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.

A U-Net for Automated msTBI Lesion Segmentation

Jon André Ottesen[1], Qinghui Liu[1], Bradley MacIntosh[1], Kyrre Eeg Emblem[1], and Atle Bjørnerud[1]

[1]Oslo University Hospital, Norway
jonakri@uio.no

Abstract. Clinical assessment of moderate-to-severe traumatic brain injury (msTBI) relies heavily on manual review of structural MRI by trained readers. Automated segmentation of msTBI lesions on T1-weighted MRI aims to relieve some of this burden. However, deep learning-based approaches remain challenging due to the extreme variability in lesion size, shape, and location. In this work, we describe our entry to the AIMS-TBI 2026 Challenge. We adapt a 3D U-Net (∼40M parameters) with residual connections and deep supervision initialized from a model trained on the ATLAS v3.0 dataset. The model was trained with a composite loss combining binary cross-entropy, dice loss, and a lesion-wise dice term on both exams with positive findings and no findings. Our submission achieved a balanced accuracy of 0.871 and a dice similarity coefficient of 0.531 on the held-out test set.

Keywords: Traumatic Brain Injury · Lesion Segmentation · U-Net · AIMS-TBI · Deep Learning

1 Introduction

Traumatic brain injury (TBI) poses a significant health burden and is among the most common causes of neurological disability, with risk of late-onset neurodegeneration [8,1]. Moderate-to-severe TBI (msTBI) results from external mechanical forces (e.g., traffic incidents, falls, head trauma) that cause complex primary injuries - including haematomas, haemorrhages, and contusions in addition to secondary injuries such as gliosis and encephalomalacia [8,3]. The sheer variety of primary and secondary injuries, coupled with variable size, shape, and tissue localisation, produces greatly heterogeneous structural damage.

Lesion heterogeneity makes automated segmentation a challenging task. Models trained on other domains, e.g., stroke, transfer sub-optimally to msTBI [3], and ground-truth annotations are expensive to obtain at scale. To bridge this gap, the AIMS-TBI challenge series provides a curated, multi-site dataset with 654 expert-annotated T1-weighted scans from 12 sites [3]. Prior work has tested ensemble-based methods, synthesised lesions, various architectures, and pre- and post-processing strategies [9,4,11,10].

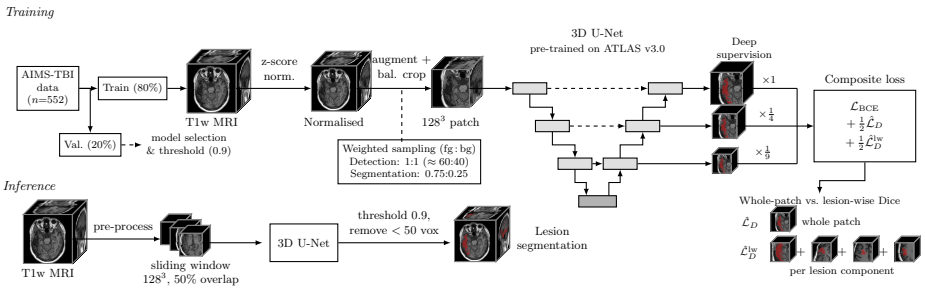

Fig. 1. Overview of the proposed pipeline. During training (top), the AIMS-TBI data were split 80/20 into training and validation sets. All T1-weighted volumes were z-score normalised, augmented, and cropped into 128^3 patches with class-balanced sampling. A 3D U-Net is initialised with pre-trained weights trained on the ATLAS v3.0 dataset. The model is trained with a composite loss combining binary cross-entropy, whole-patch dice ($\hat{\mathcal{L}}_D$), and lesion-wise dice ($\hat{\mathcal{L}}_D^{\text{lw}}$) via deep supervision. At inference (bottom), the full volume is processed by sliding-window inference, thresholded, and cleaned by removing small connected components.

Rather than adding to the growing catalogue of specialised architectures that makes utilising publicly available models challenging, we aim to configure a simple model that can achieve competitive performance. This work trains a 3D U-Net with residual connections using a composite lesion-aware loss, class-balanced sampling, and a multi-stage augmentation strategy, with weights pre-trained on the ATLAS v3.0 stroke lesion dataset [6] as a warm start. We aim to produce a strong-performing model that can easily be adopted without complex dependencies or time-consuming processing steps.

2 Methods

2.1 Dataset

This project used the AIMS-TBI 2026 Challenge dataset, consisting of 892 T1-weighted images across 12 institutions on 1.5T and 3T scanners from GE, Siemens, and Philips. T1-weighted images were collected across both chronic and acute injury phases with the corresponding binary lesion masks if applicable. All images were resampled to 1 mm^3 isotropic voxel spacing and defaced prior to release. Lesion masks were generated through a multi-rater sequential annotation process using ITK-SNAP as described in [3].

Of the 892 images, 553 were released to participants as part of the 2026 AIMS-TBI challenge, one of which was a duplicate sample and was removed, leaving 552 unique cases (n=322 positive, n=230 negative). The 552 cases constituted the training set and were randomly split into a train and validation set following an 80/20 split. Additionally, after an initial validation phase, a further 102 images (n=54 positive, n=48 negative) were made available to the participants,

bringing the total available number of cases up to 654. For simplicity and model comparison, these were added to the training set and the initial in-house split was kept.

Initial testing was performed using the initial training cases (552), whereas the model submitted to the final challenge was trained on the initial 80/20 split and the additional 102 cases that were added to the training data. Before model training, all images were cropped to remove potential zero-padded regions and reoriented to the right anterior superior (RAS) direction.

2.2 Network Architecture

To keep the architecture simple and dependency-light, we employed the 3D U-Net architecture from the nnU-Net framework [5], consisting of an encoder-decoder with skip connections and instance normalisation. Input patch size was chosen at $128 \times 128 \times 128$ to ensure low memory requirements, and the model contained approximately 39.9 million trainable parameters. Importantly, nnU-Net's self-configuring pipeline was not used. Instead, we trained the architecture directly with custom training settings.

2.3 Pre-training and Weight Transfer

Motivated by the dataset annotators, who used a model trained on the ATLAS v2.0 dataset to provide initial lesion predictions, we adopted ATLAS-based pre-training to allow for faster initial convergence. The network was pre-trained on the ATLAS v3.0 dataset [6], a curated collection of 1453 brain extracted T1-weighted MRI scans with manually segmented chronic stroke lesions, using a 60/20/20 training/validation/test split. All network layers were subsequently transferred and re-trained on the AIMS-TBI training data.

2.4 Loss Function

The model was trained with a composite loss function combining binary cross-entropy, whole-patch dice loss, and a lesion-wise dice loss:

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + \frac{1}{2}\hat{\mathcal{L}}_D + \frac{1}{2}\hat{\mathcal{L}}_D^{\text{lw}}. \tag{1}$$

The lesion-wise dice loss $\hat{\mathcal{L}}_D^{\text{lw}}$ computes dice independently per connected lesion component and averages the result across all lesions within a patch - directly penalising missed lesions regardless of size. To assess the contribution, we also trained a variant without lesion-wise dice loss. Importantly, to ensure lesion detection, lesions were weighted $100x$ compared to non-lesion tissue in the binary cross entropy loss due to class imbalance. Initial testing showed that this weight helped to stabilize initial lesion detection.

2.5 Training

All models were trained for 1000 epochs with a batch size of 2 using the AdamW optimizer lr $= 5 \times 10^{-4}$ and weight decay $= 10^{-2}$ as recommended by literature [7]. Linear learning rate warm up was used for the first 50 epochs and then annealed with a cosine schedule. Whole image normalization was performed using z-score normalization excluding zero-voxels.

The AIMS-TBI challenge consisted of two tasks: (1) patient-wise lesion detection focusing on sensitivity (lesion detection) and specificity (false positive rate), and (2) segmentation of msTBI lesions. Two models were trained with different weighting between lesion-positive and lesion-negative cases. For lesion detection, the weighting followed the data distribution, which entails approximately 60% positive and 40% negative cases; for lesion segmentation, the percentage of positive cases was increased to 75% since the task only considers positive cases. Patches from positive cases were selected to guarantee inclusion of foreground voxels to ensure the percentage of positive cases also represented the number of lesion-positive patches. An overview of the study is shown in Figure 1. Checkpoints were saved with the lowest validation loss during training.

2.6 Data Augmentation

Image augmentations were applied using spatial transforms before patch extraction and intensity and flip transforms with the MONAI framework [2]. Intensity and flip transforms were applied after patch extraction to reduce computational load. Image augmentation generally followed similar transforms to that seen in nnUNet with minor adjustments [5]. Table 1 lists all transforms and their parameters. At validation, only foreground cropping and padding to 128^3 were applied, along with random flipping and 90-degree rotations to verify that the model generalizes to different orientations.

2.7 Inference

At test time, the full image was processed via sliding window inference with a 128^3 patch size with 50% overlap. Predictions were thresholded based on the best performing threshold (0.9) for the validation set with all connected regions smaller than 50 voxels being removed. The detection model was thresholded at 0.5.

2.8 Evaluation

Method evaluation is performed differently for the two challenge tasks. For the detection task, evaluation is performed patient-wise, based on whether the model could discern any msTBI for the patient. A patient was counted as positive if any predicted lesion voxels remained after post-processing. The resulting binary predictions (lesion present yes/no) were scored against the reference labels using sensitivity, specificity, and balanced accuracy (BA).

Table 1. Data augmentation pipeline applied during training. All transforms are from the MONAI library. Prob. denotes the probability of applying the transform.

Stage	Transform	Parameters	Prob.
Spatial	BorderPad	5 voxels each side	Always
	Random Zoom	Scale 0.85–1.25, trilinear/nearest	0.25
	Random Rotation	±0.8 rad all axes, trilinear/nearest	0.25
	Foreground Crop	Source: brain mask, margin 2 voxels	Always
	Spatial Pad	Target size 128^3	Always
Intensity & Flip	Random Flip	Axis 0	0.50
	Random Flip	Axis 1	0.50
	Random Flip	Axis 2	0.50
	Random Rot90	Axes (0,1), (1,2), (0,2)	0.50 each
	Gaussian Noise	$\sigma = 0.1$	0.25
	Histogram Shift	5–15 control points	0.25
	Std Shift	Factor 0.1, nonzero voxels	0.25
	Scale Intensity	Factor ±0.5	0.25

For segmentation, only lesion-positive cases were evaluated. Segmentation quality was quantified using the dice similarity coefficient (DSC), measuring volumetric overlap, and the 95th percentile Hausdorff distance (HD95), capturing boundary agreement.

3 Results

3.1 Validation Performance

Table 2 reports performance on the held-out test split for the final loss configurations with and without lesion-wise dice, in addition to the final method submitted to the detection and segmentation task. Note, the two models tested with varying losses were trained without the additional 102 cases made available later in the challenge. The inclusion of lesion-wise dice increased the sensitivity from 0.825 to 0.875 without reducing the specificity. DSC and balanced accuracy results from the validation set are not reported here since they can be influenced by model checkpointing and general randomness during training. Instead, focus is placed on general trends seen on the validation set.

3.2 Qualitative Results

Figure 2 shows representative validation examples stratified by lesion size. The model correctly delineates lesions across a wide range of sizes (top row), while the characteristic error modes are false positives in regions resembling lesioned tissue (middle row) and missed lesions that are notably small or diffuse (bottom row). The number of true positives, false negatives, and false positives aggregated

Table 2. Results from the out of sample evaluation set. The detection model was evaluated using balanced accuracy (BA), sensitivity (Sens.), and specificity (Spec.). The segmentation model was evaluated using the dice similarity coefficient (DSC) and 95th percentile Hausdorff distance (HD95).

Method	BA	Sens.	Spec.	DSC (lesion)	HD95
Regular Dice	0.8401	0.825	0.8533	NA	NA
Lesion-wise Dice	**0.8651**	**0.875**	**0.8533**	NA	NA
Final Detection	0.8714	0.8875	0.8553	NA	NA
Final Segmentation	NA	NA	NA	0.531	26.2

by lesion volume on the validation set is shown in Figure 3. The detection model was used for this analysis rather than the segmentation-optimised model.

Across the validation set (111 patients; 69 lesion-positive, 42 lesion-free), the detection model identified 221 of 328 reference lesions (lesion-wise sensitivity 0.67) while producing 2.6 false positives per patient. Both the false negative and false positive lesions were strongly size-dependent (Figure 3). Missed lesions were almost exclusively small with a sensitivity of 0.48 for lesions below 0.1 mL, but exceeded 0.90 for lesions larger than 0.5 mL. False positive lesions showed the same trend, with 77% of false positives below 0.1 mL. Importantly, false positives were rarely detected in lesion-free brains, only 1 of 286 false positives (0.3%) occurred on an all-negative scan (0.02 per patient across the 42 lesion-free cases), while the remainder arose on scans that already contained lesions (4.1 per lesion-positive patient). This indicates that the false-positive burden stems either from over-segmentation of small components adjacent to genuine pathology or from spurious detections of other abnormal-appearing tissue elsewhere in non-healthy brains, rather than from hallucinated lesions in healthy tissue.

4 Discussion

In this study, we aimed to test how far a well-configured but otherwise standard 3D U-Net can be pushed on msTBI lesion segmentation and detection. On the evaluation set the pipeline reached a balanced accuracy of 0.871 for patient-wise detection and a lesion dice of 0.531 for segmentation. This suggests that a carefully configured conventional pipeline remains a strong baseline for this task, and one that can be adopted with minimal dependencies and greater ease of use.

We point to two decisions that shaped the pipeline, of which only the second was directly benchmarked: (1) weights initialized from a model pre-trained on the ATLAS v3.0 stroke dataset, and (2) a composite loss designed to counteract the extreme size imbalance of msTBI lesions. Non-random initialisation provided a substantially better starting point than random initialisation and allowed for faster initial convergence, although its exact importance was not benchmarked in this study, and we cannot rule out that the effect may have been smaller than assumed. The lesion-wise dice term addresses a known weakness of voxel-

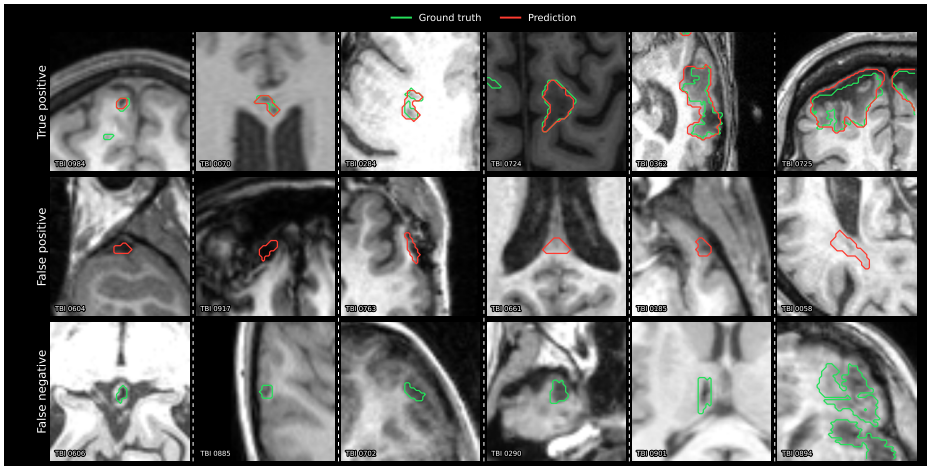

Fig. 2. Representative examples on the validation set, stratified by lesion size (small to large, left to right). Rows show true-positive, false-positive, and false-negative cases. Ground-truth lesions (green) and model predictions (red) are overlaid as contours on axial T1-weighted slices, one patient per column.

level losses, where large lesions dominate the gradient signal and leave small lesions under-penalised. In our experiments this term improved lesion sensitivity without a corresponding loss of specificity, while foreground weighting modestly improved Dice on the validation set from 0.59 to 0.61. Remaining settings were kept deliberately simple rather than extensively tuned.

The size-stratified error analysis in Figure 3 makes the lesion failures clinically interpretable. Detection is close to solved for larger lesions - sensitivity exceeds 0.90 above 0.5 mL - but falls to 0.48 for lesions below 0.1 mL. These small lesions dominate both the missed detections and the false positives. Interestingly, false positives were almost never raised in lesion-free brains, and were instead located on scans that already contained pathology, i.e., false-positive detection was less common in the 42 healthy patients than in the 69 patients with msTBIs in the internal validation set. This indicates that a focus point should be improvement in the detection and delineation of sub-0.1 mL lesions.

Several limitations temper these results. The fixed 128^3 patch size is a practical compromise given the increased memory cost during training and inference, although a larger patch size could potentially further improve the results. We report a single model per task rather than an ensemble, and the detection model was trained without the additional validation examples released later in the challenge, so its reported figures are likely conservative. Finally, the volume-stratified analysis is computed on our internal validation split and uses a fixed overlap criterion for counting detections, and should therefore be read as trends rather than exact operating points.

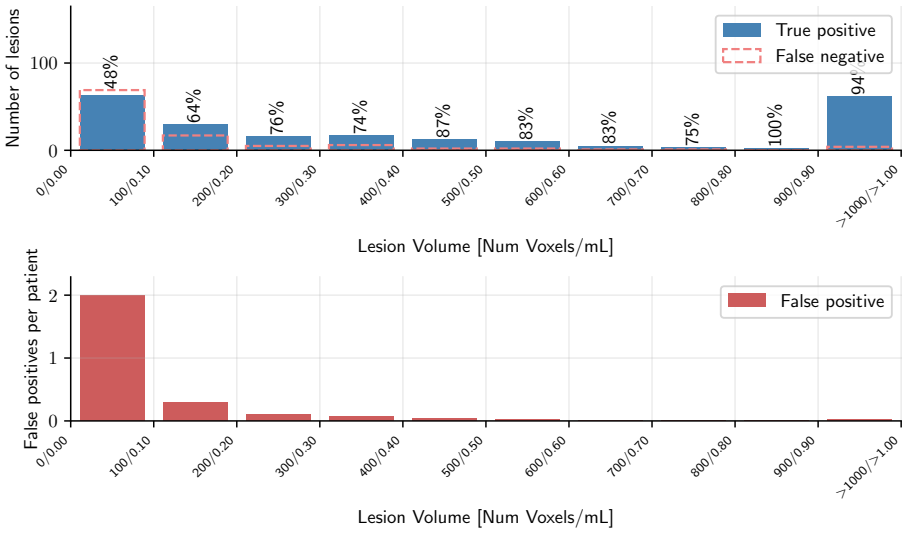

Fig. 3. Lesion-wise detection performance on the validation set as a function of lesion volume, using the detection model. Top row: reference lesions counted as detected (true positive, filled) or missed (false negative, dashed outline). A lesion is counted as detected if at least 10% of the lesion is covered by the prediction. Bottom row: false positives per patient, i.e., the number of predicted components with at most 10% overlap with the ground truth. The percentage above each column in the top row is the fraction of reference lesions in that bin that were detected. The final bin is open-ended and aggregates all lesions larger than 1 mL (58 true positives, 4 false negatives, and 3 false positives).

5 Conclusion

We described a standard 3D U-Net pipeline initialised from ATLAS v3.0 stroke weights for msTBI lesion segmentation, trained with a composite lesion-aware loss, class-balanced sampling, and a multi-stage augmentation strategy on the AIMS-TBI 2026 dataset. Without task-specific architectural additions (except weighted sampling), the pipeline achieved a detection balanced accuracy of 0.871 and a lesion dice of 0.531 on the evaluation set. Accompanying error analysis showed that failures are concentrated in sub-0.1 mL lesions while false positives are rarely raised on lesion-free scans.

Acknowledgments. The authors acknowledge the computational resources and support provided by the Research Platform at Oslo University Hospital, including access to GPU computing infrastructure used in this work.

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

References

1. Borg, J., Holm, L.W., Coronado, V.G., von Holst, H., Cassidy, J.D., Peloso, P.M., Kraus, J.F., Carroll, L.: Incidence, risk factors and prevention of mild traumatic brain injury: results of the who collaborating centre task force on mild traumatic brain injury. Journal of rehabilitation medicine **36**, 28–60 (2004)
2. Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., et al.: Monai: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022)
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: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 273–281. Springer (2025)
4. 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: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 292–300. Springer (2025)
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. Liew, S.L., Lo, B.P., Donnelly, M.R., Zavaliangos-Petropulu, A., Jeong, J.N., Barisano, G., Hutton, A., Simon, J.P., Juliano, J.M., Suri, A., et al.: A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. Scientific data **9**(1), 320 (2022)
7. Loshchilov, I., Hutter, F., et al.: Fixing weight decay regularization in adam. arXiv preprint arXiv:1711.05101 **5**(5), 5 (2017)
8. Maas, A.I., Menon, D.K., Manley, G.T., Abrams, M., Åkerlund, C., Andelic, N., Aries, M., Bashford, T., Bell, M.J., Bodien, Y.G., et al.: Traumatic brain injury: progress and challenges in prevention, clinical care, and research. The Lancet Neurology **21**(11), 1004–1060 (2022)
9. Marinov, Z., Kleesiek, J., Stiefelhagen, R.: Not all ensembles are equal: A short study on moderate to severe traumatic brain injury segmentation methods. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 282–291. Springer (2025)
10. 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: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 311–321. Springer (2025)
11. Zhu, Y., Bai, X., Jiang, H., Chen, G.: Blsegmamba: An optimized segmamba framework for mstbi lesion segmentation in mri. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 301–310. Springer (2025)