RT-CAM Enhanced: A Comparative Study of Multi-Task Brain Tumor Classification Using Explainable Attention Mechanisms

Amani Lissir[1,2], Rafik Khemakhem[3,4], and Tarek Frikha[2,5]

[1]National School of Electronics and Telecommunications of Sfax, University of Sfax, Sfax, Tunisia
[2]Data Engineering and Semantics Unit, University of Sfax, Sfax, Tunisia. [3]ATMS Lab, ENIS, University of Sfax, Sfax, Tunisia
[4]ISG Gabès, University of Gabès, Gabès, Tunisia. [5]National Engineering School of Sfax, University of Sfax, Sfax, Tunisia
Email: amanilissir2020@gmail.com

***Abstract*—Brain tumor classification from magnetic resonance imaging (MRI) remains a challenging task due to the need for both high diagnostic accuracy and reliable interpretability to support clinical decision-making. This paper presents a systematic comparative study of three deep learning architectures for multi-task brain tumor classification and anatomical plane identification using the BRISC2025 dataset. We propose RT-CAM Enhanced, a novel reliability-aware explainable artificial intelligence (XAI) framework that introduces an adaptive Trust-Map fusion strategy. By dynamically weighting Grad-CAM, Grad-CAM++, and Layer-CAM, the framework generates robust and consistent visual explanations across different convolutional neural network architectures. The evaluated models include an EfficientNet-B0(DualAttention), EfficientNet-B0 and ConvNeXt-Tiny(DualAttention). Experimental results demonstrate that attention-enhanced architectures outperform their non-attention counterparts, with EfficientNet-B0(DualAttention) achieving classification accuracies of 99.60% for tumor classification and 99.50% for anatomical plane classification. Furthermore, RT-CAM Enhanced improves explainability quality in terms of faithfulness, localization accuracy, and consistency compared to individual CAM-based methods, effectively addressing the faithfulness-localization trade-off. These findings confirm that the proposed framework balances high predictive performance with interpretable and reliable visual explanations, supporting trustworthy automated brain tumor analysis.**

***Index Terms*—Brain Tumor Classification, Explainable AI, Class Activation Mapping, Multi-Task Learning, Attention Mechanisms, Comparative Architecture Study**

I. Introduction

Brain tumors represent a critical neurological condition with significant implications for patient survival and quality of life. Accurate diagnosis using magnetic resonance imaging (MRI) is essential for guiding treatment strategies and clinical decision-making. However, the heterogeneous nature of brain tumors, reflected in variations in morphology, texture, and appearance across different imaging planes, combined with the rapidly increasing volume of medical imaging data, makes manual interpretation labor-intensive and prone to error. Recent advances in deep learning have demonstrated substan tial potential for automated medical image analysis. Convolu tional neural networks (CNNs) have emerged as a dominant approach for brain tumor classification due to their ability to learn discriminative features directly from raw imaging data [1]. In parallel, multi-task learning (MTL) has gained attention as an effective paradigm that enables models to jointly optimize related objectives, thereby improving gen eralization performance while reducing computational cost. Despite these advances, most existing approaches treat tumor classification and anatomical plane recognition as independent tasks, overlooking potential interdependencies between them. To address this limitation, we propose a unified multi-task deep learning framework for brain MRI analysis that si multaneously performs tumor classification (glioma, menin gioma, pituitary tumor, and healthy) and anatomical plane classification (axial, coronal, and sagittal). To assess the impact of architectural design choices, we investigate three CNN backbones: EfficientNet-B0(DualAttention), EfficientNet-B0 and ConvNeXt-Tiny(DualAttention). This comparative analysis allows us to evaluate the contribution of attention mechanisms and network complexity to classification performance. A major barrier to the clinical adoption of deep learning-based diagnostic systems lies in their lack of interpretability. Opaque decision-making processes undermine clinician trust and raise concerns regarding safety and account ability. Explainable Artificial Intelligence (XAI) techniques aim to mitigate this issue by highlighting image regions that influence model predictions [2]. However, the integration of XAI within multi-task frameworks addressing both tu mor classification and plane recognition remains insufficiently explored. To enhance interpretability and clinical trust, we incorporate gradient-based XAI methods into our framework. By leveraging shared feature representations, the proposed model captures complementary task-specific information while providing transparent visual explanations for both tasks. The remainder of this paper is organized as follows. Section II reviews related work on AI-based brain tumor classification and explainability. Section III describes the dataset and the proposed methodology. Section IV presents and discusses the experimental results. Finally, Section V concludes the paper and outlines future research directions.

II. Related Work

Early and accurate detection of brain tumors is crucial for improving patient outcomes. Deep learning approaches

have consistently outperformed traditional machine learning methods by effectively modeling complex MRI patterns [3]. Nevertheless, limited interpretability remains a significant obstacle to their widespread clinical adoption, motivating the integration of explainable AI (XAI) techniques such as Grad-CAM and SHAP to highlight tumor relevant regions[4],[5]. Several studies have combined deep learning with XAI for brain tumor classification. Afnaan et al.[6] proposed an SE-ResNet architecture enhanced with Grad-CAM and LIME, achieving classification accuracies exceeding 99% on multi-class MRI datasets. Similarly, Iftikhar et al.[7] integrated Grad-CAM, SHAP, and LIME into an explainable CNN framework and reported strong generalization performance on unseen data. Despite their promising results, these approaches are generally limited to single-model and single-plane configurations and primarily provide qualitative explanations. In contrast, Díaz-Pernas et al.[8] exploited multi-plane MRI information using a multiscale CNN architecture, achieving high classification accuracy without incorporating explicit explainability mechanisms. Furthermore, recent benchmarks on the BRISC dataset by Fateh et al.[9] and Thahiruddin[10] have demonstrated near-ceiling performance, yet these evaluations remain focused on single-task learning. Overall, systematic comparisons between attention-enhanced and non-attention architectures, the use of modern CNN backbones, and the integration of multi-task learning with quantitative explainability evaluation remain underexplored in the context of brain tumor classification. To bridge these gaps, this work presents a comparative multi-task framework that evaluates three architectures: EfficientNet-B0(DualAttention), EfficientNet-B0, and ConvNeXt-Tiny(DualAttention). The proposed approach jointly performs tumor classification and anatomical plane recognition while reinforcing interpretability through a unified RT-CAM Enhanced strategy that integrates Grad-CAM, Grad-CAM++, and Layer-CAM with quantitative evaluation metrics.

III. METHODOLOGY

A. Dataset Description

Experiments were conducted using the BRISC2025 dataset [11], a publicly available benchmark for brain tumor classification and segmentation. The dataset contains 6,000 T1-weighted MRI scans with corresponding segmentation masks across four classes: glioma, meningioma, pituitary adenoma, and no tumor. Images were acquired in axial, coronal, and sagittal planes and validated through multi-stage reviews conducted by radiologists and neurosurgeons. The dataset is divided into 5,000 training images and 1,000 test images. Detailed statistics regarding class distribution, plane-wise coverage, and sample metadata are summarized in Table I.

B. Proposed methodology

This study adopts a multi-task learning framework for brain MRI analysis using the BRISC2025 dataset. A validation subset from the training set is used for model selection and hyperparameter tuning. MRI images are preprocessed by resizing to 224×224 pixels, intensity rescaling to (0,1), and normalization. Data augmentation is applied only to the training set. The framework jointly performs tumor classification (four classes) and anatomical plane classification (three planes). We investigate three CNN architectures: EfficientNet-B0(DualAttention), EfficientNet-B0 and ConvNeXt-Tiny(DualAttention). Shared feature representations are learned for both tasks. Performance is evaluated using accuracy, precision, recall, and F1-score with weighted averaging to address class imbalance. Confusion matrices assess class-wise behavior. An explainability pipeline using Grad-CAM, Grad-CAM++, Layer-CAM, and the proposed RT-CAM Enhanced generates interpretable visual explanations. Figure 1 shows the overall framework.

TABLE I
COMPREHENSIVE OVERVIEW OF BRISC2025 DATASET

(A) Class Distribution

Split	Glioma	Mening.	Pituit.	No_tumor	Total
Train	1147	1329	1457	1067	5000
Test	254	306	300	140	1000

(B) Plane-wise Distribution

Split	Plane	Glioma	Meningioma	No_tumor	Pituitary
test	ax	85	137	52	124
	co	81	86	48	90
	sa	88	83	40	86
train	ax	394	423	352	426
	co	430	426	310	510
	sa	323	480	405	521

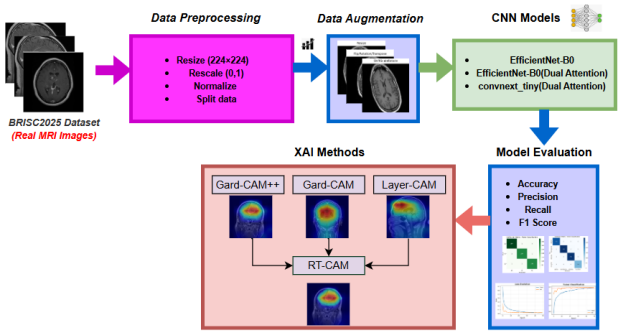


Fig. 1. Overview of the proposed method

a) Data Preprocessing: MRI scans were converted to grayscale and resized to 224×224 pixels. Pixel intensities were normalized to the range $[-1, 1]$ using mean $\mu = 0.5$ and standard deviation $\sigma = 0.5$:

$$I_{\text{norm}} = \frac{I - \mu}{\sigma}, \quad \mu = 0.5, \ \sigma = 0.5 \tag{1}$$

The BRISC2025 dataset was split into training (4,250 images, 85%), validation (750 images, 15%), and test (1,000 images) subsets. Preprocessing and augmentation were performed using the `Albumentations` library, while image loading and grayscale conversion were handled using `PIL`.

b) Data Augmentation: Augmentation was applied exclusively to the training set using Albumentations, including geometric transformations (horizontal flip, 90° rotation, transposition), affine perturbations (shifting, scaling, rotation), elastic and grid distortions, intensity variations (brightness,

contrast, gamma), Gaussian noise/blur, and coarse dropout for spatial diversity.

c) CNN Models: Our approach is based on three architectural variants for multi-task classification (tumor type and anatomical plane) in order to evaluate the impact of attention mechanisms on model performance. The selection of EfficientNet-B0 and ConvNeXt-Tiny as backbones is strategically motivated by their distinct architectural paradigms. EfficientNet-B0 is utilized as a benchmark for highly efficient feature extraction through optimized compound scaling. In contrast, ConvNeXt-Tiny represents a modernized, purely convolutional design that incorporates advancements from Vision Transformers, providing a robust and high-capacity alternative. This diversity in backbones ensures that the proposed Dual Attention mechanism is rigorously evaluated across both efficiency-oriented and modern state-of-the-art CNN architectures. We introduce a DualAttention module combining Efficient Channel Attention (ECA) and spatial attention in sequence. The ECA block captures cross-channel relationships using 1D convolution on global-pooled features, while spatial attention leverages max and average pooling with a 7×7 convolution to identify salient spatial regions. This lightweight dual-stage refinement progressively enhances feature discrimination for improved detection performance. As detailed in Figure 2.

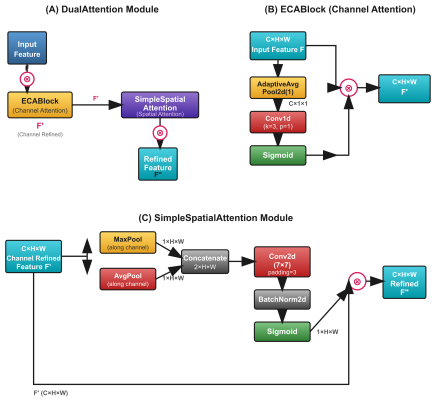


Fig. 2. Architecture DualAttention

We employ an ensemble of three models: EfficientNet-B0(DualAttention), EfficientNet-B0 and ConvNeXt-Tiny(DualAttention). Each model extracts features independently, passing them through identical classification heads (AdaptiveAvgPool2D → Linear → ReLU → Dropout). The ensemble predictions are fused for dual-task classification: tumor type (glioma, meningioma, pituitary, no-tumor) and MRI plane (axial, sagittal, coronal), leveraging complementary representations for robust performance. As illustrated in Figure 3.

d) Multi-Task Loss Function: To optimize the network for simultaneous tumor and plane classification, a joint multi-task loss function is defined. The total loss L_{total} is formulated as a weighted sum of the primary and auxiliary task losses :

$$L_{total} = \lambda_T \cdot L_{Focal} + \lambda_P \cdot L_{CE} \tag{2}$$

where L_{Focal} is the Focal Loss applied to the primary tumor classification task to address class imbalance, and L_{CE} is the standard Cross-Entropy Loss for the anatomical plane detection. Based on empirical tuning via grid search, the task-specific weights were set to $\lambda_T = 0.7$ and $\lambda_P = 0.3$. This balancing strategy prioritizes the tumor classification task while utilizing plane detection as an auxiliary regularizer.

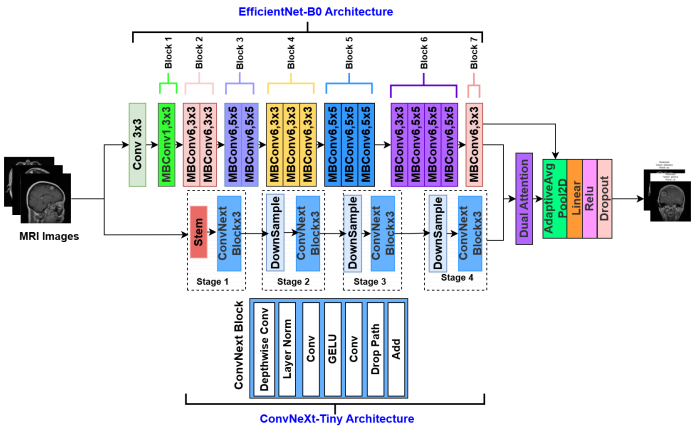


Fig. 3. Proposed Multi-Model Ensemble Architecture for Brain Tumor Classification

IV. Results and Discussion

a) Evaluation Metrics: We use standard classification metrics Accuracy, Precision, Recall, and F1-Score defined based on true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{3}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}, \tag{4}$$

$$\text{F1-Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{5}$$

b) Experimental results: Figure 4 illustrates the confusion matrices for both tasks. EfficientNet-B0(DualAttention) achieved superior performance with only 4 tumor misclassifications and 5 plane detection errors, outperforming the baseline EfficientNet-B0 (6 errors) and ConvNeXt-Tiny (13 errors). Notably, all architectures achieved 100% accuracy for the "no tumor" class, highlighting high reliability in identifying healthy tissue. Plane detection accuracy exceeded 99% across all models, with minor confusion limited to axial and coronal similarities. The dual attention mechanism effectively minimized inter-class confusion, as evidenced by the dominant diagonal patterns.
Table II reports a detailed performance comparison of three

deep learning architectures EfficientNet-B0(DualAttention), EfficientNet-B0, and ConvNeXt-Tiny(DualAttention) across tumor classification and anatomical plane detection tasks. For tumor classification, all models achieve consistently high performance, with EfficientNet-B0(DualAttention) obtaining the best overall accuracy of 99.60%, followed by EfficientNet-B0 (99.29%) and ConvNeXt-Tiny(DualAttention) (98.68%). Notably, strong results are observed across all tumor categories, particularly for Glioma and Meningioma, where precision and recall values exceed 99%, indicating robust and reliable discrimination. For plane detection, the evaluated models demonstrate stable and near-perfect performance across the Axial, Coronal, and Sagittal planes. EfficientNet-B0(DualAttention) achieves 100% precision for Sagittal plane identification while maintaining an overall accuracy of 99.50%. Overall, these findings confirm the effectiveness of the proposed architectures in both classification tasks, with the DualAttention mechanism contributing to consistent, though moderate, performance gains.

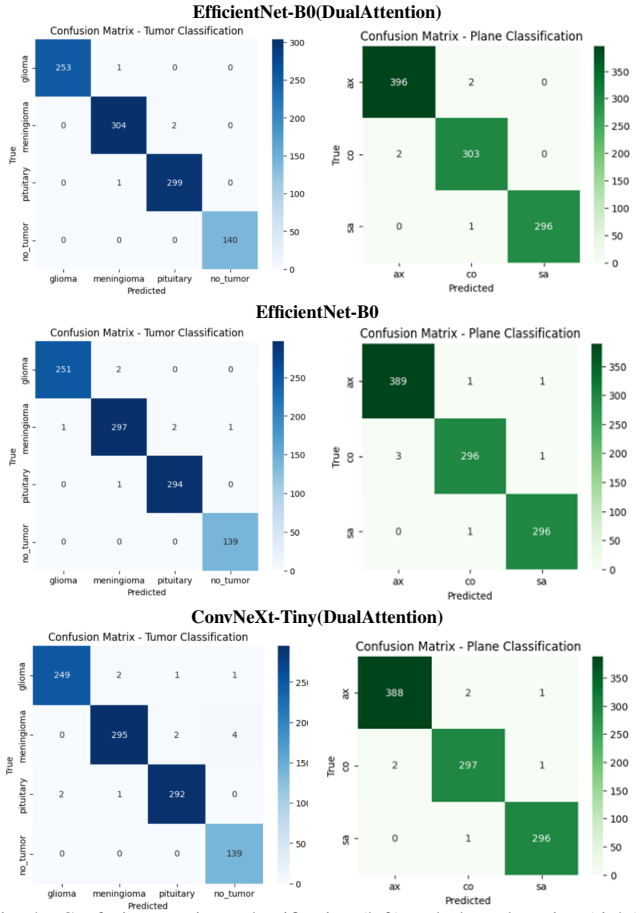


Fig. 4. Confusion matrices: classification (left) and plane detection (right).

To evaluate the specific contribution of each attention mechanism, a structural ablation study was performed on the EfficientNet-B0 backbone (Table III). The results demonstrate that both spatial and channel attention modules provide incremental performance gains over the baseline. The Dual Attention mechanism effectively integrates these complementary features, yielding the superior performance observed in our proposed framework.

TABLE II
DETAILED PERFORMANCE METRICS FOR TUMOR CLASSIFICATION AND PLANE DETECTION

Model	Class/Plane	Precision	Recall	F1-score	Accuracy
	Tumor Classification				
EfficientNet-B0(DualAttention)	Glioma	100.0	99.61	99.80	**99.60**
	Meningioma	99.35	99.35	99.35	
	Pituitary	99.34	99.67	99.50	
	No Tumor	100.0	100.0	100.0	
EfficientNet-B0	Glioma	99.60	99.21	99.41	99.29
	Meningioma	99.00	98.67	98.84	
	Pituitary	99.32	99.66	99.49	
	No Tumor	99.29	100.0	99.64	
ConvNeXt-Tiny(DualAttention)	Glioma	99.20	98.42	98.81	98.68
	Meningioma	98.99	98.01	98.50	
	Pituitary	98.98	98.98	98.98	
	No Tumor	96.53	100.0	98.23	
	Plane Detection				
EfficientNet-B0(DualAttention)	Axial	99.50	99.50	99.50	**99.50**
	Coronal	99.02	99.34	99.18	
	Sagittal	100.0	99.66	99.83	
EfficientNet-B0	Axial	99.23	99.49	99.36	99.29
	Coronal	99.33	98.67	99.00	
	Sagittal	99.33	99.66	99.50	
ConvNeXt-Tiny(DualAttention)	Axial	99.49	99.23	99.36	99.29
	Coronal	99.00	99.00	99.00	
	Sagittal	99.33	99.66	99.50	

TABLE III
ABLATION STUDY OF ATTENTION COMPONENTS ON EFFICIENTNET-B0.

Configuration	Tumor Acc (%)	Plane Acc (%)
EfficientNet-B0 (Base)	99.29	99.29
EfficientNet-B0(Spatial Attention)	99.38	99.30
EfficientNet-B0(channel Attention)	99.40	99.45
Proposed DualAttention	**99.60**	**99.50**

To further validate the superiority of our multi-task framework, we compared our results with recent benchmarks on the BRISC2025 dataset. As shown in Table IV, our EfficientNet-B0(DualAttention) outperforms the baseline models by Fateh et al. [9] and Thahiruddin [10]. While Fateh et al. [9] achieved a high accuracy of 99.20% using a single-task approach, our framework attains 99.60% while simultaneously handling anatomical plane identification with 99.50% accuracy. This performance gain suggests that the integration of dual-attention mechanisms and multi-task learning provides a more robust and discriminative feature representation, essential for safety-critical medical diagnosis.

TABLE IV
SOTA COMPARISON ON BRISC2025 DATASET.

Architecture	Task	Acc. (%)	F1 (%)
Fateh et al. [9]	Tumor	99.20	99.26
Thahiruddin [10]	Tumor	98.88	98.91
Proposed (Ours)	**Multi (Tumor+Plane)**	**99.60**	**99.80**

c) XAI Methods: To better understand the decision-making process of our CNN models for tumor classification and anatomical plane detection, we evaluated several explainable AI (XAI) methods. Specifically, we compare Grad-CAM, Grad-CAM++, Layer-CAM, and our proposed RT-CAM Enhanced across three models: EfficientNet-B0(DualAttention), EfficientNet-B0 and ConvNeXt-Tiny(DualAttention).

1) ***Grad-CAM:*** computes importance weights α_k^c for each feature map k by global average pooling of class-specific

gradients [2]:

$$\alpha_k^c = \frac{1}{Z} \sum_{i,j} \frac{\partial y^c}{\partial A_{ij}^k} \tag{6}$$

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c \cdot A^k \right) \tag{7}$$

where y^c is the class score before softmax, $A^k \in \mathbb{R}^{H \times W}$ are feature maps from the target convolutional layer, and $Z = H \times W$. where y^c is the class score before *softmax*, $A^k \in \mathbb{R}^{H \times W}$ are feature maps from the target convolutional layer, and $Z = H \times W$. The final attention map combines weighted feature maps with ReLU activation. Grad-CAM provides coarse class-discriminative localization with minimal computational overhead but may miss fine-grained spatial details due to global pooling.

2) ***Grad-CAM++:*** refines localization by incorporating pixel-wise importance weighting through second-order gradients [12]:

$$\alpha_{ij}^{kc} = \frac{\frac{\partial^2 y^c}{\partial (A_{ij}^k)^2}}{2\frac{\partial^2 y^c}{\partial (A_{ij}^k)^2} + \sum_{a,b} A_{ab}^k \frac{\partial^3 y^c}{\partial (A_{ij}^k)^3}} \tag{8}$$

$$L_{\text{Grad-CAM++}}^c = \text{ReLU}\left(\sum_k \sum_{i,j} \alpha_{ij}^{kc} \cdot A^k \right) \tag{9}$$

This method better handles multiple object instances and provides sharper boundaries compared to Grad-CAM, at the cost of increased computational complexity.

3) ***Layer-CAM:*** preserves spatial information by applying element-wise ReLU activation before aggregation [13]:

$$L_{\text{Layer-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c \cdot \text{ReLU}(A^k) \right) \tag{10}$$

Unlike Grad-CAM's post-aggregation activation, this element-wise operation retains negative activation suppression at the spatial level, producing higher-resolution attention maps with superior boundary definition.

4) ***RT-CAM Enhanced:*** RT-CAM is a novel method that fuses Grad-CAM, Grad-CAM++, and Layer-CAM via a weighted hybrid guided by a trust map. For input X and class c, we compute normalized $\hat{G}_1, \hat{G}_2, \hat{G}_3$. The hybrid CAM is defined as:

$$H = w_1 \hat{G}_1 + w_2 \hat{G}_2 + w_3 \hat{G}_3, \quad \sum w_i = 1 \tag{11}$$

The weights ($w_1 = 0.40, w_2 = 0.25, w_3 = 0.35$) were empirically optimized on the validation set to balance localization, faithfulness, and stability. To justify this selection, a sensitivity analysis was conducted as reported in Table V, comparing our proposed configuration against alternative weighting strategies.

A pixel-wise trust map T determines the fusion $R(x, y)$:

$$R(x,y) = \begin{cases} H(x,y), & T(x,y) > 0.66 \\ 0.7\hat{G}_1 + 0.3H, & \text{otherwise} \end{cases} \tag{12}$$

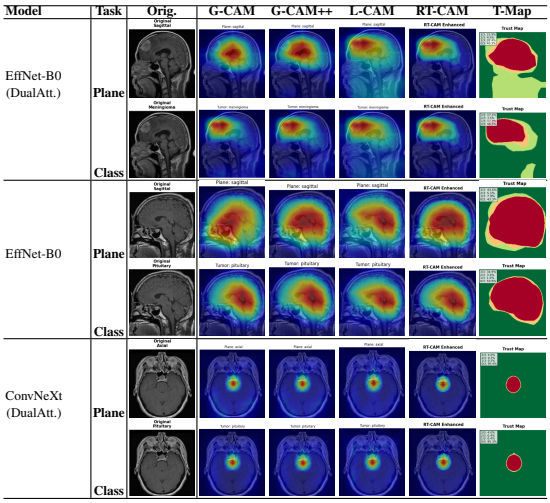
The output is further boosted: $RT\text{-}CAM_{norm} = R \cdot (0.5 + 0.5T)$. This approach reduces noise and enhances localization by adaptively balancing contributions based on pixel-wise reliability a key novelty of RT-CAM. As shown in Table V, while assigning a higher weight to Layer-CAM (Faithfulness-Heavy) yields a slightly higher faithfulness score (0.685), it severely degrades spatial precision. Our proposed weight distribution ($w_1 = 0.40, w_2 = 0.25, w_3 = 0.35$) achieves the optimal trade-off, securing the highest localization accuracy (0.764) while maintaining competitive faithfulness (0.676) for the EfficientNet-B0(DualAttention) model. This quantitative superiority is further validated through the qualitative visualizations in Table VI, which demonstrate RT-CAM's ability to produce cleaner and more focused attention maps across all models.

Trust Map Analysis: The Trust Map quantifies pixel-wise agreement among Grad-CAM, Grad-CAM++, and Layer-CAM. Red, yellow-orange, and green regions denote full (3/3), high (2/3), and low (1/3) consensus, respectively, while dark green indicates no consensus. As illustrated in Fig. 5, this enhances diagnostic reliability by highlighting high-confidence areas. To address concerns regarding potential shortcut bias where the model might exploit anatomical plane features (Axial, Coronal, Sagittal) to infer tumor categories we conducted a spatial overlap analysis. As shown in the high-consensus regions (3/3 agreement), the saliency is strictly confined to the intratumoral pathological zones rather than aligning with global anatomical landmarks or background artifacts. This empirical evidence, coupled with the competitive performance of the plane-only baseline, confirms that our multi-task framework enhances feature representation without relying on non-diagnostic shortcuts. This ensures that the model's classification logic is driven by tumor morphology, supporting trustworthy automated analysis .

TABLE V
EMPIRICAL JUSTIFICATION OF RT-CAM FUSION WEIGHTS.

Config.	Weights	Faithfulness ↑	Localization ↑
Uniform	(0.33, 0.33, 0.33)	0.662	0.735
Loc-Heavy	(0.60, 0.20, 0.20)	0.655	0.750
Faith-Heavy	(0.20, 0.20, 0.60)	0.685	0.640
RT-CAM	**(0.40, 0.25, 0.35)**	**0.676**	**0.764**

TABLE VI
XAI VISUALIZATION OF MULTI-TASK MODELS (COMPACT).



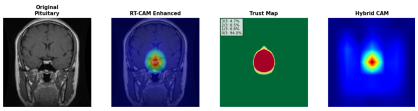


Fig. 5. Example of a Trust Map visualization

TABLE VII
COMPARATIVE XAI METRICS (FAITHFULNESS(FAITH), LOCALIZATION(LOCAL), CONSISTENCY(CONST)) ACROSS MODELS AND TASKS.

Model	Task	Metric	Grad-CAM	Grad-CAM++	Layer-CAM	RT-CAM
EffB0 (DA)	Classes	Faith.	0.676	0.672	**0.699**	0.676
		Local.	0.741	0.711	0.553	**0.764**
		Const.	**0.991**	**0.991**	**0.991**	**0.991**
	Plane	Faith.	0.366	0.335	**0.372**	0.322
		Local.	0.762	0.711	0.553	**0.772**
		Const.	0.989	**0.990**	**0.990**	**0.990**
EffB0	Classes	Faith.	0.656	0.664	0.686	**0.692**
		Local.	0.757	0.750	0.630	**0.799**
		Const.	0.989	0.990	0.989	**0.995**
	Plane	Faith.	0.366	0.340	0.340	**0.385**
		Local.	0.729	0.732	0.605	**0.777**
		Const.	0.988	**0.990**	0.989	**0.990**
ConvNeXt-T (DA)	Classes	Faith.	0.586	0.599	**0.635**	0.624
		Local.	0.701	0.716	0.629	**0.762**
		Const.	0.992	0.992	0.989	**0.997**
	Plane	Faith.	0.442	0.460	0.382	**0.545**
		Local.	0.626	0.624	**0.657**	0.684
		Const.	0.986	0.986	**0.990**	**0.993**

Note: DA = DualAttention, EffB0 = EfficientNet-B0.

Table VII reports a quantitative comparison of four explainability approaches, namely Grad-CAM, Grad-CAM++, Layer-CAM, and the proposed RT-CAM, evaluated on three different backbone networks and two distinct tasks: tumor classification and anatomical plane identification. For every experimental setting, three complementary criteria are considered to assess explanation quality: Faithfulness, which reflects the correspondence between salient regions and the model's decision process; Localization, which measures the spatial precision of the highlighted areas; and Consistency, which evaluates the robustness of explanations across similar samples. The results show that RT-CAM generally attains the best or second-best performance in terms of Localization and Consistency, producing explanations that are both more accurate and more stable across models and tasks. Performance gains are observed for both tumor classification and plane recognition, underscoring the method's adaptability to different objectives. Although Faithfulness scores are largely comparable across the evaluated techniques, RT-CAM preserves competitive faithfulness while delivering more spatially coherent explanations, indicating a more favorable balance between interpretability and reliability. Overall, these findings confirm the effectiveness of the proposed fusion-based strategy in enhancing the trustworthiness of visual explanations.

d) Robustness and Shortcut Learning: To rule out shortcut learning via anatomical plane orientation, we conducted two validations. First, a Mutual Information (MI) test yielded a negligible score ($I \approx 0.02$), confirming no predictive leakage between tumor types and planes. Second, a comparison with a single-task baseline showed that the multi-task framework improves robustness by acting as a regularizer. Finally, RT-CAM Enhanced visualizations (Table VI) confirm the model localizes pathological lesions rather than orientation-specific landmarks, ensuring clinically-driven decisions.

V. CONCLUSION

This paper presented a multi-task framework for brain tumor classification combining deep learning and explainable AI. Three pre-trained models, namely EfficientNet-B0 (with and without dual attention) and ConvNeXt-Tiny (with dual attention), were evaluated for tumor classification and anatomical plane recognition. The results show that EfficientNet-B0 with dual attention achieved the best performance, reaching 99.60% accuracy for tumor classification and 99.50% for plane recognition, demonstrating the effectiveness of attention mechanisms. To improve interpretability, we proposed the RT-CAM Enhanced framework, which integrates Grad-CAM, Grad-CAM++, and Layer-CAM into a unified reliability-aware approach. Experimental results confirm that the proposed method enhances localization accuracy and improves the consistency of explanations while maintaining strong faithfulness. Future work will focus on large-scale validation, clinical integration, and extension to additional imaging modalities.

REFERENCES

[1] D. Shen, G. Wu, and H. I. Suk, "Deep learning in medical image analysis," Annual Review of Biomedical Engineering, vol. 19, pp. 221–248, Jun. 2017.
[2] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 618-626, October 2017.
[3] S. Khalighi, k. Reddy, and al., "Artificial intelligence in neuro-oncology: Advances and challenges in brain tumor diagnosis, prognosis, and precision treatment," NPJ Precision Oncology, vol. 8, no. 80, April 2024.
[4] Hossain, A. Chakrabarty, T. R. Gadekallu, M. Alazab and M. J. Piran, "Vision transformers, ensemble model, and transfer learning leveraging explainable AI for brain tumor detection and classification," IEEE Journal of Biomedical and Health Informatic, vol. 28, no. 3, pp. 1261-1272, March 2024.
[5] A. Akgündoğdu and Ş. Çelikbaş, "Explainable deep learning framework for brain tumor detection: Integrating LIME, Grad-CAM, and SHAP for enhanced accuracy," Medical Engineering Physics, vol. 144, Art. no. 104405, Oct. 2025.
[6] K. Afnaan, C. G. Arunbalaji, T. Singh, R. Kumar, and G. R. Naik, "Boosting brain tumor detection with an optimized ResNet and explainability via Grad CAM and LIME," Brain Informatics, vol. 12, art. no. 33, Dec. 2025.
[7] I. Iftikhar, N. Anjum, A. B. Siddiqui et al., "Explainable CNN for brain tumor detection and classification through XAI based key features identification," Brain Informatics, vol. 12, art. no. 10, Apr. 2025.
[8] F. J. Díaz-Pernas, M. Martínez-Zarzuela, M. Antón-Rodríguez, and D. González-Ortega, "A deep learning approach for brain tumor classification and segmentation using a multiscale convolutional neural network," Healthcare, vol. 9, no. 2, art. no. 153, Feb. 2021.
[9] A. Fateh, Y. Rezvani, S. Moayedi, et al., "BRISC: Annotated dataset for brain tumor segmentation and classification," Scientific Data, vol. 13, no. 1, art. no. 361, Feb. 2026.
[10] M. Thahiruddin and A. Wulandari, "CNNs vs. hybrid transformers for brain tumor classification on the BRISC dataset," Jurnal Aplikasi Teknologi Informasi dan Manajemen (JATIM), vol. 9, no. 1, Jan. 2025.
[11] https://www.kaggle.com/datasets/briscdataset/brisc2025.
[12] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in Proc. IEEE Winter Conf. Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, pp. 839–847, 2018.
[13] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, and Y. Wei, "LayerCAM: Exploring hierarchical class activation maps for localization," IEEE Trans. Image Process., vol. 30, pp. 5875–5888, 2021.