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.

Task-Specific T1 MRI Pipelines for TBI: Knowledge-Distilled Segmentation and ICV-Cascaded Detection

Dahyun Kim and Jongjun Won

Neurophet Inc., Seoul, Republic of Korea
{dahyun, wjj910}@neurophet.com

Abstract. We present task-specific deep learning pipelines for the AIMS-TBI 2026 Challenge, addressing binary lesion detection and voxel-level segmentation from T1-weighted MRI with separate models optimized for each task objective. For segmentation, a MultiTalentV2-pretrained teacher is fine-tuned on original T1 volumes without skull stripping and distilled into a lightweight student network. For detection, we apply an ICV masking cascade in which the detection network operates on ICV-masked T1 volumes, improving specificity on lesion-free scans. The lesion segmentation and detection models use combined Dice and cross-entropy losses, with the segmentation model additionally incorporating a soft-target distillation objective. On the 103-case official AIMS-TBI validation set (55 lesion-positive, 48 lesion-free), the segmentation model achieves overall Dice 0.692 [95% CI: 0.621, 0.759] and lesion Dice 0.551 [0.468, 0.627] (HD95 25.7 mm, ASSD 6.6 mm); the detection cascade achieves balanced accuracy 0.879 [0.818, 0.935] (sensitivity 0.800, specificity 0.958). On the final held-out test set, the segmentation model achieves lesion Dice 0.5243 (HD95 29.5 mm, ASSD 9.3 mm) and the detection cascade achieves balanced accuracy 0.8482 (sensitivity 0.894, specificity 0.803).

Keywords: Traumatic Brain Injury · Lesion Segmentation · Detection · nnU-Net · Knowledge Distillation · Transfer Learning

1 Introduction

Moderate to severe traumatic brain injury (msTBI) is a leading cause of mortality and long-term disability worldwide, arising from external mechanical forces that induce primary injuries—haematomas, hemorrhages, and contusions—followed by secondary cascades including gliosis and encephalomalacia [7,8]. msTBI lesions are uniquely heterogeneous: focal or diffuse, spanning orders of magnitude in volume, with perilesional hemorrhages extending into subdural, epidural, and intraventricular compartments [1].

The AIMS-TBI 2026 Challenge benchmarks two tasks on T1-weighted MRI from the ENIGMA Consortium [10]: (i) scan-level binary detection and (ii) voxel-level segmentation. The two tasks impose fundamentally different requirements:

segmentation benefits from retaining the original T1 volume without skull stripping to avoid normalization artifacts, while detection benefits from restricting the input to the intracranial volume to suppress false positives. We therefore design separate pipelines optimized for each objective.

Skull stripping is standard preprocessing in brain MRI [9,3,6], but may introduce normalization artifacts specific to msTBI. nnU-Net [5] applies per-image ZScore normalization within the provided brain mask; when the skull-strip mask source varies across training samples or at inference—as may occur with perilesional hemorrhages and multi-site acquisition variability—the normalization statistics may shift, potentially suppressing lesion signal. We empirically found that training directly on original T1 volumes without skull stripping yields superior lesion Dice, which we hypothesize is due to more consistent per-image normalization statistics across training samples.

Transfer learning from large pretrained models and knowledge distillation [2] together enable high-accuracy segmentation within the GPU memory constraints of challenge inference environments.

2 Method

2.1 Dataset

The AIMS-TBI 2026 training set comprises 552 T1-weighted MRI scans from msTBI patients at 12 international sites via the ENIGMA Consortium: Baylor College of Medicine, UCLA, Pennsylvania State University, Kennedy Krieger Institute, Loma Linda University, Nationwide Children's Hospital, Murdoch Children's Research Institute, Deakin University, UT Houston, Kessler Foundation, VA Palo Alto, and the University of Oslo. Patient age ranges from 6.1 to 85 years (mean 29.2±19.0, median 18.6); 337 (61.1%) are male. Scanners include Siemens (49.3%), Philips (30.3%), and GE (20.5%). Of 552 training cases, 322 contain annotated lesions and 230 are lesion-free. The validation set contains 103 cases (55 with lesion, 48 without). Evaluation uses T1w MRI only.

All volumes are at approximately 1 mm isotropic resolution. Lesion volumes span four orders of magnitude (Table 1); the official segmentation task applies a 50-voxel ($\approx$50 mm^3) minimum lesion floor. We prepare two training variants: original T1 volumes (without skull stripping) for the segmentation model, and ICV-masked volumes from a dedicated nnU-Net ICV segmentation model for the detection model.

2.2 Task-Specific Pipelines

Segmentation Pipeline. We train an nnU-Net 3D full-resolution model directly on original T1 volumes without skull stripping. nnU-Net computes per-image ZScore statistics within the provided brain mask; when the mask source varies across training samples or differs at inference, the normalization statistics may shift and suppress lesion signal. We empirically found that training on

Table 1. Lesion volume distribution in 552 training cases (322 with lesion). Volume categories reflect quartile-based thresholds used for stratified fold assignment during training; validation performance is reported using clinically interpretable size bins in Section 3.

Category	Volume range (mm^3)	Cases
No lesion	–	230
Tiny	$\leq$100	26
Small	100–407	55
Medium	407–9,169	160
Large	$>$ 9,169	81
Total	–	552

volumes without skull stripping yields superior lesion Dice, which we hypothesize is due to more consistent per-image normalization statistics. Connected components smaller than 50 voxels are removed in post-processing.

Detection Pipeline. For scan-level detection we use a two-stage ICV masking cascade: (i) an nnU-Net model segments the intracranial volume (ICV), which is used to mask the input T1; (ii) a lesion segmentation network is applied to the masked volume, followed by connected-component-based scan-level classification ($\geq$10 voxels).

The two pipelines reflect complementary objectives: segmentation maximizes lesion Dice by training on original T1 volumes without skull stripping; detection maximizes balanced accuracy by restricting input to the intracranial volume to suppress false positives on lesion-free scans. We empirically compared the ICV masking cascade and detection without skull stripping: the approach without skull stripping achieves sensitivity 0.945 but specificity only 0.688 (balanced accuracy 0.817), while the ICV masking cascade achieves sensitivity 0.800 and specificity 0.958 (balanced accuracy 0.879). The ICV masking cascade is therefore adopted for detection to prioritize specificity and overall balanced accuracy.

Figure 1 illustrates the ICV masking effect on perilesional hemorrhages extending into CSF compartments adjacent to the skull.

2.3 Network Architecture and Training Strategy

Both task-specific pipelines use the nnU-Net 3D full-resolution framework [5], which automatically configures patch size, encoder depth, and feature map counts from dataset fingerprints, though with different model configurations suited to their respective objectives.

Detection model. The detection cascade employs a standard convolutional encoder-decoder trained from random initialization on ICV-masked T1 volumes, using the standard nnU-Net training pipeline.

Segmentation model: pretraining and distillation. The segmentation pipeline uses pretrained initialization, task-specific fine-tuning, and knowledge distillation. A large residual encoder-decoder (ResEncUNet-L, teacher) is initialized from the MultiTalentV2 Challenge Edition pretrained model [11]—a 4000-epoch ResEncUNet-L checkpoint trained with z-score normalization across diverse T1-weighted segmentation tasks—and fine-tuned on the msTBI training set. The teacher provides a high-capacity pretrained representation for task-specific adaptation; knowledge distillation then transfers this into a compact architecture satisfying the challenge inference constraints. Using the open-source distillation module from Fast-nnUNet [4] extended within the nnU-Net framework, we distill the teacher into a compact ResEncUNet-S student (SPlans128, decoder channels halved) [2]. The student is also initialized from the same MultiTalentV2 weights and trained with a combined task loss and soft-target distillation objective.

The challenge inference environment (NVIDIA T4, 16 GB VRAM, $\leq$7 min per case) cannot accommodate the ResEncUNet-L teacher ($\approx$25 GB peak VRAM even without test-time mirroring), precluding its direct submission. The distilled ResEncUNet-S student requires only $\approx$1.25 GB VRAM and completes inference in $\approx$20 s per case, comfortably within the deployment budget. Crucially, the student achieves higher Dice scores than the teacher on the official AIMS-TBI validation set (lesion Dice 0.551 vs. 0.479; overall Dice 0.692 vs. 0.673; Table 3).
Segmentation model: final model selection. Two student variants were evaluated: the standard student (S128) and a hard-split variant (S128-hs). Although S128-hs achieves marginally higher lesion Dice (0.565 vs. 0.551, Δ=0.014), S128 yields substantially better boundary metrics (HD95 25.7 mm vs. 28.3 mm; ASSD 6.6 mm vs. 8.3 mm). As HD95 and ASSD are primary ranking criteria in AIMS-TBI, S128 was selected as the final submission.

2.4 Training Details

Both the detection and segmentation teacher models use combined Dice and cross-entropy loss at equal weight ($\mathcal{L}_{\text{task}} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{CE}}$); the student additionally incorporates a soft-target distillation loss. AdamW with AMSGrad ($\text{lr}_0 = 10^{-4}$, weight decay 3×10^{-5}), and polynomial decay over $E_{\max} = 2{,}000$ epochs (250 iter/epoch). Gradients accumulate over 4 steps. Deep supervision applies geometrically decaying weights across decoder stages. Standard nnU-Net augmentation and test-time mirroring (TTA) are used.

3 Results

3.1 Evaluation Protocol

Task 1 (Detection) reports sensitivity, specificity, and balanced accuracy using a 10-voxel threshold. **Task 2 (Segmentation)** reports overall Dice, lesion Dice, HD95, and ASSD on cases with ground-truth lesion volume $\geq$50 voxels. Confidence intervals are computed via bootstrap resampling ($n = 2{,}000$).

Table 2. Main validation results (103 cases). Segmentation on 55 lesion-positive cases ($\geq$50 voxels); 95% CI by bootstrap ($n = 2{,}000$).

Task	Method	Metric	Value [95% CI]
Segmentation	Pretrained distilled network (original T1)	Overall Dice	0.692 [0.621, 0.759]
		Lesion Dice	0.551 [0.468, 0.627]
		HD95 (mm)	25.7
		ASSD (mm)	6.6
Detection	ICV masking cascade	Balanced Acc	0.879 [0.818, 0.935]
		Sensitivity	0.800
		Specificity	0.958

Table 3. Teacher vs. distilled student on the official AIMS-TBI validation set (103 cases, 55 lesion-positive). Teacher is non-deployable ($\approx$25 GB VRAM); student fits within the 16 GB challenge limit and achieves higher Dice.

Model	Overall Dice	Lesion Dice	HD95 (mm)	ASSD (mm)
Teacher (large)	0.673	0.479	26.3	**5.6**
Student (distilled, submitted)	**0.692**	**0.551**	**25.7**	6.6

3.2 Validation Results

Tables 2–5 report performance on the 103-case validation set (55 lesion-positive, 48 lesion-free). Table 3 compares the teacher and distilled student on the official AIMS-TBI validation set. Lesion size categories for the validation analysis are defined as: Tiny ($<$1,000 mm^3), Small (1,000–5,000 mm^3), Medium (5,000–20,000 mm^3), and Large ($>$20,000 mm^3).

3.3 Key Findings

Training on original T1 achieves strong segmentation performance. The segmentation model trained on original T1 volumes (without skull stripping) achieves lesion Dice 0.551 [0.468, 0.627] and overall Dice 0.692 [0.621, 0.759], which we hypothesize is due to more consistent per-image normalization statistics in the absence of skull-strip variability. We note that this attribution is a hypothesis; a matched controlled ablation isolating normalization consistency is deferred to future work.

Performance is monotonically increasing with lesion size. Both segmentation Dice and detection sensitivity increase monotonically with lesion size (Tables 4–5). For detection, sensitivity reaches 1.000 for medium and large lesions; the overall sensitivity of 0.800 was largely driven by tiny lesions ($<$1,000 mm^3, sensitivity 0.429), not by skull-stripping artifacts. This is consistent with the intrinsic difficulty of detecting very small lesions from T1 contrast alone.

Complementary pipeline design justified by empirical comparison. The detection approach without skull stripping achieves higher sensitivity (0.945)

Table 4. Segmentation lesion Dice by lesion size on 55 lesion-positive validation cases.

Size category	Volume (mm^3)	n	Lesion Dice
Tiny	$< 1{,}000$	14	0.390
Small	1,000–5,000	15	0.471
Medium	5,000–20,000	15	0.641
Large	$> 20{,}000$	11	0.743
All	–	55	0.551

Table 5. Detection sensitivity by lesion size on 55 lesion-positive validation cases.

Size category	Volume (mm^3)	n	Sensitivity
Tiny	$< 1{,}000$	14	0.429
Small	1,000–5,000	15	0.800
Medium	5,000–20,000	15	1.000
Large	$> 20{,}000$	11	1.000
All	–	55	0.800

but substantially lower specificity (0.688, balanced accuracy 0.817). The ICV masking cascade achieves lower sensitivity on tiny lesions but high specificity (0.958) on lesion-free scans, yielding a higher balanced accuracy (0.879). The two pipelines are therefore designed around their respective primary objectives: Dice maximization for segmentation (without skull stripping) and balanced accuracy maximization for detection (ICV-masked).

Distillation satisfies the deployment budget and improves accuracy. The ResEncUNet-L teacher (≈25 GB VRAM) cannot run on the challenge's NVIDIA T4 (16 GB, ≤7 min per case). Knowledge distillation into the compact ResEncUNet-S student reduces VRAM to ≈1.25 GB (≈20 s per case). The student achieves higher Dice scores than the teacher on the official AIMS-TBI validation set (lesion Dice 0.551 vs. 0.479; Table 3), while satisfying the challenge deployment constraints.

4 Conclusion

We presented task-specific nnU-Net pipelines for AIMS-TBI 2026. The central finding is that training on original T1 volumes without skull stripping empirically improves lesion segmentation in nnU-Net; we hypothesize this is due to more consistent per-image normalization statistics, though a controlled ablation is deferred to future work, and skull stripping is therefore avoided for segmentation. For detection, ICV masking is instead adopted to improve specificity; the lower overall detection sensitivity was largely driven by tiny lesions ($<1{,}000\,mm^3$), with medium and large lesions detected at 100%. Initializing from MultiTalentV2 pretrained weights and compressing via knowledge distillation into a compact

ResEncUNet-S student (1.25 GB VRAM, ≈20 s per case) satisfies the T4 deployment constraint while the student achieves higher Dice scores than the teacher on the official AIMS-TBI validation set. On the final held-out test set, the segmentation model achieves lesion Dice 0.5243, HD95 29.5 mm, and ASSD 9.3 mm; the detection cascade achieves balanced accuracy 0.8482 (sensitivity 0.894, specificity 0.803).

References

1. Covington, N.V., Duff, M.C.: Neuropsychological and neuroimaging outcomes following complicated mild, moderate, and severe traumatic brain injury: A systematic review and meta-analysis. JAMA Neurology **78**(11), 1313–1321 (2021)
2. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
3. Iglesias, J.E., Liu, C.Y., Thompson, P.M., Tu, Z.: Robust brain extraction across datasets and comparison with publicly available methods. IEEE Transactions on Medical Imaging **30**(9), 1617–1634 (2011)
4. Inc., X.F.: Fast-nnUNet: Knowledge distillation module for nnu-net. `https://github.com/77even/Fast-nnUNet` (2024), distillation module released under Apache License 2.0
5. Isensee, F., Jaeger, P.F., Kohl, S.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)
6. Isensee, F., Schell, M., Pflueger, I., Brugnara, G., Bonekamp, D., Neuberger, U., Wick, A., Schlemmer, H.P., Heiland, S., Wick, W., et al.: Automated brain extraction of multisequence mri using artificial neural networks. Human Brain Mapping **40**(17), 4952–4964 (2019)
7. Maas, A.I., Stocchetti, N., Bullock, R.: Moderate and severe traumatic brain injury in adults. The Lancet Neurology **7**(8), 728–741 (2008)
8. McKee, A.C., Daneshvar, D.H.: The neuropathology of traumatic brain injury. Handbook of Clinical Neurology **127**, 45–66 (2015)
9. Smith, S.M.: Fast robust automated brain extraction. Human Brain Mapping **17**(3), 143–155 (2002)
10. Thompson, P.M., Stein, J.L., Medland, S.E., et al.: The ENIGMA consortium: large-scale collaborative analyses of neuroimaging and genetic data. Brain Imaging and Behavior **8**(2), 153–182 (2014)
11. Ulrich, C., et al.: MultiTalent V2 Challenge Edition pretrained weights. `https://zenodo.org/records/13753413` (2024)

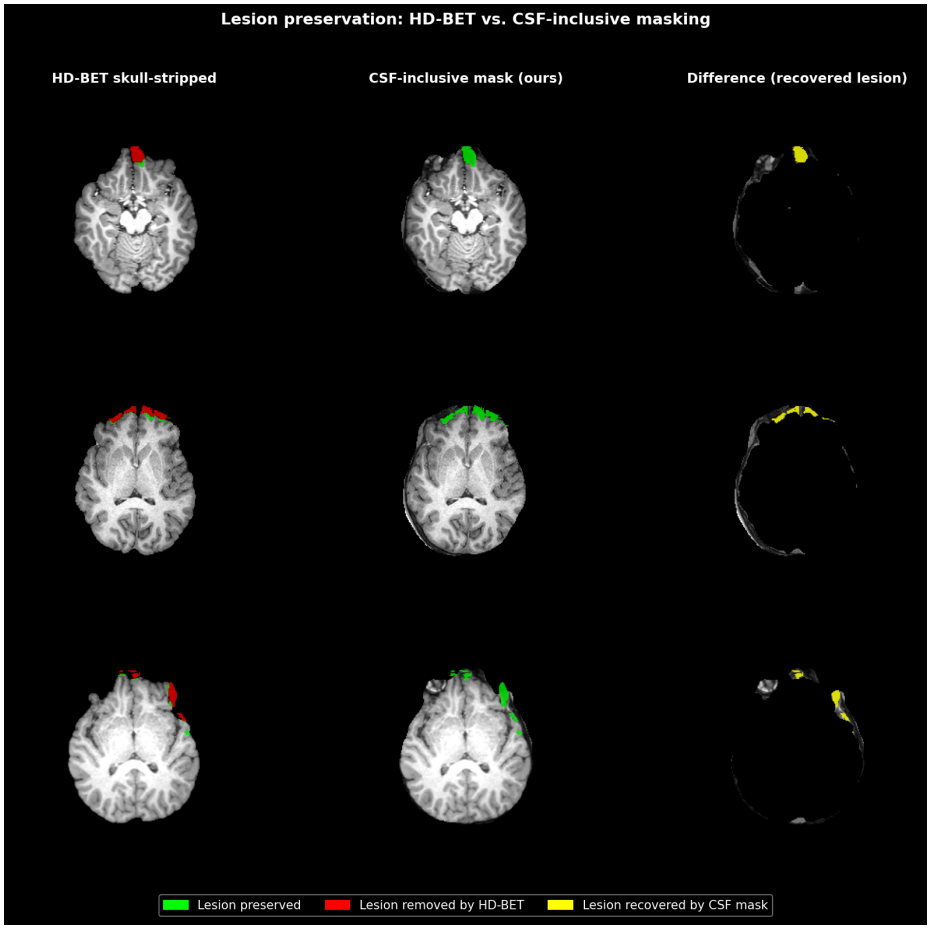

Fig. 1. Skull-stripping comparison on three representative msTBI cases, motivating task-specific preprocessing. **Left**: Standard HD-BET removes substantial perilesional lesion volume (85.2%, 73.7%, 65.3% of annotated lesion volume, top to bottom), motivating original T1 (without skull stripping) for the segmentation pipeline. **Centre**: ICV-masked volume with extended CSF boundary (used for the detection pipeline) preserves full lesion extent. **Right**: Yellow — perilesional hemorrhage recovered by ICV masking but removed by HD-BET.