evaluated on the hidden test set. The boundary-optimized strategy, while improving metrics on our validation set, proved too aggressive—the volume floor of 100 voxels and morphological closing removed small but real lesions, reducing Dice from an earlier minimal-filtering submission to 0.496 on the final leaderboard.

Fig. 5 shows the strong volume-dependence of Dice: lesions below 0.1 mL frequently achieve Dice=0, explaining why volume floors harm performance when the hidden test set contains small lesions. Fig. 6 shows the sensitivity–specificity trade-off at different component size thresholds.

Table 1. Method progression from baseline to final submissions.

Stage	Configuration	Dice	Sens.	Spec.	Bal. Acc
5-fold cross-validation (552 training cases)					
1	3d_fullres (5-fold)	0.630	0.925	0.616	0.866
1	ResEncM (5-fold)	0.643	0.915	0.687	0.873
1	2D (5-fold)	0.542	—	—	0.841
2	Ensemble (3d + ResEncM)	0.626	0.863	0.970	0.917
Validation set (100 cases)					
—	V1 5-fold, no mirror†	0.554	0.850	—	—
4	V2 fold_0, +mirror‡	0.637	0.944	—	—
Hidden test set					
6	V1 5-fold → Detection	—	0.775	0.947	0.861
6	V2 fold_0 → Segmentation	0.496	—	—	—

†Truly held-out: V1 trained only on 552 training cases.
‡Not held-out: V2 was fine-tuned on Dataset002 (552+100 cases). These metrics illustrate the effect of contrast augmentation but should not be interpreted as independent validation. The hidden test set results (bottom) provide the unbiased comparison.

Table 2. Post-processing impact on hidden test set (segmentation task).

Strategy	Test Dice	Notes
Boundary-optimized (Gaussian, closing, min_vol=100)	0.496	Final submission

4 Discussion

Why V1 performed better for detection. The baseline 5-fold ensemble outperformed the augmented model on detection. Probability averaging across 5 folds naturally suppresses uncertain regions (specificity: 0.947), while retaining only high-confidence predictions. The augmented V2 model's increased sensitivity also increased false positives, reducing detection score despite better segmentation Dice.

Segmentation gap analysis. Our segmentation Dice (0.496) trails the leading submission (0.557) primarily in voxel-level overlap rather than spatial localization—our HD95 (32.26 mm) is comparable to the leader's (32.56 mm). Cross-validation reveals the root cause: 10.6% of lesion-positive cases achieve Dice=0 (complete misses, median volume 0.12 mL), and 19.3% fall below 0.3. These small-lesion failures disproportionately drag down the mean. Cases with lesions >5 mL achieve mean Dice 0.75, while those <0.1 mL average only 0.38. Using a single-fold model for segmentation (V2) sacrifices the boundary-smoothing effect of multi-fold ensembling, and while the contrast augmentations improved detection of diverse lesion types, they did not sufficiently improve delineation precision on the smallest lesions.

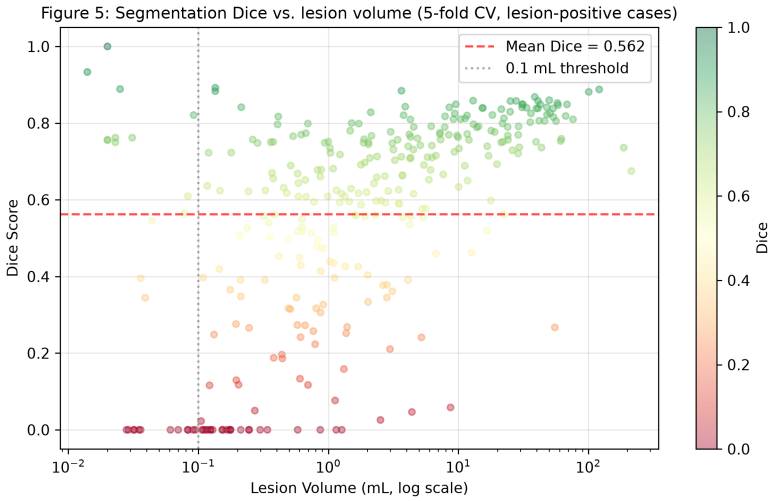

Fig. 5. Per-case Dice vs. lesion volume (5-fold CV, N=321 lesion-positive cases). Sharp degradation below 0.1 mL explains why aggressive volume floors harm hidden test set results.

Generalization gap. The 21% relative drop from CV Dice (0.630) to hidden test set Dice (0.496) suggests the hidden test set contains more small or ambiguous lesions. Post-processing sensitivity amplifies this: volume floors that help on validation remove real lesions on a harder test distribution.

Key lessons. (1) Conservative post-processing generalizes better—aggressive filtering removes real lesions. (2) Task-specific model selection is essential: ensembles excel at detection while augmented models handle segmentation diversity. (3) The combination of additional training data, contrast augmentations, and fine-tuning improved sensitivity from 85% to 94.4%; however, as these changes were introduced simultaneously, we cannot isolate the individual contribution of each factor. (4) For future work, a multi-fold augmented ensemble could combine V2's sensitivity with ensemble-driven boundary precision, potentially closing the gap on small-lesion segmentation. We also explored a gated ensemble, hysteresis thresholding, and multi-metric component filtering, but found them either too complex for the inference time budget or sensitive to parameter choice.

Computational requirements. All training was performed on 2× NVIDIA H100 80GB GPUs with DDP and `torch.compile`. V1 5-fold training ran 1000 epochs per fold at ∼10 s/epoch; V2 fine-tuning completed in approximately 2 hours. Inference takes seconds per case on a single H100 GPU.

5 Conclusion

We presented a staged approach to TBI lesion segmentation and detection: a fine-tuned model with custom contrast augmentations for segmentation (Dice

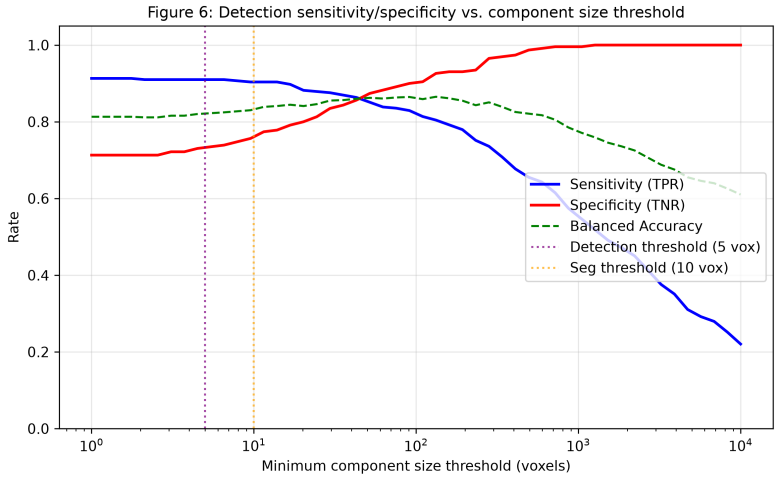

Fig. 6. Sensitivity/specificity vs. minimum component size threshold. Detection uses 5 voxels; segmentation uses 10. Larger thresholds trade sensitivity for specificity.

0.4957) and a conservative 5-fold baseline ensemble for detection (Score 0.8612). The key insight is that segmentation benefits from augmentation-driven sensitivity to diverse lesion appearances, while detection benefits from ensemble-driven specificity that suppresses false positives.

References

1. Isensee, F., et al.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
2. Isensee, F., et al.: nnU-Net revisited: A call for rigorous validation in 3D medical image segmentation. arXiv:2404.09556 (2024)
3. Maas, A.I., et al.: Traumatic brain injury: integrated approaches to improve prevention, clinical care, and research. The Lancet Neurology **16**(12), 987–1048 (2017)
4. Monteiro, M., et al.: Multiclass semantic segmentation and quantification of traumatic brain injury lesions on head CT using deep learning. J. Med. Imaging **7**(5), 056001 (2020)
5. AIMS-TBI 2026 Challenge. https://aims-tbi.grand-challenge.org/