

Relational Explainability of Brain Tumor MRI: Linking Classification Attention to Tumor Masks

Ripa Akter

International University of Business Agriculture and Technology

Md. Abdul Awal

`abdul.awal@iubat.edu`

International University of Business Agriculture and Technology

Research Article

Keywords: Deep learning, allowbreakGrad-CAM++, allowbreakBrain Tumor Classification, allowbreakTumor Segmentation, allowbreakExplainable AI, allowbreakBrain tumor MRI, allowbreakAttention map validation, allowbreakSpatial correspondence analysis.

Posted Date: October 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11241855/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Relational Explainability of Brain Tumor MRI: Linking Classification Attention to Tumor Masks

Ripa Akter¹ and Md. Abdul Awal^{1*†}

¹Department of Computer Science and Engineering, International
University of Business Agriculture and Technology, Dhaka, Bangladesh.

*Corresponding author(s). E-mail(s): abdul.awal@iubat.edu;

Contributing authors: ripa65468@gmail.com;

[†]These authors contributed equally to this work.

Abstract

Attention-based explanations (e.g., Grad-CAM++ (Gradient-weighted Class Activation Mapping Plus)) are commonly used in medical imaging, but their anatomical validity is rarely evaluated quantitatively. In this work we propose a novel relational explainability framework which goes beyond the conventional qualitative attention visualization by quantitatively validating the spatial correspondence between the classifier attention and the segmented tumor regions. Brain tumor segmentation is performed using DeepLabV3+ with EfficientNet-B4 encoder and tumor classification is performed using EfficientNet-B3. The Grad-CAM++ is employed to produce class-discriminative attention maps and the Exact Top-20% attention region is derived for quantitative analysis. The framework is evaluated on brain MRI datasets of gliomas, meningiomas, pituitary tumors, and non-tumor cases. The proposed method achieved a Dice coefficient of 87.5% for tumor segmentation and classification accuracy of 99.3% on the BRISC 2025 dataset. The classification accuracy on external evaluation was 93.5%. The quantitative validation of the explainability was performed by comparing the Exact Top-20% Grad-CAM++ attention region with the segmented tumor region in terms of Relational IoU, Tumor Coverage, and CAM Precision. The proposed framework achieved an overall Relational IoU of 0.0712, Tumor Coverage of 77.09%, and CAM Precision of 7.16%. These findings demonstrate class-dependent spatial alignment between classification attention and delineated tumor regions, with substantial tumor coverage but markedly lower spatial precision, underscoring the complementary rather than identical roles of discriminative attention and anatomical segmentation. The framework also identified the complementary functions of classification attention and anatomical segmentation, offering a more comprehensive understanding of the model's behaviour. The

proposed framework is an architecture-independent validation strategy providing an objective and quantitative approach to evaluate attention-based explanations, contributing to improving the transparency, interpretability and trustworthiness of artificial intelligence systems for medical image analysis.

Keywords: Deep learning, Grad-CAM++, Brain Tumor Classification, Tumor Segmentation, Explainable AI, Brain tumor MRI, Attention map validation, Spatial correspondence analysis.

1 Introduction

Brain tumors are one of the most critical disorders of the central nervous system, and an accurate diagnosis is necessary for effective treatment planning and prognosis [1]. Magnetic resonance imaging (MRI) is widely used for brain tumor detection and characterization because of its detailed structural information on brain tissues. However, manual interpretations of MRI scans are time-consuming and can suffer from variability in terms of accuracy, consistency and efficiency [2]. These challenges have resulted in the development of reliable computational approaches for automated brain tumor analysis. Accurate localization of tumor type and boundaries and early detection are essential for treatment planning and patient management [3]. Therefore, the use of machine learning and artificial intelligence techniques has been increasing for assisting automated tumor diagnosis and assessment.

Machine learning (ML) [4] and deep learning (DL) models have been quite successful in medical image analysis. In particular, convolutional neural networks (CNNs) can learn hierarchical feature representations from large-scale datasets automatically, and are therefore very powerful for different medical imaging applications [5]. These models have been extensively used in brain tumor classification and segmentation tasks [6]. Classification models usually classify between glioma, meningioma, pituitary tumor, and non-tumor cases [7]. Segmentation models provide the exact tumor boundaries at the pixel level. Although deep learning models have demonstrated remarkable performance, they are frequently criticized as “black-box” systems because their decision-making process is difficult to interpret [8]. This opacity can restrict clinical trust and impede the deployment of automated diagnostic systems in practical healthcare settings [9].

Explainable artificial intelligence (XAI) methods have been proposed to understand the decision-making behavior of deep learning models to address this problem [10]. Of these, gradient-based saliency methods, such as Grad-CAM++, are commonly employed to visualize the class-discriminative regions that affect model predictions [11]. Such explanations are often used as evidence of model interpretability, but their evaluation is often qualitative. Hence, it remains unclear whether the highlighted regions really correspond to anatomically relevant tumor structures. This may lead to a model that is very accurate in classification but also depends on irrelevant image features or spurious correlations.

To overcome this limitation, we propose a segmentation-guided relational explainability framework that quantifies the spatial correspondence between classification attention maps and tumor regions. We present a framework consisting of two individual deep learning pipelines. The classification pipeline utilizes EfficientNet-B3 [12] and aims to classify brain MRI images into glioma, meningioma, pituitary tumor, and non-tumor categories, as well as generate Grad-CAM++ attention maps to highlight regions influencing model predictions [13]. The segmentation pipeline is based on DeepLabV3+ with an EfficientNet-B4 encoder [14] to provide detailed pixel-wise tumor masks. As opposed to multitask architectures that learn classification and segmentation jointly, using independent pipelines allows objective post-hoc validation of classifier attention against tumor regions obtained through segmentation.

We anticipate that the anatomical classification decisions will have quantifiable spatial correspondence between the Grad-CAM++ attention maps and the tumor regions supplied by the segmentation model. This relationship is quantified using Relational IoU, Tumor Coverage and CAM Precision, as well as qualitative attention-tumor visualizations [15]. 754 well-aligned MRI images shared between the classification and segmentation tasks are analyzed relationally. We conduct a Top- k sensitivity analysis with activation thresholds of 10%, 15%, 20%, 25% and 30%, to check the robustness of the chosen Top-20% operating point.

The main contributions of this work can be summarized as below:

- We propose an architecture-agnostic relational explainability framework guided by segmentation integrating EfficientNet-B3, Grad-CAM++, and DeepLabV3+ with an EfficientNet-B4 encoder for independent post-hoc validation.
- We quantitatively evaluate the attention-tumor correspondence using Relational IoU, Tumor Coverage and CAM Precision on 754 paired MRI images. We perform Top- k sensitivity analysis for robustness and justification of Top-20% operating point.
- We conduct internal and external validation and introduce a quantitative explainability-auditing strategy for more transparent and trustworthy medical imaging.

2 Related Work

In recent studies focusing on the analysis of brain tumors using MRI, several deep learning architectures have been investigated for classification and segmentation. Lv et al. proposed a multi-model framework that used various segmentation networks including U-Net, nnU-Net, TransUNet, DeepLab, SwinUNet, and MedFormer, as well as classification models like ResNet, DenseNet, and GoogleNet. However, their method obtained promising classification and segmentation performances, but did not incorporate explainability techniques, thus unable to assess whether the classifier focused on clinically meaningful tumor regions [16].

Kumar et al. proposed a CNN based framework for brain tumor classification and segmentation from T1 weighted MRI images. The model achieved classification accuracies of 86.23% using Adam and 81.6% with SGDM, highlighting the efficiency of CNN-based techniques in automated tumor diagnosis [17]. Akter et al. proposed a

CNN-based classification framework with a U-Net segmentation model for automatic brain tumor detection. The system achieved classification accuracies of 98.7% and 98.8%, demonstrating the potential of combining classification and segmentation for clinical applications [18].

Several works are dedicated to improving the classification performance. Gundogan et al. [26] used XGBoost and Grad-CAM for brain tumor classification and visual interpretation. However, their framework did not include a segmentation component, and they could not quantitatively evaluate if the classifier attention matched the tumor regions [19]. Lavanyadevi et al. [20] proposed a probabilistic neural network framework for MRI-based tumor diagnosis by combining GLCM-based feature extraction, PCA-based dimensionality reduction, and K-means segmentation. Kaldera et al. combined CNN-based classification and Faster R-CNN segmentation for glioma and meningioma detection, providing reliable region localization and excellent classification performance [21]. Vinta et al. proposed an Ensemble Deep Learning Network (EDLNet) that combined the Modified Faster R-CNN and DRCNN architectures and reported high classification accuracy on benchmark brain tumor datasets [22]. Madhavi et al. explored transfer learning based classification and segmentation using VGG16, ResNet50, EfficientNetB0 and U-Net architectures showing the effectiveness of transfer learning for MRI based tumor analysis [23]. Kanchanamala et al. proposed a hybrid QDCNNs-DMN framework by combining quantum dilated CNNs and deep maxout networks for automated brain tumor classification, which achieved strong diagnostic performance [24]. Sadr et al. proposed a deep learning framework using a 26-layer CNN and transfer learning with VGG16 and Xception models, achieving high classification accuracy on several benchmark datasets [25].

Although the classification, segmentation, and explainable artificial intelligence of brain tumors have advanced significantly, the existing research has largely focused on the predictive accuracy and segmentation quality, with limited attention to the reliability of the generated explanations. In particular, Grad-CAM-based explanations are often regarded as visual evidence of model interpretability, but their spatial correspondence with anatomically localized tumor regions is rarely evaluated quantitatively. Thus, satisfactory predictive performance does not necessarily mean that evidence used by a classifier is spatially related to the tumor.

This gap motivates a relational view of explainability, in which classifier attention is measured against a separately obtained tumor localization reference, rather than interpreted in isolation. To this end, this work adds quantitative attention—tumor correspondence as another layer of model evaluation. We examine this spatial relationship at different attention-selection thresholds to assess the consistency of classifier attention with tumor anatomy. This offers a complementary view to traditional classification and segmentation evaluation and enhances the evaluation of explanation reliability in brain tumor MRI analysis.

3 Methods

We present a framework for explainable analysis of brain tumors that combines architecturally independent classification and segmentation models [26] with a post-hoc

comparison step to see if classification attention matches tumor regions [27], as illustrated in Figure 1.

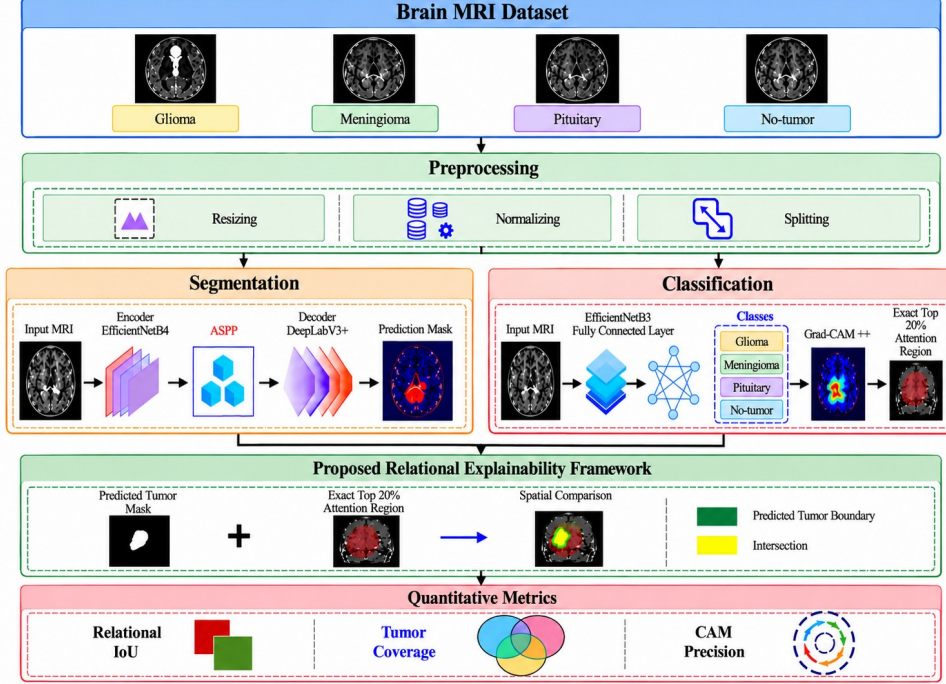


Fig. 1 Overview of the proposed explainable brain tumor analysis framework, comprising EfficientNet-B3 classification, DeepLabV3+ with EfficientNet-B4 segmentation, and post-hoc relational explainability. Grad-CAM++ attention is compared with tumor masks using Relational IoU, Tumor Coverage, and CAM Precision, with Top-k sensitivity analysis across 10–30% activation levels.

3.1 Dataset

In this study, we used the publicly available dataset BRISC 2025[28] obtained from Kaggle (<https://www.kaggle.com/datasets/briscdataset/brisc2025>) that provides MRI data for brain tumor classification and segmentation tasks. The dataset has four classes for the classification task, namely glioma, meningioma, non-tumor and pituitary tumor. Training set: 5,000 images for training. These are 1147 images of glioma, 1329 images of meningioma, 1067 images of non-tumor and 1457 images of pituitary. The original test set consists of 1,000 images including 254 glioma, 306 meningioma, 140 non-tumor and 300 pituitary.

We chose this dataset because it has both classification labels and segmentation masks. It simultaneously allows to evaluate classification and segmentation tasks. The flexibility is a significant advantage, especially when studying different types of tumors and performing comprehensive model evaluations. For the segmentation task,

the dataset contains T1-weighted MRI images and real-life tumor masks. The segmentation subset contains 3933 training images with masks and 860 test images with masks.

To ensure no data leakage between training and test sets, duplicate images were removed from the entire dataset. We found 106 duplicate images in the training and test sets and removed these from the test sets. After duplicate removal, the sizes of the classification test set were reduced from 1,000 to 894 images and the sizes of the segmentation test set were reduced from 860 to 754 images. This removed detected image-level duplicates between the training and test partitions, thereby decreasing the risk of direct image-level leakage.

For a relational analysis between classification attention maps and segmentation masks, we used a group of 754 images from the test sets of both tasks. This subset was extracted from the segmentation test set and matched with the corresponding samples of the classification test set in order to test both models on the same MRI scans.

To test the classification model on unseen data, Brain Tumor MRI Dataset[29] (Masoud Nickparvar, Kaggle) and BRISC 2025 dataset were used. This dataset comprises 1600 MRI images of different types of tumors and used only for the purpose of checking the accuracy of external classification. The model was not trained on the external dataset.

3.1.1 Preprocessing

For the classification task, all images in the data set were resized to 224×224 pixels before training the model. For the segmentation task, all images and the corresponding masks were resized to 256×256 pixels[30]. To perform classification and segmentation, data augmentation was performed by randomly flipping horizontally, rotating, and adjusting color jitter and contrast [31]. All the images were normalized with standard mean and standard deviation values to make the learning process more stable and speed up the convergence. The segmentation training set was divided 90-10 to form a validation set for tuning the model. We used 90% of the training images for model training, and the remaining 10% was used as a validation set during the tuning phase. Finally, the segmentation model is evaluated on a testing set to obtain the final performance. Moreover, the classification model employed 5-fold cross-validation for robust performance evaluation.

3.2 Classification Task

We used a pre-trained EfficientNet-B3 architecture [32] for multiclass tumor classification. The model was chosen for its high performance and effectiveness for multiclass image classification. Figure 2 shows the architecture of the pre-trained EfficientNet-B3 classification model.

Five-fold cross-validation was applied to the training set in the experiments[33]. The classification training set was split into five equal folds for this procedure. Five iterations of the model were trained, using four folds for training and one fold for model tuning each time. The test set was left unaltered for the final assessment at that time. The results were averaged across all five folds to determine the final performance

metrics, and the model’s stability was indicated by the corresponding mean $\pm$ standard deviation values. To measure the statistical reliability of the results obtained, 95% confidence intervals (CI) were also computed.

After completing the fivefold cross-validation, we ensured that the model’s architecture was tuned for maximum generalization, minimizing the risk of overfitting to any specific portion of the dataset. Then the final model was evaluated on the unseen test dataset to report the final performance. This evaluation technique decreases the dependence on a single training-validation split and enhances the estimation of the variability in the model performance. This is why we used five fold cross-validation to get a precise result, which helps in detecting the exact tumor type from the brain tumor classification.

An external validation experiment was carried out utilizing the independent Brain Tumor MRI Dataset (Masoud Nickparvar, Kaggle) to assess the generalization performance of the classification model. This dataset was solely utilized for classification evaluation and did not participate in model training.

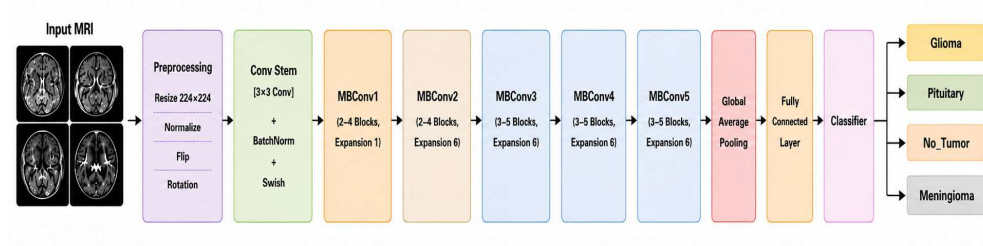


Fig. 2 Architecture of the EfficientNet-B3-based brain tumor classification model. The figure illustrates the EfficientNet-B3 backbone used for multi-class brain tumor classification, highlighting feature extraction, global pooling, and the final softmax-based tumor type prediction.

3.3 Segmentation Task

Tumor segmentation was conducted using DeepLabV3+ with a pre-trained EfficientNet-B4 encoder [34]. We chose EfficientNet-B4 due to its strong feature extraction ability providing features to the decoder for pixel-wise tumor segmentation. Figure 3 illustrates the architecture of the DeepLabV3+ model with the EfficientNet-B4 encoder.

The DeepLabV3+ decoder uses dilated convolutions and an Atrous Spatial Pyramid Pooling (ASPP) module to capture contextual information at multiple scales [35]. This allows the model to recognize tumour areas of different dimensions and shapes. The features extracted are gradually upsampled and fused with low level encoder features to preserve spatial details. Finally, a 1×1 convolution layer outputs the pixel-wise tumor mask [36].

The segmentation performance was evaluated using the Dice coefficient and Intersection over Union (IoU) metrics [37]. In applications of medical imaging, it has been demonstrated that the combination of EfficientNet-B4 and DeepLabV3+ can

produce accurate tumor masks . The Dice score is a popular metric that considers the class imbalance. The intersection over the union of the predicted tumor and the ground-truth tumor is a more stringent metric called the IoU.

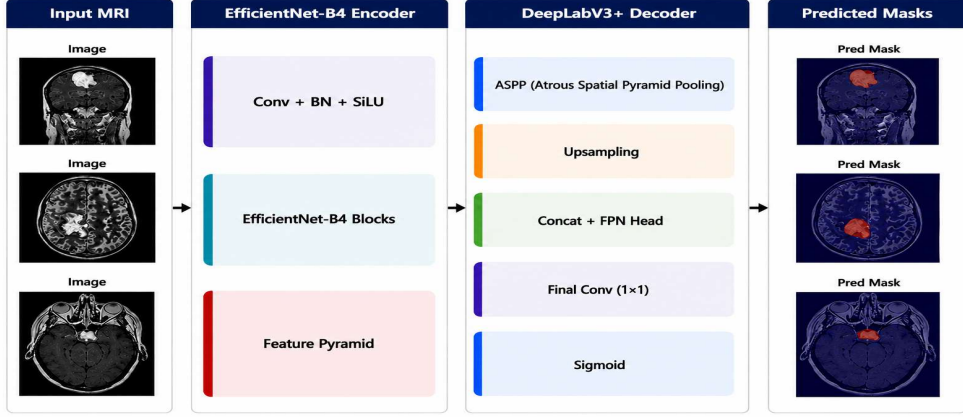


Fig. 3 Architecture of the DeepLabV3+-based brain tumor segmentation model. The segmentation framework employs an EfficientNet-B4 encoder for feature extraction and a DeepLabV3+ decoder with atrous spatial pyramid pooling (ASPP) to generate pixel-wise tumor masks from MRI images.

3.4 Explainable AI

The classification network predicts the tumour category of an input brain MRI image. However, the decision-making process of the classification network is inherently difficult to interpret, because deep learning models operate as black-box systems. Thus, the model does not directly inform which anatomical regions contribute most to the prediction. For all the predicted categories, including glioma, meningioma, pituitary tumor and non-tumor cases, we used Gradient-weighted Class Activation Mapping++ (Grad-CAM++) to improve the interpretability of the proposed classification framework.

Grad-CAM++ generates class-discriminative localization maps by exploiting the positive gradients of the predicted class score w.r.t. the feature maps of the last convolutional layer. The obtained gradient weights are then combined to generate a heatmap highlighting the regions of the image that contribute strongly to the predicted class. Compared to conventional Grad-CAM, Grad-CAM++ can give better localization of multiple discriminative regions and generate more accurate visual explanations. It is thus particularly suited to complex medical image classification tasks.

The resulting Grad-CAM++ attention maps provide intuitive visual explanations of the classification process by highlighting the most influential image regions for each prediction. However, these heatmaps can only be interpreted qualitatively, and do not objectively show whether the highlighted regions represent the actual tumor location. The generated attention maps are further used in the proposed quantitative relational analysis framework to measure the spatial correspondence to the tumor regions identified by the segmentation model.

3.5 Proposed Relational Explainability Framework

While Grad-CAM++ gives intuitive visual explanations of the classification process, it does not objectively determine whether the highlighted regions correspond to the actual tumor location. Thus, qualitative visualization is not enough to evaluate the spatial validity of attention-based explanations in brain tumor classification. To address this limitation, a quantitative relational explainability framework is proposed to assess the spatial correspondence between the classification attention and the tumor regions detected by the segmentation model.

For each MRI image, the trained classification model first produces a class-specific Grad-CAM++ attention map associated with the predicted tumor category. The resulting attention map is normalized to the range $[0,1]$. After that, only the top 20% pixels with the highest activation are kept to form an Exact Top-20% attention region. This strategy reliably retains the most discriminative regions while attenuating low-activation responses, which allows a robust and reproducible spatial comparison across all samples.

The resulting exact Top-20% attention region is then compared spatially with the predicted tumor mask generated by the segmentation model. This comparison allows us to objectively assess whether the discriminative regions detected by the classification model are spatially corresponding to the tumor regions localized by the segmentation model.

To quantitatively describe this spatial relationship, three complementary evaluation metrics are adopted, namely Relational IoU, Tumor Coverage, and CAM Precision. Together, these metrics measure the degree of spatial overlap between the classification attention and the segmented tumor region, the fraction of the tumor region that is covered by the attention map, and the fraction of the attention region that lies within the predicted tumor area.

The mathematical formulations of these evaluation metrics are given in the Performance Evaluation section. The proposed framework enhances conventional visual explainability with an objective quantitative evaluation of the spatial consistency between classification attention and segmentation-based tumor localization, which allows for a more thorough interpretation of a model’s decision-making process.

To assess the robustness of the chosen attention threshold, an additional sensitivity analysis was carried out on the predicted tumor masks provided for the 754 matched images. Grad-CAM++ attention regions were evaluated at five activation thresholds (10 %, 15 %, 20 %, 25 % and 30 %) where the pixels with the highest activation at each threshold were retained. For each threshold we calculated Relational IoU, Tumor Coverage and CAM Precision in relation to the corresponding predicted tumor mask. This analysis quantified the sensitivity of the relational measures to the choice of activation threshold and offered empirical support for the selection of the Top-20% operating point.

3.6 Performance Evaluation

The performance of classification, segmentation, and relational analysis is determined by using different metrics.

Classification performance was evaluated by 5-fold cross-validation, and accuracy, precision, recall, F1 score, standard deviation, and 95% confidence interval of the final model. These results are calculated and reported with the Confusion matrix . We also present Receiver Operating Characteristic curves (ROC curves), showing the classification performance of each class.

Accuracy is defined as.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Precision shows the proportion of samples that are genuinely positive. Precision is calculated as:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

Recall indicates the proportion of actual positive samples that are correctly predicted and is calculated as:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

The F1 score is defined as the harmonic average of the precision and recall scores. The F1 score is denoted as:

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

Standard deviation quantifies the dispersion of performance metrics around their mean across multiple cross-validation folds. The standard deviation is calculated as:

$$\sigma = \sqrt{\frac{1}{K-1} \sum_{i=1}^K (x_i - \mu)^2} \quad (5)$$

95% confidence interval is denoted as:

$$\text{CI}_{95\%} = \mu \pm t_{0.975, n-1} \frac{\sigma}{\sqrt{n}} \quad (6)$$

where μ is the mean and σ is standard deviation.

The color-coded Grad-CAM++ maps show the contribution intensity of different spatial regions to the model prediction. Regions with high activation and high influence on the predicted class are shown with warmer colors (red, yellow) and regions with lower contribution are shown with cooler colors (green to blue). The clustering of warm regions over tumor regions is a good indication that the proposed model is focused on clinically relevant structures.

Training optimization was carried out with Dice Loss as the objective. The performance of the segmentation was evaluated using the Dice coefficient and Intersection over Union (IoU).

$$\text{DiceLoss} = 1 - \frac{2 \sum_i p_i g_i}{\sum_i p_i + \sum_i g_i + \epsilon} \quad (7)$$

$$\text{IoU} = \frac{\sum_i p_i g_i}{\sum_i p_i + \sum_i g_i - \sum_i p_i g_i + \epsilon} \quad (8)$$

where p_i and g_i denote the predicted and ground truth labels, respectively, and ϵ is a small constant added to avoid division by zero.

We employ three complementary evaluation metrics, namely Relational IoU, Tumor Coverage and CAM Precision, to quantitatively evaluate the spatial overlap between classification attention and tumor regions detected by segmentation model. In the following equations, A is the Exact Top-20% Grad-CAM++ attention region and B is the predicted tumor mask generated by the segmentation model.

The relational IoU computes the spatial intersection between the attention region and the predicted tumor mask defined as:

$$\text{Relational IoU} = \frac{|A \cap B|}{|A \cup B|} \quad (9)$$

where a larger Relational IoU denotes the more spatial consistency between the classification attention and the segmented tumor region.

Tumor Coverage is the percentage of the predicted tumor region covered by the attention region, defined as

$$\text{Tumor Coverage} = \frac{|A \cap B|}{|B|} \times 100 \quad (10)$$

Higher Tumor Coverage indicates that the classification model looks at a larger percentage of the segmented tumor region.

CAM Precision is calculated as the percentage of the attention region over the predicted tumor mask, i.e.,

$$\text{CAM Precision} = \frac{|A \cap B|}{|A|} \times 100 \quad (11)$$

Higher CAM Precision indicates more discriminative attention is focused in the predicted tumor region.

These three complementary metrics give a complete quantitative assessment of the spatial concordance of classification attention and segmentation-based tumor localization.

3.7 Experimental Setup

The experiments in this study were executed using Python and PyTorch in a GPU-accelerated Google Colab environment. We used classification experiments with a starting learning rate of 1e-4, a batch size of 16, and 30 epochs of training. The learning rate for segmentation was 1e-4, and there were 30 epochs with a batch size 16.

4 Results and Discussion

4.1 Classification

The classification model obtained an accuracy of 0.993 on the internal test dataset, which means a very low misclassification rate. High values of precision, recall and F1-score were also observed, indicating effective discrimination between the four classification categories. Figure 4 shows the training and validation accuracy and loss curves, demonstrating stable convergence during training.

The results obtained from the 5-fold cross validation are shown in Table 1. The performance was similar across the folds, suggesting the model did not rely too heavily on a specific data split. This suggests that the feature representation learned was stable during the cross-validation process.

Table 2 shows the mean $\pm$ standard deviation of the cross-validation performance and the corresponding 95% confidence intervals. The small standard deviations and tight confidence intervals indicate little variation across folds and support the stability of the classification model.

Table 3 lists the class-wise precision, recall and F1-score values that were obtained from the independent internal test set. The results show balanced performance for glioma, meningioma, pituitary tumor and non tumor categories which indicates that the model was able to distinguish between different tumor types with high accuracy.

The trained EfficientNet-B3 classifier was evaluated on an independent Brain Tumor MRI dataset without any further training or fine-tuning for better understanding of model generalization. The external dataset consisted of 1,600 MRI images of four classes such as glioma, meningioma, pituitary tumor and non-tumor. The results of external validation are reported in Table 4 for each class.

The model achieved an overall external validation accuracy of 0.935. Pituitary tumors were the best performing class, with a precision of 0.957, a recall of 0.993, and an F1-score of 0.974. The non-tumor class also performed well with a precision of 0.957, recall of 0.943 and F1-score of 0.950. The model attained a recall of 0.988 and an F1-score of 0.924 for meningioma. The precision of the glioma class was still high (0.973) but the recall was relatively low (0.818), which implied that part of the glioma samples were misclassified to other tumor classes.

The confusion matrices for the internal and external evaluation are shown in Figure 5. The internal results show a good classification performance on a class-wise basis, while the external evaluation shows a relatively higher number of misclassifications, especially for the glioma samples.

The internal test accuracy was reached as 0.993, however, the external validation accuracy dropped to 0.935. In cross-dataset evaluation, such performance differences are expected and may be attributed to differences in MRI acquisition protocols, scanner characteristics, image quality and tumor appearance. These factors can lead to distribution shifts between datasets and affect the model predictions on unseen data.

The ROC curves for both internal and external evaluations are presented in Fig. 6 and also show the discriminative performance of the classifier for the four categories. In summary, the classification results demonstrate that the proposed model achieves good performance on both internal and external datasets. The high internal accuracy,

stable cross validation results and competitive external validation performance indicate that the model can well discriminate different brain tumor categories and has good generalization capability.

Table 1 The proposed classification model’s five-fold cross-validation performance shows constant accuracy, precision, recall, and F1-score across folds, demonstrating the model’s stability throughout training and assessment.

Fold	Accuracy	Precision	Recall	F1-score
1	0.989	0.9893	0.9885	0.9889
2	0.979	0.9793	0.9792	0.9792
3	0.989	0.9889	0.9891	0.9890
4	0.969	0.9692	0.9704	0.9690
5	0.986	0.9872	0.9859	0.9865

Table 2 Overall evaluation of the model’s stability and robustness across various evaluation metrics is provided by summary statistics of the five-fold cross-validation results, such as mean, standard deviation, and 95% confidence intervals.

Metric	Mean	Std	95% CI (Lower)	95% CI (Upper)
Accuracy	0.9824	0.0076	0.9757	0.9891
Precision	0.9828	0.0077	0.9760	0.9895
Recall	0.9826	0.0070	0.9764	0.9888
F1-score	0.9825	0.0076	0.9758	0.9892

Table 3 Using the independent test dataset, the proposed model’s class-wise classification performance showed excellent generalization ability by preserving high precision, recall, and F1-score for every tumor class.

Class	Precision	Recall	F1-score	Support
Glioma	1.000	0.996	0.998	254
Meningioma	0.987	0.997	0.992	207
No Tumor	0.993	1.000	0.996	140
Pituitary	1.000	0.990	0.995	293
Macro Average	0.995	0.996	0.995	894
Weighted Average	0.995	0.995	0.995	894
Accuracy		0.993		894

Table 4 External validation performance of the proposed classification model on the Brain MRI dataset (1600 images).

Class	Precision	Recall	F1-score	Support
Glioma	0.973	0.818	0.889	400
Meningioma	0.868	0.988	0.924	400
No Tumor	0.957	0.943	0.950	400
Pituitary	0.957	0.993	0.974	400
Macro Avg	0.939	0.935	0.934	1600
Weighted Avg	0.939	0.935	0.934	1600
Accuracy		0.935		1600

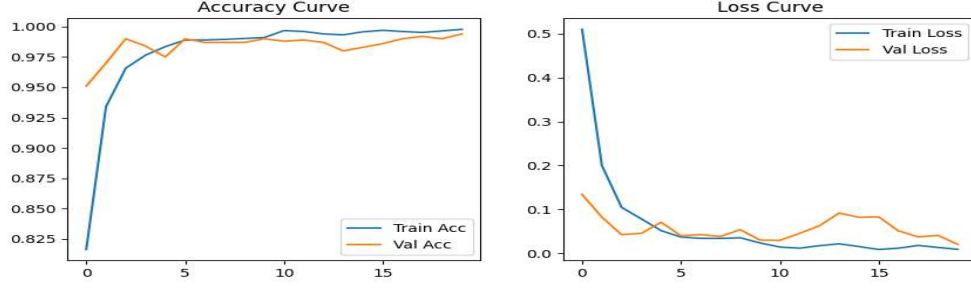


Fig. 4 Training and validation performance curves for the classification model. Accuracy and loss curves demonstrate stable convergence and consistent learning behavior during training.

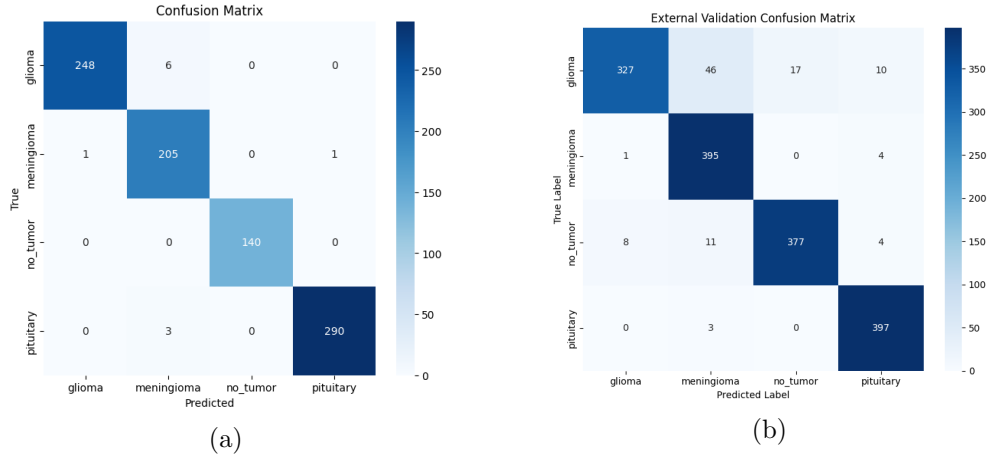


Fig. 5 Confusion matrices for the proposed brain tumor classification model. (a) Performance on the internal test dataset. (b) Performance on the independent external validation dataset. Rows denote ground-truth labels and columns denote predicted classes for glioma, meningioma, no-tumor, and pituitary tumors.

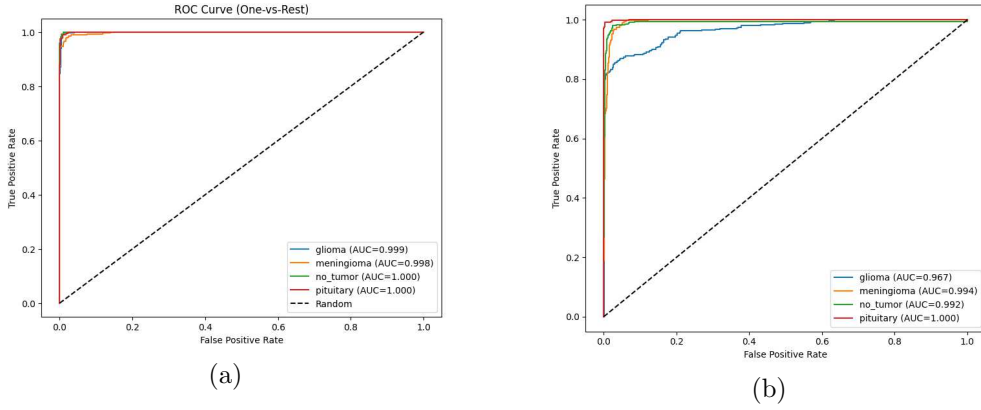


Fig. 6 ROC curves for the proposed one-versus-rest brain tumor classification model. (a) The internal test dataset's ROC curves. (b) ROC curves derived from the external, independent validation dataset. For glioma, meningioma, no-tumor, and pituitary tumor, the curves show the model's class-wise discriminative performance; higher AUC values denote superior classification abilities.

4.1.1 Visual Explanation

We used Grad-CAM++ to make visualizations of representative samples from each tumor class to look at the parts of the image that affected the classifier's predictions. The activation maps show areas of the MRI images that are close to the tumor, with little response seen in the surrounding background tissue.

The attention maps stay focused on structures related to lesions in all categories, but their spatial distributions depend on the class. In some cases, there is a lot of activity close to the tumor core, while in others, there is activity that spreads across larger areas of the lesion. Despite these differences, the activations remain restricted to anatomically relevant areas associated with the predicted class.

Figure 7 shows representative Grad-CAM++ visualization for the four classification categories. Attention responses are mostly clustered around lesion-relevant areas, with some variation in their spatial distributions across classes.

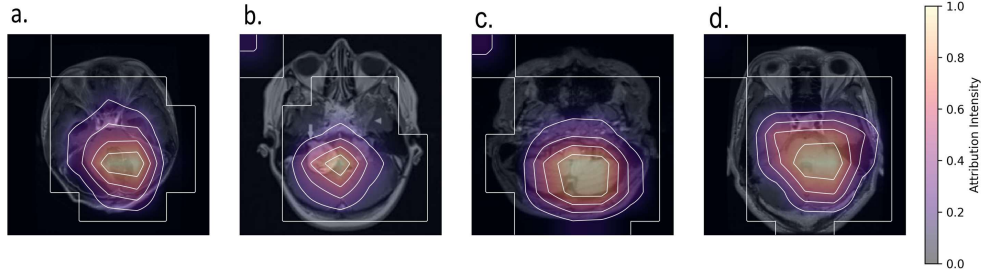


Fig. 7 Grad-CAM++ heatmaps show the areas of space that help the classification model make its predictions. For (a) glioma, (b) meningioma, (c) no-tumor, and (d) pituitary tumor, warmer colors mean that the predicted class has a higher activation intensity.

4.2 Segmentation

The segmentation model achieved 78% IoU score and 87.5% Dice score which shows the predicted masks are in good agreement with the ground-truth tumor regions. Such results demonstrate that the model was able to capture the tumor boundaries reasonably well across different MRI samples.

The qualitative results were consistent with the quantitative results. By visual inspection, the predicted masks followed the tumor regions and were able to identify most tumor boundaries, even in cases with irregular shapes and varying image contrast. The segmentation quality was generally acceptable, though a few difficult cases revealed some discrepancies in boundaries.

The results of the segmentation are stable over the dataset. Representative MRI images with ground-truth and predicted masks are shown in Figure 8 to demonstrate the ability of the model to capture the tumor regions. The training and validation Dice and IoU curves are shown in Figure 9 which shows that the learning is stable. The predicted masks were then used as spatial references in the proposed explainability analysis based on relations.

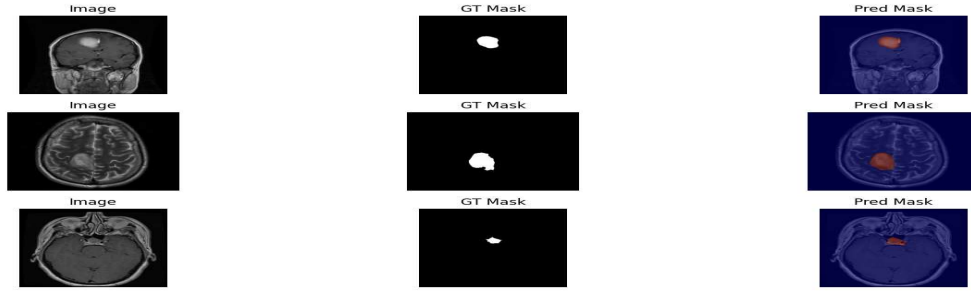


Fig. 8 Qualitative segmentation results on brain MRI images. Representative examples of input MRI scans, corresponding ground-truth masks, and predicted tumor segmentation outputs demonstrating the model’s ability to delineate tumor boundaries.

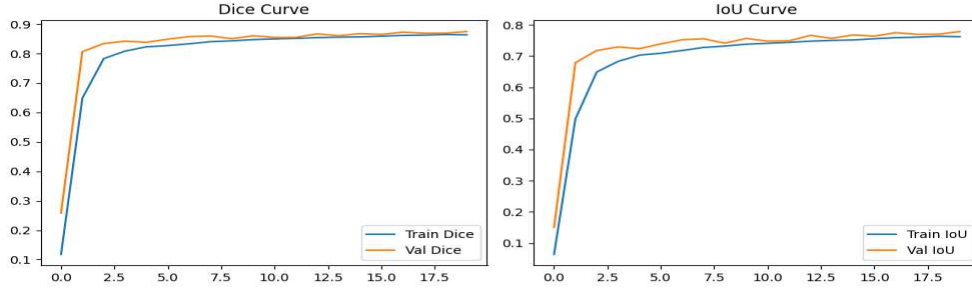


Fig. 9 Validation and training Curves for the segmentation model for the Dice and IoU. The curves illustrate the learning progression and stability of the DeepLabV3+-based segmentation model during training.

4.3 Relational Analysis Performance

A relational analysis was performed to investigate the spatial relationship between classification decisions and segmentation-based tumor localization using 754 exactly matched MRI images shared between the classification and segmentation tasks. The predicted tumor segmentation mask was compared with the discriminative region obtained by Grad-CAM++ for each image. We kept the highest-activation 20% of Grad-CAM++ pixels for each image in an effort to define classifier attention in a consistent and reproducible way. We used three complementary metrics to quantify spatial concordance: Relational IoU, Tumor Coverage, and CAM Precision.

The outcome for each class is summarized in Table 5. Meningioma obtained the highest Relational IoU (0.1078 ± 0.0792) and CAM Precision ($10.81 \pm 8.01\%$), followed by glioma (0.0953 ± 0.0787 and $9.60 \pm 7.93\%$, respectively). Pituitary tumors showed significantly lower values for both metrics (0.0244 ± 0.0255 and $2.48 \pm 2.59\%$). Conversely, Tumor Coverage was high for meningioma ($99.24 \pm 5.94\%$) and glioma ($93.40 \pm 15.80\%$), but low and more variable for pituitary tumors ($47.37 \pm 42.34\%$).

Figure 10 shows the class-wise Relational IoU, Tumor Coverage, and CAM Precision values, and Figure 11 shows representative examples of the spatial relationship between classifier attention and tumor regions. These visual findings further demonstrate the variation in spatial concordance between the different tumor classes.

The overall mean Relational IoU, Tumor Coverage and CAM Precision of the 754 matched MRIs were 0.0712, 77.09% and 7.16%, respectively. These overall values were derived from sample-level measurements and the class-wise results are presented separately as mean $\pm$ standard deviation. The lower concordance for pituitary tumors may partly be due to the small lesion regions which may have a great influence on overlap based metrics when small spatial differences exist. In general, the high Tumor Coverage but relatively low Relational IoU and CAM Precision for glioma and meningioma suggests that classifier attention often covers the predicted tumor regions, while extending beyond their exact segmentation boundaries. Thus, these metrics describe complementary aspects of spatial concordance and should not be taken as direct measures of segmentation accuracy.

Each spatial concordance metric was assessed separately with the Kruskal-Wallis test. As shown in Table 6, Relational IoU ($H = 320.2205$, $p = 2.92 \times 10^{-70}$), Tumor Coverage ($H = 297.5393$, $p = 2.46 \times 10^{-65}$), and CAM Precision ($H = 316.3329$, $p = 2.04 \times 10^{-69}$) had significant class-wise differences, indicating significant differences in spatial concordance between classification attention and predicted tumor localization across tumor classes.

Table 5 Class-wise spatial concordance between predicted tumor segmentation masks and Exact Top-20% Grad-CAM++ attention regions. Values are reported as mean $\pm$ standard deviation.

Tumor Class	N	Relational IoU	Tumor Coverage (%)	CAM Precision (%)
Glioma	254	0.0953 ± 0.0787	93.40 ± 15.80	9.60 ± 7.93
Meningioma	207	0.1078 ± 0.0792	99.24 ± 5.94	10.81 ± 8.01
Pituitary	293	0.0244 ± 0.0255	47.37 ± 42.34	2.48 ± 2.59

Class-wise Quantitative Evaluation of Spatial Concordance

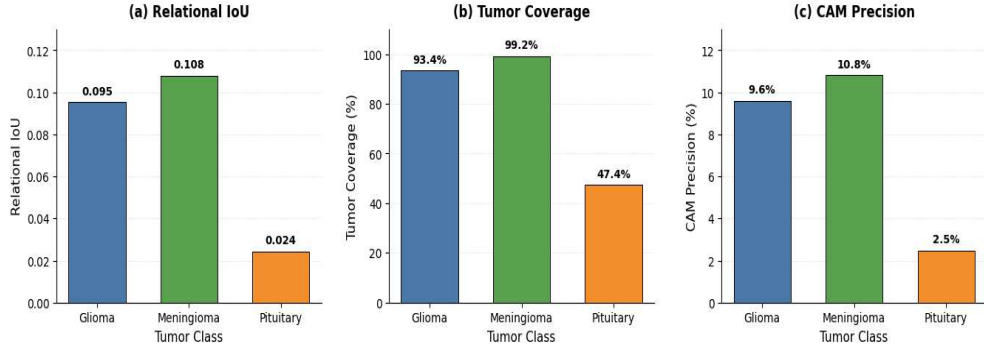


Fig. 10 Class-wise quantitative evaluation of the spatial relationship between the Exact Top-20% Grad-CAM++ attention regions and the predicted tumor masks using (a) Relational IoU, (b) Tumor Coverage, and (c) CAM Precision.

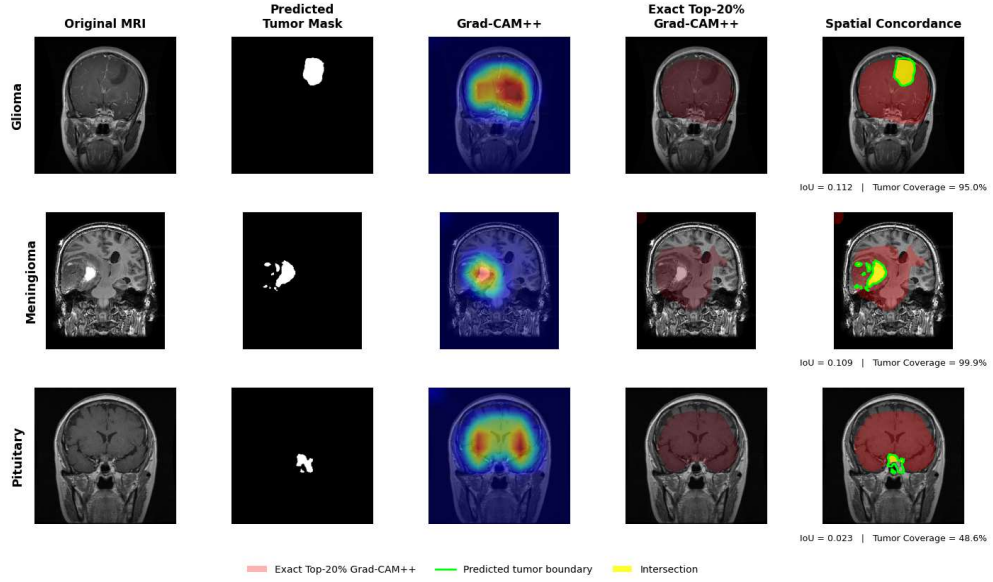


Fig. 11 Representative qualitative visualization of spatial concordance between classification attention and segmentation-based tumor localization for glioma, meningioma, and pituitary tumors. The visualization presents the original MRI, predicted tumor mask, Grad-CAM++ activation, Exact Top-20% attention region, and their spatial concordance.

Table 6 Kruskal–Wallis test results evaluating class-wise differences in the spatial concordance metrics (Relational IoU, Tumor Coverage, and CAM Precision) between the Exact Top-20% Grad-CAM++ attention regions and the predicted tumor masks.

Metric	H Statistic	p-value	Significance
Relational IoU	320.2205	2.92×10^{-70}	Significant
Tumor Coverage (%)	297.5393	2.46×10^{-65}	Significant
CAM Precision (%)	316.3329	2.04×10^{-69}	Significant

4.3.1 Top-k Sensitivity Analysis

A sensitivity analysis on the 754 matched MRI images was conducted to evaluate the robustness of the Top-20% Grad-CAM++ threshold selected. For each threshold, the top 10%, 15%, 20%, 25% and 30% activation pixels were kept and Relational IoU, Tumor Coverage and CAM Precision were computed against the respective predicted segmentation masks.

The overall results are summarized in Table 7. The Relational IoU dropped from 0.1110 ± 0.1185 at 10% to 0.0556 ± 0.0478 at 30%, and the CAM Precision from $11.65 \pm 12.71\%$ to $5.57 \pm 4.82\%$. Conversely, Tumor Coverage increased from $61.50 \pm 37.61\%$ to $97.86 \pm 9.38\%$. The framework obtained a Relational IoU of 0.0769 ± 0.0687 , Tumor Coverage of $89.98 \pm 20.24\%$ and a CAM Precision of $7.76 \pm 6.99\%$ at the selected Top-20% threshold.

These trends are shown in Figure 12. When the retained activation percentage increases, pixels of lower activation are considered, resulting in increased tumor coverage but decreased spatial precision and overlap. This shows a trade-off between anatomical coverage and attention localization at the tested thresholds. The consistency of the directional trends across the five thresholds suggests that the observed relational patterns are not unique to a particular threshold. The Top-20% setting thus constitutes a trade-off between tumor coverage and spatial concentration.

Table 7 Sensitivity analysis of relational explainability metrics against Top- k Grad-CAM++ activation thresholds using predicted segmentation masks on 754 matched MRI images. Values are expressed as mean $\pm$ standard deviation.

Top- k (%)	Relational IoU	Tumor Coverage (%)	CAM Precision (%)
10	0.1110 ± 0.1185	61.50 ± 37.61	11.65 ± 12.71
15	0.0924 ± 0.0888	79.39 ± 29.95	9.41 ± 9.13
20	0.0769 ± 0.0687	89.98 ± 20.24	7.76 ± 6.99
25	0.0650 ± 0.0562	95.56 ± 13.14	6.53 ± 5.68
30	0.0556 ± 0.0478	97.86 ± 9.38	5.57 ± 4.82

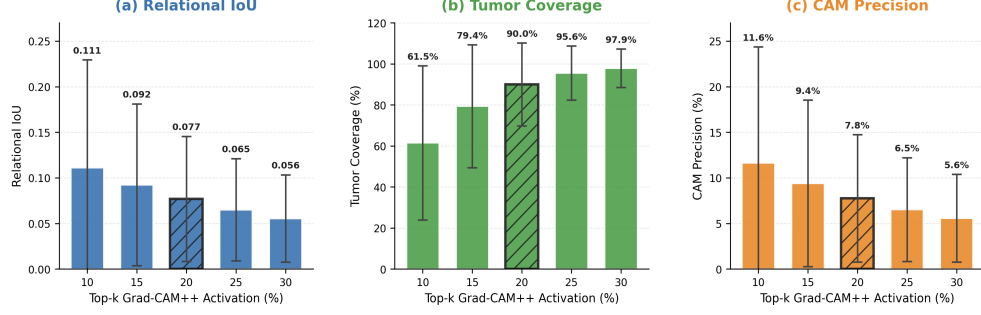


Fig. 12 Top-k sensitivity analysis of the predicted segmentation mask over 754 matched MRI images. We report Relational IoU, Tumor Coverage and CAM Precision for 10%, 15%, 20%, 25% and 30% Grad-CAM++ activation thresholds. Bars are means $\pm$ standard deviation; hatched bars are the selected Top-20% threshold.

5 Conclusion

In this work we address an important limitation of attention-based explainability in brain tumors MRI analysis: A visually convincing attention map does not prove that the classifier is focusing on anatomically meaningful tumor regions. To fill this gap, we propose a relational explainability framework that integrates independently trained classification and segmentation models into a unified post-hoc validation pipeline. The framework not only uses Grad-CAM++ as qualitative visual evidence but also quantitatively examines the spatial relationship of Grad-CAM++ with tumor localization by means of Relational IoU, Tumor Coverage, and CAM Precision. The classification and segmentation models exhibited excellent predictive and localization performance, which provided a solid basis for this relational assessment.

The experimental results demonstrated spatial correspondence between classifier attention and tumor regions that are measurable yet class-dependent. Importantly, the analysis showed that high tumor coverage does not always translate into high spatial precision, highlighting a fundamental difference between discriminative attention for classification and full anatomical delineation by segmentation. This result shows the value of assessing explanations beyond visual inspection and provides a more objective way to study which spatial evidence contributes to a model’s predictions. We propose a framework that separates the validation of explanations from the underlying model architectures, offering an architecture-agnostic and reproducible strategy that can be applied to other medical imaging tasks where the anatomical fidelity of AI explanations is a critical requirement.

However, the framework was evaluated on public datasets and 2D MRI slices, and the generated attention maps were not validated by experienced radiologists in the clinical setting. Future work will therefore investigate multi-institutional and 3D volumetric data and integrate clinician-guided evaluations to further assess the robustness, clinical relevance, and practical applicability of quantitative relational explainability in medical imaging.

References

- [1] Khalighi, S., Reddy, K., Midya, A., Pandav, K.B., Madabhushi, A., Abedalthagafi, M.: Artificial intelligence in neuro-oncology: advances and challenges in brain tumor diagnosis, prognosis, and precision treatment. *NPJ precision oncology* **8**(1), 80 (2024)
- [2] Khalifa, M., Albadawy, M.: Ai in diagnostic imaging: revolutionising accuracy and efficiency. *Computer Methods and programs in biomedicine update* **5**, 100146 (2024)
- [3] Amin, J., Sharif, M., Haldorai, A., Yasmin, M., Nayak, R.S.: Brain tumor detection and classification using machine learning: a comprehensive survey. *Complex & intelligent systems* **8**(4), 3161–3183 (2022)
- [4] Janiesch, C., Zschech, P., Heinrich, K.: Machine learning and deep learning. *Electronic markets* **31**(3), 685–695 (2021)
- [5] Krichen, M.: Convolutional neural networks: A survey. *Computers* **12**(8), 151 (2023)
- [6] Yao, X., Wang, X., Wang, S.-H., Zhang, Y.-D.: A comprehensive survey on convolutional neural network in medical image analysis. *Multimedia Tools and Applications* **81**(29), 41361–41405 (2022)
- [7] Alsajri, A., Hacimahmud, A.V.: Review of deep learning: Convolutional neural network algorithm. *Babylonian Journal of Machine Learning* **2023**, 19–25 (2023)
- [8] Biratu, E.S., Schwenker, F., Ayano, Y.M., Debelee, T.G.: A survey of brain tumor segmentation and classification algorithms. *Journal of Imaging* **7**(9), 179 (2021)
- [9] Dhar, T., Dey, N., Borra, S., Sherratt, R.S.: Challenges of deep learning in medical image analysis—improving explainability and trust. *IEEE Transactions on Technology and Society* **4**(1), 68–75 (2023)
- [10] Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., Hussain, A.: Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation* **16**(1), 45–74 (2024)
- [11] Apostolopoulos, I.D., Athanasoula, I., Tzani, M., Groumpos, P.P.: An explainable deep learning framework for detecting and localising smoke and fire incidents: evaluation of grad-cam++ and lime. *Machine Learning and Knowledge Extraction* **4**(4), 1124–1135 (2022)
- [12] Batool, A., Byun, Y.-C.: Lightweight efficientnetb3 model based on depthwise separable convolutions for enhancing classification of leukemia white blood cell

- images. IEEE access **11**, 37203–37215 (2023)
- [13] Hussain, T., Shouno, H.: Explainable deep learning approach for multi-class brain magnetic resonance imaging tumor classification and localization using gradient-weighted class activation mapping. Information **14**(12), 642 (2023)
 - [14] JS, N., *et al.*: Brain tumor segmentation using multi-scale attention u-net with efficientnetb4 encoder for enhanced mri analysis. Scientific Reports **15**(1), 1–20 (2025)
 - [15] Murrugarra-Llerena, J., Kirsten, L.N., Zeni, L.F., Jung, C.R.: Probabilistic intersection-over-union for training and evaluation of oriented object detectors. IEEE Transactions on Image Processing **33**, 671–681 (2024)
 - [16] Lv, C., Shu, X.-J., Liang, Q., Qiu, J., Xiong, Z.-C., Ye, J., Li, S., Liu, C.Q., Niu, J.Z., Chen, S.-B., *et al.*: Braintumnet: multi-task deep learning framework for brain tumor segmentation and classification using adaptive masked transformers. Frontiers in Oncology **15**, 1585891 (2025)
 - [17] Kumar, S., Kumar, D.: Human brain tumor classification and segmentation using cnn. Multimedia Tools and Applications **82**, 7599–7620 <https://doi.org/10.1007/s11042-022-13713-2>
 - [18] Akter, A., Nosheen, N., Ahmed, S., Hossain, M., Yousuf, M.A., Almoayad, M.A.A., Hasan, K.F., Moni, M.A.: Robust clinical applicable cnn and u-net based algorithm for mri classification and segmentation for brain tumor. Expert Systems with Applications **238**, 122347
 - [19] Gundogan, E.: A novel hybrid deep learning model enhanced with explainable ai for brain tumor multi-classification from mri images. Applied Sciences **15**(10), 5412 (2025)
 - [20] Lavanyadevi, R., Machakowsalya, M., Nivethitha, J., Kumar, A.N.: Brain tumor classification and segmentation in mri images using pnn. In: Proceedings of the IEEE International Conference on Electrical, Instrumentation and Communication Engineering (ICEICE), pp. 1–6. IEEE
 - [21] Kaldera, H.N.T.K., Gunasekara, S.R., Dissanayake, M.B.: Brain tumor classification and segmentation using faster r-cnn. In: Proceedings of the Advances in Science and Engineering Technology International Conferences (ASET), pp. 1–6. IEEE
 - [22] Vinta, S.R., Chintalapati, P.V., Babu, G.R., Tamma, R., Sai Chaitanya Kumar, G.: Edlnet: ensemble deep learning network model for automatic brain tumor classification and segmentation. Journal of Biomolecular Structure and Dynamics **43**(15), 8129–8141

- [23] Madhavi, K.R., Soujanya, K.L.S., Inaganti, R., Swamy, K.A., Yalavarthi, S., Munagala, M.: Brain tumor classification and segmentation using transfer learning from mri images. *International Journal of Computer Information Systems and Industrial Management Applications* **17**, 14–14
- [24] Kanchanamala, P., Kuppusamy, V., Ganesan, G.: Qdcnn-dmn: A hybrid deep learning approach for brain tumor classification using mri images. *Biomedical Signal Processing and Control* **101**, 107199
- [25] Sadr, H., Nazari, M., Yousefzade-Chabok, S., Emami, H., Rabiei, R., Ashraf, A.: Enhancing brain tumor classification in mri images: A deep learning-based approach for accurate classification and diagnosis. *Image and Vision Computing*, 105555
- [26] Mzoughi, H., Njeh, I., BenSlima, M., Farhat, N., Mhiri, C.: Vision transformers (vit) and deep convolutional neural network (d-cnn)-based models for mri brain primary tumors images multi-classification supported by explainable artificial intelligence (xai). *The Visual Computer* **41**(4), 2123–2142 (2025)
- [27] Rasool, N., Bhat, J.I.: A critical review on segmentation of glioma brain tumor and prediction of overall survival. *Archives of Computational Methods in Engineering* **32**(3), 1525–1569 (2025)
- [28] Fateh, A., Rezvani, Y., Moayedi, S., Rezvani, S., Fateh, F., Fateh, M.: Brisc: Annotated dataset for brain tumor segmentation and classification with swin-hafnet. *arXiv preprint arXiv:2506.14318* (2025)
- [29] Nickparvar, M.: Brain Tumor MRI Dataset. Kaggle (2026). <https://doi.org/10.34740/KAGGLE/DSV/14832123> . <https://www.kaggle.com/dsv/14832123>
- [30] Santos, C., Aguiar, M., Welfer, D., Dias, M., Pereira, A., Ribeiro, M., Belloni, B.: A new approach for fundus lesions instance segmentation based on mask r-cnn x101-fpn pre-trained architecture. *IEEE Access* **11**, 43603–43618 (2023)
- [31] Bajaj, S., Bala, M., Angurala, M.: A comparative analysis of different augmentations for brain images. *Medical & Biological Engineering & Computing* **62**(10), 3123–3150 (2024)
- [32] Ramakrishna, M.T., Pothanaicker, K., Selvaraj, P., Khan, S.B., Venkatesan, V.K., Alzahrani, S., Alojail, M.: Leveraging efficientnetb3 in a deep learning framework for high-accuracy mri tumor classification. *Computers, Materials & Continua* **81**(1) (2024)
- [33] Rastogi, D., Johri, P., Tiwari, V., Elngar, A.A.: Multi-class classification of brain tumour magnetic resonance images using multi-branch network with inception block and five-fold cross validation deep learning framework. *Biomedical Signal Processing and Control* **88**, 105602 (2024)

- [34] Lamba, K., Rani, S., Shabaz, M.: Synergizing advanced algorithm of explainable artificial intelligence with hybrid model for enhanced brain tumor detection in healthcare. *Scientific Reports* **15**(1), 20489 (2025)
- [35] Polat, H.: A modified deeplabv3+ based semantic segmentation of chest computed tomography images for covid-19 lung infections. *International Journal of Imaging Systems and Technology* **32**(5), 1481–1495 (2022)
- [36] Chowdary, G.J., G, S., M, P., Yogarajah, P.: Nucleus segmentation and classification using residual se-unet and feature concatenation approach incervical cytopathology cell images. *Technology in Cancer Research & Treatment* **22**, 15330338221134833 (2023)
- [37] Montazerolghaem, M., Sun, Y., Sasso, G., Haworth, A.: U-net architecture for prostate segmentation: The impact of loss function on system performance. *Bioengineering* **10**(4), 412 (2023)