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.

# Single-Model Lesion Detection and Segmentation for Moderate–Severe Traumatic Brain Injury on T1-weighted MRI

Ujjwal Baid[1], Bhakti Baheti[1], and Anant Madabhushi[1,2]

[1] Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, USA
[2] Atlanta Veterans Affairs Medical Center, Atlanta, GA, USA
ujjwalbaid0408@gmail.com

**Abstract.** Moderate-to-severe traumatic brain injury (msTBI) produces lesions that are focal or diffuse, span grey matter, white matter and cerebrospinal fluid, and vary over four orders of magnitude in volume. The AIMS-TBI 2026 challenge scores lesion detection and lesion segmentation on two separate leaderboards, on T1-weighted MRI only. We submitted a single self-configuring 3D residual-encoder U-Net that serves both tasks: the segmentation mask is the network output, and the detection decision is derived from it by connected component analysis at the challenge's own 10-voxel floor. Three design choices were settled experimentally. First, we enlarged the training cohort by merging the 2026 training release with the labelled 2025 validation release, giving 652 cases. Second, on cases no constituent model had seen, no five-fold ensemble beat the best single fold, so we shipped the cheaper single fold, though the margin is schedule-dependent. Third, removing predicted components below the challenge's 50-voxel scoring floor raised specificity but degraded lesion-wise detection rate, Dice and count error, so predictions are submitted unfiltered. On a 131-case internal held-out split the model reaches a mean Dice of 0.577 on lesion-containing scans and a balanced accuracy of 0.746; on the official Test Phase 1 data it reaches a Dice of 0.5154 and a detection score of 0.8331 (sensitivity 0.7188, specificity 0.9474). Residual error is concentrated in small lesions: mean Dice falls steeply with ground-truth lesion volume, and below roughly 500 voxels the network frequently predicts an empty mask.

**Keywords:** Traumatic brain injury · Lesion segmentation · Lesion detection · nnU-Net · T1-weighted MRI.

## 1 Introduction

Moderate-to-severe traumatic brain injury (msTBI) arises when external forces move the brain rapidly within the skull, producing primary injuries such as haematoma, haemorrhage and contusion and a cascade of secondary injuries including gliosis and encephalomalacia [11, 14]. Each patient accumulates a unique

combination of these processes, so msTBI lesions are heterogeneous to a degree considered a hallmark of the condition [1]: unlike stroke or tumour lesions they may be focal or diffuse, occur in homologous regions of both hemispheres, and extend across multiple tissue types. That heterogeneity propagates into the pipelines on which large-scale neuroimaging research depends, introducing local and global errors in registration, normalization and parcellation [5], and current practice compensates by ignoring lesions, excluding patients with large lesions, or segmenting them by hand, which respectively biases results, limits generalizability, and fails to scale. The ENIGMA Brain Injury working group [16] therefore needs an automatic lesion identification tool operating on the one modality acquired consistently across sites, T1-weighted (T1w) MRI, which is what the AIMS-TBI challenge series targets [3, 2]. The 2026 edition introduces two changes that shape our design: detection and segmentation are scored on separate leaderboards with different lesion-size floors ($> 10$ and $> 50$ voxels), and additional modalities are released for *training* while the test data remain T1w-only.

We report a deliberately conservative submission. Rather than propose a new architecture, we take a strong self-configuring baseline [7, 9] and spend the available compute on design decisions we could measure. We contribute (i) a single network serving both leaderboards, detection being derived from the segmentation by connected component analysis (Section 2.3); (ii) a cohort expansion merging the 2026 training release with the labelled 2025 validation release for an 18% larger training set (Section 2.1); (iii) an evidence-based rejection of two conventional defaults: five-fold ensembling did not beat a single fold on unseen data, and component filtering at the 50-voxel floor traded three ranked detection metrics for two (Sections 3.1, 3.2); (iv) two reusable pitfalls of this benchmark, a pair of defects in a re-implemented metric that between them invert the post-processing decision, and a reference container template that silently discards voxel geometry (Section 2.4); and (v) a size-resolved error analysis localizing the remaining error in lesions below $\sim$500 voxels.

*Related work.* Submissions to the 2025 edition were dominated by U-Net–style encoder–decoders, most built on nnU-Net [7]. Reported strategies include ensembling several U-Net variants and tuning post-processing thresholds [12], augmenting training with synthetic lesions pasted into lesion-free brains [6], replacing whole-volume z-scoring with brain-mask–adaptive normalization [15], state-space (Mamba) backbones with hybrid losses [18], quasiseparable mixing blocks [13], dual-encoder networks with morphology-aware losses [10], V-Net with connected component filtering [17], and cascades refining a U-Net output with slice-level and radiomics classifiers [4]. The best 2025 model achieved an overall Dice of 0.637 and 0.515 on lesion-containing scans. Component filtering is widely reported as helpful in that literature, and ensembling is generally assumed to be. We re-tested both on held-out data and, under the 2026 metric definitions, reached the opposite conclusion for each.

## 2 Methods

### 2.1 Data and quality control

The 2026 training release contains 552 subjects contributed by 13 sites through the ENIGMA Pediatric and Adult msTBI working groups, acquired on GE, Siemens and Philips scanners at 1.5 T and 3 T, mostly at 1 mm isotropic resolution and defaced. Each subject provides a T1w volume, an expert binary lesion mask and demographics; 236 also provide diffusion MRI. An automated quality-control pass surfaced three integrity problems that would each have caused silent data loss: a truncated gzip lesion mask (case excluded); a T1w volume named `.ni.gz` rather than `.nii.gz`, which a naive glob would skip; and a stray SPM tissue-segment file, quarantined. This leaves 551 usable cases, 230 lesion-free and 321 lesion-containing. Lesion volumes are strongly right-skewed (median 1,785 voxels) and about 5% of lesion-containing cases fall below the 50-voxel segmentation floor (Fig. 2a).

*Cohort expansion.* Because the test data are T1w-only and diffusion covers fewer than half the subjects, we prioritized quantity of T1w supervision over modality breadth, merging the 551 curated cases with the 101 usable labelled cases of the AIMS-TBI 2025 validation release into a 652-case cohort (375 lesion-containing, 277 lesion-free). The two releases differ systematically: the 2025 cases contain no lesion below 50 voxels and have a median lesion volume of 4,569 voxels against 1,785. We therefore drew a single 80/20 split *stratified jointly on cohort of origin and lesion-volume band* (none, $\leq$500, $\leq$5000, $>$ 5000 voxels), giving 521 training and 131 held-out cases. The held-out set contains 76 cases with a lesion above the 10-voxel detection floor, 73 above the 50-voxel segmentation floor, and 55 lesion-free cases. Stratifying on lesion band keeps the small-lesion regime, the dominant failure mode of Section 3.3, measurable on the held-out side rather than absorbed into training.

### 2.2 Network, training and inference

We use nnU-Net v2 with the ResEnc-L planner and the `3d_fullres` configuration [7, 9]. The planner self-configures a 3D residual-encoder U-Net with 6 resolution stages, 32$\rightarrow$64$\rightarrow$128$\rightarrow$256$\rightarrow$320$\rightarrow$320 features, $[1, 3, 4, 6, 6, 6]$ residual blocks per encoder stage and one convolution per decoder stage, $3^3$ kernels, instance normalization and leaky ReLU. Images are resampled to $1\times1\times1$ mm and z-score normalized within the non-zero (brain) region; training uses $192\times256\times160$ patches, batch size 2 and 33% guaranteed foreground sampling, optimizing the nnU-Net default sum of soft Dice and cross-entropy with deep supervision by SGD with Nesterov momentum (initial learning rate $10^{-2}$, polynomial decay, weight decay $3\times10^{-5}$, 250 iterations per epoch). Lesion-free brains are kept in training throughout: they are the only signal teaching the network to emit an empty mask, and they are 42% of the cohort.

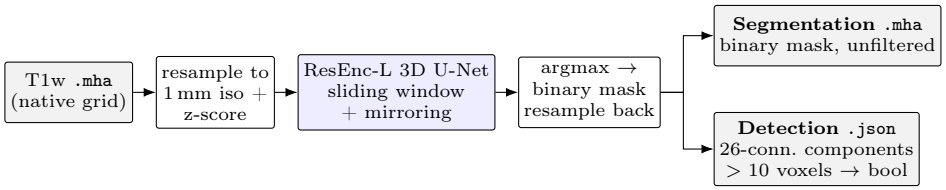


**Fig. 1.** Single-network, dual-leaderboard inference. One set of weights produces the segmentation mask; the detection decision is derived from that mask by connected component analysis at the challenge's 10-voxel floor. Voxel spacing, origin and direction are carried through and copied back onto the output.

We train for 250 epochs rather than nnU-Net's default 1000. This was a measured choice rather than a compute compromise: on unseen cases the shorter schedule scored higher (Table 2) at a quarter of the cost, which made the remaining configuration experiments affordable. One fold takes ≈6 h at 85 s/epoch on a single data-centre GPU. Inference uses nnU-Net's Gaussian-weighted sliding window with test-time mirroring.

### 2.3 Deriving detection from segmentation

The two tasks are scored separately but describe the same object, so we train one segmentation network and derive both outputs from it (Fig. 1). The segmentation container writes the binary mask directly. The detection container runs 26-connectivity component labelling on that mask, discards components of 10 voxels or fewer to match the detection floor, and reports lesion presence as a Boolean. This halves the compute of a two-model submission and keeps the two leaderboards mutually consistent. A dedicated detection classifier was held in reserve in case internal scoring showed detection lagging; it did not.

### 2.4 Evaluation and containerization

The organizers' official scoring script was unpublished during the development window, so we re-implemented the published metrics: DSC, 95th-percentile Hausdorff distance (HD95) and average symmetric surface distance (ASSD) over cases whose ground truth contains a lesion above 50 voxels; and lesion-wise detection rate, absolute lesion-count difference, sensitivity, specificity and balanced accuracy at the 10-voxel floor. Two details proved consequential. Applying the size floor to the *ground truth* but not to the prediction makes prediction-side filtering look artificially expensive, because it deletes true positives the challenge never scores; and defining case-level presence as "any non-zero voxel" rather than "any component above the floor" lets one stray voxel flip a case to positive, understating unfiltered specificity (0.436 rather than the correct 0.545 here). Correcting both cut the apparent benefit of component filtering by ≈86% and reversed the decision of Section 3.1.

**Table 1.** Shipped configuration on the 131-case held-out split, with and without removal of predicted components at or below the 50-voxel segmentation floor. Segmentation $n = 73$; detection $n = 131$ (76 lesion-positive, 55 lesion-free).

| **Segmentation** | unfilt. | filt. | **Detection** | unfilt. | filt. |
|---|---|---|---|---|---|
| DSC mean ↑ | **0.577** | 0.572 | sensitivity ↑ | **0.947** | 0.908 |
| DSC median ↑ | **0.653** | 0.644 | specificity ↑ | 0.545 | **0.800** |
| HD95 (mm) ↓ | **25.97** | 28.66 | balanced acc. ↑ | 0.746 | **0.854** |
| ASSD (mm) ↓ | **7.06** | 8.05 | detection rate ↑ | **0.665** | 0.543 |
| zero-DSC ↓ | **6** | 9 | $\lvert\Delta\text{count}\rvert$ ↓ | **1.489** | 1.618 |

Both algorithms ship as Grand Challenge containers built on PyTorch 2.8 with CUDA 12.8 and nnU-Net 2.8.1, matching the training environment; weights are mounted as a separate artefact rather than baked into the image. The reference template invites one error that we flag here: its I/O helpers return bare NumPy arrays and so discard voxel spacing, origin and direction. nnU-Net resamples to the spacing recorded in its plan, so following the template literally yields predictions at the wrong physical scale. We keep the SimpleITK image throughout, pass true spacing to the predictor, and copy the geometry back onto the output. Single-fold inference measured 10.5 s per case at a 7.2 GB VRAM peak, inside the 16 GB budget of the challenge's target NVIDIA T4.

## 3 Results

### 3.1 Held-out performance and post-processing

Table 1 reports the shipped configuration on the 131 held-out cases: mean DSC 0.577 (median 0.653), HD95 25.97 mm, ASSD 7.06 mm, and detection sensitivity 0.947 at specificity 0.545.

Removing predicted components at or below the 50-voxel scoring floor is standard practice in this challenge series, and has the expected effect: specificity rises from 0.545 to 0.800 and balanced accuracy from 0.746 to 0.854. But it simultaneously degrades sensitivity, lesion-wise detection rate and count error (three of the five ranked detection metrics), while costing DSC, HD95 and ASSD and raising zero-DSC cases from 6 to 9. At this setting the filter discards 237 of 483 predicted components. A sweep over cutoffs $\{0, 5, 10, 15, 20, 30, 50\}$ on two independent test sets showed the same monotone trade in both, and under the challenge's case-wise rank aggregation the unfiltered predictions ranked best on both. We therefore submit unfiltered predictions and leave the filtering to the challenge's own floors. Tuning the probability threshold offers no further gain: lowering it from 0.5 to 0.05 changes DSC by +0.004 and recovers 2 of 72 empty predictions, because on the cases it misses the network emits almost no foreground probability mass at all rather than hesitating near the boundary.

**Table 2.** Configuration selection on the 101 labelled 2025 validation cases, unseen by every model in the comparison. Segmentation $n = 54$; detection $n = 101$. "ens" denotes the five-fold ensemble.

| Configuration | DSC ↑ | DSC med. ↑ | HD95 ↓ | ASSD ↓ | bal. acc. ↑ |
|---|---|---|---|---|---|
| single fold, 250 ep | **0.526** | 0.638 | 31.17 | 9.68 | 0.845 |
| single fold, 1000 ep | 0.476 | 0.595 | 29.35 | **7.59** | 0.848 |
| ens, 250 ep | 0.436 | 0.546 | 34.49 | 9.61 | 0.838 |
| ens, 1000 ep | 0.487 | **0.649** | **29.46** | 8.27 | **0.868** |

### 3.2 Configuration selection

Choosing between training schedules, and between a single fold and a five-fold ensemble, requires predictions on cases that no constituent model has seen; a cross-validation ensemble cannot supply this, because four of the five folds were trained on any given case. We therefore scored all four candidate configurations of the 551-case model on the 101 labelled 2025 validation cases, unseen by every model in that comparison (Table 2).

Neither conventional default helped. Training four times longer did not improve DSC (0.476 versus 0.526), consistent with a residual error made of small lesions that more gradient steps on the same distribution do not address. More surprisingly, no ensemble beat the best single fold: at a matched 250-epoch schedule the five-fold ensemble lost 0.090 DSC, and the 1000-epoch ensemble stayed 0.039 below the best single fold, although it did take the best balanced accuracy. We ship one fold, which is also five times cheaper at inference; with no configuration showing a gain over it, the cheaper option carried the default.

We stop short of concluding that ensembling is harmful on this task. The comparison rests on the $n = 54$ lesion-containing cases of one cohort, and the effect is not sign-stable across schedules: at a matched 1000-epoch budget the ordering reverses (0.487 for the ensemble against 0.476 for the single fold). One mechanism would be consistent with the size-resolved analysis of Section 3.3: per-voxel probability averaging attenuates exactly the low-confidence, low-volume predictions this task depends on, because a small lesion found by one fold and missed by four cannot survive averaging above 0.5. That would predict the largest ensembling penalty in the sub-500-voxel regime where our errors concentrate. Testing it requires per-fold agreement statistics we did not compute, so we record it as a hypothesis, not a finding.

A third choice follows from the framework's own defaults: nnU-Net validates with the final weights unless told otherwise, so every number here describes `checkpoint_final`, and that is what we export. Shipping `checkpoint_best` would have deployed weights that were never scored.

### 3.3 Where the error is

Performance is a strong function of lesion size (Fig. 2b). On the held-out split, mean DSC is 0.738 above 5,000 voxels ($n = 26$, no failures), 0.528 between

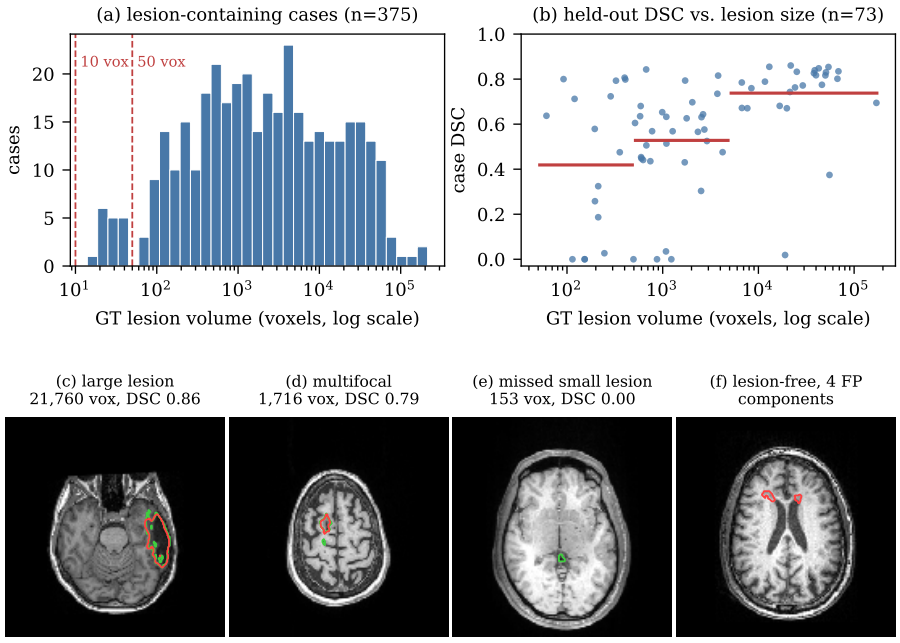


**Fig. 2.** **(a)** Ground-truth lesion volume over the 375 lesion-containing cases of the merged cohort, with the challenge's 10- and 50-voxel scoring floors. **(b)** Per-case DSC against lesion volume on the 73 scoreable held-out cases; red bars are band means for 50–500, 500–5000 and $>$ 5000 voxels. **(c–f)** Held-out predictions (red) against ground truth (green) on the axial slice of maximum lesion extent: (c) a large lesion, accurately delineated; (d) a multifocal lesion; (e) a 153-voxel lesion for which the network predicts nothing; (f) a lesion-free brain on which four spurious components survive the 10-voxel floor.

500 and 5,000 ($n$ = 30, 2 failures) and 0.419 between 50 and 500 ($n$ = 17, 4 failures). The six DSC=0 cases have a median ground-truth volume of 324 voxels against 2,519 for the rest. An earlier cross-validation over all 551 cases showed the same pattern at larger scale: the 70 zero-DSC cases had a median lesion volume of 150 voxels versus 3,481 for the rest, and 66 of them predicted an entirely empty mask. The failure is one of detection sensitivity to small lesions, not of boundary delineation, and it is not explained by scanner vendor, site or fold. The complementary error is false positives on lesion-free brains: 25 of 55 lesion-free held-out cases receive at least one component above 10 voxels, though almost always only one or two (Fig. 2f). This is the direct cost of shipping unfiltered predictions, and the reason held-out specificity is 0.545.

**Table 3.** Official AIMS-TBI 2026 Test Phase 1 results for the submitted containers.

| **Segmentation** | | **Detection** | |
|---|---|---|---|
| Score (DSC, lesion scans) | 0.5154 | Score (balanced acc.) | 0.8331 |
| HD95 (mm) | 33.24 | Sensitivity | 0.7188 |
| ASSD (mm) | 12.94 | Specificity | 0.9474 |

### 3.4 Official Test Phase 1 results

Table 3 gives the scores returned by the organizers on the hidden Test Phase 1 data; the segmentation score is the mean DSC over lesion-containing scans and the detection score the balanced accuracy. The test Dice of 0.5154 is within 0.001 of the best lesion-only Dice reported for the 2025 edition (0.515), although the two are computed on different cohorts under different task definitions and are not a like-for-like comparison. No direct benchmark against contemporary methods is available to us: the 2026 leaderboard was not public at the time of submission, and the 2025 figures were computed under different lesion-size floors. The detection operating point moved substantially between our held-out estimate (sensitivity 0.947, specificity 0.545) and the official set (0.719 and 0.947): our split did not predict the balance of errors on the hidden data.

## 4 Discussion

Our central finding is a negative one. Two moves that the literature on this task treats as near-automatic—five-fold ensembling and small-component filtering—neither helped under the 2026 metric definitions when evaluated on data no constituent model had seen. Both results are conditioned on those metric definitions and on our cohorts, and neither is a claim about the two techniques in general. Neither is visible from cross-validation alone, and the filtering result was in fact inverted before we corrected two defects in our own metric implementation.

The gap between our held-out detection operating point and the official one is large: sensitivity and specificity essentially exchanged places. We cannot resolve the cause without the hidden data; plausible contributors are the smaller Phase 1 cohort, a lesion-size and lesion-free mix different from our stratified split, and any divergence between our inferred rule for case-level presence and the organizers' official one. The practical consequence is that an unfiltered submission chosen to maximize detection rate met a test set on which unfiltered specificity was already 0.947, leaving the filter's principal benefit little room to act.

*Practical implications of the size-resolved error.* For the ENIGMA use case that motivates the challenge, the size bands of Section 3.3 translate into a usage rule. Above ∼5,000 voxels the masks are reliable enough (DSC 0.738, no failures on the held-out split) to support lesion-aware registration and parcellation quality control; between 500 and 5,000 voxels they are marginal (DSC 0.528); below

∼500 voxels (DSC 0.419, 4 of 17 cases at DSC 0) they should not carry lesion-burden quantification. Because the failure mode is missed detection rather than poor delineation, a study can screen on predicted lesion volume instead of reviewing every case by hand: an empty or near-empty prediction flags the scans that need a reader, and those concentrate at small ground-truth volumes.

*Limitations.* The submission uses a single fold rather than all trained weights; that was an evidence-based choice, but a thin one. It rests on a single 80/20 split and, for DSC, on the 54 lesion-containing cases of the 2025 release; we did not repeat the split under other seeds, and the ordering of single fold and ensemble reverses at a 1000-epoch budget. The shift in detection operating point between our held-out split and the official test set is direct evidence that the split does not represent the hidden distribution, and that detection thresholds cannot be calibrated against it. All internal numbers come from a metric re-implementation validated against a second independent implementation, not against the organizers' script. The merged cohort mixes two release years whose lesion-size distributions differ, handled by stratification but still a source of covariate shift. The released diffusion MRI is unused in the final model: we built and quality-controlled a registration pipeline (brain-extracted rigid registration [8] of the mean $b_0$ to T1w, then resampling of fractional anisotropy and mean diffusivity onto the T1w grid, successful for 227 of 236 cases), but the test set is T1w-only, so diffusion can only act as a training-time signal and there was no time to validate a distillation setup. Finally, only two submissions are permitted in total, limiting how much of the design space can be probed on the leaderboard.

*Future work and conclusion.* All remaining headroom lies in lesions below ∼500 voxels, where the network usually predicts nothing at all, and post-hoc remedies are exhausted. We propose no remedy here; establishing that the residual error is one of detection rather than delineation is what rules the post-hoc options out and moves the problem to training time. The remaining levers are at training time: region- or blob-based losses that do not vanish on tiny structures, oversampling of small-lesion patches, and a lesion-focused patch configuration. The multimodal release additionally offers a privileged-information route, a T1w+diffusion teacher distilled into the T1w-only student the challenge evaluates, though that subset is not cohort-representative (in our release no GE scan carries diffusion data), so we expect gains to be modest. In summary, a single residual-encoder nnU-Net serves both leaderboards from one set of weights, reaching a Dice of 0.5154 and a detection score of 0.8331 (sensitivity 0.7188, specificity 0.9474) on the official Test Phase 1 data, with residual error concentrated in small lesions.

## References

1. Covington, N.V., Duff, M.C.: Heterogeneity is a hallmark of traumatic brain injury, not a limitation: a new perspective on study design in rehabilitation research. American Journal of Speech-Language Pathology **30**(2S) (2021)

2. Dennis, E.L., et al.: AIMS-TBI – automated identification of moderate-severe traumatic brain injury lesions (2025). https://doi.org/10.5281/zenodo.15084120
3. Deutscher, E., Tustison, N.J., Onicas, A., Wilde, E.A., Pease, M., Bakas, S., Dennis, E.L.: Automated identification of moderate-to-severe traumatic brain injury lesions (AIMS-TBI) 2025 MICCAI challenge. In: Bakas, S., et al. (eds.) Segmentation, Classification, and Synthesis for Brain Tumors and Traumatic Brain Injuries. MICCAI 2025 Challenges: BraTS-Lighthouse 2025 and AIMS-TBI 2025, Part II. Lecture Notes in Computer Science, vol. 16377, pp. 273–281. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-16370-7_24
4. Dhamat, G., Sundaresan, V.: Traumatic brain injury segmentation using an ensemble of encoder-decoder models. In: Bakas, S., et al. (eds.) Segmentation, Classification, and Synthesis for Brain Tumors and Traumatic Brain Injuries. MICCAI 2025 Challenges: BraTS-Lighthouse 2025 and AIMS-TBI 2025, Part II. Lecture Notes in Computer Science, vol. 16377, pp. 341–349. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-16370-7_31
5. Diamond, B.R., Donald, C.L.M., et al.: Optimizing the accuracy of cortical volumetric analysis in traumatic brain injury. MethodsX **7** (2020)
6. Hamadache, R.E., Lisazo, C., Yalcin, C., Abramova, V., Lal-Trehan Estrada, U.M., Cartaya Lathulerie, A., Rivas Díaz, M., Casamitjana, A., Oliver, A., Lladó, X.: nnU-Net and synthetic lesions for automated segmentation of moderate-to-severe traumatic brain injury. In: Bakas, S., et al. (eds.) Segmentation, Classification, and Synthesis for Brain Tumors and Traumatic Brain Injuries. MICCAI 2025 Challenges: BraTS-Lighthouse 2025 and AIMS-TBI 2025, Part II. Lecture Notes in Computer Science, vol. 16377, pp. 292–300. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-16370-7_26
7. 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)
8. Isensee, F., Schell, M., Pflueger, I., Brugnara, G., Bonekamp, D., Neuberger, U., Wick, A., Schlemmer, H.P., Heiland, S., Wick, W., Bendszus, M., Maier-Hein, K.H., Kickingereder, P.: Automated brain extraction of multisequence MRI using artificial neural networks. Human Brain Mapping **40**(17), 4952–4964 (2019)
9. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K.H., Jaeger, P.F.: nnU-Net revisited: A call for rigorous validation in 3D medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention (MICCAI). Springer (2024)
10. Kuk, H., Abbas, Z., Lee, S.W.: DeaMNet: A dual-encoder network with hybrid morphology-aware loss for precise TBI lesion segmentation. In: Bakas, S., et al. (eds.) Segmentation, Classification, and Synthesis for Brain Tumors and Traumatic Brain Injuries. MICCAI 2025 Challenges: BraTS-Lighthouse 2025 and AIMS-TBI 2025, Part II. Lecture Notes in Computer Science, vol. 16377, pp. 350–358. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-16370-7_32
11. Maas, A.I.R., Stocchetti, N., Bullock, R.: Moderate and severe traumatic brain injury in adults. The Lancet Neurology **7**(8), 728–741 (2008)
12. Marinov, Z., Kleesiek, J., Stiefelhagen, R.: Not all ensembles are equal: A short study on moderate to severe traumatic brain injury segmentation methods. In: Bakas, S., et al. (eds.) Segmentation, Classification, and Synthesis for Brain Tumors and Traumatic Brain Injuries. MICCAI 2025 Challenges: BraTS-Lighthouse 2025 and AIMS-TBI 2025, Part II. Lecture Notes in Computer Science, vol. 16377, pp. 282–291. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-16370-7_25

13. Mazher, M., Niederer, S.A., Qayyum, A.: NeuroNetMix: A 3D encoder-decoder with quasiseparable mixing for robust lesion segmentation in traumatic brain injury. In: Bakas, S., et al. (eds.) Segmentation, Classification, and Synthesis for Brain Tumors and Traumatic Brain Injuries. MICCAI 2025 Challenges: BraTS-Lighthouse 2025 and AIMS-TBI 2025, Part II. Lecture Notes in Computer Science, vol. 16377, pp. 330–340. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-16370-7_30
14. McKee, A.C., Daneshvar, D.H.: The neuropathology of traumatic brain injury. Handbook of Clinical Neurology **127**, 45–66 (2015)
15. Son, I., Lee, G., Sim, S., Uhm, K.H.: Lesion segmentation in moderate to severe traumatic brain injury: An nnU-Net based approach with adaptive normalization in the AIMS-TBI 2025 challenge. In: Bakas, S., et al. (eds.) Segmentation, Classification, and Synthesis for Brain Tumors and Traumatic Brain Injuries. MICCAI 2025 Challenges: BraTS-Lighthouse 2025 and AIMS-TBI 2025, Part II. Lecture Notes in Computer Science, vol. 16377, pp. 322–329. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-16370-7_29
16. Thompson, P.M., et al.: ENIGMA and global neuroscience: a decade of large-scale studies of the brain in health and disease across more than 40 countries. Translational Psychiatry **10**(1), 100 (2020)
17. Zavadil, B., Lenkovics, A., Szabó, B., Szlávecz, Á., Benyó, B.: Hybrid traumatic brain injury lesion segmentation using voxel-based V-Net model and connected component filtering. In: Bakas, S., et al. (eds.) Segmentation, Classification, and Synthesis for Brain Tumors and Traumatic Brain Injuries. MICCAI 2025 Challenges: BraTS-Lighthouse 2025 and AIMS-TBI 2025, Part II. Lecture Notes in Computer Science, vol. 16377, pp. 311–321. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-16370-7_28
18. Zhu, Y., Bai, X., Jiang, H., Chen, G.: BLSegMamba: An optimized SegMamba framework for msTBI lesion segmentation in MRI. In: Bakas, S., et al. (eds.) Segmentation, Classification, and Synthesis for Brain Tumors and Traumatic Brain Injuries. MICCAI 2025 Challenges: BraTS-Lighthouse 2025 and AIMS-TBI 2025, Part II. Lecture Notes in Computer Science, vol. 16377, pp. 301–310. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-16370-7_27