

A multi-level attention CNN-transformer based framework for the detection of brain tumor using regional dual-score explainability

Original

A multi-level attention CNN-transformer based framework for the detection of brain tumor using regional dual-score explainability / Kaprekar, A., Gudigar, A., Raghavendra, U., Inamdar, M.A., Salvi, M., Priyanka, N., Acharya, U.R.. - In: SCIENTIFIC REPORTS. - ISSN 2045-2322. - 16:1(2026). [10.1038/s41598-026-52658-6]

Availability:

This version is available at: 11583/3014047 since: 2026-08-06T10:55:20Z

Publisher:

Springer Nature

Published

DOI:10.1038/s41598-026-52658-6

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)



OPEN

A multi-level attention CNN-transformer based framework for the detection of brain tumor using regional dual-score explainability

Aryaman Kaprekar¹, Anjan Gudigar^{1✉}, U. Raghavendra¹, Mahesh Anil Inamdar¹, Massimo Salvi², Priyanka³ & U. R. Acharya^{4,5}

Deep learning models for brain tumor diagnosis often lack interpretability beyond qualitative visual heatmaps. Clinicians require not only tumor localization but also quantitative assessment of explanation quality and diagnostic relevance, capabilities absent in conventional explainability methods. This paper introduces a classification-explainability framework addressing these limitations. The Multi-Level Hybrid Network (MLHnet) integrates CNN-Transformer components with multi-level attention for robust feature learning. The key novelty lies in the Dual-Score Regional XAI framework, which: (1) identifies tumor regions using hybrid saliency maps combining gradient-based and activation-based information; (2) quantifies geometric tumor characteristics via a Shape Score capturing size, circularity, and saliency concentration; and (3) evaluates explanation faithfulness using regional perturbation analysis, measuring prediction sensitivity to tumor-specific versus non-tumor regions. The framework is evaluated on a publicly available brain Magnetic Resonance Imaging (MRI) dataset containing 7,023 images across four classes. Using five-fold cross-validation, the proposed model achieves an average test accuracy of 99.30% (best fold: 99.47%). Expert radiologist evaluation confirms 87.64% explanation correctness, with the Dual-Score metrics successfully differentiating tumor classes based on morphological and saliency patterns. Overall, the proposed framework offers a lightweight, high-performing, and interpretable solution for reliable brain tumor diagnosis from MRI scans on the evaluated public dataset.

Keywords Brain tumor classification, Magnetic Resonance Imaging, Hybrid deep learning, Explainable artificial intelligence, Medical image analysis

Brain tumors are one of the most life-threatening neurological abnormalities, often leading to critical conditions if not diagnosed and treated at an early stage^{1,2}. Magnetic Resonance Imaging (MRI) is the most widely used non-invasive imaging modality for brain tumor diagnosis. Manual interpretation of MRI scans is time-consuming, cognitively demanding, and prone to human error.

Artificial Intelligence-based medical image analysis has gained significant attention over the past few decades. Traditional Machine Learning (ML) approaches rely on prepared features and classical classifiers, which often struggle to achieve high accuracy³. In the past decade, with the easy availability of sophisticated computational resources, Deep Learning (DL) techniques have displayed better performance by learning hierarchical features directly from image data. Pretrained Convolutional Neural Networks (CNNs) can be fine-tuned and used to perform brain tumor-related tasks such as classification and segmentation accurately⁴. Vision Transformers (ViT)⁵ have been explored in recent years for medical image diagnosis due to their attention-based representations⁶. These advances highlight the potential of DL to serve as reliable support tools for radiologists, with the capability to identify complex patterns, and overcome human bias and error. However, effectively

¹Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal 576104, India. ²Biolab, PolitoBIOMed Lab, Department of Electronics and Telecommunications, Politecnico di Torino, Corso Duca degli Abruzzi 24, Turin 10129, Italy. ³Department of Medical Imaging Technology, Manipal College of Health Professions, Manipal Academy of Higher Education, Udipi, Manipal, India. ⁴School of Mathematics, Physics, and Computing, University of Southern Queensland, Springfield, QLD 4300, Australia. ⁵Centre for Health Research, University of Southern Queensland, Springfield, Australia. ✉email: anjan.gudigar@manipal.edu

capturing tumor characteristics across multiple spatial scales while maintaining robust generalisation remains a challenging problem.

Despite their impressive performance, DL models retain a black-box nature, posing critical challenges in deployment for clinical practice, where transparency, trust and accountability are essential⁷. Explainable Artificial Intelligence (XAI)⁸ addresses these concerns by providing interpretable explanations to model predictions. In image classification, XAI techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM)⁹, Occlusion¹⁰ and Local Interpretable Model-agnostic Explanations (LIME)¹¹ have been widely adopted to highlight salient regions of the input. Using such methods, practitioners can verify if the model predicts based on meaningful tumor regions rather than random correlations. Such expert-AI collaboration can make medical diagnosis faster, more reliable and less laborious while improving accuracy, maintaining trust and transparency. However, most existing approaches provide pixel-level explanations and lack mechanisms to assess whether the highlighted regions correspond to clinically meaningful abnormalities.

In this study, we propose a unified deep learning framework that integrates brain tumor classification with abnormal region localization and explainability. We compare the proposed methodology with various state-of-the-art models and XAI methods. The contribution of this study is as follows:

- Multi-Level Hybrid Network (MLHnet), a CNN-Transformer hybrid that leverages mid-level and final-level feature representations to capture tumor texture and high-level semantic information.
- Refinement of multi-level features using feature enhancing modules like Convolution Block Attention Module (CBAM)¹² and Generalised Mean (GeM) Pooling¹³. These components improve feature discrimination and robustness.
- Dual-Score Regional XAI, which uses hybrid saliency to localize abnormal tumor regions and evaluates explanation faithfulness, through a perturbation-based test.

The remainder of the paper is organized into six sections. Section 2 lists and reviews current research and state-of-the-art literature on brain tumor classification and XAI. Section 3 explains the architecture and process of the proposed methodology. Section 4 presents the experimentation setup, results and the ablation study. Section 5 is the discussion where we explain and analyse the results, compare the proposed method with current state-of-the-art studies and explore the limitations and future directions. Section 6 summarises and concludes the proposed research.

Literature review

Brain tumor classification from MRI scans has seen rapid development with the advent of deep learning, particularly convolutional neural networks (CNNs). Early CNN-based approaches focused on local regions, textures and edges, with studies showing that tumor-focused data augmentation and deeper architectures with small convolutional kernels improved classification performance¹⁴. As CNNs became dominant, transfer learning gained popularity such as¹⁵ where Deepak & Ameer fine-tuned pretrained CNNs with standard classifiers and reported strong multi-class performance for that time. Afshar et al.¹⁶ proposed a modified CapsNet which used coarse tumor boundary information as an input to reduce sensitivity to background and improve brain tumor classification performance. Beyond CNNs, DDFC¹⁷ proposes an attention-integrated architecture based on Gated Recurrent Units (GRUs), a lightweight recurrent neural network variant designed to model sequential dependencies through update and reset gates, in order to enhance feature selectivity. Brain Tumor Detection Network (BTDN)¹⁸ proposes a secure data transmission method for MRI-based diagnosis while achieving highly accurate classification performance. Despite their strong performance, CNN-based models primarily learn local texture, edge, and morphological patterns, and may therefore have limited ability to capture long-range contextual dependencies that can also be important for distinguishing tumor subtypes.

Dosovitskiy et al.⁵ introduced vision transformers, demonstrating that a patch-based self-attention can be effective for image recognition. In medical imaging, a comparative study¹⁹ evaluated multiple ViT variants for brain tumor image classification. It reported strong results demonstrating that ViTs are effective backbones when pretrained weights are available. Hierarchical transformer designs such as Swin Transformer²⁰ restrict attention to shifted local windows while still allowing cross window interactions, also introducing relative position bias. Several segmentation-assisted pipelines have been explored to localize tumor regions prior to classification, including transformer-based segmentation models²¹, object-detection-driven approaches²², and explainability-oriented segmentation frameworks such as Neuro-XAI²³. While Transformer-based methods improve global context modelling through self-attention, they typically involve higher computational complexity and often benefit from large-scale pretraining, which may limit their practicality in data-constrained or resource-constrained clinical settings.

Hybrid architectures combine CNNs' local features with Transformers' global context. For example, CAFNet²⁴ integrates a CNN and a ViT via a Cross-Attention Fusion (CAF) module. Napa et al.²⁵ propose BRAIN-META, an ensemble of ten hybrid CNN-ViT models. Each base model fuses a pretrained CNN with a ViT block then stacks the softmax outputs into meta-features, and trains an XGBoost meta-learner. Swaminathan et al.²⁶ present a DenseNet and Custom-CNN hybrid, a DenseNet121 backbone concatenated with a tailor-made CNN head for binary tumor detection. EFFResNet-ViT²⁷ is a fusion approach that combines EfficientNetB0 and Resnet50 to extract features, followed by a ViT module, which yields a high accuracy of 99.31% in multiclass brain tumor classification. However, many existing hybrid methods rely on single-level or late-stage feature aggregation and do not explicitly exploit the complementary information contained in intermediate and high-level representations through dedicated multi-level fusion and refinement mechanisms.

Recent Deep Learning research has integrated XAI in its workflow. GradCAM heatmaps, used to explain image regions driving predictions, are the most commonly used XAI method^{25,28,23,29,18,27}. SHAP (SHapley Additive exPlanations)³⁰, based on cooperative game theory, is another popular XAI method used by³¹ and ³².

Current XAI methods highlight tumor-relevant regions; however, they do not adequately explain why the model predicts a specific class based on those regions. Moreover, most existing approaches lack quantitative measures to assess the faithfulness of the highlighted regions. These gaps motivate the proposed approach, in which MLHnet focuses on holistic learning from brain tumor scans and the Dual-Score Regional XAI demonstrates both model-agnostic and class-agnostic behavior.

Proposed methodology

The proposed model is a unified framework for classification and explainability. The MLHnet serves as the classification module and is based on a Hybrid CNN-Transformer architecture. It incorporates a Multi-Level Attention Block (MAB) and a fusion classifier head to capture discriminative features. Predictions made by MLHnet are then explained using the Dual Score Regional XAI module, which uses hybrid saliency maps to localize tumor regions. The faithfulness of the localized regions is evaluated through a regional perturbation test. The complete model architecture and workflow are illustrated in Fig. 1.

Multi-level hybrid network (MLHnet)

CNN-Transformer hybrid models have gained attention in medical image analysis for their ability to combine local feature extraction with contextual representation. In brain tumor diagnosis, both regional feature discrimination and accurate localisation of abnormalities are required, often under limited positive samples, making hybrid representations particularly effective.

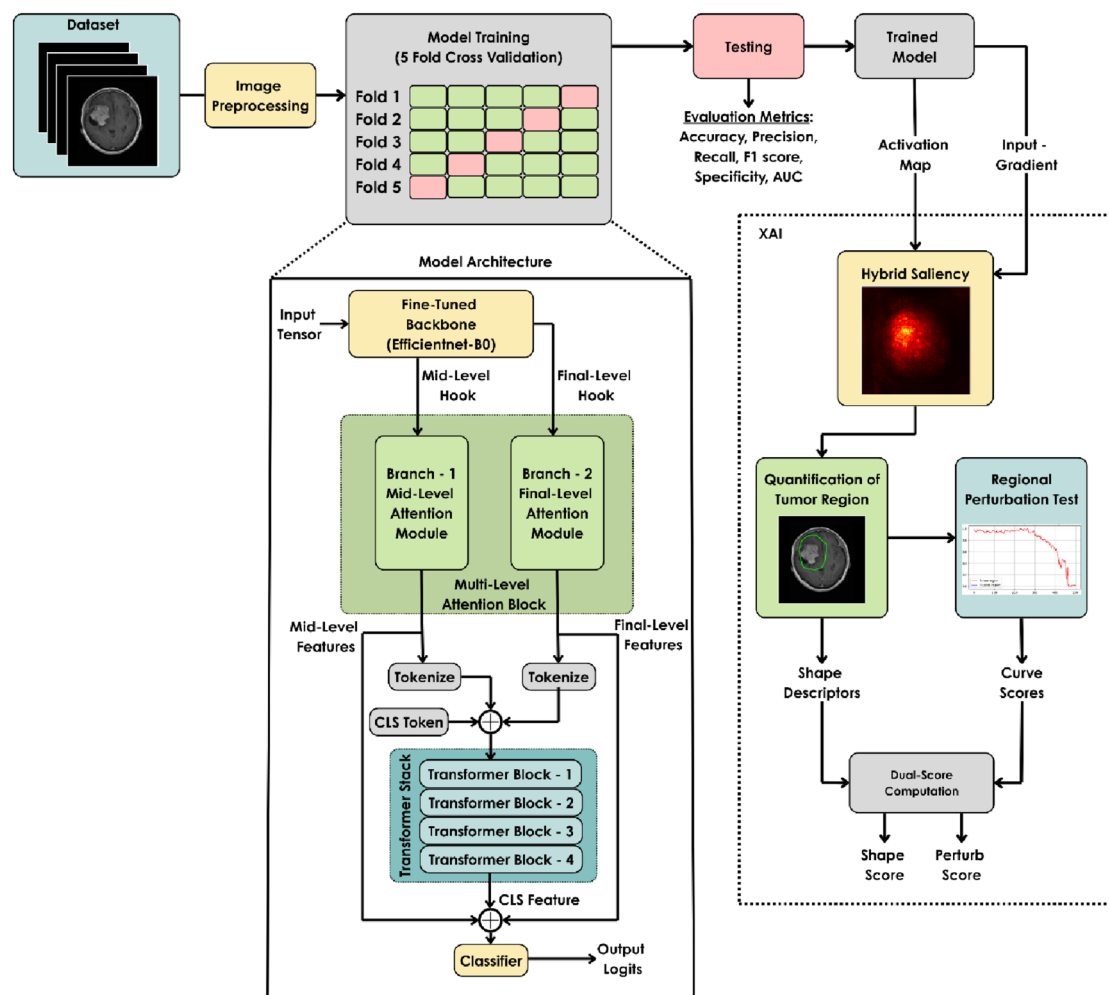


Fig. 1. Overview of the proposed methodology. Brain MRI images are preprocessed and used to train the MLHnet model through 5-fold cross-validation. The best-performing model is selected based on validation metrics and subsequently employed for Dual-Score Regional XAI analysis on the test set.

This paper proposes MLHnet, a CNN-Transformer hybrid that utilises multi-level attention modules to enhance feature representation across classes. A CNN backbone extracts features at intermediate and final stages which are processed in the Multi-level Attention Block and passed to a 4-block transformer. The classifier then concatenates all extracted features and outputs the prediction logits.

Backbone- EfficientNet-B0

The CNN backbone used for feature extraction is the EfficientNet-B0³³. EfficientNet-B0 is the baseline EfficientNet model proposed by Tan and Le. It achieves high accuracy with very low memory requirements and computational costs. It uses Mobile Inverted Bottleneck Convolutions (MBConv), Squeeze and Excitation (SE) modules, depth-wise separable convolutions and Swish activation functions. With only 5.3 million parameters and 7 blocks, EfficientNet-B0 was selected as the backbone feature extractor. This choice was motivated by its favourable balance between representational capacity and computational efficiency. This is particularly suitable for medical imaging settings where lightweight models with strong discriminative performance are desirable.

We input an image denoted as $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$ to a sequence of EfficientNet-B0 blocks where each block produces feature maps passed on to the next block:

$$\mathbf{F}_l = \mathcal{E}_l(\mathbf{I}), l = 1, \dots, 7 \quad (1)$$

where $\mathcal{E}_l(\cdot)$ represents the transformation performed by the l -th EfficientNet block; H , W and C are the height, width and channel (3) of the input image. Instead of using only the final level to extract features, we additionally extract features from an intermediate stage in order to combine complementary representations at different semantic levels. Intermediate features preserve richer spatial and texture-related information, while deeper features capture more abstract semantic patterns that are useful for tumor subtype discrimination. Based on this rationale, two feature maps of different resolutions are extracted: one from the 4th block ($\mathbf{F}_{\text{mid}}$) and one from the 7th block ($\mathbf{F}_{\text{final}}$). The 4th block was selected as a representative intermediate stage with sufficient spatial detail, while the 7th block provides the final high-level semantic representation.

The mid-level feature map is extracted after the first four stages of EfficientNet-B0, corresponding to:

$$\mathbf{F}_{\text{mid}} = \mathcal{E}_{1:4}(\mathbf{I}), \quad (2)$$

This feature map has: $\mathbf{F}_{\text{mid}} \in \mathbb{R}^{C_m \times H_m \times W_m}$, where the channel dimension $C_m = 80$.

The final-level feature map is extracted after all seven stages of EfficientNet-B0, corresponding to:

$$\mathbf{F}_{\text{final}} = \mathcal{E}_{1:7}(\mathbf{I}), \quad (3)$$

This feature map has: $\mathbf{F}_{\text{final}} \in \mathbb{R}^{C_f \times H_f \times W_f}$, where the channel dimension $C_f = 1280$.

Multilevel Attention Block (MAB)

The MAB consists of two branches: one for the intermediate level and the other for the final level (see Fig. 2). Each MAB has a Convolutional Block Attention Module (CBAM), which sequentially applies channel attention followed by spatial attention.

Given an input feature map $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$, channel attention aggregates global spatial information using both average pooling and max pooling:

$$\mathbf{z}_{\text{avg}} = \text{GAP}(\mathbf{F}), \mathbf{z}_{\text{max}} = \text{GMP}(\mathbf{F}) \quad (4)$$

Where GAP stands for Global Average Pooling and GMP stands for Global Max Pooling.

These are passed to a Multi-Layer Perceptron (MLP) with sigmoid function (σ) to generate channel attention weights:

$$\mathbf{M}_c = \sigma(\text{MLP}(\mathbf{z}_{\text{avg}}) + \text{MLP}(\mathbf{z}_{\text{max}})) \quad (5)$$

The feature is refined via channel-wise multiplication:

$$\mathbf{F}_c = \mathbf{M}_c \odot \mathbf{F} \quad (6)$$

Spatial attention on the contrary, emphasizes salient regions by aggregating channel information. Given $\mathbf{F}_c$, spatial descriptors are computed as:

$$\mathbf{s}_{\text{avg}} = \text{AvgPool}_{\text{channel}}(\mathbf{F}_c), \mathbf{s}_{\text{max}} = \text{MaxPool}_{\text{channel}}(\mathbf{F}_c) \quad (7)$$

These are concatenated and convolved to produce a spatial attention map:

$$\mathbf{M}_s = \sigma(f^{7 \times 7}([\mathbf{s}_{\text{avg}}; \mathbf{s}_{\text{max}}])) \quad (8)$$

where $f^{7 \times 7}$ is the 2D convolution operation with 7×7 kernel size.

The feature map from CBAM is:

$$\mathbf{F}_{\text{att}} = \mathbf{M}_s \odot \mathbf{F}_c \quad (9)$$

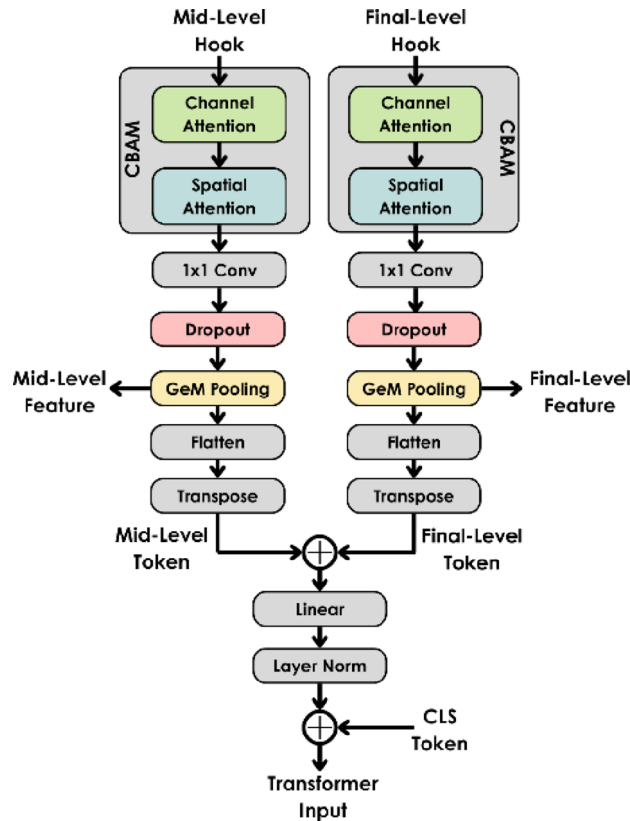


Fig. 2. Architecture of the MAB. The block consists of two parallel branches that process mid-level and final-level features. These features are then tokenized and concatenated with a learnable class token before being fed to the transformer encoder.

Channel attention focuses on “What features are important?” while Spatial attention answers “Where are the important features located?”.

The two branches of MAB for mid and final level features have different spatial dimensions, so each branch is projected into a common embedding space using 1×1 convolution. A dropout rate³⁴ of 0.3 is applied after the 1×1 convolution layers for regularisation and to prevent overfitting.

$$\hat{\mathbf{F}} = (\text{Dropout}_{0.3}(\text{Conv}_{1 \times 1}(\mathbf{F}_{\text{att}}))) \quad (10)$$

The GeM pooling is adopted in place of GAP and GMP to reduce over-smoothing and sensitivity to noise. It uses a learnable parameter (p) where $p = 1$ means GeM acts as GAP while $p \rightarrow \infty$ means GeM acts as GMP. GeM pooling computes for each channel ($\mathbf{g}_c$) as:

$$\mathbf{g}_c = \left(\frac{1}{HW} \sum_{i=1}^H \sum_{j=1}^W \left(\max(\hat{\mathbf{F}}_{c, i, j}, \epsilon) \right)^p \right)^{\frac{1}{p}}, c = 1, \dots, d \quad (11)$$

Where $\hat{\mathbf{F}}_{c, i, j}$ is the refined feature at spatial location $\{i, j\}$ of the channel c , p is a learnable pooling parameter and ϵ ensures numerical stability.

This adaptive pooling strategy is beneficial because the resulting representations are used both for token generation in the Transformer module and for the downstream classification stage.

Transformer

The transformer enables the model to capture long-range/ global context. It has four blocks stacked sequentially with each block following the architecture illustrated in Fig. 3.

The input received from the MAB is tokenized by flattening, linear projection and normalisation. A learnable class token ($\mathbf{t}_{\text{cls}}$) is added to form transformer output features.

$$\mathbf{Z} = [\mathbf{t}_{\text{cls}}; \mathbf{T}_{\text{mid}}; \mathbf{T}_{\text{final}}] \quad (12)$$

where $\mathbf{T}_{\text{mid}}$ and $\mathbf{T}_{\text{final}}$ are the mid and final level token obtained from tokenization of mid and final features respectively.

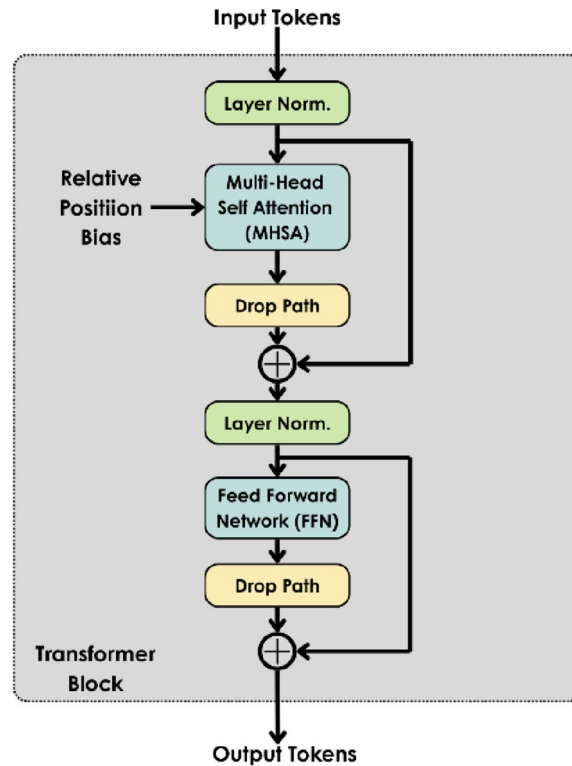


Fig. 3. Architecture of a single transformer block. The input token sequence undergoes Multi-Head Self Attention, followed by a Feed-Forward Network.

These tokens are then sent to the Multi-Head Self-Attention (MHSA) blocks. The Relative Position bias²⁰ encodes the relative distance between tokens which preserves spatial context after flattening. It is added to the scaled dot product attention, after which the attention weights are softmax-normalised. Each transformer block includes a Feed-Forward Network (FFN) implemented as a two-layer MLP with Gaussian Error Linear Unit (GELU) activation. Layer Normalization (LN) and residual connections are applied around both the MHSA and FFN modules:

$$\mathbf{Z}' = \mathbf{Z} + \text{DropPath}(\text{MHSA}(\text{LN}(\mathbf{Z}))) \quad (13)$$

$$\mathbf{Z}_{out} = \mathbf{Z}' + \text{DropPath}(\text{FFN}(\text{LN}(\mathbf{Z}')))) \quad (14)$$

DropPath (stochastic depth)³⁵ of 0.1 is used to randomly drop residual branches, helping in regularisation. Here, $\mathbf{Z}_{out}$ denotes the output obtained from a single block, whereas $\mathbf{Z}_{cls}$ represents the output derived from the complete four-layer stacked architecture.

Fusion Classifier

The final prediction is obtained from a fusion classifier, which combines the CNN Encoder mid-level ($\mathbf{g}_{mid}$), final-level CNN ($\mathbf{g}_{final}$) and transformer ($\mathbf{Z}_{cls}$) features to represent texture, semantic and contextual information, respectively. These are concatenated to form a unified feature vector.

$$\mathbf{f} = [\mathbf{g}_{mid}; \mathbf{g}_{final}; \mathbf{Z}_{cls}] \in \mathbb{R}^{3d} \quad (15)$$

The Classifier Head consists of linear layers for dimensionality reduction, ReLU activations for non-linearity and dropout for regularisation.

$$\hat{\mathbf{y}} = \text{FC}_2(\text{Dropout}(\text{ReLU}(\text{FC}_1(\mathbf{f})))) \quad (16)$$

where $\hat{\mathbf{y}} \in \mathbb{R}^C$ denotes the predicted class logits, FC denotes fully connected dense neural network and C is the number of target classes.

Dual-Score Regional XAI

The Dual-Score Regional XAI framework provides a post-hoc interpretability mechanism for predictions made by the MLHnet. Our method explains the model output after prediction through region-level analysis targeting clinically relevant abnormality regions. A tumor region is defined using hybrid saliency, followed by a Most

Relevant First (MoRF)-like perturbation test³⁶. The result is a class-agnostic shape score and a perturb score to quantify faithfulness. The proposed XAI workflow is illustrated in Fig. 4.

Hybrid Saliency

The hybrid saliency map is generated by combining pixel-level saliency with high-level semantic maps, providing a more stable and robust representation. An input image ($\mathbf{I}$) is passed through the trained model. Pixel-level sensitivity $\mathbf{S}_{\text{grad}}$ (pixel wise gradient saliency map) is computed by backpropagation of the prediction corresponding to the target class.

$$\mathbf{S}_{\text{grad}} = \max_{k \in \{1,2,3\}} \left| \frac{\partial f_c(\mathbf{I})}{\partial \mathbf{I}_k} \right| \quad (17)$$

where f_c is the trained model outputs (logits) and k are the RGB channels. Activation based saliency maps ($\mathbf{S}_{\text{cam}}$) is hooked from the second branch of the MAB containing the CBAM refined attention map.

$$\mathbf{S}_{\text{cam}} = \frac{1}{C} \sum_{c=1}^C \mathbf{F}_{\text{att}} \quad (18)$$

The map is up-sampled to the input size using bilinear interpolation. Both maps are min-max normalised to $[0,1]$. The final hybrid saliency map ($\mathbf{S}_{\text{hybrid}}$) is computed using elementwise multiplication, which enhances regions that are important across both saliency maps:

$$\mathbf{S}_{\text{hybrid}} = \mathbf{S}_{\text{grad}}^{(1-\alpha)} \odot \mathbf{S}_{\text{cam}}^{\alpha} \quad (19)$$

where $\alpha \in [0,1]$ controls the effect of each saliency map. The fused map is normalised again to maintain consistency in the quantitative range.

Quantification of Tumor Region

Rather than relying on predefined manual annotations, the tumor region is inferred from the hybrid saliency map. A binary mask is constructed by applying a high-percentile threshold:

$$\mathbf{M}(x, y) = \begin{cases} 1, & \mathbf{S}_{\text{hybrid}}(x, y) > \tau_p \\ 0, & \text{otherwise} \end{cases} \quad (20)$$

where τ_p corresponds to the p -th percentile. Morphological operation i.e., opening is applied on the binary mask to suppress noisy artifacts. Using connected component analysis, the largest connected region is selected as the abnormality region. Convex hull reconstruction and morphological closing are used for visualising a smoothened circular region. Several shape descriptors are calculated from the detected tumor region such as:

$$\text{Area} : A = \sum_{(x,y) \in \Omega} \mathbf{R}_{\text{tumor}}(x, y) \quad (21)$$

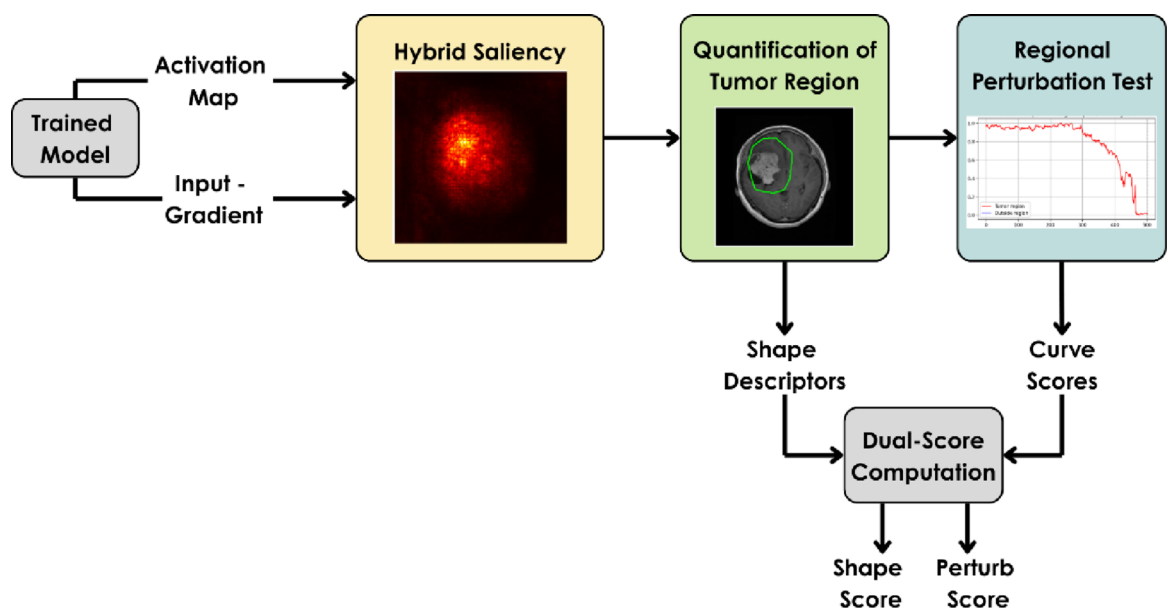


Fig. 4. Proposed XAI workflow: Input Gradient and Activation map from the CBAM of the final level attention block are collected from the trained model and fed to the proposed framework.

where $\mathbf{R}_{\text{tumor}}$ represent pixels pertaining to the detected tumor region, and x, y are the pixel coordinates.

$$\text{Normalized area : } A_{\text{norm}} = \frac{A}{|\Omega|} \quad (22)$$

where $|\Omega|$ denotes the total number of image pixels.

Saliency-weighted area:

$$A_{\text{weighted}} = \sum_{(x,y) \in \Omega} R_{\text{tumor}}(x,y) S_{\text{hybrid}}(x,y) \quad (23)$$

Saliency-weighted Centroid:

$$x_c = \frac{\sum_{(x,y) \in \mathbf{R}_{\text{tumor}}} x \cdot \mathbf{S}_{\text{hybrid}}(x,y)}{\sum_{(x,y) \in \mathbf{R}_{\text{tumor}}} \mathbf{S}_{\text{hybrid}}(x,y)} \quad (24)$$

$$y_c = \frac{\sum_{(x,y) \in \mathbf{R}_{\text{tumor}}} y \cdot \mathbf{S}_{\text{hybrid}}(x,y)}{\sum_{(x,y) \in \mathbf{R}_{\text{tumor}}} \mathbf{S}_{\text{hybrid}}(x,y)}$$

Mean Saliency:

$$S_{\text{mean}} = \frac{1}{A} \sum_{(x,y) \in \mathbf{R}_{\text{tumor}}} \mathbf{S}_{\text{hybrid}}(x,y) \quad (25)$$

Maximum Saliency:

$$S_{\text{max}} = \max_{(x,y) \in \mathbf{R}_{\text{tumor}}} S_{\text{hybrid}}(x,y) \quad (26)$$

Normalized Saliency-weighted Area:

$$A_{\text{weighted}}^{\text{norm}} = \frac{A_{\text{weighted}}}{\sum_{(x,y) \in \Omega} \mathbf{S}_{\text{hybrid}}(x,y)} \quad (27)$$

$$\text{Circularity : } \hat{C} = \frac{4\pi A}{P^2} \quad (28)$$

where P denotes the perimeter of $\mathbf{R}_{\text{tumor}}$.

Regional Perturbation Test

To evaluate the faithfulness of the generated explanations, we perform a MoRF-like region-aware Perturbation test. This test evaluates if the salient pixels, especially those within the computed region, are influential to the model output. Given the saliency map $\mathbf{S}_{\text{hybrid}}$, all pixels are ranked in descending order of saliency. The top $K\%$ most salient pixels are selected for perturbation. The selected top- K pixels are divided into two groups:

1. Pixels inside the tumor region: Pixels lying within the boundary of the generated tumor region.
2. Pixels outside the tumor region: Pixels lying outside the generated tumor region, within and outside the brain.

Perturbations are applied sequentially, first to tumor-region pixels and then to outside pixels. This allows the demonstration of faithfulness of the tumor region. Each pixel is then perturbed by replacing its value with a value corresponding to the dataset mean (μ). After each step, the predicted score for the target class is recorded. The resulting score sequence is normalized to the range $[0, 1]$ to enable comparison across samples.

Dual-score computation

The framework uses two scores to quantify the XAI representation. The shape-based score (T_{shape}) captures localized spatial characteristics in a class-agnostic manner, yielding different value ranges for different classes under the same scoring logic. This enables to quantify how the model differentiates between classes based on geometry. The perturbation-based score (T_{perturb}) helps quantify the faithfulness of the model explanations.

The shape score is computed as a weighted combination of metrics derived from the quantification of the tumor region. It is defined as:

$$T_{\text{shape}} = w_1 \hat{C} + w_2 S_{\text{mean}} + w_3 A_{\text{weighted}}^{\text{norm}} + w_4 A_{\text{norm}} \quad (29)$$

where $\{w_i\}$ are empirically determined weights satisfying $\sum_i w_i = 1$. The resulting score is clipped to the range $[0, 1]$.

From the regional perturbation test, we obtain two confidence curves:

- C_{tumor} : scores after perturbing tumor-region pixels,
- C_{outside} : scores after perturbing non-tumor pixels.

The Area Under Curve (AUC) is computed and normalized:

$$\text{Drop}_{\text{tumor}} = 1 - \text{AUC}(\mathbf{C}_{\text{tumor}}) \quad (30)$$

$$\text{Drop}_{\text{outside}} = 1 - \text{AUC}(\mathbf{C}_{\text{outside}}) \quad (31)$$

A separation term is used to emphasize the tumor region. It is defined as:

$$\text{Separation} = \frac{\text{Drop}_{\text{tumor}}}{\text{Drop}_{\text{tumor}} + \text{Drop}_{\text{outside}}} \quad (32)$$

The perturbation-based trust score is then computed as:

$$T_{\text{perturb}} = b \cdot \text{Drop}_{\text{tumor}} + (1 - b) \cdot \text{Separation} \quad (33)$$

where $b \in [0,1]$ controls the effect of Tumor Drop on the perturb score.

The complete Dual-Score Regional XAI algorithm is given below:

Input: Image I , trained model f_c , percentile threshold τ_p , weights w_i , hybrid saliency factor α , perturbation fraction K , dataset mean pixel value μ .

Output: Shape score T_{shape} , Perturbation score T_{perturb}

Step 1: Compute Hybrid Saliency:

- 1) Pixel-level saliency (gradient-based):

$$S_{\text{grad}} = \max_{k \in \{R, G, B\}} \left| \frac{\partial f_c(I)}{\partial I_k} \right|$$

Activation-based saliency (CBAM attention from the final layer of Backbone):

$$S_{\text{cam}} = \frac{1}{C} \sum_{c=1}^C F_{\text{att}}$$

Combine into a hybrid saliency:

$$S_{\text{hybrid}} = S_{\text{grad}}^{1-\alpha} \odot S_{\text{cam}}^{\alpha}$$

Step 2: Extract Tumor Region

- 1) Divide pixels into two groups using a percentile threshold :

$$M(x, y) = \begin{cases} 1 & S_{\text{hybrid}}(x, y) > \tau_p \\ 0 & \text{otherwise} \end{cases}$$

- 2) Select the largest connected component as the tumor region.
- 3) Compute region descriptors: Area, Normalized area, Saliency-weighted centroid, Mean saliency, Max saliency, Normalized saliency-weighted area, Circularity

Step 3: Regional Perturbation Test

- 1) Rank all pixels in descending order of hybrid saliency.
- 2) Select top $K\%$ pixels, split into tumor and outside regions.
- 3) Perturb sequentially by replacing pixel values with the dataset mean.
- 4) record predicted scores: C_{tumor} , C_{outside}
- 5) Compute AUC-based drops:

$$\text{Drop}_{\text{tumor}} = 1 - \text{AUC}(C_{\text{tumor}}), \text{Drop}_{\text{outside}} = 1 - \text{AUC}(C_{\text{outside}})$$

$$\text{Separation} = \frac{\text{Drop}_{\text{tumor}}}{\text{Drop}_{\text{tumor}} + \text{Drop}_{\text{outside}}}$$

Step 4: Compute Dual Scores

- 1) Shape score (geometry): $T_{\text{shape}} = w_1 \hat{C} + w_2 S_{\text{mean}} + w_3 A_{\text{weighted}}^{\text{norm}} + w_4 A_{\text{norm}}$
- 2) Perturbation score (faithfulness): $T_{\text{perturb}} = b \cdot \text{Drop}_{\text{tumor}} + (1 - b) \cdot \text{Separation}$

Algorithm: Dual-Score Regional XAI

Class	Training (Train + Validation)	Testing
Glioma	1,321	300
Meningioma	1,339	306
No Tumor	1,595	405
Pituitary	1,457	300
Total	5,712	1,311

Table 1. Dataset used in this study.

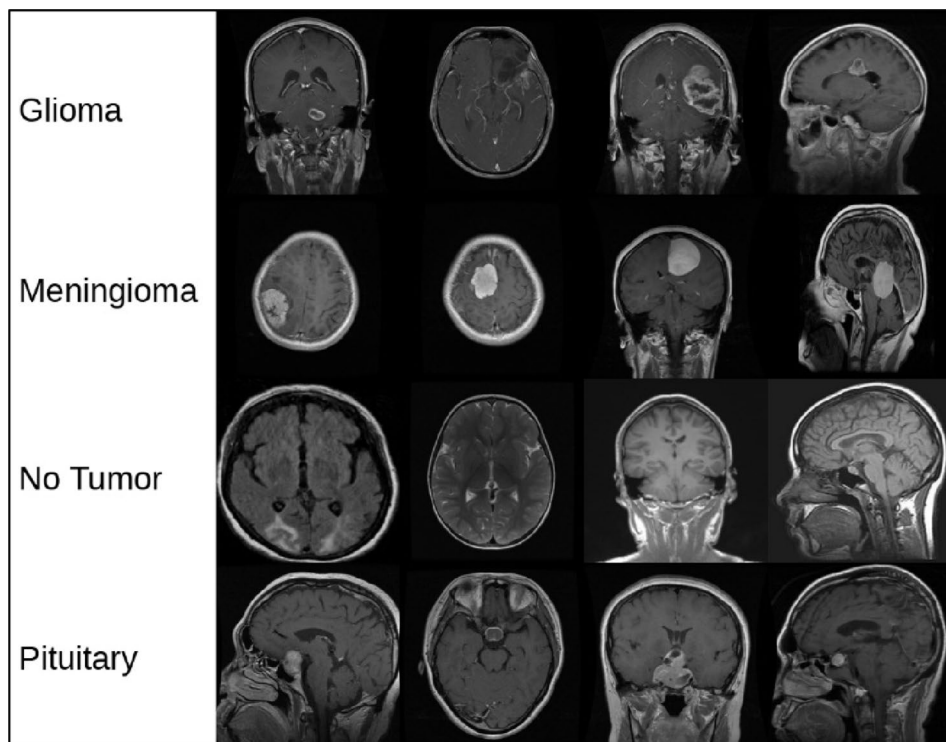


Fig. 5. Representative MRI samples from the dataset. Each row shows the three anatomical views (axial, coronal, sagittal) for one tumor class. From top to bottom: glioma, meningioma, no tumor (healthy), and pituitary tumor. The dataset exhibits significant variability in tumor appearance, size, and contrast across different views and classes.

Results

Training and evaluation of the proposed framework were performed on a system equipped with two Nvidia Tesla T4 GPUs having 29GB RAM and 4 CPU cores. Deep Learning was implemented using Python (v3.12.12) and Pytorch (v2.8.0+cu126) with supporting libraries such as Torchvision (v0.23.0+cu126), NumPy(v2.0.2), Scikit-learn (v1.6.1), Pillow (v11.3.0), SciPy (v1.15.3) and Matplotlib (v3.10.0). This setup enabled faster training and evaluation with moderate batch sizes (batch_size = 32). To ensure reproducibility, fixed random seeds (random_seed(42)) were applied for all processes therefore all randomized operations could be replicated. All experiments were conducted on the same hardware setup, ensuring fair comparison.

Dataset and preprocessing

The Dataset used for all experiments is publicly available on Kaggle³⁷. This dataset contains 7023 images of human brain MRI images which are classified into 4 classes: glioma, meningioma, no tumor and pituitary (Table 1). The training subset contains 5,712 images and the testing subset consists of 1,311 images. All images are stored in JPEG format. The MRI images are in Axial, Coronal and Sagittal views as shown in Fig. 5.

For the proposed MLHnet model, 5- fold stratified cross validation was used to ensure reliable estimation of metrics. For each fold, 80–20 split was taken for the train and validation sets. After validation, the trained model for each fold was tested on the testing subset of 1311 images. The XAI experiments were performed on the testing subset only. The ablation study was performed using the training, validation and test split used in fold 3 which yielded the highest result for the proposed model.

All MRI images are converted to RGB format and resized to 224 by 224 pixels to ensure compatibility with CNN backbones. Pixel values were normalized to [0,1]. Data augmentation was applied online to the training set

to improve model robustness, meaning that different transformed variants of the original images were generated dynamically during training. The augmentation pipeline included colour jittering with brightness variation up to 0.6, contrast up to 1.0, saturation up to 0.3, and hue up to 0.1. In addition, images were randomly rotated within ± 90 degrees and horizontally and vertically flipped with a probability of 0.5.

Training and testing setup

The model was trained using 5-fold stratified cross validation for 50 epochs per fold. The cross-entropy loss function was used to optimize the multi-class objective. AdamW³⁸ optimization was implemented to update the model parameters with a learning rate of 1×10^{-4} . It combines Adaptive Momentum learning with weight regularisation (weight_decay=0.001), which helps in model generalisation. An early stopping strategy was applied where the training loop was terminated when the validation loss did not improve for a defined number of epochs (15 epochs). This technique helps to cutoff the training before it overfits the training subset.

After training and validation, the trained model was evaluated on the testing subset. Test metrics were computed for each fold, and final reported results were obtained by averaging the metrics, providing a reliable estimate of model performance. The metrics used to evaluate were accuracy, precision, recall, F1-score, area under the ROC curve, and confusion matrices. Macro-averaging was used to obtain multi-class results. Let the confusion matrix have True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN). Metrics were calculated by the following formulas:

Accuracy = $\frac{TP + TN}{TP + TN + FP + FN}$ (34)

Precision = $\frac{TP}{TP + FP}$ (35)

Recall = $\frac{TP}{TP + FN}$ (36)

F1-score = $2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$ (37)

Specificity = $\frac{TN}{TN + FP}$ (38)

AUC = $\int_0^1 TPR(FPR) d(FPR)$ (39)

where:

$TPR = \frac{TP}{TP + FN}$ (True Positive Rate) $FPR = \frac{FP}{FP + TN}$ (False Positive Rate).

After all 5 folds were completed, these result metrics were compared and averaged to evaluate model consistency.

Explainability setup

The Explainability experiment was conducted on the test set, excluding the no tumor class, to focus on regions relevant to tumor detection. The trained weights of the best performing model were used. The Perturb value is the mean pixel value of the validation subset. The shape weights were calculated empirically from the validation set. The validation subset is used to determine dataset related constants to reduce model bias. Perturbation curves were recorded for each image and then averaged across classes to evaluate the model’s reaction to perturbing salient regions. Class-wise Box plots were calculated for the shape score to view inter-class variation. The XAI hyperparameters used are given in Table 2.

5-fold cross-validation

The training and validation metrics are given in Table 3. The model performs consistently well with an average accuracy of 99.13% across folds. The model does not appear to be overfitted, as the average validation accuracy is 98.92%, which is very close to the training performance. From the Table 3, we can also infer that the model of fold 3 is highly robust, wherein the validation accuracy even exceeds the training accuracy. Across all folds,

Parameter	Default Value	Description
α (Saliency Map constant)	0.7	Balance the weight between gradient-based and activation-based saliency maps
Threshold_percentile (τ_p)	90	Percentile cutoff for extracting the tumor region from the saliency map
Perturb_value	[0.15, 0.15, 0.15]	Mean RGB pixel intensity of the validation set used for perturbation
Top_K_percent	1%	Fraction of most salient pixels analyzed in perturbation test
shape_weights	[0.0002, 0.0569, 0.7818, 0.1611]	Weights for circularity, mean saliency, weighted area, and normalized area in Shape Score
b (Tumor Drop constant)	0.6	Balance the parameter between Tumor Drop and Separation in Perturb Score

Table 2. Hyperparameters used in the Dual-Score Regional XAI framework.

Fold	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Specificity (%)
1	Train	99.10	99.08	99.07	99.07	99.69
	Val	98.78	98.72	98.82	98.77	99.58
2	Train	99.30	99.28	99.28	99.28	99.76
	Val	99.13	99.12	99.08	99.10	99.64
3	Train	99.26	99.26	99.24	99.25	99.74
	Val	99.74	99.72	99.73	99.73	99.89
4	Train	98.82	98.80	98.78	98.79	99.54
	Val	98.34	98.31	98.24	98.27	99.32
5	Train	99.17	99.14	99.14	99.14	99.67
	Val	98.60	98.58	98.59	98.59	99.46

Table 3. Train and validation performance across 5-fold cross-validation.

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Specificity (%)	AUC (%)
1	99.31	99.29	99.27	99.28	99.77	99.97
2	99.08	99.05	99.01	99.03	99.70	99.98
3	99.47	99.44	99.42	99.42	99.83	99.93
4	99.39	99.38	99.34	99.36	99.80	99.99
5	99.24	99.19	99.19	99.19	99.75	99.97
Average	99.30 ± 0.14	99.27 ± 0.15	99.25 ± 0.15	99.26 ± 0.15	99.77 ± 0.04	99.97 ± 0.02

Table 4. Test results for each fold.

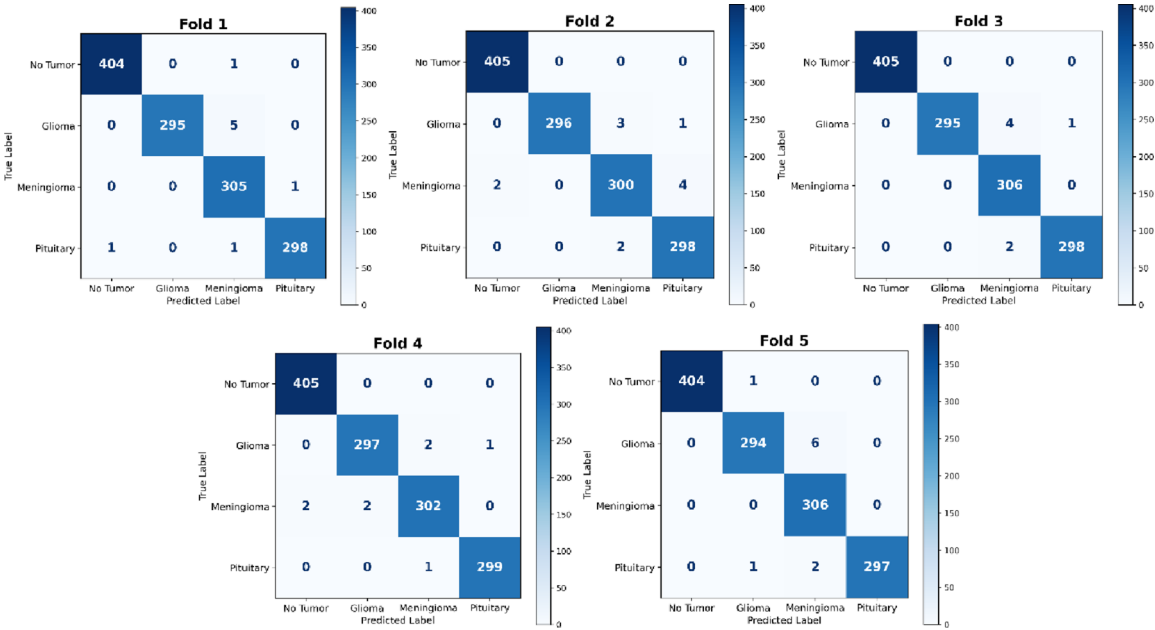


Fig. 6. Confusion matrices for all five folds on test set (1311 images).

the training subset achieved average precision, recall, F1-score, and specificity values above 99%, while the corresponding validation metrics remained consistently high at approximately 98.9%.

Table 4 shows performance metrics on the test set for all the folds. With this data we can judge which model has learnt the best and has good generalisation capability. All folds display similarly high scores exceeding 99% across most metrics, with low variance, highlighting the model's consistently strong performance. The Confusion Matrices of each fold can be visualised in Fig. 6. It is observed that healthy class images have a near perfect prediction record. The Glioma class is often misclassified as the Meningioma class because of their similar appearance. The Pituitary class has a fairly accurate prediction record, though it was confused with meningioma in a few cases.

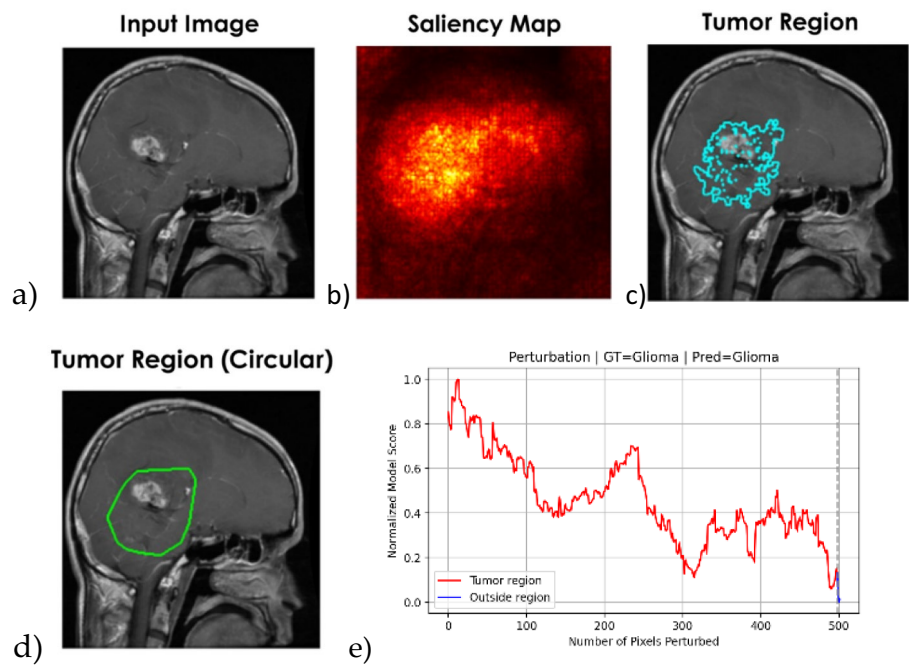


Fig. 7. Step-by-step visualization of the Dual-Score Regional XAI pipeline for a glioma sample. (a) Original MRI input image; (b) Hybrid saliency map combining gradient and activation information; (c) Raw tumor region extracted via percentile thresholding ($\tau_p = 90$); (d) Refined tumor region after morphological post-processing; (e) Regional perturbation curves showing the drop in model confidence as salient pixels within (red) and outside (blue) the tumor region is progressively masked.

True Label	Confidence	Shape Score	Perturb Score	Tumor Drop	Separation	Area	Norm Area	Circularity	Mean Saliency	Weighted Area
Glioma	0.994 ± 0.039	0.253 ± 0.043	0.446 ± 0.100	0.504 ± 0.123	0.359 ± 0.105	3692.65 ± 507.71	0.074 ± 0.010	0.074 ± 0.032	0.489 ± 0.056	1790.126 ± 234.016
Meningioma	0.998 ± 0.015	0.333 ± 0.040	0.479 ± 0.142	0.564 ± 0.183	0.352 ± 0.081	4127.49 ± 305.50	0.082 ± 0.006	0.127 ± 0.040	0.400 ± 0.053	1646.400 ± 203.017
Pituitary	0.998 ± 0.022	0.144 ± 0.059	0.442 ± 0.123	0.493 ± 0.157	0.365 ± 0.088	2208.21 ± 976.46	0.044 ± 0.019	0.057 ± 0.028	0.567 ± 0.057	1213.468 ± 479.315

Table 5. Class-wise statistics of Dual-Score Regional XAI metrics on the test set (excluding no tumor class, N=906 images).

Among the 5 folds, fold 3 has a superior validation and test performance with an accuracy of 99.74% and 99.47%, respectively. Therefore, the weights from fold 3 were used in the proposed XAI framework.

Dual-score Regional XAI

The purpose of Dual Score in the proposed XAI is to access and quantify the faithfulness of the explanations as well as how different the explanations for each class are. As shown in Fig. 7, the input image is first processed to generate a hybrid saliency map. This saliency map is then used to localize a raw tumor region, which is further refined through post-processing to obtain the final tumor region. A region-wise perturbation test is subsequently applied to evaluate the impact of the localized region on the model’s prediction.

The metrics obtained from the proposed XAI method are shown in Table 5. Model confidence represents the certainty associated with the predicted class. The proposed Dual Score consists of two components: the Shape Score and the Perturb Score.

The Shape Score quantifies class-wise differences in explanations based on the shape and structural characteristics of the localized tumor region. It includes circularity, which reflects how closely the region resembles a circular shape; mean saliency, defined as the average saliency intensity within the tumor region; weighted area, which combines region size with saliency importance, and normalized area. A low standard deviation across these shape metrics indicates consistent and distinguishable class-specific explanations.

The Perturb Score measures the model’s sensitivity to perturbations applied within the localized tumor region. Tumor Drop captures the reduction in prediction confidence when the identified tumor region is perturbed, providing a quantitative measure of explanation faithfulness.

The Kernel Density Estimate (KDE) plots in Fig. 8 demonstrate the class-agnostic scoring mechanism.

Meningioma exhibits the highest mean confidence. It also has the largest area with a more circular shape than other classes. A higher tumor drop score indicates that perturbing the tumor region leads to a significant drop in model confidence. Pituitary tumors show high confidence while showing small tumor regions based on the

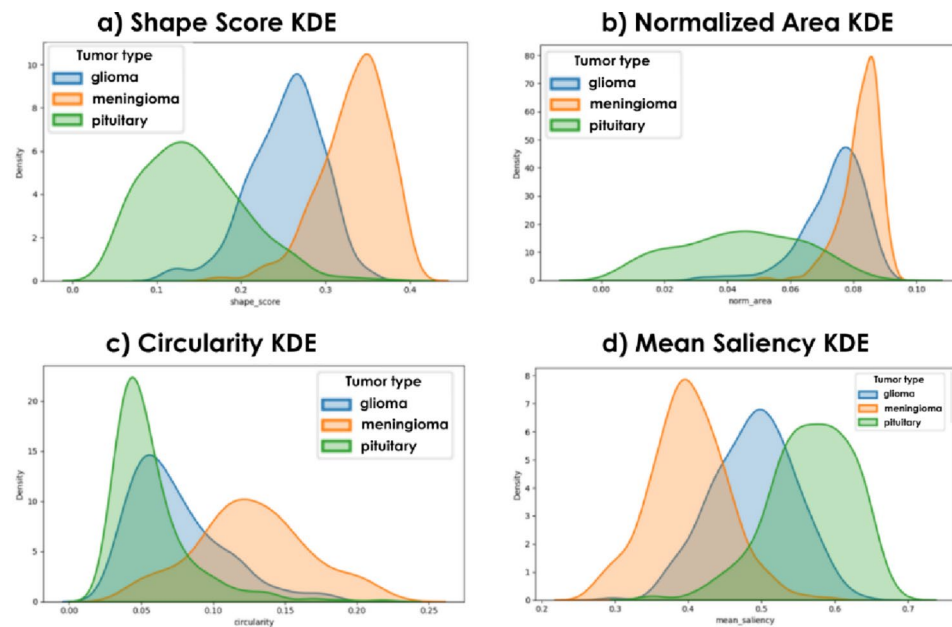


Fig. 8. Class-wise Kernel Density Estimate plots of (a) Shape Score, (b) Normalised Area, (c) Circularity, (d) Mean Saliency.

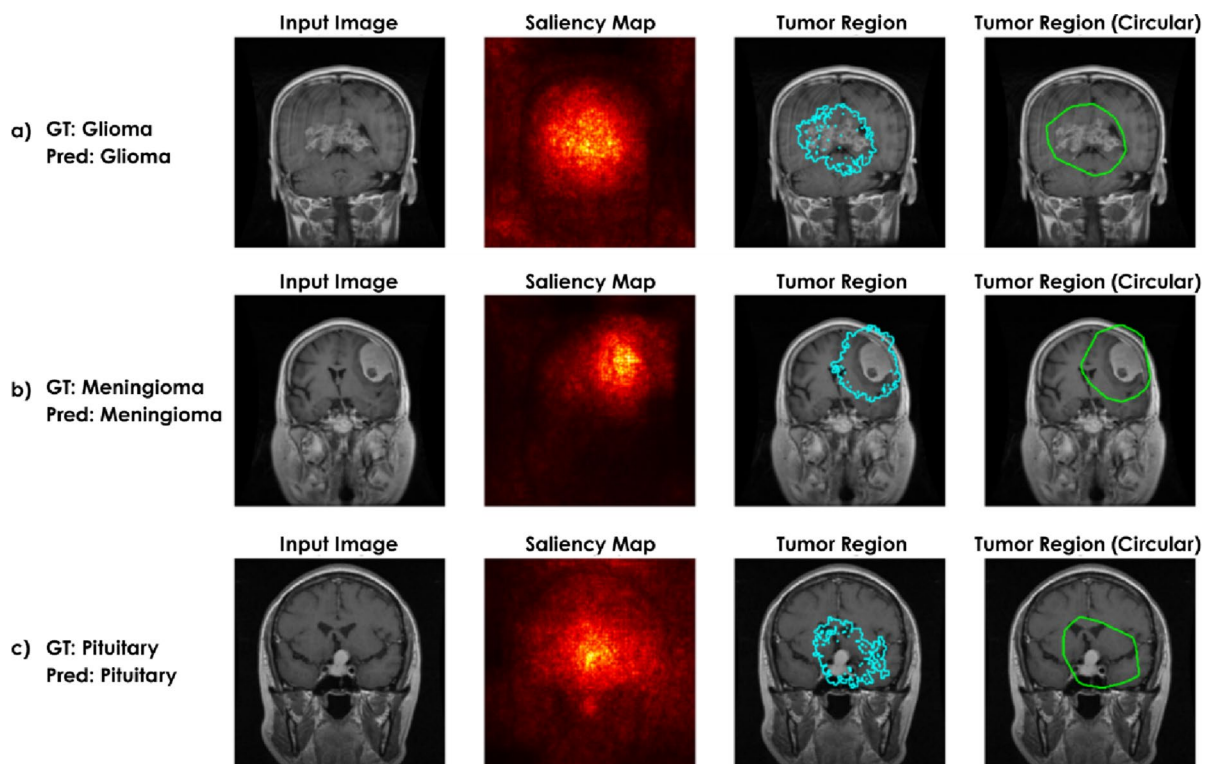


Fig. 9. Examples of correctly localized tumor regions validated by a radiologist. Each row represents a different tumor type (top to bottom: glioma, meningioma, pituitary). The columns are the input MRI image, the saliency map generated by Hybrid Saliency, the raw tumor region correctly discovered and the smoothed tumor region for better visualisation.

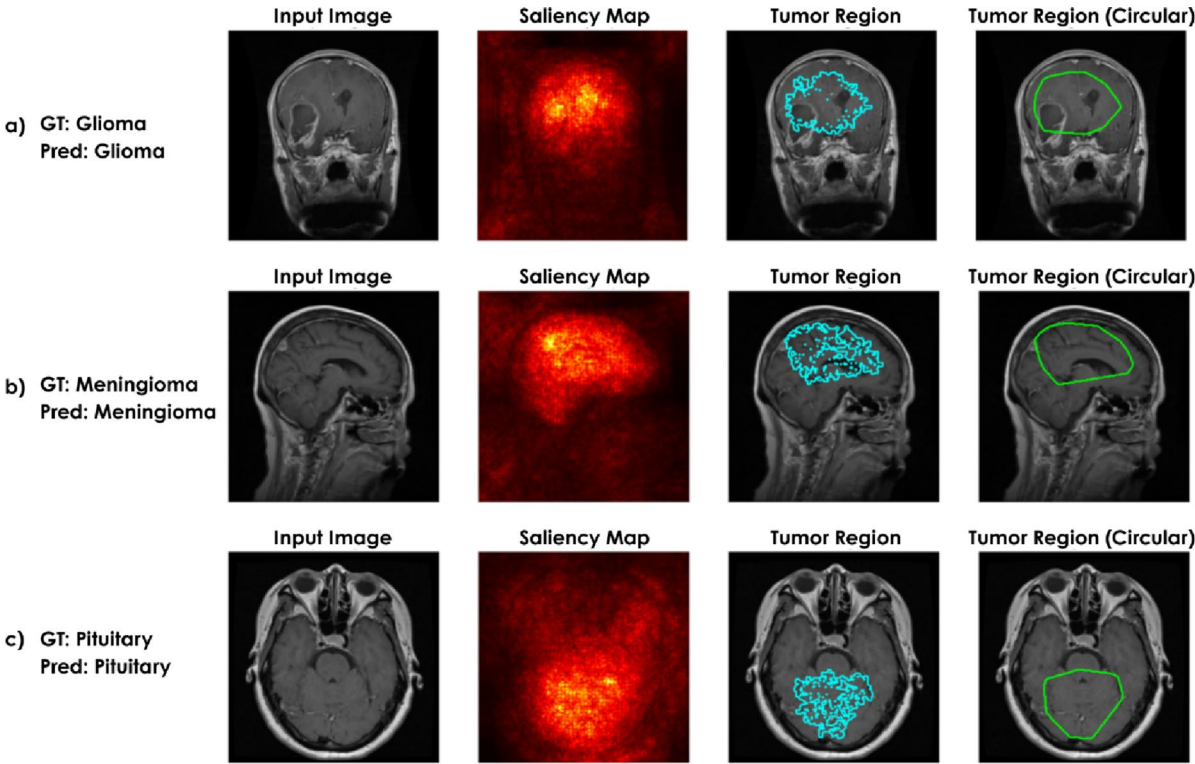


Fig. 10. Examples of incorrectly localized tumor regions. The XAI framework struggles when saliency is diffusely distributed across large areas or exhibits multiple local maxima.

Tumor Type	Correctly Explained	Incorrectly Explained	Total	% Correct
Glioma	267	33	300	89.00%
Meningioma	304	2	306	99.35%
Pituitary	223	77	300	74.33%
Overall	794	112	906	87.64%

Table 6. Expert validation of XAI-generated tumor localizations by a board-certified radiologist. A total of 906 test images (excluding the no tumor class) were evaluated. An explanation is considered correct if the identified region corresponds to the clinically relevant tumor location.

shape score. A high mean saliency indicates how the model focuses on a small specific salient region. The model finds pituitary regions the hardest to localize, as indicated by the lowest perturb score among the classes. Glioma has lower confidence with more irregular tumor boundaries. It has a moderate score across all metrics, except for a high weighted area, which shows emphasis of the model on the tumor region.

The clinical validity of the generated explanations was assessed by a radiologist's evaluation. Representative examples of correctly localized tumor regions are shown in Fig. 9, demonstrating the framework's ability to identify clinically relevant areas with high fidelity, particularly for meningiomas and gliomas. However, failure cases illustrated in Fig. 10 reveal limitations when saliency is diffusely distributed or contains multiple competing maxima, particularly in pituitary tumor images. These cases account for the lower explanation correctness rate (74.33%) observed for pituitary tumors compared to other classes.

Table 6 presents the expert evaluation of the XAI-generated tumor localization results. In the present study, the explanation maps were reviewed by a single experienced board-certified radiologist, who assessed whether the highlighted regions corresponded to clinically relevant tumor areas. A total of 906 test images, excluding the no-tumor class, were examined. An explanation was considered correct when the identified region corresponded to the clinically relevant tumor location. Overall, the proposed XAI framework achieved an explanation correctness of 87.64%. Among all tumor classes, meningioma achieved the highest explanation correctness (99.35%), whereas pituitary tumor showed the lowest correctness (74.33%), indicating greater difficulty in localizing this category.

Efficient-NetB0	Transformer	RPB	MLFE	GeM	CBAM	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
✓	✗	✗	✗	✗	✗	99.01	98.98	98.97	98.97
✓	✓	✗	✗	✗	✗	98.93	98.93	98.88	98.90
✓	✓	✓	✗	✗	✗	98.86	98.79	98.83	98.80
✓	✓	✓	✓	✗	✗	99.24	99.23	99.17	99.20
✓	✓	✓	✓	✓	✗	99.08	99.05	99.03	99.04
✓	✓	✓	✓	✓	✓	99.47	99.44	99.42	99.42

Table 7. Results of ablation study. RPB: Relative Position Bias, MLFE: Multi Level Feature Extraction.

Fusion Strategy	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Concatenation	99.47	99.44	99.42	99.42
Attention-based ³⁹	99.24	99.24	99.17	99.21
Bilinear Fusion ⁴⁰	99.31	99.30	99.28	99.29
Weighted Fusion ⁴¹	99.31	99.28	99.27	99.28

Table 8. Comparison of fusion strategies.

State of the Art Models	Year	XAI	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
A-GRU [17]	2024	✗	99.32	100	99.12	99.01
VGG16-Resnet50-CDBN fusion [42]	2024	✗	98.98	97.95	97.82	97.87
BDTN [18]	2025	✗	95.33	94.89	94.78	94.91
EFFResNet-ViT [27]	2025	✓	99.31	99.26	99.25	99.26
PDSCNN-RRELM [32]	2025	✓	99.5	100	100	100
BRAIN-META [25]	2026	✓	97.1	97	96.92	96.94
ShallowMRI with Self Attention Learning [28]	2026	✗	98.24	98.17	98.1	98.13
Proposed Model	2026	✓	99.47	99.44	99.42	99.42

Table 9. Comparison with state-of-the-art methods on various brain MRI datasets available on Kaggle.

Ablation Study

The proposed architecture is designed to capture complex brain tumor characteristics with correct context-based decision making. While each component of the proposed model may not add much to the performance itself, together they synergize with each other improving overall performance and reliability as shown in Table 7. The Efficientnet-B0 serves as a strong baseline; however, it lacks any contextual understanding necessary to scale up to complex cases. To address this, a transformer module is added in series. When used without well-structured tokens or large-scale training data, the transformer alone leads to a performance drop. Similarly, the inclusion of relative positional bias, while beneficial for spatial awareness in principle, introduces additional noise in the absence of robust token representations, further degrading performance. To address the problem of poor token representation, we integrated Multi-Level Feature Extraction, resulting in a substantial improvement in accuracy. This indicates the benefit of combining low-level texture information with high-level semantic information. It was seen that GAP used in the MLFE component, treated all spatial locations uniformly. GeM was added to concentrate pooling only on highly salient features using a learnable parameter. This component alone did not improve performance, as it tended to prioritize visually dominant regions rather than clinically relevant ones. Finally, CBAM is integrated to explicitly guide the GeM mechanism toward tumor-relevant features, effectively making clinically important regions the most salient. With CBAM, the model learns to prioritize diagnostically meaningful features rather than visually dominant patterns. This final addition completes the proposed MLHnet architecture, in which all components synergistically collaborate to achieve an accuracy of 99.47%, demonstrating that each module contributes uniquely to the overall performance.

To further justify the adopted fusion design, we additionally compared the proposed concatenation strategy with attention-based³⁹, bilinear⁴⁰, and weighted fusion⁴¹ alternatives under identical architectural and training settings. The comparison results are reported in Table 8.

Comparison with state-of-the-art

Table 9 summarizes the performance of the proposed model alongside recent state-of-the-art (SOTA) models evaluated on the publicly available Brain MRI Kaggle dataset. Although all studies were conducted on the MRI brain datasets, differences in preprocessing, augmentation, data partitioning, and training protocols may affect strict comparability of the reported results. Within this context, the proposed framework demonstrates strong performance, achieving an accuracy of 99.47% with consistently high precision, recall, and F1-score, while also

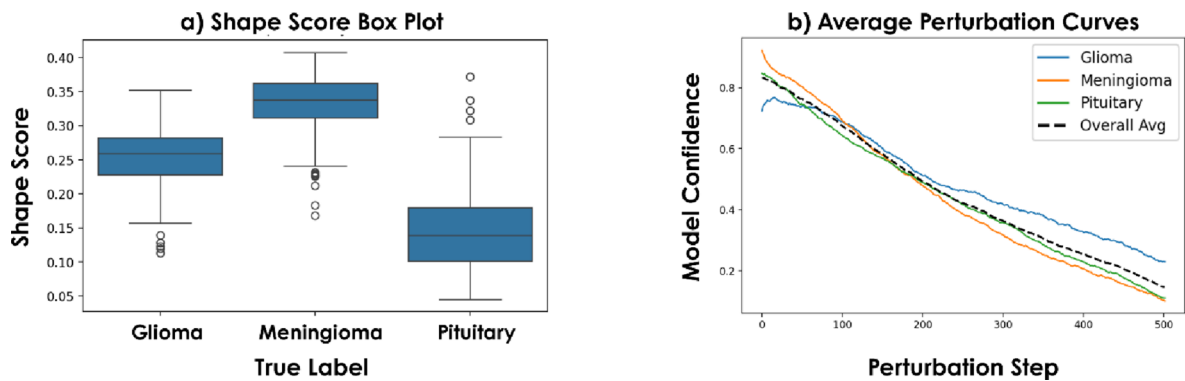


Fig. 11. Quantitative analysis of XAI explanations across tumor classes. (a) Box plot of Shape Scores showing class-specific geometric characteristics (b) Average perturbation curves demonstrating the drop in model confidence as the top 1% most salient pixels are progressively masked. The consistent downward trend across all classes validated that the model relies on the identified tumor regions for prediction.

XAI Method	Mean MoRF AUC (↓)	Mean Computation Time (s) (↓)
Saliency	27.4530	0.9362
Input x Gradient	24.6613	0.9910
Integrated Gradients	25.1077	3.3984
Guided Backpropagation	22.0926	0.9313
Gradient-weighted Class Activation Mapping (Grad-CAM)	17.3712	0.8707
Guided Grad-CAM	19.5394	0.9743
Deep Learning Important Features (DeepLIFT)	25.2155	0.7949
Occlusion	22.4891	17.2085
Local Interpretable Model-agnostic Explanations (LIME)	21.7401	17.1666
SmoothGrad Integrated Gradients	32.5691	23.6708
Hybrid_Saliency	25.3667	0.9070

Table 10. Quantitative comparison of XAI methods using Most Relevant First (MoRF) perturbation analysis.

providing an interpretable XAI component for model explanation. In addition, the model is lightweight, with a size of 21.27 MB, 5.31 million learnable parameters, and a computational cost of 1.61 GFLOPs using float32 precision. These characteristics indicate its potential suitability for efficient deployment in resource-constrained settings.

Discussion
Impact of the Dual-Score Regional XAI

The results of Dual-Score Regional XAI for the entire test set are aggregated and averaged, as illustrated by Fig. 11. The Shape Score Box Plot (Fig. 11 (a)) shows the proposed XAI to differentiate tumor classes based on shape score. The interquartile ranges of all classes are apart from each other, whereas the whiskers and outliers overlap with each other. This highlights the complexity of the brain tumor classification problem. It is seen that Meningioma is often higher than other classes of tumor. The interquartile ranges of glioma and meningioma are close to each other, suggesting similarity in the shape characteristics of these two tumor classes. The Pituitary tumor class is smaller than the rest of the classes, as suggested by its low position on the plot. The Average Perturbation Curve by class (Fig. 11 (b)) shows a steady decrease across all classes, indicating a drop in model confidence when the pixels within the generated tumor region are perturbed. The proposed XAI method was conducted exclusively on tumor-positive samples. It does not provide interpretability for healthy classes or false negative predictions.

To verify the competency of Hybrid Saliency used for Dual-Score Regional XAI, a comparative study is done, taking inspiration from Kakogeorgiou and Karantzalos’s paper on XAI⁴³. While experimenting, two metrics were calculated MoRF AUC and Computation Time. The perturbation value was taken as 0.15 and the perturbation was done pixelwise for 500 iterations. Final MoRF curves were normalised for better readability. Lower MoRF AUC values indicate better identification of truly salient regions by the XAI method.

Table 10 displays the performance of XAI methods, and Fig. 12 illustrates the MoRF curves calculated. Saliency⁴⁴ is the baseline model as it is simple with acceptable scores across both metrics. GradCAM⁹ achieves the lowest MoRF AUC with the fastest computation time indicating its popularity. However, as the perturbation increases, the model starts to detect artifacts, artificially increasing the model score as seen in the plot. Guided GradCAM⁹ also performs competitively but at a higher computation time. LIME¹¹ is highly susceptible to

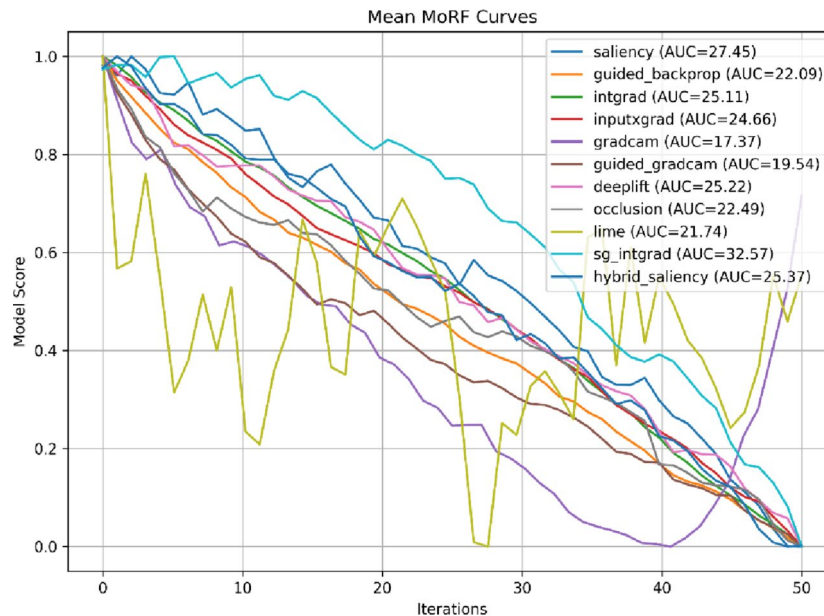


Fig. 12. Comparison of XAI methods using Most Relevant First (MoRF) perturbation analysis. The curves show normalized model confidence as the most salient pixels are progressively masked.

artifacts. Occlusion¹⁰ and SmoothGrad-Integrated Gradients⁴⁵ have very high computational time with poor trade-off, as MoRF AUC is subpar as well. Deeplift⁴⁶ has a high MoRF AUC and the lowest computation time; however, it depends on baseline selection, causing attribution bias. Guided-Backpropagation⁴⁷, InputxGradient⁴⁶ and Integrated Gradients⁴⁸ have a favourable tradeoff between MoRF AUC and Computation time where Integrated Gradients has more computation time. To make the proposed XAI framework, high-performance and low complexity are achieved using Hybrid Saliency, which gives a mean MoRF AUC of 25.3667 and a mean computation time of 0.9070 s.

Advantages of the proposed system

- The proposed model achieves strong performance on a widely used public brain MRI dataset and compares favorably with recent reported methods.
- The proposed architecture benefits from rich hierarchical feature representations, tumor localization and global-local contextual learning.
- The proposed explainability framework provides interpretable region-based visualizations and quantitative scores that support analysis of model behavior.
- The proposed model maintains a balanced trade-off performance and computational efficiency, enabling potential real-time clinical deployment.

Limitations and Future Directions

The proposed model was trained and evaluated on a single publicly available dataset. Although this dataset is widely used, reliance on a single source may limit the generalizability of the findings across institutions, scanners, acquisition protocols, patient populations, and real-world clinical distributions. Therefore, external validation on independent cohorts is necessary before making broader claims regarding robustness, clinical reliability, or practical applicability.

In addition, although the proposed XAI framework provided encouraging results, its radiologist-based evaluation indicates that there remains room for improvement in the clinical fidelity of the generated explanations. The current explainability pipeline also relies on empirically defined thresholds, hyperparameters, and scoring criteria, which should be further validated across different datasets and clinical settings. In particular, the framework may fail to precisely localize the tumor region when saliency is broadly distributed over large areas, even if local maxima are still present within the true abnormal region. Future work will therefore investigate improved hyperparameter tuning and more effective post-processing strategies to better isolate tumor-relevant regions. Beyond this, the proposed architecture could be extended to process entire MRI volumes rather than individual 2D images, while the XAI module could be adapted to generate and analyze 3D saliency maps. The framework may also be explored in other medical imaging domains, including ultrasound and retinal imaging, to assess its broader applicability.

Conclusion

In this study, we propose the MLHnet, a CNN-Transformer hybrid architecture that leverages both mid-level and final-level feature extraction to capture fine-grained texture details and high-level semantic information.

The proposed architecture achieves state-of-the-art performance, achieving an accuracy of 99.47% on a publicly available brain MRI Kaggle dataset. To address the black-box nature of deep learning models, we also introduce the Dual-Score Regional XAI framework, which localizes tumor-relevant regions and quantifies explanation faithfulness. The resulting system is lightweight, computationally efficient, and clinically interpretable.

The robustness and effectiveness of the proposed system are validated through extensive experiments, including 5-fold cross-validation, radiologist evaluation, ablation studies, and comparative analyses with existing methods. Nevertheless, since the study was conducted on a single public dataset, further validation on independent external datasets and real-world clinical cohorts is required before drawing broader conclusions regarding generalizability, robustness, or clinical applicability.

Data availability

The datasets generated and/or analysed during the current study are available in the Kaggle repository, <https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset> (Version 1).

Received: 6 March 2026; Accepted: 6 May 2026

Published online: 21 May 2026

References

1. Ijaz, M. et al. Diagnostics of brain tumor in the early stage: current status and future perspectives. *Apr 09*. <https://doi.org/10.1039/d4bm01503g> (2025). Royal Society of Chemistry.
2. Society, A. C., Causes, A. B. T., Factors, R. & Prevention *American Society of Clinical Oncology*, Accessed: Feb. 10, 2026. [Online]. (2026). Available: https://www.cancer.org/cancer/types/brain-spinal-cord-tumors-adults/causes-risks-prevention.html?utm_source=chatgpt.com
3. Zacharakis, E. I. et al. Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme. *Magn. Reson. Med.* **62** (6), 1609–1618. <https://doi.org/10.1002/mrm.22147> (2009).
4. Asiri, A. A. et al. Jul., Brain Tumor Detection and Classification Using Fine-Tuned CNN with ResNet50 and U-Net Model: A Study on TCGA-LGG and TCIA Dataset for MRI Applications., **13**, 7, (2023). <https://doi.org/10.3390/life13071449>
5. Dosovitskiy, A. et al. Jun., An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, [Online]. (2021). Available: <http://arxiv.org/abs/2010.11929>
6. Reddy, C. K. K. et al. A fine-tuned vision transformer based enhanced multi-class brain tumor classification using MRI scan imagery. *Front. Oncol.* **14** <https://doi.org/10.3389/fonc.2024.1400341> (2024).
7. Salvi, M. et al. Explainability and uncertainty: Two sides of the same coin for enhancing the interpretability of deep learning models in healthcare. *Int. J. Med. Inf.* **197** <https://doi.org/10.1016/j.ijmedinf.2025.105846> (May 2025).
8. Samek, W., Wiegand, T. & Müller, K. R. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models, Aug. [Online]. (2017). Available: <http://arxiv.org/abs/1708.08296>
9. Selvaraju, R. R. et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *Dec* <https://doi.org/10.1007/s11263-019-01228-7> (2019).
10. Zeiler, M. D. & Fergus, R. LNCSS 8689 - Visualizing and Understanding Convolutional Networks, (2014).
11. Ribeiro, M. T., Singh, S. & Guestrin, C. ‘Why should i trust you?’ Explaining the predictions of any classifier, in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Aug. pp. 1135–1144. (2016). <https://doi.org/10.1145/2939672.2939778>
12. Woo, S., Park, J., Lee, J. Y. & Kweon, I. S. CBAM: Convolutional Block Attention Module.
13. Radenović, F., Tolias, G. & Chum, O. Fine-tuning CNN Image Retrieval with No Human Annotation, Jul. [Online]. (2018). Available: <http://arxiv.org/abs/1711.02512>
14. Cheng, J. et al. Enhanced performance of brain tumor classification via tumor region augmentation and partition. *PLoS One.* **10** (10). <https://doi.org/10.1371/journal.pone.0140381> (Oct. 2015).
15. Deepak, S. & Ameer, P. M. Brain tumor classification using deep CNN features via transfer learning. *Comput. Biol. Med.* **111** <https://doi.org/10.1016/j.combiomed.2019.103345> (Aug. 2019).
16. Afshar, P., Plataniotis, K. N. & Mohammadi, A. CAPSULE NETWORKS FOR BRAIN TUMOR CLASSIFICATION BASED ON MRI IMAGES AND COARSE TUMOR BOUNDARIES (IEEE, 2018).
17. Saboor, A. et al. DDFC: deep learning approach for deep feature extraction and classification of brain tumors using magnetic resonance imaging in E-healthcare system. *Sci. Rep.* **14** (1). <https://doi.org/10.1038/s41598-024-56983-6> (Dec. 2024).
18. Mathivanan, S. K. et al. A secure hybrid deep learning framework for brain tumor detection and classification, *J. Big Data*, vol. 12, no. 1, Dec. (2025). <https://doi.org/10.1186/s40537-025-01117-6>
19. Asiri, A. A. et al. Advancing Brain Tumor Classification through Fine-Tuned Vision Transformers: A Comparative Study of Pre-Trained Models. *Sensors* **23** (18). <https://doi.org/10.3390/s23187913> (Sep. 2023).
20. Liu, Z. et al. Aug., Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, [Online]. (2021). Available: <http://arxiv.org/abs/2103.14030>
21. Chen, J. et al. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Med. Image Anal.* **97** <https://doi.org/10.1016/j.media.2024.103280> (Oct. 2024).
22. Almufareh, M. F., Imran, M., Khan, A., Humayun, M. & Asim, M. Automated Brain Tumor Segmentation and Classification in MRI Using YOLO-Based Deep Learning. *IEEE Access.* **12**, 16189–16207. <https://doi.org/10.1109/ACCESS.2024.3359418> (2024).
23. Saeed, T. et al. Neuro-XAI: Explainable deep learning framework based on deeplabV3+ and bayesian optimization for segmentation and classification of brain tumor in MRI scans. *J. Neurosci. Methods.* **410** <https://doi.org/10.1016/j.jneumeth.2024.110247> (Oct. 2024).
24. Jayaraman, G. et al. A hybrid CNN–ViT framework with cross-attention fusion and data augmentation for robust brain tumor classification. *Sci. Rep.* **15** (1). <https://doi.org/10.1038/s41598-025-28636-9> (Dec. 2025).
25. Kumar Napa, K., Murugan, S., Murugan, J. S. & Jayanthi, A. BRAIN-META: A reproducible CNN–vision transformer meta-ensemble pipeline for explainable brain tumor classification. *MethodsX* **16** <https://doi.org/10.1016/j.mex.2025.103769> (Jun. 2026).
26. Swaminathan, A. D. et al. Dec., Hybrid Deep Learning for Brain Tumor Detection: Combining DenseNet and Custom CNN for Enhanced Accuracy, *Biomed. Eng. Comput. Biol.*, vol. 16, (2025). <https://doi.org/10.1177/11795972251395954>
27. Hussain, T. et al. EFRResNet-ViT: A Fusion-Based Convolutional and Vision Transformer Model for Explainable Medical Image Classification. *IEEE Access.* **13**, 54040–54068. <https://doi.org/10.1109/ACCESS.2025.3554184> (2025).
28. Rehman Khan, S. U. et al. ShallowMRI: A novel lightweight CNN with novel attention mechanism for Multi brain tumor classification in MRI images. *Biomed. Signal. Process. Control.* **111** <https://doi.org/10.1016/j.bspc.2025.108425> (Jan. 2026).

29. Anitha, A., Nair, A. & Kamaraj, B. Brain Tumor Detection and Classification Using Deep Neural Network and Interpretation Using XAI Techniques, in *Proceedings – 2024 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems: Harmonizing Signals, Data, and Energy: Bridging the Digital Future, SPICES 2024*, Institute of Electrical and Electronics Engineers Inc., (2024). <https://doi.org/10.1109/SPICES62143.2024.10779829>
30. Lundberg, S. & Lee, S. I. A Unified Approach to Interpreting Model Predictions, Nov. [Online]. (2017). Available: <http://arxiv.org/abs/1705.07874>
31. Ahmed, S., Nobel, S. N. & Ullah, O. An Effective Deep CNN Model for Multiclass Brain Tumor Detection Using MRI Images and SHAP Explainability, in *3rd International Conference on Electrical, Computer and Communication Engineering, ECCE*, Institute of Electrical and Electronics Engineers Inc., 2023., Institute of Electrical and Electronics Engineers Inc., 2023. (2023). <https://doi.org/10.1109/ECCE57851.2023.10101503>
32. Nahiduzzaman, M. et al. A hybrid explainable model based on advanced machine learning and deep learning models for classifying brain tumors using MRI images. *Sci. Rep.* **15** (1). <https://doi.org/10.1038/s41598-025-85874-7> (Dec. 2025).
33. Tan, M. & Le, Q. V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, Sep. [Online]. (2020). Available: <http://arxiv.org/abs/1905.11946>
34. Srivastava, N., Hinton, G. E., Krizhevsky, A., Hinton, G. & Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **15** (1), 1929–1958. <https://doi.org/10.5555/2627435.2670313> (2014).
35. Gao Huang, D., Sedra, Y. S., ZhuangLiu & Weinberger, K. Q. Deep Networks with Stochastic Depth. *Springer Verlag*. <https://doi.org/10.1007/978-3-319-46493-0> (2016).
36. Samek, W., Binder, A., Montavon, G., Bach, S. & Müller, K. R. Evaluating the visualization of what a Deep Neural Network has learned, Sep. [Online]. (2015). Available: <http://arxiv.org/abs/1509.06321>
37. Nickparvar, M. Brain Tumor MRI Dataset, Kaggle, 2026. Version 1. <https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset>.
38. Loshchilov, I. & Hutter, F. Decoupled Weight Decay Regularization, Jan. [Online]. (2019). Available: <http://arxiv.org/abs/1711.05101>
39. Dai, Y., Gieseke, F., Oehmcke, S., Wu, Y. & Barnard, K. Attentional feature fusion. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 3560–3569). (2021).
40. Lin, T. Y., RoyChowdhury, A. & Maji, S. Bilinear CNN models for fine-grained visual recognition. In *Proceedings of the IEEE international conference on computer vision* (pp. 1449–1457). (2015).
41. Tan, M., Pang, R. & Le, Q. V. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10781–10790). (2020).
42. Zebari, N. A. et al. A deep learning fusion model for accurate classification of brain tumours in Magnetic Resonance images. *CAAI Trans. Intell. Technol.* **9** (4), 790–804. <https://doi.org/10.1049/cit2.12276> (Aug. 2024).
43. Kakogeorgiou, I. & Karantzas, K. Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing. *Int. J. Appl. Earth Obs. Geoinf.* **103** <https://doi.org/10.1016/j.jag.2021.102520> (Dec. 2021).
44. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, Apr. [Online]. (2014). Available: <http://arxiv.org/abs/1312.6034>
45. Smilkov, D., Thorat, N., Kim, B., Viégas, F. & Wattenberg, M. SmoothGrad: removing noise by adding noise, Jun. [Online]. (2017). Available: <http://arxiv.org/abs/1706.03825>
46. Shrikumar, A., Greenside, P. & Kundaje, A. Learning Important Features Through Propagating Activation Differences, [Online]. (2017). Available: <http://goo.gl/RM8jvH>
47. Springenberg, J. T., Dosovitskiy, A., Brox, T. & Riedmiller, M. Striving for Simplicity: The All Convolutional Net, Apr. [Online]. (2015). Available: <http://arxiv.org/abs/1412.6806>
48. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks, **70**, (2017). <https://doi.org/10.5555/3305890.3306024>

Acknowledgements

The work was conducted in the Advanced Healthcare Lab, School of Electrical Engineering, Manipal Institute of Technology (MIT), Manipal Academy of Higher Education (MAHE), Karnataka, India. We thank MIT and MAHE for their support and cooperation.

Author contributions

Aryaman Kaprekar: Writing—original draft, Software, Methodology, Conceptualization. Anjan Gudigar: Writing—review & editing, Supervision. U.Raghavendra: Writing—review & editing, Supervision. Mahesh Anil Inamdar: Validation, Formal analysis. Massimo Salvi: Writing—review & editing, Validation. Priyanka: Validation, Formal analysis. U. R. Acharya: Writing—review & editing, visualization.

Funding

Open access funding provided by Manipal Academy of Higher Education, Manipal. The authors received NO FUNDING for this work.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2026