Received: 12 September 2025 | Revised: 24 March 2026 | Accepted: 10 April 2026 | Published online: 18 May 2026

RESEARCH ARTICLE

Journal of Computational and Cognitive Engineering
2026, Vol. 00(00) 1–13
DOI: 10.47852/bonviewJCCE62027647

BON VIEW PUBLISHING

Explainable AI Framework for Automated Lesion Segmentation and Severity Assessment in Neonatal Hypoxic-Ischemic Encephalopathy

Athira Chandran[1] **and Lekshmi Chandrika Reghunath**[1,*]

[1] *School of Artificial Intelligence, Amrita Vishwa Vidyapeetham-Coimbatore, India*

Abstract: Neonatal hypoxic-ischemic encephalopathy (HIE) is a critical brain injury in newborns resulting from perinatal oxygen deprivation, often leading to long-term neurological impairments and developmental delays. Early and accurate diagnosis is essential for improving clinical outcomes. In this work, we propose a multimodal framework for automated HIE lesion segmentation and severity classification from magnetic resonance imaging scans. A SegResNet architecture was employed to segment lesions from apparent diffusion coefficient (ADC) and Z-scored ADC maps. In the subsequent stage, an early fusion strategy was adopted to integrate features derived from the predicted lesion masks with clinical metadata. The fused representation was classified into mild, moderate, and severe HIE categories using an XGBoost classifier. The Boston Neonatal Brain Injury Dataset for Hypoxic-Ischemic Encephalopathy was used for the experiments. The severity classification received an accuracy of 88.64%, and the segmentation model received a Dice score of 0.758 ± 0.29. Explainable AI with Gradient-weighted Class Activation Mapping was also used to improve the clinical dependability. These results demonstrate the effectiveness of the proposed system in both precise lesion localization and clinically meaningful severity stratification, highlighting its potential to support early clinical decision-making in neonatal care.

Keywords: hypoxic-ischemic encephalopathy (HIE), magnetic resonance imaging (MRI), medical image segmentation, SegResNet, multimodal fusion

1. Introduction

Hypoxic-ischemic encephalopathy (HIE) is a severe neurological disorder resulting from a lack of oxygen and blood flow to the brain during the perinatal period. It is one of the leading causes of neonatal mortality and long-term neurodevelopmental impairments. It is an important challenge in global health and the main cause of neonatal mortality. Neonatal encephalopathy is considered the third major contributor to death in children under five years of age, according to recent analysis of the Global Burden of Disease (GBD) 2021 study. This study reveals that 2–8 infants per 1000 live births around the world are affected by neonatal encephalopathy, resulting in 600,000 neonatal deaths in 2021 alone [1]. Accurate and timely assessment of HIE severity is crucial for guiding therapeutic interventions and improving clinical outcomes.

Magnetic resonance imaging (MRI) is widely regarded as the gold standard for identifying structural abnormalities in the neonatal brain. However, manual interpretation of MRI scans is labor-intensive, requires highly trained radiologists, and is subject to inter-observer variability. Murphy et al. [2] exhibited this limitation quantitatively in a study, where the same set of 20 MRI scans with neonatal HIE was individually identified by two expert clinicians. According to this study, the two experts' similarity was only indicated by a median F1-score of 0.56. This finding highlights the necessity for an automated solution to increase diagnosis accuracy and demonstrates the diversity in manual detection.

Deep learning methods have demonstrated great promise in medical image analysis in recent years, especially in problems involving segmentation and classification. In brain MRI segmentation, convolutional neural networks (CNNs) and their sophisticated variations have demonstrated excellent performance. However, it is still difficult to identify minor lesion patterns and deal with class imbalance, particularly when differentiating between mild and more severe cases of HIE. HIE lesions are usually widespread and small. According to Bao et al. [3], about half of their patients have lesions that account for less than 1% of the entire brain volume. Standard models perform worse as a result of this significant class imbalance. For the baseline U-Net, the Dice score was roughly 0.5 [3]. A similar issue was noted in another study, which also suggested new loss functions to deal with the issue of tiny lesions [4]. Medical images segmentation and multi-scale feature extraction have found success with SegResNet, a hybrid architecture that combines residual connections

***Corresponding author:** Lekshmi Chandrika Reghunath, School of Artificial Intelligence, Amrita Vishwa Vidyapeetham-Coimbatore, India. Email: cr_lekshmi@cb.amrita.edu

[5] with U-Net-like encoder–decoder architectures [6]. However, small, low-contrast lesions, which are essential for detecting mild cases, are frequently difficult for traditional methods to detect.

This paper presents a novel deep learning architecture for automated HIE assessment in order to overcome these issues. The detection of modest abnormalities is improved by incorporating a class-weighted loss function to increase sensitivity toward underrepresented lesion classes. An XGBoost-based classifier is used for severity classification after segmentation, utilizing both radiomics-inspired properties and deep feature representations. Also, a dual-explainability technique is employed to guarantee the model's therapeutic relevance and to go beyond the straightforward accuracy measures. Following explainable AI (XAI) analysis of the SegResNet, a tree-based model is used to make the severity classifier interpretable. Experimental results show better segmentation accuracy, robust detection of mild instances, and improved severity classification when the suggested approach is tested on the publicly available BONBID dataset. This work advances the development of AI-driven systems for infant brain damage evaluation that are clinically dependable.

1.1. Problem statement

Due to the difficulties of precisely segmenting brain lesions and categorizing severity levels, neonatal HIE presents a significant problem in medical imaging. The problem is defined as a two-stage task: (i) lesion segmentation from MRI modalities and (ii) severity classification using both imaging and clinical data.

Let $\mathbb{R}$ denote the set of real numbers. Given an ordered pair (X, C), where $X \in \mathbb{R}^{HxWxD}$ represents the input MRI scan (with H, W, and D representing the spatial dimensions of height, width, and depth) and $C \in \mathbb{R}^m$ represents a vector of m associated clinical attributes, the objective is to learn a mapping system f that produces an ordered tuple (Y_s, Y_c):

$$f: (X, C) \rightarrow (y_s, y_c)$$

The functional dependencies of the system are defined as follows:

Segmentation Output (Y_s): $Y_S \in \{0,1\}^{H\times W\times D}$ denotes the binary lesion segmentation mask. Formally, for each voxel coordinate (i, j, k) in the 3D spatial domain, where i, j, and k correspond to the dimensions of height (H), width (W), and depth (D), respectively—the term $Y_s(i, j, k)$ represents the value of the segmented voxel. A value of $Y_s(i, j, k) = 1$ indicates the presence of a hypoxic-ischemic lesion at that specific location, while $Y_s(i, j, k) = 0$ represents healthy brain tissue. This voxel-wise binary classification provides a precise spatial localization of the injury within the 3D MRI volume [3].

Classification Output (Y_C): $Y_C \in$ {mild, moderate, severe} represents the severity class. This output is a function of both the clinical vector C and the imaging features derived from the segmented mask Y_S, such that $Y_C = f_{cls}(C, \phi(Y_S))$, where ϕ denotes a feature extraction operation.

The framework aims to minimize segmentation error for Y_S while maximizing classification accuracy for Y_C.

1.2. Major contributions

The major contribution of this work is the development of an integrated framework for automated lesion segmentation, multimodal severity classification, and XAI analysis for neonatal HIE. The prior works mainly focused on these problems independently. Our framework provides a unified pipeline that combines lesion segmentation, severity classification, and explainability of the model.

The main contributions of this integrated system are:

1) We propose a novel two-stage framework that first segments HIE lesions and subsequently classifies their severity using multi-sequence neonatal MRI scans. The segmentation stage leverages a SegResNet architecture with a 2.5D input strategy to effectively capture both spatial and contextual information.
2) We use a hybrid binary cross-entropy (BCE) + Dice loss, which functions as a reliable class-weighted loss and balances per-pixel accuracy with the severe class imbalance.
3) An early fusion strategy is adopted for constructing a rich feature representation by integrating segmentation-derived imaging features with available clinical data, thereby bridging radiological and clinical perspectives.
4) For severity assessment, an XGBoost classifier is trained on the fused feature set to categorize lesions into clinically relevant classes of *mild, moderate,* and *severe*, demonstrating improved accuracy and robustness.
5) We confirm the framework's clinical reliability with quantitative and qualitative XAI analysis using Gradient-weighted Class Activation Mapping (Grad-CAM). We provide a reliable quantitative proof of the model's faithfulness using Insertion and Deletion Area Under the Curve (AUC) scores.

2. Literature Review

Research on neonatal HIE has increasingly leveraged neuroimaging and machine learning (ML) to enhance lesion detection, outcome prediction, and diagnosis. Early works mainly explored MRI biomarkers for prognosis. Zheng et al. [7] showed that perfusion imaging using pulsed arterial spin labeling could forecast motor and language outcomes in neonates.

With the rapid progress of artificial intelligence, deep learning has become central to automated neonatal brain image analysis. Raurale et al. [8] employed CNNs on Electroencephalogram (EEG) with quadratic time–frequency distributions, enabling automated HIE grading. Similarly, Tian et al. [9] developed an MRI-based deep learning model that surpassed conventional diagnostic workflows. Another important step was the introduction of the BONBID dataset by Bao et al. [3], which provided standardized imaging and clinical data to support reproducible research.

Several domain-specific models have since been proposed. Bao et al. [10] introduced ParadiseNet, a segmentation network combining global-local learning and uncertainty refinement for accurate lesion segmentation. Kumar et al. [11] developed a recurrent neural network-based framework that integrated texture features, although low-contrast regions remained challenging. Usman et al. [12] compared multiple loss functions for 3D U-Net segmentation of HIE lesions, showing that compound losses improved performance. Further advancements in segmentation for addressing the class imbalance were done by Rao et al. [4]. They proposed a 2D U-Net ++ architecture with a novel Tversky, Log-Hausdorff, and Focal losses (TLHF) loss function, particularly to reduce the effects of class imbalance and to improve the identification of small, low differentiation boundaries. Beyond imaging, Ashoori et al. [13] extracted cerebral oxygenation features to predict short-term outcomes using XGBoost, while Lewis et al. [14] leveraged radiomic and geometric MRI features for predicting 18-month neurodevelopmental outcomes. More recently, Lew et al. [15] introduced the AI-OPiNE framework, which combined artificial intelligence with clinical data to predict

neurodevelopmental outcomes, underscoring the clinical utility of such automated pipelines.

Traditional ML models are common in the HIE literature for classification tasks such as severity grading from EEG signals [8, 13]. However, the use of ML models in lesion segmentation from MRI is not well-established, as they require complex, pixel-wise feature engineering for every voxel in the scan. Due to this reason, end-to-end deep learning models that can learn spatial features are used for the segmentation task instead of ML models. Therefore, the related performance benchmarks for our SegResNet, as discussed, are state-of-the-art deep learning architectures [4, 10, 12] rather than traditional ML techniques.

Despite these advancements, existing systems still face several limitations that hinder their widespread clinical adoption. Many deep learning models struggle with detecting subtle or small HIE lesions, particularly mild cases, due to class imbalance and low lesion contrast in MRI scans. Additionally, the majority of research uses single-modality data, which makes it difficult to obtain complementary data from advanced MRI sequences, EEG, and clinical characteristics. The inability of models trained on a particular dataset to function consistently across various imaging methods and scanner modifications presents another difficulty. Furthermore, outcome prediction's clinical relevance for making decisions in real time is diminished because it is often delayed or based on a small number of indicators. These drawbacks emphasize the need for thorough, multimodal, and clinically understandable frameworks that can offer precise segmentation and trustworthy severity classification in neonatal HIE.

An important challenge in HIE segmentation is the severe class imbalance, due to which small, diffuse lesions are not identified [3]. Previous works have tried to address this in many ways. Some studies explored weighted loss functions, such as focal loss, which has a theoretical background of focusing more on challenging boundary pixels instead of the easy background pixels [12]. Usman et al. [12] and Rao et al. [4] experimented with region-based loss functions like the Dice loss, which is a more common and stable approach. In case of imbalanced datasets, Dice loss is theoretically better as it measures only how well the predicted lesion regions overlap with the ground truth and avoids the per-pixel accuracy, which measures the correctly predicted background pixels.

Another important limitation faced in this area is the use of private, single-center datasets in earlier studies. This approach reduces the generalizability of the models and violates the FAIR (Findable, Accessible, Interoperable, Reusable) data principles. The recently released BONBID-HIE dataset [3], which we use in this study, addresses this gap by providing a well-annotated, public benchmark dataset for research purposes.

Furthermore, the literature review shows that most existing studies focus mainly on unimodal data or single tasks like outcome or severity prediction using non-imaging data like EEG [8, 13] or lesion segmentation using only MRI images [4, 10, 12]. As per our knowledge, there is a lack of an integrated system that combines lesion segmentation and severity classification. This helps to connect the lesion characteristics with clinical outcomes.

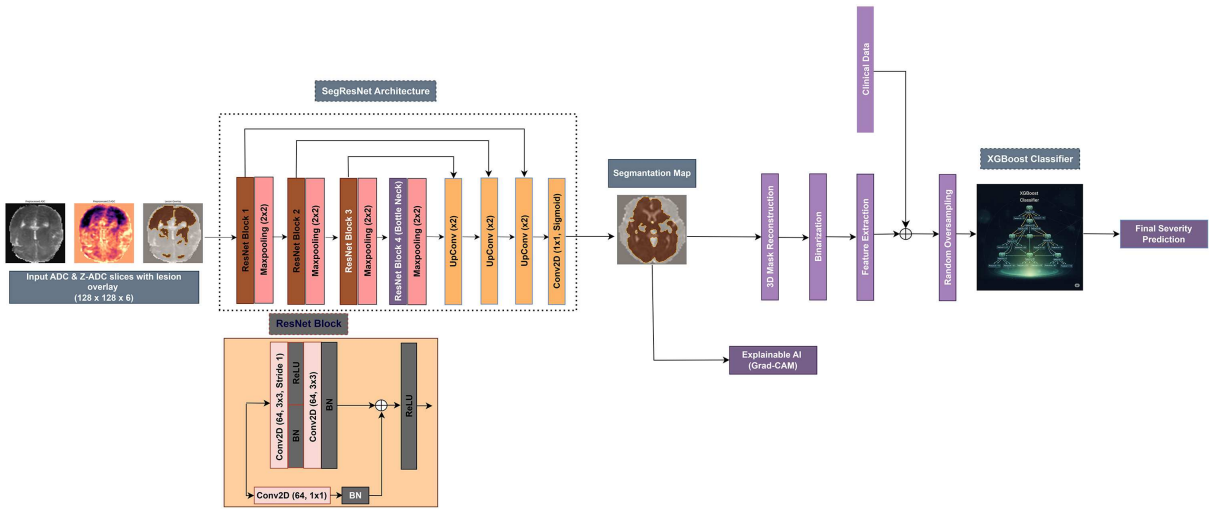
Our methodological decisions have been affected by broader trends in medical image analysis, which extend beyond particular HIE applications. Recent research has shown that multi-scale convolutions and deep learning ensembles are effective for a variety of neurological and pediatric diagnostic tasks [16–19]. Additionally, a lot of recent literature has stressed the need for XAI to guarantee transparency and reliability in healthcare applications [20–24]. Together, these studies demonstrate the trend toward models that are both accurate and comprehensible, an idea backed by multiple reviews on XAI in MRI analysis [25–29]. Our suggested framework, which combines intrinsic and post hoc explainability with robust segmentation, is motivated by this context.

3. System Description

In order to reduce the gap between segmentation and clinical severity, the suggested methodology uses a unique two-stage workflow for neonatal HIE assessment. An overview of the proposed methodology is illustrated in Figure 1.

First, a SegResNet model with 2.5D inputs from apparent diffusion coefficient (ADC) and Z-scored ADC (Z-ADC) maps is used to segment lesions. Accurate identification of brain lesions and the extraction of quantitative measurements like lesion volume and form are made possible by this phase. A thorough XAI analysis utilizing Grad-CAM is also carried out to evaluate the segmentation model's clinical usefulness and explainability.

Figure 1
The proposed methodological framework



The second step involves extracting quantitative data, including lesion volume, from the predicted lesion masks. An early fusion method is then used to merge these imaging features with clinical attributes to create a full feature set. This multimodal data is then used to classify HIE severity into mild, moderate, and severe cases using an XGBoost classifier. Gradient boosting efficiently handles organized, tabular input, whereas deep convolutional networks can handle the complex image segmentation process because of this task separation. The system is more clinically reliable due to its modular design, which also guarantees robust lesion identification and increases the accuracy of severity categorization.

3.1. Feature extraction

Features are extracted for the severity classification task using a four-step procedure after the segmentation stage.

1) Prediction: The trained 2.5D SegResNet model processes an entire patient's 3D volume by iterating through it slice-by-slice. For each 2.5D input slice (128 × 128 × 6), the model outputs a corresponding 2D (128 × 128 × 1) voxel-wise probability map.
2) Reconstruction: These individual 2D probability maps are then stacked back together in their correct anatomical order. This step converts the 2D predictions into a single, unified 3D probability volume for the entire brain, where each voxel contains a continuous value from 0.0 to 1.0 representing the likelihood of it being a lesion.
3) Binarization: The 3D probability map is binarized into a 3D lesion mask using an optimal threshold. A sensitivity analysis was conducted on the validation set by evaluating several individual thresholds (e.g., 0.3, 0.4, 0.5, 0.6, 0.7) for choosing the optimal threshold. It was found that a threshold of 0.5 gave the most stable and balanced result.
4) Extraction: Finally, from this single, unified 3D binary mask, the quantitative, clinically meaningful 3D shape descriptors are calculated.

From the reconstructed 3D mask, several quantitative features are computed to capture lesion morphology and extent of injury:

1) Lesion volume: Total count of lesion voxels, reflecting the overall size of the brain injury.
2) Bounding box volume: The volume of the smallest 3D box that surrounds the lesion and gives details about its direction and geographical placement.
3) Equivalent diameter: A useful metric for differentiating between compact and irregular lesion shapes. It is defined as the diameter of a sphere with the same volume as the lesion.
4) Extent: The degree of lesion spread, indicated by the ratio of lesion volume to bounding box volume.

During the feature extraction stage, advanced radiomic or shape features such as sphericity, intensity histograms, or gray-level co-occurrence matrices were eliminated. This was an intentional methodological decision. Our project's main objective is to guarantee clinical interpretability. While the straightforward, first-order variables we chose (lesion volume, extent, etc.) are clinically significant and simple for a physician to understand, the higher-order radiomic features are complicated and challenging to interpret.

An early fusion method, which combines diverse data sources into a single feature vector before classification, is then used to integrate these imaging-derived features with each patient's clinical information. This improves severity classification over late fusion methods by enabling the model to identify complex nonlinear correlations between imaging biomarkers and clinical indications.

3.2. Model architecture

Two specialized models are used in the proposed framework: an XGBoost classifier for severity prediction utilizing the retrieved multimodal data and a SegResNet for lesion segmentation.

SegResNet (segmentation model): A SegResNet design, which combines the advantages of ResNet [5] and U-Net [6], is used to segment lesions. In order to capture multi-scale contextual characteristics, the encoder gradually downsamples the input using residual blocks and max-pooling in the encoder–decoder architecture of the network. While skip connections combine encoder characteristics with decoder outputs to maintain spatial details, the decoder uses transposed convolutions to symmetrically upsample these representations. By resolving the vanishing gradient issue and guaranteeing robust feature learning, residual blocks allow for deeper and more stable training. In the end, voxel-wise probability maps of the lesions are generated via a sigmoid activation after a 1 × 1 convolution. With 21 convolutional layers and about 8.06 million parameters, the model has enough capacity to capture the nuanced characteristics of HIE lesions. Table 1 displays the SegResNet model's comprehensive design. It follows an encoder–decoder structure with ResNet blocks, skip connections, and a bottleneck stage, enabling effective feature extraction and accurate lesion boundary reconstruction.

Table 1
SegResNet model architecture for HIE lesion segmentation

Layer	Output shape	Filters/units
Input	(128, 128, 6)	–
Encoder Path		
ResNet Block 1	(128, 128, 64)	64
Max Pooling	(64, 64, 64)	–
ResNet Block 2	(64, 64, 128)	128
Max Pooling	(32, 32, 128)	–
ResNet Block 3	(32, 32, 256)	256
Max Pooling	(16, 16, 256)	–
Bridge (*Bottleneck*)		
ResNet Block 4	(16, 16, 512)	512
Decoder Path (with Skip Connections)		
Up-conv+ Concat	(32, 32, 512)	256
ResNet Block 5	(32, 32, 256)	256
Up-conv+ Concat	(64, 64, 256)	128
ResNet Block 6	(64, 64, 128)	128
Up-conv+ Concat	(128, 128, 128)	64
ResNet Block 7	(128, 128, 64)	64
Output Layer		
1 × 1 Convolution (Sigmoid)	(128, 128, 1)	1

XGBoost (severity classification model): XGBoost (eXtreme Gradient Boosting), a potent ensemble-based algorithm well known for its effectiveness and resilience on structured data, is

used to carry out the final severity classification. XGBoost creates a sequential ensemble of decision trees, each of which iteratively fixes the mistakes of its predecessors. The model is able to capture intricate, nonlinear interactions inside the multimodal feature space because of this boosting technique. In order to handle the multi-class classification problem in this work, the classifier was set up with 200 estimators, generating class probabilities for mild, moderate, and severe HIE. Because of its exceptional performance on tabular data and its capacity to simulate complex feature interactions while preserving computational efficiency, XGBoost was selected.

4. Performance Evaluation

4.1. Dataset and preprocessing

The Boston Neonatal Brain Injury Dataset for Hypoxic-Ischemic Encephalopathy (BONBID-HIE), which includes MRI images of 133 neonates, is used in this study. The dataset comprises clinical data, brain atlases, and ADC and Z-ADC maps with expert-annotated lesion masks. It is split into 85 training, 44 testing, and 4 validation cases. Z-ADC maps indicate outlier locations in comparison to a normative atlas, whereas ADC maps show diffusion aberrations associated with acute damage. Maternal and newborn characteristics are included in clinical records, offering supplementary data for subsequent prediction.

The preprocessing pipeline used was as follows:

1) Brain cropping: Otsu's thresholding on ADC maps was used to eliminate non-brain regions, and then the brain mask's bounding box was cropped. This guaranteed attention to pertinent anatomy while lowering computational overhead.
2) Resizing: To provide consistent model input, cropped volumes were shrunk to 128 × 128 pixels in the axial plane.
3) Normalization: To lessen intensity bias from nearby structures, Z-score normalization was used within the brain mask on both ADC and Z-ADC maps.
4) 2.5D slice stacking: To create a six-channel input (128 × 128 × 6), each input was created by stacking the target slice with its neighboring slices from both ADC and Z-ADC maps. This maintained 2D efficiency while providing local 3D context.
5) Data augmentation: During training, geometric augmentations (flips, 90° rotations) and photometric changes (brightness and contrast modifications) were used to improve robustness.

The preprocessing stage confirms that the inputs are consistent and appropriate for segmentation. The normalized ADC map, the Z-ADC map produced by intensity standardization, and the lesion overlay with ground truth masks are examples of the results shown in Figure 2. ADC preserves structural detail, while Z-ADC enhances abnormal regions, providing complementary information for accurate lesion detection.

4.2. Training strategy

SegResNet: The segmentation model was trained for up to 50 epochs with an initial learning rate of 1×10^{-4} using the Adam optimizer and a batch size of 16. Early stopping with a patience of 15 epochs was applied to prevent overfitting and for robust training. Training converged after 30 epochs as validation loss plateaued. Learning rate reduction on a plateau with a patience of 5 was used to fine-tune the learning rate. Model checkpointing was used for selecting the best model based on the Dice coefficient. A hybrid loss combining BCE and Dice loss was employed to balance voxel-wise accuracy with overlap-based performance. This hybrid loss function was used to address the challenges of this dataset. Stable gradients and reliable voxel-level accuracy are ensured by the BCE, while the Dice score serves as an effective class-weighting mechanism for handling the severe class imbalance. This combination contributes to both per-pixel stability and reliable performance on small, critical lesion regions.

XGBoost: Severity labels were derived using validated clinical thresholds (e.g., Apgar score, lowest pH) in combination with extracted lesion volume. To address class imbalance, **random oversampling** was applied to increase the representation of minority classes, and sample weights were incorporated during training. The XGBoost classifier was then trained on this balanced dataset to perform final severity classification. The XGBoost parameters, such as n_estimators = 200, were selected based on preliminary tuning on the validation set to optimize the multi-class log-loss (mlogloss) without overfitting.

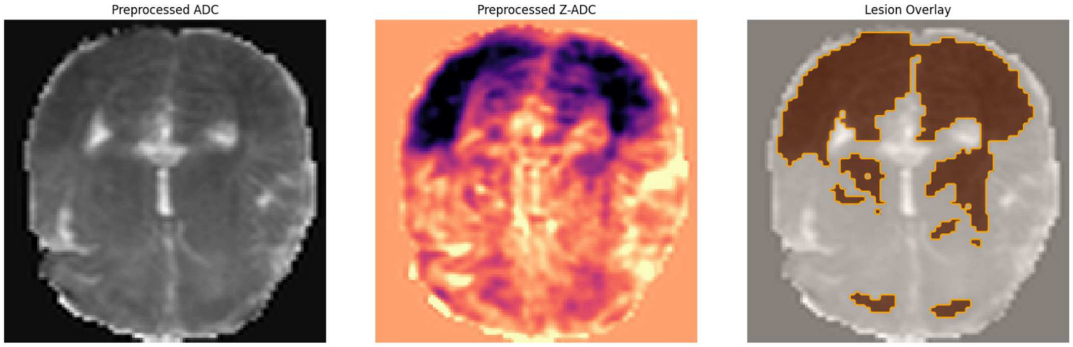
4.3. Evaluation procedure

Standard quantitative indicators were used to assess the system independently for segmentation and severity categorization.

4.3.1 Segmentation evaluation:

1) Dice coefficient: Measures spatial overlap between predicted and ground truth masks [10].
2) Sensitivity (Recall): Ability to correctly detect lesion pixels.
3) Specificity: Ability to correctly identify healthy tissue.

Figure 2
Visualization of slices after preprocessing



4) MASD (Mean Average Surface Distance): Average boundary distance between predicted and ground truth masks (lower is better).
5) NSD (Normalized Surface Distance): Percentage of boundary voxels within a tolerance (e.g., 2 mm) of the ground truth (higher is better).

4.3.2. Classification evaluation:

1) Accuracy: Overall proportion of correct severity predictions [12].
2) Precision: Fraction of predicted positives that are correct.
3) Recall: Fraction of actual positives correctly detected.
4) F1-score: Harmonic mean of precision and recall, balancing sensitivity and precision.

Overall, a thorough examination of lesion segmentation and severity categorization is guaranteed by this evaluation approach. The success of the system in differentiating between mild, moderate, and severe cases is captured by the classification metrics, whereas Dice, MASD, and NSD directly quantify the spatial and boundary accuracy of the segmentation task. When taken as a whole, these measures offer a fair assessment of the model's dependability in both image-level analysis and clinical decision assistance.

4.4. Explainable AI for model interpretability

We used a post hoc explainability technique to go beyond overall performance indicators and achieve a thorough knowledge of the predictive behavior of the model. An advanced attribution technique called Grad-CAM was used to create visual explanations for the SegResNet model's segmentation results. The main aim was to generate saliency maps that highlight the specific pixel regions in the input image that were critical in the model's decision to classify that area as a lesion.

The earlier layers in the network contain only very fine details like edges, textures, etc., while the deeper layer captures high-level information about the lesion but lacks enough spatial details. In order to balance these, the final convolutional layer of the SegResNet encoder was used in this analysis. It consists of the richest semantic features crucial for identifying the lesion, along with adequate spatial resolution for meaningful visualization.

Grad-CAM was chosen over other associated methods, such as saliency maps, for some reasons. First, as a secondary analysis, the gradient-based method is framework-independent and can be applied to the trained SegResNet model without the need for any modifications in the architecture or re-training [30]. Second, it has high performance, and only a single backward pass is required for generating a heatmap. Lastly, when compared with other simpler saliency methods, which are noisy in appearance, Grad-CAM produces smooth, class-differentiating heatmaps that are easily interpretable by humans in a clinical context.

The Grad-CAM algorithm is mathematically defined as follows: for a given class output y^c (representing the lesion class) and the activation maps A^k of the target convolutional layer, the gradients of the score for class c with respect to the feature map activations A^k_{ij} are determined using backpropagation. Global average pooling is applied on gradients across their spatial dimensions (i, j) to obtain the neuron importance weights α^c_k, which represent the contribution of each feature map k to the final decision:

$$\alpha^c_k = \left(\frac{1}{Z}\right) \sum_i \sum_j \frac{\partial y^c}{\partial A^k_{ij}} \tag{1}$$

where Z is the total number of pixels in the feature map.

The final Grad-CAM heatmap, $L^c_{Grad-CAM}$, is a weighted linear combination of the forward-pass activation maps, followed by a Rectified Linear Unit (ReLU) activation. The ReLU is applied to ensure that only the features that have a positive influence on the class of interest are visualized:

$$L^c_{Grad-CAM} = ReLU\left(\sum_k \alpha^c_k A^k\right) \tag{2}$$

The obtained low-resolution heatmap is then resampled to the original input image dimensions and superimposed on the input slice to provide a transparent and understandable visualization of the model's focus. One of the main advantages of this technique is its applicability to any CNN-based architecture without requiring modifications or updates.

5. Results and Analysis

The performance of the proposed two-stage system was evaluated on the unseen test set. Both the lesion segmentation (Stage I) and the severity classification (Stage II) were analyzed using standard quantitative metrics and qualitative visualizations.

5.1. Segmentation performance

Table 2 presents the quantitative comparison of our segmentation model against the benchmark method reported by Bao et al. [10]. The proposed SegResNet achieved a mean Dice coefficient of 0.758, indicating strong overlap between the predicted and ground truth lesion masks, but with a standard deviation of 0.29, which suggests inconsistent performance. The sensitivity of 0.802 highlights the model's ability to capture true lesion regions, which is particularly important in clinical settings to avoid missing affected tissue. Although the benchmark model achieved a slightly lower MASD score, our approach outperformed it in Dice score, sensitivity, and NSD, confirming improved boundary alignment and lesion detection. While the model shows a good volumetric overlap, the MASD result of 3.64 ± 7.58 mm indicates that the lesion delineation was less reliable.

Table 2
Comparison of segmentation performance

Metric	Our model	Bao et al. [10]
Dice coefficient	0.758 ± 0.29	0.61
Sensitivity (recall)	0.802 ± 0.29	0.64
Specificity	0.993 ± 0.012	0.9989
MASD (mm)	3.64 ± 7.58	3.48
NSD @ 2 mm	0.79 ± 0.27	0.72

Figure 3 shows the training and validation curves. The training loss decreases steadily, while validation loss stabilizes after around 10 epochs. The Dice coefficient for training steadily improves, whereas validation Dice plateaus around 0.70, demonstrating effective learning with early stopping preventing overfitting.

Qualitative examples in Figures 4 and 5 further confirm that the model accurately localizes lesions and captures fine boundaries, although some small regions show slight over-segmentation.

The MASD result of 3.64 ± 7.58 represents the direct quantitative measure of the model's inconsistency in boundary regions. The high standard deviation indicates the model's inconsistency in boundary regions. This inconsistency is visually confirmed in

Figure 3
Training and validation curves for loss and dice coefficient

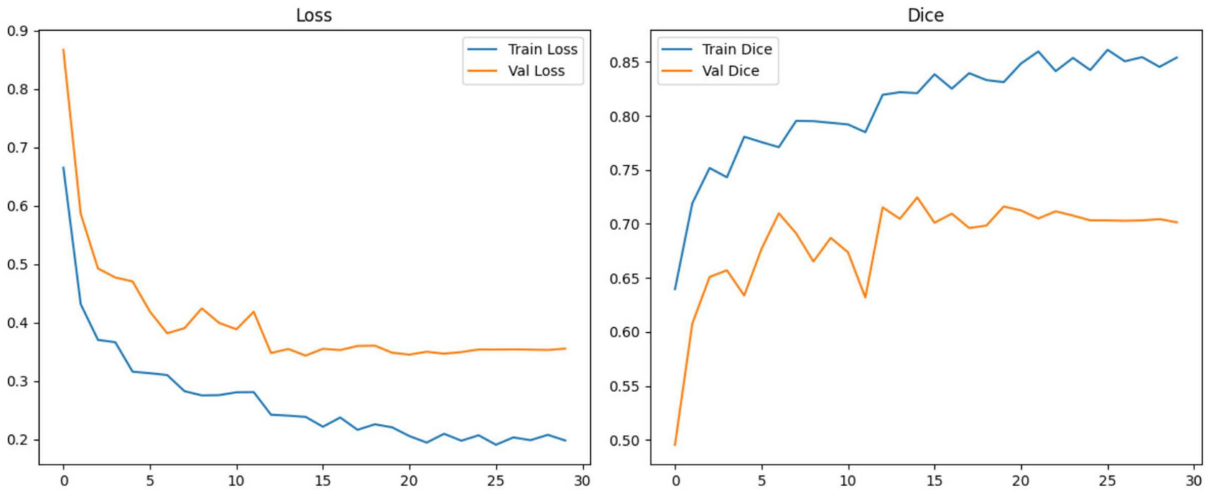


Figure 4
Qualitative results of lesion segmentation: input ADC slice, ground truth mask, and predicted mask

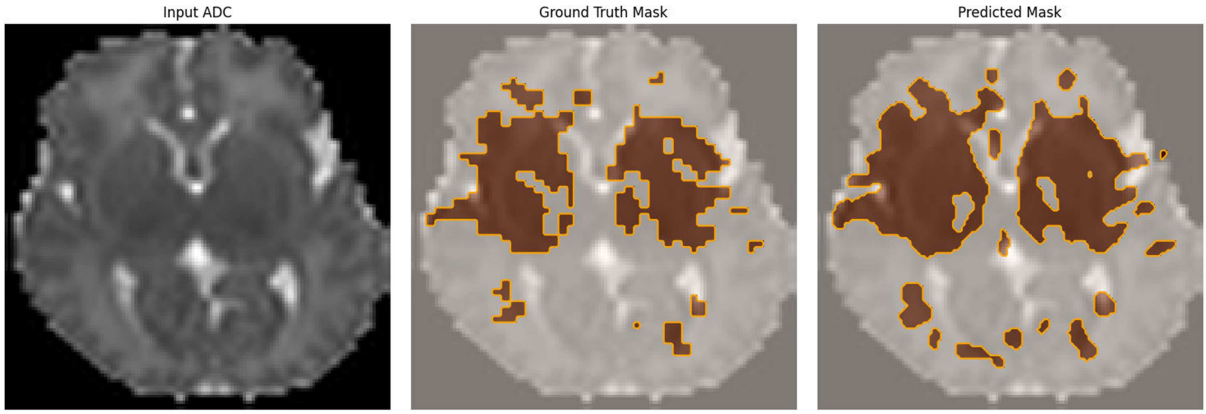
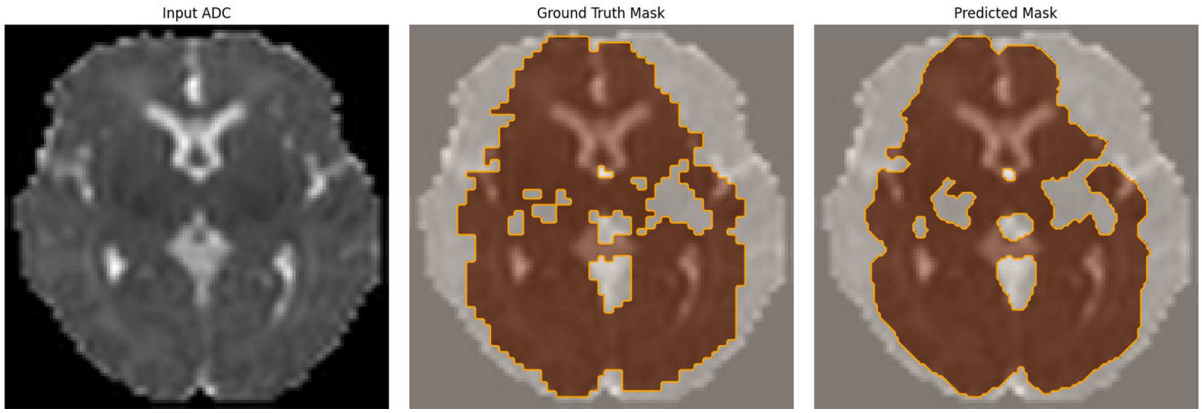


Figure 5
Additional visualization of segmentation performance



the error maps given in Figure 6. These error maps highlight the over-segmented regions (red regions) and under-segmented regions (blue regions), which are the primary cause of high MASD variance. The correctly segmented areas are represented using green color.

5.2. Classification performance

The second stage of the framework integrates segmentation-derived lesion features with clinical attributes for HIE severity grading using XGBoost. Table 3 reports the classification results,

Figure 6
Over-segmented and under-segmented lesion regions

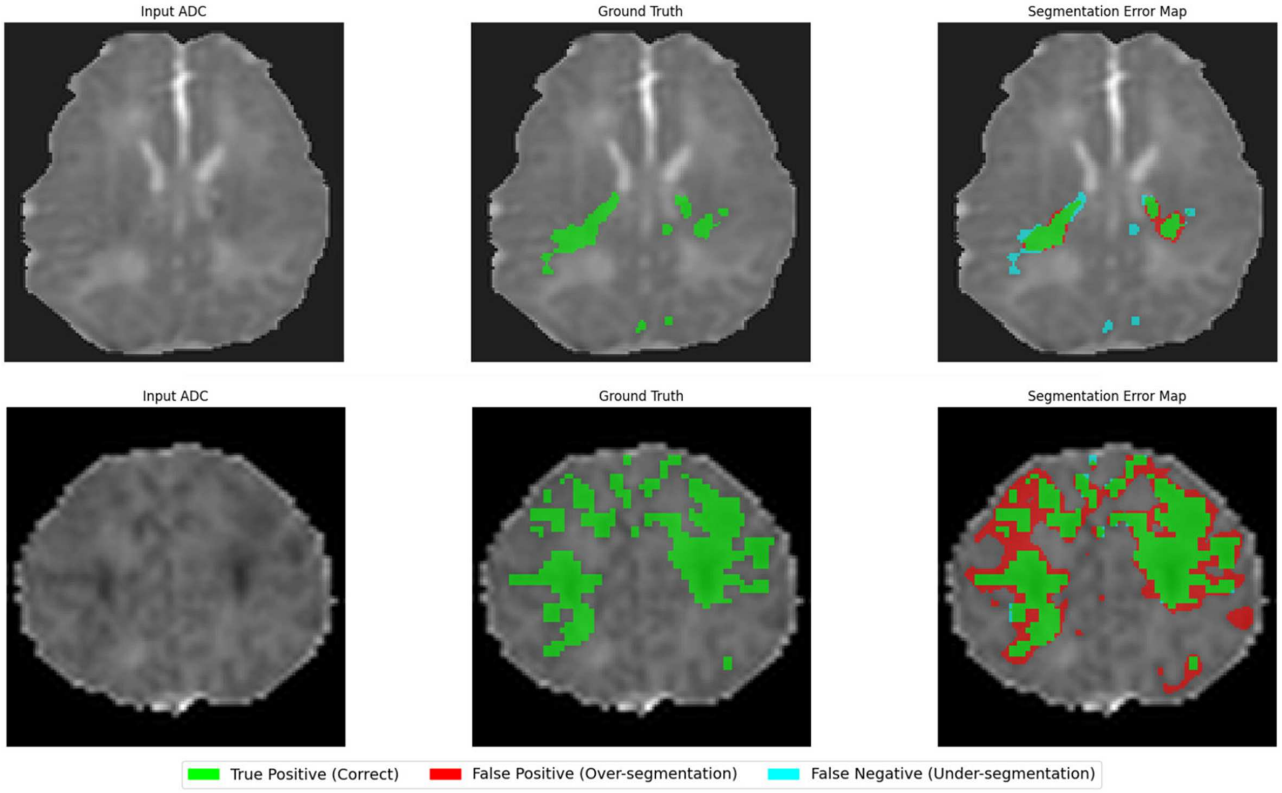


Table 3
Classification report for test set

Class	Precision	Recall	F1-score	Support
Mild (0)	0.38	1.00	0.55	3
Moderate (1)	1.00	0.76	0.86	21
Severe (2)	1.00	1.00	1.00	20
Accuracy			0.89	44

while Figure 7 shows the confusion matrix. The classifier achieved an overall accuracy of 88.64%, demonstrating strong performance on moderate and severe cases.

The classifier achieved perfect precision and recall for severe cases, which are clinically the most critical. Moderate cases were classified with an F1-score of 0.86, though some were misclassified as mild. While mild cases achieved a recall of 1.00, the precision was lower (0.38), mainly due to the limited number of samples in this category, leading to false positives.

The model perfectly classified all severe cases. Its main difficulty is in distinguishing between mild and moderate cases. Misclassifying moderate instances as mild is a serious mistake that can result in infants who need therapeutic treatment receiving insufficient care.

5.3. Explainable AI (XAI)

The model's predictive behavior was thoroughly demonstrated by the use of Grad-CAM. This was validated for clinical reliability using both quantitative and qualitative analysis.

1) Qualitative analysis of model behavior and clinical implications: We were able to gain a thorough knowledge of the model's decision-making process by visually examining Grad-CAM heatmaps on the test set. The predicted segmentation map and associated heatmap are shown in Figure 8. Three representative cases from the test set were thoroughly examined in order to further evaluate the model's interpretability.

a. True positives (successful detections): The visualization of the XAI analysis for a true positive case is shown in Figure 9. The heatmaps consistently showed high-intensity activations that were spatially associated with the lesion boundaries for the lesions that were accurately segmented. Over the diseased tissue, the areas of highest activations were clearly identified. This strongly suggests that rather than relying on irrelevant correlations or deceptive artifacts in the image background, the model has successfully learned the significant disease signs in accordance with radiological procedures.

Clinical implications: This overlap shows that the model can be a trustworthy tool for identifying areas of interest, perhaps lowering supervision errors, and streamlining diagnostic procedures.

Figure 7
Confusion matrix of severity classification on the test set

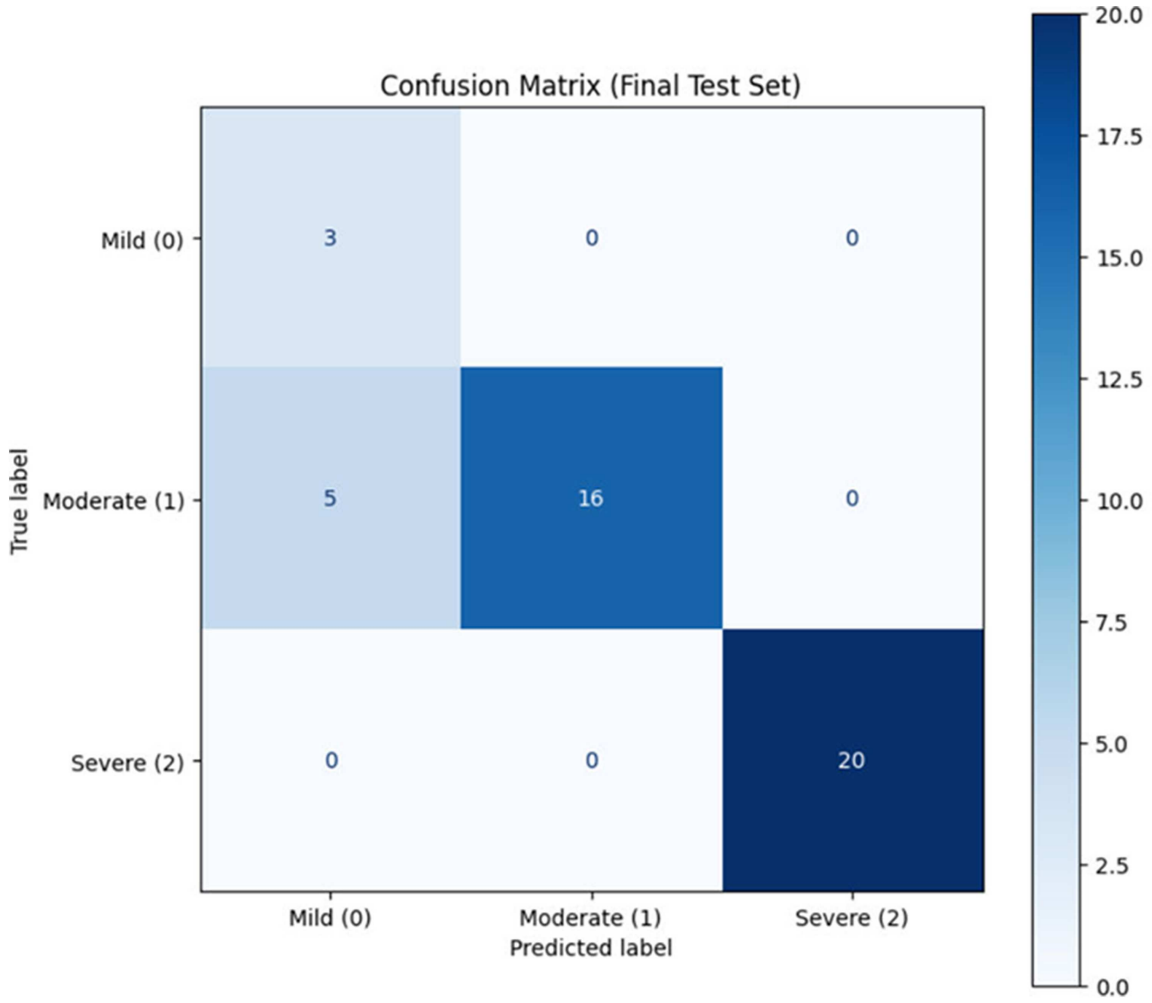


Figure 8
Visualization of the predicted mask and the corresponding heatmap

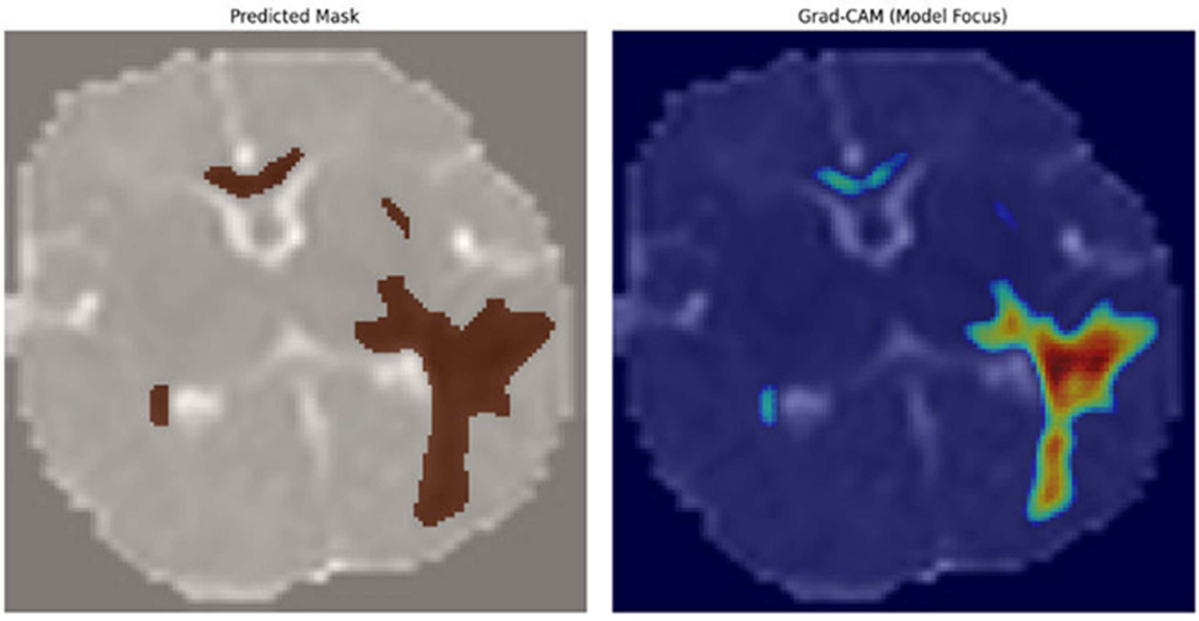
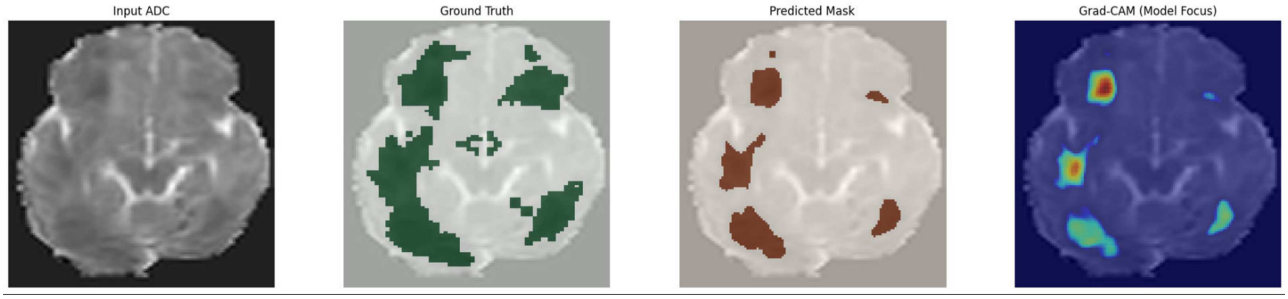


Figure 9
Explainable AI visualization of a true positive case: input ADC slice, ground truth lesion, predicted mask, and Grad-CAM heatmap



b. False negatives (missed lesions): Figure 10 shows the XAI visualization of a false negative case. The heatmaps displayed low-intensity, latent activations in the appropriate anatomical region when the model was unable to identify a clinically verified lesion. This important realization implies that such failures typically result from a lack of confidence, where the extracted features did not surpass the learned activation threshold, rather than a total failure in the model's attention.

Clinical implications: These findings suggest the sensitivity limits of the model, usually for small or low-contrast lesions. It suggests that the diagnostic accuracy could be enhanced by a human-in-the-loop system in which medical professionals examine regions with weak activity.

c. False positives (incorrect detections): The XAI analysis of a false positive case is visualized in Figure 11. The heatmap proved to be an invaluable diagnostic tool in situations where the model misclassified healthy tissue, precisely pinpointing the anatomical areas or patterns that confused the model.

Clinical implication: This finding allows for the early identification of systematic model biases. Understanding which pathological structures are being misidentified can help future data preparation and augmentation strategies even better, and it helps to improve the specificity of the model and reduce the rate of false positives.

2) Quantitative validation of faithfulness: Insertion and deletion AUC analysis [31] was conducted to provide a more reliable measure of the model's faithfulness. The insertion metric measures the increase in probability as the most important pixels are repeatedly added to a null image. The deletion metric measures the reduction in the model's probability as the most important pixels are repeatedly removed. A high insertion AUC and a low Deletion AUC are desired.

The results of the faithfulness analysis are shown in Table 4. The model achieved a mean insertion AUC of 0.954 ± 0.01

Figure 10
Explainable AI visualization of a false negative case: input ADC slice, ground truth lesion, predicted mask, and Grad-CAM heatmap

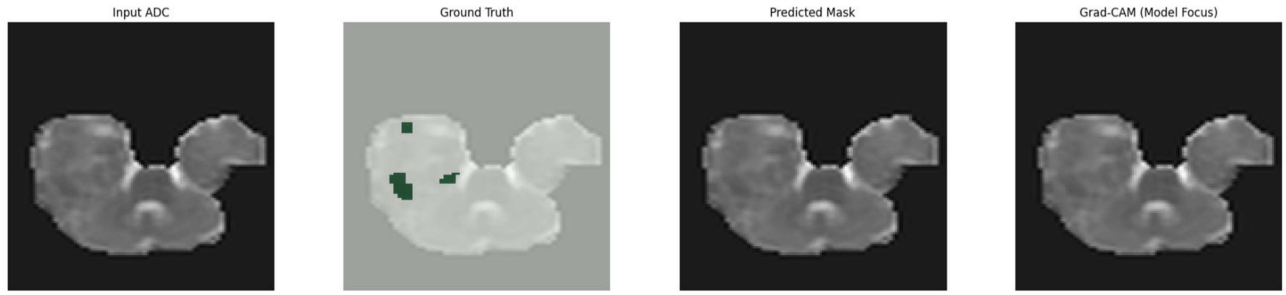


Figure 11
Explainable AI visualization of a false positive case: input ADC slice, ground truth lesion, predicted mask, and Grad-CAM heatmap

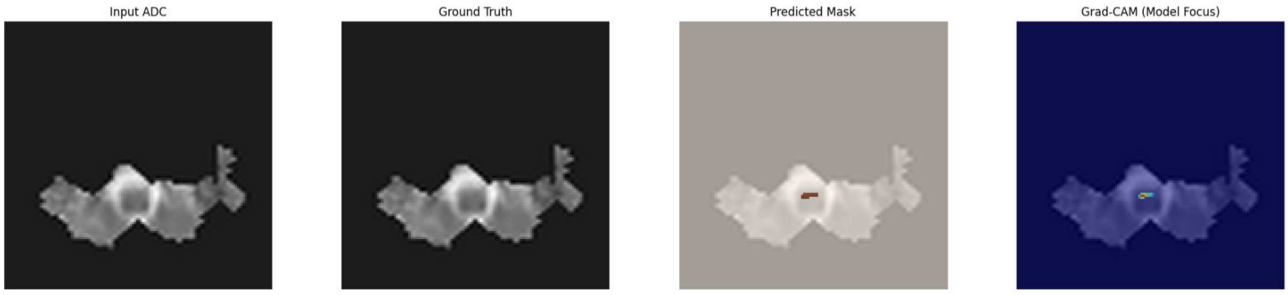


Table 4
XAI faithfulness analysis (insertion/deletion AUC)

Sample index	Insertion AUC	Deletion AUC
105	0.9317	0.1931
156	0.9549	0.2955
460	0.9644	0.1969
712	0.9544	0.2015
1245	0.9659	0.1387
Mean (± std dev)	0.954 ± 0.01	0.205 ± 0.06

and a mean deletion AUC of 0.205 ± 0.06. These quantitative results demonstrate the faithfulness of the model. This ensures that the identified pixels are sufficient for the model to make predictions.

3) Quantitative validation of model focus: In addition to qualitative evaluation, quantitative assessment is also important for an objective and reproducible analysis of the model's attention mechanism. We quantitatively measured the spatial alignment between the model's focus and ground truth by calculating the Dice coefficient. For isolating the more salient regions, the continuous Grad-CAM heatmap was binarized by setting a threshold at the 75th percentile of its non-zero activation values. The results for a random subset of test cases are presented in Table 5.

Table 5
Quantitative analysis of model focus using Grad-CAM

Sample index	Focus vs. ground truth (Dice)	Focus vs. prediction (Dice)
895	0.5396	0.4017
401	0.3115	0.4095
597	0.4245	0.4015
1244	0.6962	0.4008
215	0.4073	0.4005
Mean (± std dev)	0.476 ± 0.130	0.403 ± 0.003

4) XAI threshold sensitivity analysis: A sensitivity analysis was conducted to ensure the reliability of the focus vs. ground truth analysis. The mean Dice score between the model's focus and ground truth was calculated at multiple thresholds. The results of the sensitivity analysis are presented in Table 6. It is observed that the Dice score shows a steady increase when the threshold becomes stricter. It shows that our original 75th percentile result (Dice 0.152) is part of a reliable and meaningful pattern. Furthermore, the higher Dice score achieved for the 90th percentile confirms that the model's most confident features are accurately aligned with the ground truth pathology. This analysis confirms that our XAI findings are reliable and not a result of threshold-specific bias.

Table 6
XAI threshold sensitivity analysis

Percentile threshold	Mean focus vs GT Dice	Std focus vs GT Dice
50^{th}	0.088	0.075
75^{th}	0.152	0.134
90^{th}	0.284	0.261

5) Clinical implications of XAI analysis: The quantitative and qualitative analyses provide a clear connection between interpretability of the model and clinical reliability. The clinical applications of this XAI analysis are:

   a. Diagnostic confirmation and trust building: The heatmap of true positive analysis indicates that the model's predictions are based on the correct pathological tissue. This makes it easier for the physicians to verify that the model isn't generating predictions based on unrelated patterns. Building clinician trust is aided by this alignment between the model's focus and the recognized disease.
   b. Error analysis and bias detection: In case of false positives, this analysis aids in determining if the model misidentified another anatomical structure as a lesion or in determining why the model failed. This ensures that doctors are aware of the situations in which the model fails and aids in the detection of model biases.

The study reveals a mean Dice score of 0.476 (SD = 0.130), which indicates a strong, although somewhat erratic alignment between the ground truth lesions and the model's focus. Focus vs. prediction's extremely low SD (±0.003) attests to outputs that are extremely stable and consistently driven by high-activation zones. Reliable attention alignment with actual lesion locations is demonstrated by the strong agreement between the two Dice scores.

The overall framework, which used XGBoost for severity grading and SegResNet for segmentation, worked well, reaching 88.64% accuracy in severity classification and precise lesion localization. Despite the fact that class imbalance reduces mild-class precision, clinically significant cases were routinely identified. The XAI results reinforce the model's promise as a reliable decision-support tool by confirming that its decisions are clear, clinically significant, and well-aligned with the objective of neonatal HIE lesion detection.

6. Conclusion

The proposed strategy shows great promise as a decision-support system for neonatal HIE by combining SegResNet for lesion segmentation with an XGBoost classifier for severity grading. While the classification stage achieved an overall accuracy of 88.64%, especially in the recognition of moderate and severe instances, which are the most clinically urgent, the segmentation model showed excellent sensitivity, ensuring reliable detection of injured brain regions. In addition, this study demonstrated the clinical reliability of the SegResNet model through an in-depth XAI analysis.

The primary contribution of the work is the validation of a two-part XAI framework. Post hoc XAI in Stage 1 verifies that model attention and ground truth are strongly aligned both qualitatively and quantitatively. Because the XGBoost severity classifier is intrinsically interpretable in Stage 2, physicians can audit predictions using a limited number of transparent features.

Segmentation quality drives all downstream operations, according to an analysis of the entire three-stage pipeline. Lesion volume is distorted by mistakes like under- or over-segmentation (Figure 6), which can lead to some intermediate cases being incorrectly categorized as severe. Even though XAI results (Tables 5 and 6) demonstrate loyal and strong attention, segmentation accuracy ultimately determines the clinical reliability of

Grad-CAM; therefore, improved border delineation (MASD) is crucial for better grading and credibility.

There are still issues with differentiating mild cases: because there are few mild samples and a significant class imbalance, the classifier frequently mistakes moderate with mild. To improve the distinction between these classes, future research will employ imbalance-handling methods like Synthetic Minority Oversampling Technique (SMOTE) or Adaptive Synthetic Sampling Approach for Imbalanced Learning (ADASYN) on the tabular features.

Although the BONBID dataset offers a helpful benchmark, its single-center design and class imbalance may restrict its generalizability. The solution needs to be validated on bigger multicenter datasets prior to deployment. In order to improve robustness and facilitate clinical application as a clinician-in-the-loop tool for interpretable and accurate early HIE assessment, future research will concentrate on such validation.

Ethical Statement

This study is a secondary analysis of the BONBID-HIE, a publicly available dataset consisting of fully de-identified neonatal MRI scans, expert-annotated lesion masks, and associated clinical metadata. The dataset was originally collected with appropriate institutional ethics approval and informed parental consent by the data providers. Since no new data were collected and all data were anonymized prior to access, additional ethical approval was not required for this study.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in the Zenodo repository at https://zenodo.org/records/10602767.

Author Contribution Statement

Athira Chandran: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Lekshmi Chandrika Reghunath:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration.

References

[1] Tan, S., Mageziyi, S., Long, L., Dilimulati, N., Hui, Z., & Rejiafu, N. (2025). Analysis of global, regional, and national burdens of neonatal encephalopathy from 1990 to 2021: Insights from the Global Burden of Disease Study 2021. *Frontiers in Public Health*, *13*, 1627448. https://doi.org/10.3389/fpubh.2025.1627448

[2] Murphy, K., van der Aa, N. E., Negro, S., Groenendaal, F., de Vries, L. S., Viergever, M. A., . . . , & Išgum, I. (2017). Automatic quantification of ischemic injury on diffusion-weighted MRI of neonatal hypoxic ischemic encephalopathy. *NeuroImage: Clinical*, *14*, 222–232. https://doi.org/10.1016/j.nicl.2017.01.005

[3] Bao, R., Song, Y., Bates, S. V., Weiss, R. J., Foster, A. N., Jaimes, C., . . . , & Ou, Y. (2025). BOston neonatal brain injury data for hypoxic ischemic encephalopathy (BONBID-HIE): I. MRI and lesion labeling. *Scientific Data*, *12*(1), 53. https://doi.org/10.1038/s41597-024-03986-7

[4] Rao, A., Shukla, A., Bhargava, J., Ou, Y., & Bao, R. (2026). Spatial prior-guided boundary and region-aware 2D lesion segmentation in neonatal hypoxic ischemic encephalopathy. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025: 28th International Conference*, 520–530. https://doi.org/10.1007/978-3-032-04965-0_49

[5] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

[6] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015: 18th International Conference*, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

[7] Zheng, Q., Martin-Saavedra, J. S., Saade-Lemus, S., Vossough, A., Zuccoli, G., Gonçalves, F. G., . . . , & Hwang, M. (2020). Cerebral pulsed arterial spin labeling perfusion weighted imaging predicts language and motor outcomes in neonatal hypoxic-ischemic encephalopathy. *Frontiers in Pediatrics*, *8*, 576489. https://doi.org/10.3389/fped.2020.576489

[8] Raurale, S. A., Boylan, G. B., Mathieson, S. R., Marnane, W. P., Lightbody, G., & O'Toole, J. M. (2021). Grading hypoxic-ischemic encephalopathy in neonatal EEG with convolutional neural networks and quadratic time–frequency distributions. *Journal of Neural Engineering*, *18*(4), 046007. https://doi.org/10.1088/1741-2552/abe8ae

[9] Tian, T., Gan, T., Chen, J., Lu, J., Zhang, G., Zhou, Y., . . . , & Zhu, W. (2023). Graphic intelligent diagnosis of hypoxic-ischemic encephalopathy using MRI-based deep learning model. *Neonatology*, *120*(4), 441–449. https://doi.org/10.1159/000530352

[10] Bao, R., Weiss, R. J., Bates, S. V., Song, Y., He, S., Li, J., . . . , & Ou, Y. (2025). PARADISE: Personalized and regional adaptation for HIE disease identification and segmentation. *Medical Image Analysis*, *102*, 103419. https://doi.org/10.1016/j.media.2024.103419

[11] Kumar, P. M., Korani, W., Shahwar, T., & C, G. (2025). Detection of sub-acute brain injury and hypoxic-ischemic encephalopathy using I2C2-WGO and CO-GW-RNN. *Current Medical Imaging*, *21*, e15734056352573. https://doi.org/10.2174/0115734056352573241118122026

[12] Usman, A., Haseeb, A., & Syed, T. (2025). *Multimodal HIE lesion segmentation in neonates: A comparative study of loss functions*. *arXiv Preprint: 2502.09148*

[13] Ashoori, M., O'Toole, J. M., Garvey, A. A., O'Halloran, K. D., Walsh, B., Moore, M., . . . , & McDonald, F. B. (2024). Machine learning models of cerebral oxygenation (rcSO2) for brain injury detection in neonates with hypoxic-ischaemic encephalopathy. *The Journal of Physiology*, *602*(22), 6347–6360. https://doi.org/10.1113/JP287001

[14] Lewis, J. D., Miran, A. A., Stoopler, M., Branson, H. M., Danguecan, A., Raghu, K., . . . , & Kalish, B. T. (2025). Automated neuroprognostication via machine learning in neonates

with hypoxic-ischemic encephalopathy. *Annals of Neurology*, *97*(4), 791–802. https://doi.org/10.1002/ana.27154

[15] Lew, C. O., Calabrese, E., Chen, J. V., Tang, F., Chaudhari, G., Lee, A., . . . , & Li, Y. (2024). Artificial intelligence outcome prediction in neonates with encephalopathy (AI-OPiNE). *Radiology: Artificial Intelligence*, *6*(5), e240076. https://doi.org/10.1148/ryai.240076

[16] Prakash, J. A., Ravi, V., Sowmya, V., & Soman, K. P. (2023). Stacked ensemble learning based on deep convolutional neural networks for pediatric pneumonia diagnosis using chest X-ray images. *Neural Computing and Applications*, *35*(11), 8259–8279. https://doi.org/10.1007/s00521-022-08099-z

[17] Nair, S. S., Sowmya, V., Gopalakrishnan, E. A., Gupta, S., & Khanna, N. (2024). Effect of segmentation on skin lesion classification using multi-scale convolution neural networks. In *2024 IEEE 21st India Council International Conference*, 1–6. https://doi.org/10.1109/INDICON63790.2024.10958450

[18] Ganeshkumar, M., Ravi, V., Sowmya, V., Gopalakrishnan, E. A., Soman, K. P., & Chakraborty, C. (2022). Identification of intracranial haemorrhage (ICH) using ResNet with data augmentation using CycleGAN and ICH segmentation using SegAN. *Multimedia Tools and Applications*, *81*(25), 36257–36273. https://doi.org/10.1007/s11042-021-11478-8

[19] Madan, Y., Veetil, I. K., Sowmya, V., Gopalakrishnan, E. A., & Soman, K. P. (2022). Synthetic data augmentation of MRI using generative variational autoencoder for Parkinson's disease detection. In *Evolution In Computational Intelligence: Proceedings of the 9th International Conference on Frontiers in Intelligent Computing: Theory and Applications*, 171–178. https://doi.org/10.1007/978-981-16-6616-2_16

[20] Chou, J.-L., Huang, Y.-L., Hsieh, C.-Y., Huang, J.-X., & Wei, C.-S. (2022). Accelerating brain research using explainable artificial intelligence. In *2022 IEEE International Conference on Multimedia and Expo Workshops*, 1–1. https://doi.org/10.1109/ICMEW56448.2022.9859322

[21] Nyrup, R., & Robinson, D. (2022). Explanatory pragmatism: A context-sensitive framework for explainable medical AI. *Ethics and Information Technology*, *24*(1), 13. https://doi.org/10.1007/s10676-022-09632-3

[22] Yang, G., Ye, Q., & Xia, J. (2022). Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Information Fusion*, *77*, 29–52. https://doi.org/10.1016/j.inffus.2021.07.016

[23] Sakshi, & Verma, G. (2025). Explainable artificial intelligence in healthcare: Transparency and trustworthiness. In R. Agrawal, P. S. Rathore, G. G. Devarajan, & R. R. Divivedi (Eds.), *Artificial intelligence and cybersecurity in healthcare* (pp. 243–272). Wiley. https://doi.org/10.1002/9781394229826.ch10

[24] Widjaja, G., Wahid, N., Raha, S., Pande, S. D., & Manerkar, S. G. V. (2024). An exhaustive exploration of explainable AI-driven applications in healthcare, enhancing diagnostic accuracy, treatment efficacy, and patient trust. In A. Khamparia & D. Gupta (Eds.), *Explainable artificial intelligence for biomedical and healthcare applications* (pp. 230–248). CRC Press.

[25] Amin, A., Hasan, K., & Hossain, M. S. (2025). XAI-empowered MRI analysis for consumer electronic health. *IEEE Transactions on Consumer Electronics*, *71*(1), 1423–1431. https://doi.org/10.1109/TCE.2024.3443203

[26] Charaabi, H., Mzoughi, H., Hamdi, R., & Njah, M. (2023). EXplainable Artificial Intelligence (XAI) for MRI brain tumor diagnosis: A survey. In *2023 International Conference on Cyberworlds*, 171–178. https://doi.org/10.1109/CW58918.2023.00033

[27] Madapatha, S., & Fernando, P. (2024). A systematic literature review of XAI-based approaches on brain disease detection using brain MRI images. In *2024 4th International Conference on Advanced Research in Computing*, 19–24. https://doi.org/10.1109/ICARC61713.2024.10499752

[28] Nicolaou, A., Pantzaris, M., Loizou, C. P., Kakas, A., & Pattichis, C. S. (2023). An Explainable AI model in the assessment of Multiple Sclerosis using clinical data and Brain MRI lesion texture features. In *2023 IEEE EMBS International Conference on Biomedical and Health Informatics*, 1–4. https://doi.org/10.1109/BHI58575.2023.10313379

[29] Fontes, M., De Almeida, J. D. S., & Cunha, A. (2024). Application of example-based explainable artificial intelligence (XAI) for analysis and interpretation of medical imaging: A systematic review. *IEEE Access*, *12*, 26419–26427. https://doi.org/10.1109/ACCESS.2024.3367606

[30] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 618–626. https://doi.org/10.1109/ICCV.2017.74

[31] Petsiuk, V., Das, A., & Saenko, K. (2018). RISE: Randomized input sampling for explanation of black-box models. *arXiv Preprint: 1806.07421*

How to Cite: Chandran, A., & Reghunath, L. C. (2026). Explainable AI Framework for Automated Lesion Segmentation and Severity Assessment in Neonatal Hypoxic-Ischemic Encephalopathy. *Journal of Computational and Cognitive Engineering*. https://doi.org/10.47852/bonviewJCCE62027647