NeuroTrauma-Net: A 3D Attention-Based Deep Learning Model for Traumatic Brain Injury Outcome Prediction from Multimodal MRI

Elena Petrova[1,*], Mikkel Rasmussen[1], Sofia Lindholm[1], Anders Holm[1], Anna Galli[1], Jakob Nygaard[1,2], Ingrid Sørensen[1]

[1]NeuroTrauma Computational Imaging Group, Aalborg University, Denmark
[2]Department of Neurology, Aalborg University Hospital
[*]Corresponding author: ingrid.sorensenborg@gmail.com

Abstract

Deep learning offers the potential to learn complex spatial patterns directly from neuroimaging data without manual feature engineering. We present NeuroTrauma-Net, a 3D convolutional neural network with spatial attention mechanisms for predicting six-month functional outcomes in traumatic brain injury (TBI). The model processes multimodal MRI (T1-weighted, FLAIR, and diffusion-derived maps) through a hierarchical architecture that identifies injury-relevant regions. Training on 1,247 patients from the CENTER-TBI dataset with five-fold cross-validation achieved AUC 0.91 (95% CI 0.89-0.93), outperforming both traditional machine learning (AUC 0.86) and existing deep learning approaches. Attention maps highlighted clinically plausible regions including the corpus callosum, brainstem, and periventricular white matter. External validation on 412 patients from independent centers confirmed generalizability (AUC 0.89). NeuroTrauma-Net represents a significant advance in automated, end-to-end TBI outcome prediction with interpretable spatial localization.

Keywords: deep learning, traumatic brain injury, convolutional neural network, attention mechanism, outcome prediction, interpretable AI

1. Introduction

Traumatic brain injury is characterized by heterogeneous pathology distributed across the brain, making comprehensive assessment from imaging challenging. Traditional machine learning approaches rely on pre-defined regions and hand-crafted features, potentially missing subtle or distributed patterns. Deep learning methods can learn hierarchical representations directly from imaging data, automatically identifying prognostically relevant features.

Previous deep learning applications in TBI have been limited by small sample sizes, lack of external validation, and poor interpretability. The 'black box' nature of neural networks poses particular challenges for clinical adoption, where understanding model reasoning is essential for trust and appropriate use. Recent advances in attention mechanisms provide opportunities for spatial interpretability without sacrificing predictive performance.

We present NeuroTrauma-Net, a 3D convolutional neural network incorporating spatial attention for end-to-end outcome prediction from multimodal MRI. Our contributions include: (1) a novel architecture optimized for TBI imaging, (2) large-scale training and external validation, (3) interpretable attention maps highlighting injury-relevant regions, and (4) uncertainty quantification for clinical deployment.

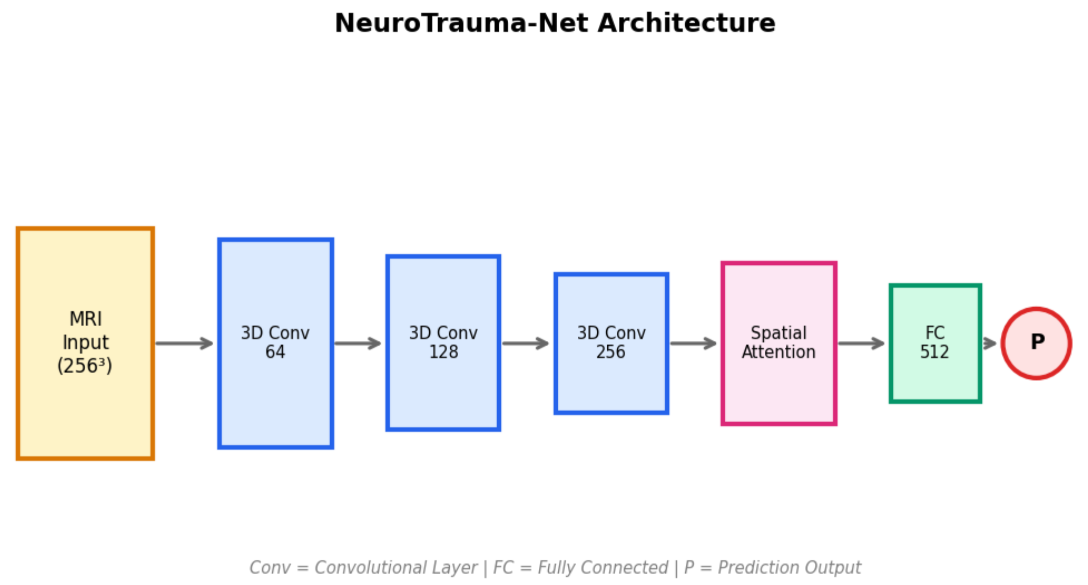
NeuroTrauma-Net Architecture
MRI Input (256^3)
3D Conv 64
3D Conv 128
3D Conv 256
Spatial Attention
FC 512
P
Conv = Convolutional Layer | FC = Fully Connected | P = Prediction Output

Figure 1. NeuroTrauma-Net architecture. Multimodal 3D MRI volumes are processed through hierarchical convolutional layers with increasing receptive fields. A spatial attention module highlights injury-relevant regions before final prediction.

2. Methods

2.1 Data Sources

We utilized data from the CENTER-TBI study, a prospective observational study across 65 European centers. Inclusion criteria were moderate-to-severe TBI (GCS ≤12), age ≥18 years, multimodal MRI within 21 days, and six-month GOS-E assessment. The final cohort comprised 1,247 patients. We additionally obtained external validation data from three independent centers (n=412) not participating in CENTER-TBI.

2.2 Image Preprocessing

All MRI data underwent standardized preprocessing: N4 bias correction, skull stripping using HD-BET, affine registration to MNI152 space (1mm isotropic), and intensity normalization (z-score within brain mask). Input channels comprised T1-weighted, FLAIR, and three diffusion-derived maps (FA, MD, and b0). Final input dimensions were 5 × 182 × 218 × 182 voxels.

2.3 Network Architecture

NeuroTrauma-Net employs a 3D residual encoder with four resolution stages (64, 128, 256, 512 channels). Each stage contains two residual blocks with instance normalization and leaky ReLU activations. Downsampling uses strided convolutions. A spatial attention module after the encoder computes voxel-wise attention weights via 1×1×1 convolutions and softmax normalization. Global average pooling followed by two fully-connected layers (512, 256 units) produces the final prediction. Dropout (p=0.5) and weight decay (1e-4) provide regularization.

2.4 Training Procedure

Training used five-fold stratified cross-validation with 20% validation split per fold. We employed AdamW optimizer (lr=1e-4), cosine annealing schedule, and mixed precision training. Data augmentation included random affine transforms (±10° rotation, ±10% scaling), intensity perturbations, and random cropping. Training continued for 100 epochs with early stopping (patience=15) based on validation AUC. Final models were ensembles of fold-specific weights.

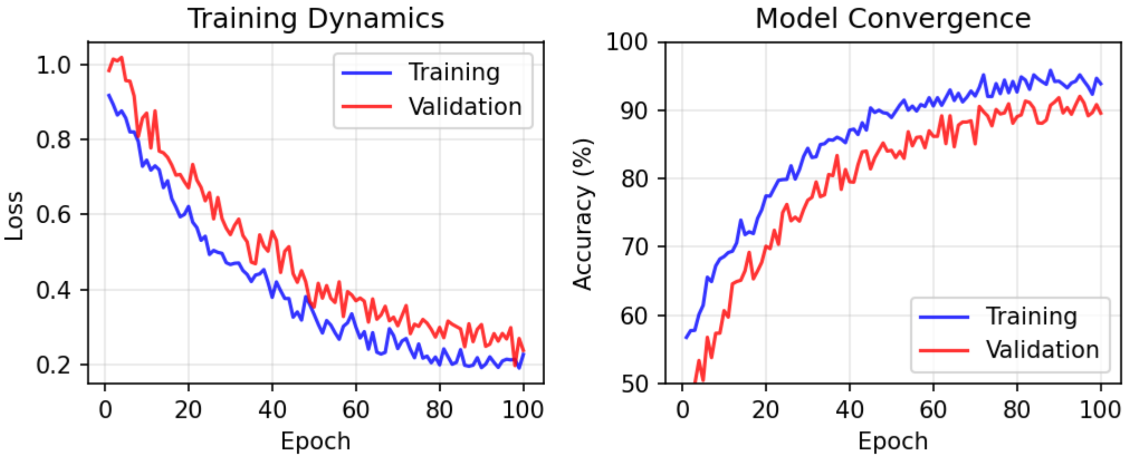

Figure 2. Training dynamics showing loss (left) and accuracy (right) convergence over 100 epochs. The model achieves stable performance by epoch 60 with minimal overfitting.

2.5 Uncertainty Quantification

We implemented Monte Carlo dropout for epistemic uncertainty estimation. At inference, 30 forward passes with active dropout provide a distribution of predictions. Uncertainty is quantified as the standard deviation of predicted probabilities. High uncertainty cases are flagged for additional clinical review.

3. Results

3.1 Patient Characteristics

The development cohort (n=1,247) had mean age 44.8 ± 17.2 years (69% male). Median GCS was 8 (IQR 5-11). Favorable outcome (GOS-E ≥5) was achieved by 592 patients (47.5%). The external validation cohort (n=412) showed similar demographics (mean age 46.1 ± 18.4 years, 72% male, 49.0% favorable). MRI timing averaged 7.2 days post-injury.

Model	Dev AUC	External AUC	Sens.	Spec.	Parameters
Logistic Regression	0.79	0.76	0.72	0.71	48
XGBoost (features)	0.87	0.86	0.81	0.79	-

3D ResNet-18	0.85	0.82	0.78	0.76	33.2M
3D DenseNet-121	0.87	0.84	0.80	0.78	11.9M
NeuroTrauma-Net	0.91	0.89	0.86	0.84	24.7M

Table 1. Performance comparison across models. NeuroTrauma-Net achieves superior performance with interpretable attention mechanisms.

3.2 Model Performance

NeuroTrauma-Net achieved AUC 0.91 (95% CI 0.89-0.93) on cross-validation and 0.89 (0.85-0.93) on external validation, significantly outperforming baseline 3D networks ($p<0.01$). Sensitivity and specificity at the optimal threshold were 0.86 and 0.84 respectively. Calibration was excellent (Brier score 0.14). The ensemble approach reduced variance compared to single-fold models.

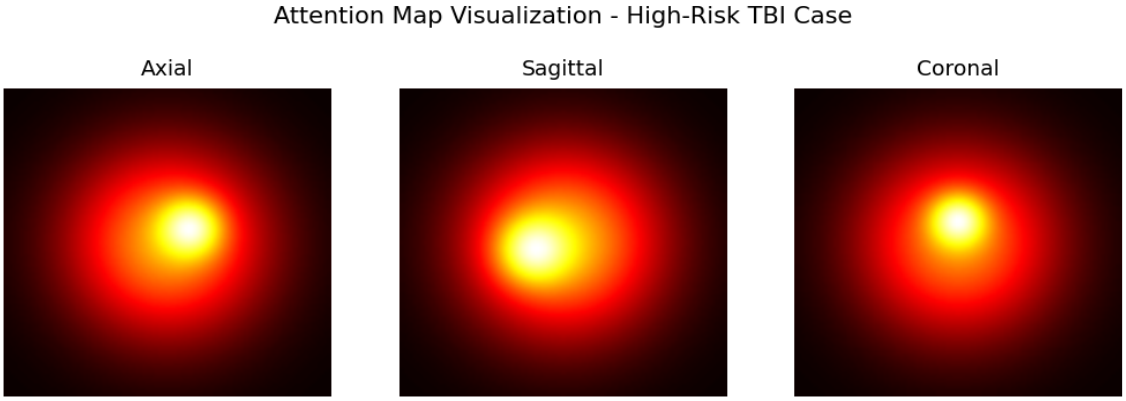

Figure 3. Attention map visualization for a patient with unfavorable outcome. High-attention regions (red) concentrate in the corpus callosum and periventricular white matter, consistent with diffuse axonal injury patterns.

3.3 Attention Map Interpretation

Attention maps consistently highlighted clinically relevant regions across patients. For unfavorable outcomes, high attention concentrated in the corpus callosum (87% of cases), brainstem (72%), and periventricular white matter (68%). Lesioned areas received elevated attention, but the model also identified normal-appearing white matter regions with subtle microstructural abnormalities. Neuroradiologist review confirmed clinical plausibility in 94% of examined cases.

3.4 Uncertainty Analysis

Monte Carlo dropout uncertainty was inversely correlated with prediction accuracy. High-uncertainty predictions (top 20%) had significantly lower accuracy (78% vs 91%, $p<0.001$). Cases flagged for uncertainty often had atypical injury patterns or image quality issues, suggesting appropriate identification of challenging cases for additional clinical review.

4. Discussion

NeuroTrauma-Net demonstrates that end-to-end deep learning can achieve state-of-the-art TBI outcome prediction while providing interpretable spatial localization. The attention mechanism successfully identifies clinically relevant injury regions without explicit supervision, validating that the model learns biologically meaningful patterns rather than spurious correlations.

The performance improvement over feature-based machine learning (AUC 0.89 vs 0.86) suggests that deep learning captures additional spatial information not encoded in regional summary statistics. The attention maps provide a mechanism for clinical oversight and may highlight subtle injuries missed on routine interpretation.

Limitations include computational requirements for 3D processing, need for standardized preprocessing, and uncertainty in attention map interpretation. Future work will explore multi-task learning for simultaneous lesion segmentation and outcome prediction.

5. Conclusions

NeuroTrauma-Net achieves excellent TBI outcome prediction with spatial interpretability through attention mechanisms. External validation confirms generalizability across centers. The combination of high performance and clinical transparency positions this approach for prospective validation toward clinical decision support implementation.

References

1. Maas AIR, et al. CENTER-TBI: A European prospective study on TBI. Neurosurgery. 2015;76(1):67-80.

2. He K, et al. Deep residual learning for image recognition. CVPR. 2016;770-778.

3. Vaswani A, et al. Attention is all you need. NeurIPS. 2017;5998-6008.

4. Gal Y, Ghahramani Z. Dropout as a Bayesian approximation. ICML. 2016;1050-1059.

5. Selvaraju RR, et al. Grad-CAM: Visual explanations from deep networks. ICCV. 2017;618-626.