

Evidence From Within: Explainable YOLOv8 for Reproducible Brain Tumour Localisation

Connor Mc Cabe¹[0000–0000–0000–0000], Dost
Muhammad^{2,3}[0000–0001–9186–6392]*, and Malika
Bendeche³[0000–0003–0069–1860]

¹ School of Computer Science, University of Galway, Galway, Ireland

² Research Ireland Centre for Research Training in Artificial Intelligence, Ireland

³ ADAPT Research Centre, School of Computer Science, University of Galway,
Galway, Ireland

Abstract. Accurate brain tumour segmentation is insufficient for clinical decision support if the model cannot show whether its prediction is based on tumour-relevant evidence. Post-hoc methods such as Grad-CAM and Score-CAM are widely used in medical imaging, but their heatmaps are not optimised for tumour-mask alignment and may be diffuse or spatially imprecise. This paper proposes an explainable YOLOv8-based framework for efficient whole tumour localisation on multi-modal BraTS2020 MRI slices. The framework combines a reproducible H5-to-YOLO conversion pipeline, YOLOv8n-seg for single-class tumour localisation, and a native evidence head attached to internal YOLOv8 feature representations. Unlike post-hoc CAM methods, the evidence head is trained to generate tumour-relevant spatial evidence directly from the model feature space.

The native evidence maps are compared with Grad-CAM and Score-CAM on tumour-positive test samples using Dice, IoU, energy inside the tumour mask, pointing-game accuracy, and Hausdorff distance. The proposed method achieves higher Dice and IoU, stronger pointing-game performance, and lower boundary disagreement than both CAM baselines. Its maximum activation falls inside the tumour mask in 84% of test cases, compared with 38% for Grad-CAM and Score-CAM. Qualitative results further show compact tumour-centred evidence maps, while the CAM baselines are less precise or more diffuse. These findings show that explanation quality in medical segmentation should be evaluated as a measurable spatial property. The proposed framework provides an efficient and reproducible explanation-by-design approach for interpretable YOLO-style brain tumour localisation.

Keywords: Applied intelligence · Explainable artificial intelligence · Medical informatics · Reproducible AI · Brain tumour localisation · YOLOv8 segmentation · Native evidence learning · BraTS2020 · Magnetic resonance imaging · Heatmap alignment

* Corresponding author: d.muhammad1@universityofgalway.ie

1 Introduction

Brain tumours require accurate localisation, delineation, and follow-up assessment for diagnosis, treatment planning, surgical guidance, radiotherapy planning, and monitoring. Magnetic resonance imaging (MRI) is central to glioma assessment because T1, contrast-enhanced T1, T2, and FLAIR sequences provide complementary structural and pathological information. In clinical practice, tumour burden is defined not only by lesion presence, but also by spatial extent, surrounding tissue relationships, and multimodal appearance [12].

The Brain Tumour Segmentation (BraTS) benchmark has standardised glioma segmentation research through multi-institutional MRI data, expert annotations, and common evaluation protocols [19]. Although 3D volumetric networks dominate BraTS-style benchmarks, efficient 2D localisation models remain useful for tumour-bearing slice identification, visual screening, lightweight deployment, and explanation-oriented analysis [19, 5, 11].

Deep learning has improved medical image segmentation, but clinical adoption remains limited by uncertainty about whether predictions are based on clinically meaningful evidence [14, 13, 14]. A segmentation mask shows what the model predicted, but not necessarily why. This is especially important in brain tumour MRI, where acquisition variability, intensity normalisation artefacts, and heterogeneous tumour appearance can affect model behaviour. While XAI has gained attention in medical imaging, much early work focused on classification rather than dense prediction. Semantic segmentation requires spatially structured explanations at pixel or region level, making explanation quality more demanding to evaluate [4, 17].

Class activation mapping methods such as Grad-CAM and Score-CAM are widely used to visualise regions influencing convolutional neural networks [16, 20]. However, these post-hoc heatmaps are not directly optimised during learning, can be coarse or threshold-sensitive, and may not align with expert annotations. For brain tumour MRI, clinically meaningful explanations should be localised around tumour tissue and quantitatively comparable with tumour masks.

In this work, we propose an explainable YOLOv8-based framework for brain tumour localisation and segmentation using multi-modal BraTS2020 MRI slices. YOLOv8 provides an efficient single-stage architecture with segmentation support, making it suitable for fast localisation and mask prediction [18]. Although recent studies have explored YOLOv8-style models for brain tumour detection and localisation [7], explanation remains external or underexplored in most YOLO-based medical imaging studies.

The proposed framework converts BraTS2020 H5 MRI slices into a YOLO-compatible segmentation format, trains YOLOv8n-seg for whole tumour localisation, and attaches a native evidence head to internal YOLOv8 feature representations. Unlike Grad-CAM and Score-CAM, which are applied after prediction, the evidence head is supervised using tumour masks and produces a feature-coupled spatial evidence map. We hypothesise that this design can generate heatmaps that are more spatially aligned with tumour annotations while preserving YOLOv8 localisation efficiency.

The main contributions of this work are as follows:

1. We propose an explainable YOLOv8-based framework for efficient brain tumour localisation and segmentation from multi-modal BraTS2020 MRI slices.
2. We introduce a native evidence head that generates model-internal heatmaps from YOLOv8 feature representations, enabling explanation-by-design rather than relying only on post-hoc visualisation.
3. We develop an H5-to-YOLO conversion and training pipeline for BraTS2020, allowing tumour masks to support both YOLOv8 segmentation and evidence supervision.
4. We compare native evidence maps with Grad-CAM and Score-CAM using qualitative visualisation and quantitative tumour-mask alignment metrics.
5. We evaluate explanation quality as a measurable spatial property using Dice, IoU, energy inside the tumour mask, pointing-game accuracy, and boundary disagreement.

2 Related Work

Brain tumour segmentation, YOLO-based medical localisation, post-hoc visual explanation, and native explainability form the main research context of this work. Existing BraTS studies have achieved strong volumetric segmentation performance using multimodal MRI and advanced encoder-decoder or hybrid architectures [9, 19, 8, 5]. However, these methods are mainly evaluated through predictive metrics and do not directly assess whether the model evidence is clinically meaningful. YOLO-based approaches offer efficient lesion localisation and have recently been adapted to medical imaging and brain tumour detection [15, 2, 7]. In parallel, Grad-CAM and Score-CAM remain common post-hoc explanation baselines [16, 20], but their heatmaps are not explicitly optimised for tumour-mask alignment. Recent work on self-explainable and biomedical XAI argues for explanation mechanisms that are integrated into model design rather than applied only after prediction [6, 3, 10]. Table 1 summarises how the proposed framework is positioned with respect to these research directions.

3 Methodology

Let $\mathcal{D} = \{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^N$ denote the multimodal BraTS2020 MRI dataset, where $\mathbf{X}_i$ is the MRI volume of subject i and $\mathbf{Y}_i$ is the corresponding voxel-level tumour annotation. Since YOLOv8 segmentation is formulated for 2D image inputs, each 3D volume is decomposed into axial slices. For slice s , FLAIR, T1ce, and T2 are stacked as a pseudo-RGB input:

$$\mathbf{x}_{i,s} = [\mathbf{I}_{i,s}^{\text{FLAIR}}, \mathbf{I}_{i,s}^{\text{T1ce}}, \mathbf{I}_{i,s}^{\text{T2}}] \in R^{H \times W \times 3}. \quad (1)$$

The original BraTS labels $\mathbf{Y}_{i,s} \in \{0, 1, 2, 4\}^{H \times W}$ are merged into a binary whole tumour mask:

$$\mathbf{M}_{i,s}(u, v) = 1 [\mathbf{Y}_{i,s}(u, v) \in \{1, 2, 4\}], \quad \mathbf{M}_{i,s} \in \{0, 1\}^{H \times W}. \quad (2)$$

Table 1. Compact summary of related work and positioning of the proposed framework.

Research area	Main focus	Remaining gap	Position of this work
BraTS tumour segmentation	Multimodal volumetric glioma segmentation using MRI sequences such as T1, T1ce, T2, and FLAIR [9, 19, 5].	Strong segmentation performance, but limited assessment of whether internal evidence is spatially meaningful.	Uses BraTS2020 for efficient tumour localisation and explicit explanation-alignment evaluation.
YOLO-based medical imaging	Fast lesion detection, localisation, and segmentation using single-stage architectures [15, 2, 7].	Most studies focus on precision, recall, mAP, or mask quality, with limited native explanation.	Extends YOLOv8n-seg with a native evidence head for tumour-centred heatmap generation.
Post-hoc CAM explanations	Visual explanation using Grad-CAM and Score-CAM after model training [16, 20].	Heatmaps can be coarse, layer-dependent, and weakly aligned with expert tumour masks.	Compares Grad-CAM and Score-CAM with native evidence maps using the same spatial metrics.
Native explainability	Explanation mechanisms integrated into the architecture or training process [6, 3, 1].	Faithfulness and spatial alignment require careful evaluation, especially in medical segmentation.	Trains a feature-coupled evidence head and evaluates explanation quality quantitatively.

The task is therefore single-class whole tumour localisation and segmentation. The model predicts a bounding region, segmentation mask, confidence score, and native evidence map:

$$f_{\theta}(\mathbf{x}_{i,s}) = \left(\hat{\mathbf{B}}_{i,s}, \hat{\mathbf{S}}_{i,s}, \hat{c}_{i,s}, \mathbf{E}_{i,s} \right). \quad (3)$$

The H5 data are converted into YOLOv8 segmentation format through modality normalisation, mask binarisation, contour extraction, and polygon-label generation. Each modality image $\mathbf{I}$ is independently normalised using percentile clipping:

$$\tilde{\mathbf{I}}(u, v) = \frac{\text{clip}(\mathbf{I}(u, v), p_1(\mathbf{I}), p_{99}(\mathbf{I})) - p_1(\mathbf{I})}{p_{99}(\mathbf{I}) - p_1(\mathbf{I}) + \epsilon}, \quad (4)$$

where p_1 and p_{99} are the first and ninety-ninth percentiles. Each binary tumour mask is converted into one or more external contours. For connected component

k , the YOLO polygon is:

$$\mathbf{P}_{i,s}^k = \left[\frac{x_1^k}{W}, \frac{y_1^k}{H}, \dots, \frac{x_{n_k}^k}{W}, \frac{y_{n_k}^k}{H} \right], \quad (5)$$

where (x_j^k, y_j^k) is the j -th polygon vertex. The corresponding YOLO label is $\mathbf{z}_{i,s}^k = [0, \mathbf{P}_{i,s}^k]$, where class index 0 denotes whole tumour. For slices with multiple disconnected tumour components, each valid component is written as a separate YOLO polygon with the same class label. Components smaller than 20 pixels are removed to suppress very small artefacts, and empty slices are retained in a controlled proportion to preserve background representation.

The segmentation model is based on YOLOv8n-seg. Given an input slice, the YOLOv8 backbone and neck extract multi-scale feature maps:

$$\left\{ \mathbf{F}_{i,s}^{(l)} \right\}_{l=1}^L = g_\theta(\mathbf{x}_{i,s}), \quad (6)$$

where g_θ denotes the feature extractor. The segmentation head predicts bounding boxes, class confidence scores, mask coefficients, and prototype masks:

$$\left(\hat{\mathbf{B}}_{i,s}, \hat{\mathbf{C}}_{i,s}, \hat{\mathbf{A}}_{i,s}, \hat{\mathbf{Q}}_{i,s} \right) = h_\theta \left(\left\{ \mathbf{F}_{i,s}^{(l)} \right\}_{l=1}^L \right). \quad (7)$$

The final mask is reconstructed by combining mask coefficients with prototype masks:

$$\hat{\mathbf{S}}_{i,s} = \sigma \left(\sum_{r=1}^R \hat{a}_{i,s}^r \hat{\mathbf{Q}}_{i,s}^r \right), \quad (8)$$

where $\hat{a}_{i,s}^r$ is the r -th mask coefficient, $\hat{\mathbf{Q}}_{i,s}^r$ is the r -th prototype, and $\sigma(\cdot)$ is the sigmoid function. YOLOv8 is trained using:

$$\mathcal{L}_{\text{YOLO}} = \lambda_{\text{box}} \mathcal{L}_{\text{box}} + \lambda_{\text{cls}} \mathcal{L}_{\text{cls}} + \lambda_{\text{dfl}} \mathcal{L}_{\text{dfl}} + \lambda_{\text{seg}} \mathcal{L}_{\text{seg}}. \quad (9)$$

To provide model-internal explainability, a native evidence head ϕ_ψ is attached to a selected YOLOv8 feature map $\mathbf{F}_{i,s}^{(\ell)}$ and maps it into a single-channel heatmap:

$$\mathbf{E}_{i,s} = \text{Up} \left(\sigma \left(\phi_\psi \left(\mathbf{F}_{i,s}^{(\ell)} \right) \right) \right), \quad \mathbf{E}_{i,s} \in [0, 1]^{H \times W}. \quad (10)$$

Here, ϕ_ψ is a lightweight convolutional module and $\text{Up}(\cdot)$ upsamples the heatmap to the input resolution. Unlike Grad-CAM and Score-CAM, the proposed evidence map is not produced using backward gradients or repeated masked forward passes. It is trained to map internal YOLOv8 features to tumour-relevant spatial evidence.

The evidence head is supervised using the whole tumour mask. Its objective combines alignment, sparsity, and smoothness:

$$\mathcal{L}_{\text{EH}} = \mathcal{L}_{\text{align}} + \lambda_{\text{sparse}} \mathcal{L}_{\text{sparse}} + \lambda_{\text{TV}} \mathcal{L}_{\text{TV}}, \quad (11)$$

where

$$\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{BCE}}(\mathbf{E}_{i,s}, \mathbf{M}_{i,s}) + \left(1 - \frac{2 \sum_{u,v} \mathbf{E}_{i,s}(u,v) \mathbf{M}_{i,s}(u,v) + \epsilon}{\sum_{u,v} \mathbf{E}_{i,s}(u,v) + \sum_{u,v} \mathbf{M}_{i,s}(u,v) + \epsilon} \right). \quad (12)$$

The sparsity term is $\mathcal{L}_{\text{sparse}} = \frac{1}{HW} \sum_{u,v} \mathbf{E}_{i,s}(u,v)$, and the total variation term encourages spatial coherence between neighbouring pixels. In this implementation, YOLOv8 is first trained using Equation 9; the evidence head is then trained on frozen YOLOv8 feature representations using Equation 11, with $\lambda_{\text{sparse}} = 0.01$ and $\lambda_{\text{TV}} = 0.005$.

The proposed evidence head should be interpreted as a feature-coupled spatial evidence module, not as proof of causal faithfulness. Because the evidence head is trained after YOLOv8 using frozen feature representations, it evaluates whether tumour-relevant spatial information is recoverable from the learned feature space. Mask-supervised evidence alignment supports spatial plausibility and annotation consistency, but it does not alone establish that the evidence map is causally responsible for the final YOLOv8 prediction.

For comparison, Grad-CAM and Score-CAM are implemented on the same trained YOLOv8 model. Given the selected feature map $\mathbf{F}^{(\ell)}$ and tumour target score y^c , Grad-CAM computes:

$$\alpha_k^c = \frac{1}{Z} \sum_u \sum_v \frac{\partial y^c}{\partial \mathbf{F}_k^{(\ell)}(u,v)}, \quad \mathbf{G}_{\text{GradCAM}}^c = \text{ReLU} \left(\sum_k \alpha_k^c \mathbf{F}_k^{(\ell)} \right). \quad (13)$$

Score-CAM replaces gradient weighting with score-based activation weighting:

$$\beta_k^c = \frac{\exp(y^c(\mathbf{x}_{i,s} \odot \text{Up}(\mathbf{A}_k)))}{\sum_j \exp(y^c(\mathbf{x}_{i,s} \odot \text{Up}(\mathbf{A}_j)))}, \quad \mathbf{G}_{\text{ScoreCAM}}^c = \text{ReLU} \left(\sum_k \beta_k^c \mathbf{A}_k \right). \quad (14)$$

All heatmaps are resized to the same spatial resolution. Score-CAM uses the top 32 activation channels to reduce computational cost while retaining the most responsive feature maps.

Explanation quality is assessed by comparing each heatmap with the ground-truth tumour mask. For method $m \in \{\text{EH}, \text{GradCAM}, \text{ScoreCAM}\}$, the heatmap $\mathbf{H}_{i,s}^m$ is normalised to $[0, 1]$ and converted into a binary support region:

$$\tilde{\mathbf{H}}_{i,s}^m = \frac{\mathbf{H}_{i,s}^m - \min(\mathbf{H}_{i,s}^m)}{\max(\mathbf{H}_{i,s}^m) - \min(\mathbf{H}_{i,s}^m) + \epsilon}, \quad \mathbf{R}_{i,s}^m(u,v) = 1 \left[\tilde{\mathbf{H}}_{i,s}^m(u,v) \geq \tau_{80}(\tilde{\mathbf{H}}_{i,s}^m) \right]. \quad (15)$$

The evaluation metrics are summarised in Table 2. These metrics quantify overlap, intensity concentration, peak localisation, and boundary agreement between each explanation and the tumour mask.

For a test set $\mathcal{T}$, each method-level score is reported as:

$$\bar{\mathcal{Q}}^m = \frac{1}{|\mathcal{T}|} \sum_{(i,s) \in \mathcal{T}} \mathcal{Q}_{i,s}^m, \quad (16)$$

Table 2. Explanation-alignment metrics used to evaluate heatmaps against tumour masks.

Metric	Formula
Dice	$\text{Dice}_{\text{exp}}^m = \frac{2 \mathbf{R}_{i,s}^m \cap \mathbf{M}_{i,s} + \epsilon}{ \mathbf{R}_{i,s}^m + \mathbf{M}_{i,s} + \epsilon}$
IoU	$\text{IoU}_{\text{exp}}^m = \frac{ \mathbf{R}_{i,s}^m \cap \mathbf{M}_{i,s} + \epsilon}{ \mathbf{R}_{i,s}^m \cup \mathbf{M}_{i,s} + \epsilon}$
Energy inside mask	$\text{EIM}^m = \frac{\sum_{u,v} \tilde{\mathbf{H}}_{i,s}^m(u,v) \mathbf{M}_{i,s}(u,v) + \epsilon}{\sum_{u,v} \tilde{\mathbf{H}}_{i,s}^m(u,v) + \epsilon}$
Pointing game	$\text{PG}^m = 1 \left[\mathbf{M}_{i,s} \left(\arg \max_{(u,v)} \tilde{\mathbf{H}}_{i,s}^m(u,v) \right) = 1 \right]$
Hausdorff distance	$\text{HD}^m = \max \left\{ \sup_{\mathbf{a} \in \partial \mathbf{R}_{i,s}^m} \inf_{\mathbf{b} \in \partial \mathbf{M}_{i,s}} \ \mathbf{a} - \mathbf{b}\ _2, \sup_{\mathbf{b} \in \partial \mathbf{M}_{i,s}} \inf_{\mathbf{a} \in \partial \mathbf{R}_{i,s}^m} \ \mathbf{b} - \mathbf{a}\ _2 \right\}$

where $\mathcal{Q}_{i,s}^m$ denotes any metric from Table 2. This enables the native evidence head, Grad-CAM, and Score-CAM to be compared under identical spatial evaluation conditions.

4 Experimental Setup

Experiments were conducted on the BraTS2020 training data in H5 format, where each file contained a 2D multimodal MRI slice and its tumour mask. The metadata files, `meta_data`, `name_mapping`, and `survival_info`, were used only for inspection and consistency checking. Each slice was converted into a three-channel pseudo-RGB image using FLAIR, contrast-enhanced T1, and T2. All non-background labels were merged into a binary whole tumour mask and converted into YOLOv8 polygon annotations.

To prevent data leakage, splitting was performed strictly at subject level before slice conversion. Patient identifiers were inferred from H5 filenames using a deterministic rule. In total, 369 subjects were identified and split with random seed 42 into 258 training, 55 validation, and 56 testing subjects, with no subject overlap across subsets. After conversion and controlled retention of background-only samples, the YOLO-formatted dataset contained 18,300 training slices, 4,028 validation slices, and 4,232 testing slices. Tumour-positive slices accounted for 15,872, 3,493, and 3,671 samples in the three subsets, respectively.

The framework was implemented in Python using PyTorch and Ultralytics YOLOv8. H5 processing used `h5py`, while image conversion, mask processing, contour extraction, and polygon generation used OpenCV. The base model was YOLOv8n-seg, initialised with pretrained weights and fine-tuned for single-class whole tumour segmentation at 640×640 resolution. After YOLOv8 training, a lightweight native evidence head was attached to internal feature layer index -2 and trained on frozen YOLOv8 feature representations using binary whole tumour masks. The evidence head was trained for 25 epochs with learning rate

1×10^{-3} , using BCE, Dice, sparsity, and total variation terms with $\lambda_{\text{sparse}} = 0.01$ and $\lambda_{\text{TV}} = 0.005$.

Grad-CAM and Score-CAM were implemented as post-hoc baselines on the same trained YOLOv8 model. Score-CAM used the top 32 activation channels. For all explanation methods, heatmaps were resized, normalised to $[0, 1]$, and binarised using the 80th percentile threshold. Explanation evaluation was performed on the same 100 tumour-positive slices selected from the patient-disjoint test set using the fixed experimental seed, enabling paired comparison across all methods. ““

5 Results

5.1 Localisation and segmentation performance

The H5-to-YOLO conversion pipeline transformed the BraTS2020 H5 data into YOLOv8 segmentation format without failed samples. In total, 26,560 slices were converted, including 18,300 training slices, 4,028 validation slices, and 4,232 test slices. The tumour-positive counts were 15,872, 3,493, and 3,671, respectively, with the remaining slices retained as background-only examples.

YOLOv8n-seg showed stable optimisation for single-class whole tumour localisation. As shown in Table 4, all training losses decreased, with the largest reduction observed for classification loss. This suggests that the model learned effective foreground-background discrimination for tumour-bearing slices.

Table 5 reports validation performance. The model achieved similar precision for box and mask outputs, indicating reliable tumour-positive predictions when a region was detected. However, recall remained moderate, which suggests that small, subtle, or low-contrast lesions remain challenging for isolated 2D YOLO-style localisation.

After YOLOv8n-seg training, the native evidence head was trained on frozen YOLOv8 feature representations. As shown in Table 6, the training loss decreased by 32.11% over 25 epochs, indicating that the evidence head learned tumour-relevant spatial evidence from the trained feature space.

5.2 Explanation alignment results

The native evidence head was compared with Grad-CAM and Score-CAM on 100 tumour-positive test slices selected from the patient-disjoint test set using the fixed experimental seed. The same slices were used for all methods to enable paired comparison. Table 7 reports mean scores with 95% confidence intervals.

The native evidence head achieved the strongest overall spatial agreement with tumour masks. It improved Dice by 94.03% over Grad-CAM and 164.65% over Score-CAM, and achieved the best IoU, pointing-game accuracy, and Hausdorff distance. Its maximum activation fell inside the tumour mask in 84% of test samples, compared with 38% for both CAM baselines. It also reduced boundary distance by 28.84% relative to Grad-CAM and 28.51% relative to Score-CAM.

Paired Wilcoxon signed-rank tests showed significant improvements for Dice, IoU, pointing-game accuracy, and HD relative to both baselines, with $p < 0.001$

Table 3. Compact experimental configuration.

Component	Setting
Dataset	BraTS2020 H5 MRI slices with image and mask arrays.
Subject split	Strictly patient-disjoint; 369 subjects split into 258 training, 55 validation, and 56 testing subjects.
Random seed	42.
Input channels	FLAIR, T1ce, and T2 as pseudo-RGB input.
Task	Single-class whole tumour localisation and segmentation.
Slice-level split	Train: 18,300 slices; validation: 4,028 slices; test: 4,232 slices.
Tumour-positive slices	Train: 15,872; validation: 3,493; test: 3,671.
Background-only slices	Train: 2,428; validation: 535; test: 561.
Base model	YOLOv8n-seg with pretrained initialisation.
Input size	640×640 .
YOLOv8 training	80 epochs, batch size 8, workers 0, SGD, learning rate 0.01, momentum 0.9.
AMP	Disabled due to local CUDA <code>torchvision</code> NMS compatibility.
Evidence layer	YOLOv8 internal feature layer index -2 .
Evidence-head training	25 epochs on frozen YOLOv8 feature maps, learning rate 1×10^{-3} .
Evidence supervision	Binary whole tumour masks with BCE, Dice, sparsity, and total variation terms.
Loss weights	$\lambda_{\text{sparse}} = 0.01$ and $\lambda_{\text{TV}} = 0.005$.
XAI baselines	Grad-CAM and Score-CAM on the same trained YOLOv8 model.
Score-CAM setting	Top 32 activation channels.
Heatmap threshold	80th percentile threshold after normalisation to $[0, 1]$.
Explanation samples	100 tumour-positive test slices selected from the patient-disjoint test set.
Segmentation metrics	Precision, recall, mAP_{50} , $\text{mAP}_{50:95}$, Dice, and IoU.
Explanation metrics	Explanation Dice, explanation IoU, energy inside mask, pointing game, and Hausdorff distance.
Environment	Windows, Jupyter Notebook, Python 3.9.18, PyTorch 2.2.2 + CUDA 11.8, Ultralytics 8.3.51.
Hardware	NVIDIA GeForce RTX 4080 GPU with 16 GB memory.
Runtime design	Native evidence requires one forward pass; Grad-CAM requires forward and backward passes; Score-CAM requires multiple forward passes.

Table 4. YOLOv8n-seg training loss convergence for whole tumour localisation.

Loss term	Initial	Final	Reduction
Box loss	1.6307	0.7673	52.95%
Segmentation loss	3.1459	1.5922	49.39%
Classification loss	2.4903	0.4439	82.18%
Distribution focal loss	1.3860	0.9316	32.79%

Table 5. Validation performance of YOLOv8n-seg for whole tumour localisation and segmentation.

Output	Precision	Recall	mAP ₅₀	mAP _{50:95}
Box	0.8092	0.5377	0.6323	0.3711
Mask	0.8056	0.5353	0.6243	0.3563

in all reported comparisons. The large rank-biserial correlations indicate consistent improvements across paired samples. Grad-CAM achieved the highest energy-inside-mask score; therefore, we do not claim that native evidence outperforms Grad-CAM on this individual metric. Rather, the proposed method provides a stronger balance across overlap, peak localisation, and boundary agreement.

5.3 Qualitative heatmap comparison

Figure 1 shows six representative examples selected from the same evaluated subset. The native evidence head produced compact and tumour-centred heatmaps, while Grad-CAM highlighted broader tumour vicinity and Score-CAM produced more diffuse activation across brain or peripheral regions. These visual patterns are consistent with the higher pointing-game accuracy and lower boundary disagreement of the proposed method.

5.4 Efficiency and runtime analysis

The native evidence head is computationally efficient because it generates heatmaps directly from internal YOLOv8 feature representations using a lightweight convolutional branch. It requires one forward pass and no gradient computation. Grad-CAM requires a forward and backward pass, while Score-CAM requires multiple masked forward passes, making it the most computationally expensive explanation baseline.

The validation run required 652.39 seconds for 4,028 validation slices on the RTX 4080 workstation, corresponding to 10.87 minutes. Table 9 summarises the computational characteristics of the explanation methods.

Overall, the results support the main hypothesis: YOLOv8n-seg provides efficient tumour localisation, and the proposed native evidence head improves spatial explanation alignment while avoiding the repeated forward-pass cost of Score-CAM.

Table 6. Native evidence-head training behaviour.

Metric	Value	Epoch
Initial training loss	0.4933	1
Final training loss	0.3349	25
Best validation loss	0.4149	3
Final validation loss	0.4307	25
Training-loss reduction	32.11%	1–25

Table 7. Explanation alignment on 100 tumour-positive test samples. Values are mean with 95% confidence intervals in brackets. Higher is better for Dice, IoU, energy, and PG; lower is better for HD.

Method	Dice		IoU		Energy		PG		HD	
Grad-CAM	0.1096	[0.0837, 0.1356]	0.0636	[0.0477, 0.0794]	0.7360	[0.6700, 0.8020]	0.38	[0.2844, 0.4756]	477.26	[467.65, 486.87]
Score-CAM	0.0804	[0.0661, 0.0946]	0.0434	[0.0354, 0.0514]	0.0323	[0.0269, 0.0377]	0.38	[0.2844, 0.4756]	475.09	[465.54, 484.63]
Native evidence	0.2127	[0.1814, 0.2440]	0.1284	[0.1074, 0.1494]	0.5582	[0.5073, 0.6091]	0.84	[0.7678, 0.9122]	339.63	[321.70, 357.57]

6 Discussion and Future Work

This study examined whether a YOLOv8-based brain tumour localisation model can be extended with a native evidence head to generate spatially meaningful explanations from internal feature representations. The results support the central hypothesis that feature-coupled evidence learning can produce more tumour-aligned explanations than commonly used post-hoc CAM methods. However, the scope of this work should be interpreted carefully. YOLOv8n-seg is not intended to replace full volumetric 3D segmentation networks. Instead, it provides an efficient 2D localisation framework in which explanation quality can be evaluated as a measurable spatial property.

The localisation and segmentation results show that YOLOv8n-seg achieved reliable precision for both bounding-box localisation and mask prediction, with box precision of 0.8092 and mask precision of 0.8056. The recall values were more moderate, with box recall of 0.5377 and mask recall of 0.5353. This suggests that the model was conservative in detecting tumour-positive regions and may miss small, subtle, low-contrast, or weakly represented lesions in individual 2D slices. From a clinical perspective, this is an important limitation. The proposed framework should therefore be viewed as an explainable localisation aid for highlighting tumour-relevant regions, not as a standalone screening or diagnostic system.

The main finding is that the native evidence head outperformed Grad-CAM and Score-CAM on the key spatial explanation metrics. It achieved the highest explanation Dice, highest explanation IoU, highest pointing-game accuracy, and lowest Hausdorff distance. The pointing-game result is particularly relevant, as the native evidence head placed its maximum activation inside the tumour mask in 84% of test samples, compared with 38% for both Grad-CAM and Score-CAM. This indicates that the strongest evidence point produced by the native branch was more likely to correspond to annotated tumour tissue.

Table 8. Paired improvements of native evidence over post-hoc CAM baselines. Positive values indicate improvement for Dice, IoU, and PG; for HD, positive values indicate reduced boundary distance. RBC denotes rank-biserial correlation from paired Wilcoxon signed-rank tests.

Comparison	Dice Δ		IoU Δ		PG Δ		HD reduction Δ	
Native vs Grad-CAM	0.1031	[0.0875, 0.1186]	0.0649	[0.0541, 0.0756]	0.4600	[0.3579, 0.5621]	137.63	[124.03, 151.23]
RBC		1.000		1.000		0.958		1.000
Native vs Score-CAM	0.1323	[0.1130, 0.1516]	0.0850	[0.0708, 0.0993]	0.4600	[0.3618, 0.5582]	135.46	[120.58, 150.34]
RBC		0.999		0.999		1.000		0.980

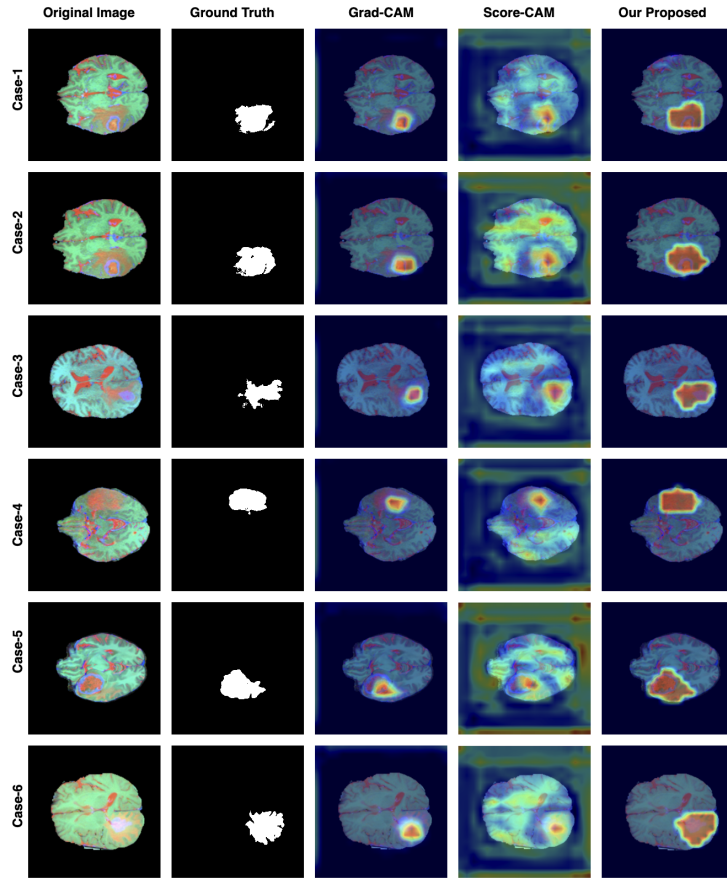


Fig. 1. Qualitative comparison of explanation heatmaps for six tumour-positive cases selected from the evaluated test subset. The native evidence head produces compact tumour-centred heatmaps, while Grad-CAM is less precise and Score-CAM shows more diffuse activation across non-tumour regions.

This improvement can be attributed to the design of the evidence head. Grad-CAM and Score-CAM are applied after model training and are not directly opti-

Table 9. Computational characteristics of the explanation methods.

Method	Forward passes	Backward pass	Relative cost
Native evidence	1	No	Low
Grad-CAM	1	Yes	Moderate
Score-CAM	Multiple	No	High

mised for tumour-mask alignment. Their heatmaps are class-discriminative, but not necessarily spatially precise. In contrast, the proposed evidence head is supervised using tumour masks and learns to transform YOLOv8 internal feature maps into compact tumour-relevant evidence. The qualitative results support this interpretation, showing more tumour-centred activation for the proposed method, while Grad-CAM was less precise and Score-CAM often produced diffuse activation across non-tumour regions.

At the same time, mask-supervised evidence alignment should not be interpreted as proof of causal faithfulness. The evidence head is trained after YOLOv8 using frozen feature representations, which means that it evaluates whether tumour-relevant spatial information is recoverable from the learned feature space. It does not prove that the evidence map is causally responsible for the final YOLOv8 prediction. Therefore, the proposed explanation should be described as feature-coupled and spatially aligned, rather than fully decision-coupled or causally faithful.

The comparison also shows that explanation quality should not be judged using a single metric. Grad-CAM achieved the highest energy-inside-mask score, but this did not translate into better overlap, peak localisation, or boundary agreement. This suggests that heatmap energy alone can overstate explanation quality if the spatial support remains broad or imprecise. For medical segmentation, overlap, pointing-game accuracy, and boundary agreement provide a more complete assessment of whether the highlighted evidence is clinically meaningful.

A further advantage of the native evidence head is computational efficiency. It generates the explanation map through a lightweight convolutional branch using internal YOLOv8 features and requires only a forward pass. Grad-CAM requires a backward pass, while Score-CAM requires multiple forward passes over activation-masked inputs. This makes the proposed evidence head more suitable for rapid and repeatable explanation generation, particularly when many MRI slices must be processed.

Future work should focus on three directions. First, the current 2D slice-based formulation should be extended to 2.5D or 3D modelling to capture volumetric tumour continuity and improve recall for subtle lesions. Second, the evidence head should be integrated directly into the YOLOv8 training objective so that localisation, segmentation, and evidence learning are optimised jointly, making the explanation pathway more decision-coupled. Third, stronger clinical and external validation is needed, including radiologist assessment and testing on independent MRI datasets from different scanners, institutions, and acquisition

protocols. These steps are necessary before the framework can be considered for real-world clinical deployment.

7 Conclusion

This work presented an explainable YOLOv8-based framework for brain tumour localisation using multi-modal BraTS2020 MRI slices. The framework converts H5 MRI data into YOLOv8 segmentation format, trains YOLOv8n-seg for whole tumour localisation, and introduces a native evidence head for generating tumour-relevant heatmaps from internal feature representations. Compared with Grad-CAM and Score-CAM, the native evidence head produced more spatially aligned explanations, with higher Dice and IoU, stronger pointing-game performance, and lower Hausdorff distance. These findings support the main claim that explanation quality in medical segmentation should be evaluated as a measurable spatial property. Overall, the proposed evidence-head design offers an efficient explanation-by-design alternative for interpretable YOLO-style brain tumour localisation.

Acknowledgements

This research was supported by Taighde Éireann, Research Ireland, under grant numbers GOIPG/2025/8471, 18/CRT/6223, 13/RC/2106/ P_2 , and 13/RC/2094/ P_2 . A CC BY public copyright licence has been applied to any Author Accepted Manuscript version arising from this submission.

Code and Data Availability

The datasets used in this study are publicly available. The source code will be made available after publication from the corresponding author upon reasonable request.

References

1. Ahmed, F., et al.: Explainable artificial intelligence in medical imaging. Scientific Reports (2026)
2. Cai, Z., et al.: A systematic review of YOLO-based object detection in medical imaging: Advances, challenges, and future directions. Computers, Materials and Continua (2025)
3. Dagnaw, G.H., Zhu, Y., Maqsood, M.H., Yang, W., Dong, X., Yin, X., Liew, A.W.C.: Explainable artificial intelligence in biomedical image analysis: A comprehensive survey (2025)
4. Gipiškis, R., Tsai, C.W., Kurasova, O.: Explainable AI (XAI) in image segmentation in medicine, industry, and beyond: A survey (2024)
5. Hashmi, S., Lugo, J., Elsayed, A., Saggurthi, D., Elseiagy, M., Nurkamal, A., Walia, J., Maani, F.A., Yaqub, M.: Optimizing brain tumor segmentation with MedNeXt: BraTS 2024 SSA and pediatrics (2024)
6. Hou, J., et al.: Self-explainable AI for medical image analysis (2024)

7. Kang, M., Ting, C.M., Ting, F.F., Phan, R.C.W.: BGF-YOLO: Enhanced YOLOv8 with multiscale attentional feature fusion for brain tumor detection. In: *Proceedings of Medical Image Computing and Computer Assisted Intervention*. vol. LNCS 15008. Springer Nature Switzerland (2024)
8. Maani, F., Hashmi, A.U.R., Aljuboory, M., Saeed, N., Sobirov, I., Yaqub, M.: Advanced tumor segmentation in medical imaging: An ensemble approach for brats 2023 adult glioma and pediatric tumor tasks. In: *International Challenge on Cross-Modality Domain Adaptation for Medical Image Segmentation*, pp. 264–277. Springer (2023)
9. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging* **34**(10), 1993–2024 (2014)
10. Muhammad, D., Bendechache, M.: Unveiling the black box: A systematic review of explainable artificial intelligence in medical image analysis. *Computational and Structural Biotechnology Journal* **24**, 542–560 (2024). <https://doi.org/10.1016/j.csbj.2024.08.005>, <https://www.sciencedirect.com/science/article/pii/S2001037024002642>
11. Muhammad, D., Bendechache, M.: Can ai be faster, accurate, and explainable? spikenet makes it happen. In: *Annual Conference on Medical Image Understanding and Analysis*. pp. 43–57. Springer (2025)
12. Muhammad, D., Bendechache, M.: More than just a heatmap: elevating xai with rigorous evaluation metrics. *Frontiers in Medical Technology* **7**, 1674343 (2025)
13. Muhammad, D., Ozisik, P.A., Keles, A., Ahmed, I., Bendechache, M.: Clinalignnet: A clinician-interactive framework for faithful and aligned explainability in mri. In: *2026 IEEE 23rd International Symposium on Biomedical Imaging (ISBI)*. pp. 1–5. IEEE (2026)
14. Muhammad, D., Salman, M., Bendechache, M.: Cracking the clinical code: A scoping review on mechanistic interpretability in medical report generation. *Computational and Structural Biotechnology Reports* p. 100066 (2025)
15. Palaniappan, D., et al.: YOLO in healthcare: A comprehensive review of object detection applications. *IEEE Access* (2025)
16. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 618–626 (2017). <https://doi.org/10.1109/ICCV.2017.74>
17. Shah, H., Muhammad, D., Almutairi, S.S., Al Moteri, M.A.: Lightweight secure communication framework with explainable artificial intelligence for trustworthy healthcare analytics. *Machine Learning with Applications* p. 100874 (2026)
18. Shivashakarappa, K., Ghimire, N., Wang, L., Dumenyo, K., Pariyar, S., Busch, W., Taheri, A.: Yolo-based high-throughput phenotyping pipeline for soybean nodulation traits in genomic research. *Frontiers in Plant Science* **17**, 1816132 (2026)
19. Correia de Verdier, M., Saluja, R., Gagnon, L., LaBella, D., Baid, U., Tahon, N.H., et al.: The 2024 brain tumor segmentation (BraTS) challenge: Glioma segmentation on post-treatment MRI (2024)
20. Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., Hu, X.: Score-CAM: Score-weighted visual explanations for convolutional neural networks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. pp. 111–119 (2020)