

Hybrid 3D CNN–transformer model for early brain tumor detection with multi-modal magnetic resonance imaging

Vivek Kumar Sharma, Gaurav Kumar Ameta

Department of Computer Science and Engineering, Parul University, Gujarat, India

Article Info

Article history:

Received Jul 23, 2025

Revised Sep 2, 2025

Accepted Sep 11, 2025

Keywords:

Attention-based fusion
Brain tumor classification
BraTS2023 GLI
Multi-modal magnetic
resonance imaging
Swin transformer

ABSTRACT

Accurate and early diagnosis of brain tumors using multi-modal magnetic resonance imaging (MRI) remains a critical challenge due to tumor heterogeneity and complex spatial representation. This study proposes a novel hybrid deep learning framework that integrates a 3D convolutional neural network (3D CNN) with swin transformer blocks and an attention-based feature fusion module (ABFFM). The model leverages multi-modal MRI inputs—T1, T1Gd, T2, and fluid-attenuated inversion recovery (FLAIR)—and features a dual-branch classification head for binary tumor detection and multi-label tumor sub-region classification: enhancing tumor (ET), tumor core (TC), and whole tumor (WT). Experiments conducted on the BraTS2023-GLI dataset demonstrate that the proposed model achieves a superior classification accuracy of 96.51%, with precision of 97.98%, recall of 97.04%, and F1-score of 97.61%, outperforming state-of-the-art methods. Furthermore, intrinsic attention weights offer interpretability by highlighting modality-specific contributions. The proposed model establishes a clinically promising approach for brain tumor analysis, with strong implications for early diagnosis and treatment planning.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Vivek Kumar Sharma
Department of Computer Science and Engineering, Parul University
Waghodia, Vadodara, Gujarat 391760, India
Email: vkshere4every1@gmail.com

1. INTRODUCTION

Brain tumors are among the most severe and life-threatening neurological disorders, posing significant challenges to timely diagnosis and treatment due to their heterogeneous nature and complex presentation in radiological imaging [1], [2]. Early detection and accurate classification of brain tumors play a pivotal role in determining treatment strategies and improving patient prognosis [3]. However, the variability in tumor morphology, location, and tissue contrast—especially in multi-modal magnetic resonance imaging (MRI)—often complicates manual diagnosis [4]. Radiologists face the arduous task of interpreting large volumes of 3D MRI data, which can be subjective, time-intensive, and prone to inter-observer inconsistencies [5], [6]. An overview of the high-level pipeline for automated brain tumor analysis using deep learning is illustrated in Figure 1.

Traditional machine learning approaches for brain tumor classification often rely on manual feature extraction [7], which demands domain expertise and limits the model's ability to generalize across diverse imaging datasets. In contrast, deep learning—particularly convolutional neural networks (CNNs)—has shown remarkable performance in automating feature extraction and learning complex data representations directly from images [8]. Numerous recent studies have explored a range of deep learning and hybrid approaches for brain tumor detection and classification using MRI data. For instance, Anantharajan *et al.* [9]

proposed a hybrid model integrating traditional preprocessing and machine learning techniques, wherein MRI scans were enhanced using adaptive contrast enhancement algorithm (ACEA), segmented using fuzzy c-means clustering, and classified using an ensemble deep neural support vector machine (EDN-SVM). This method achieved a notable accuracy of 97.93%, demonstrating the benefits of combining handcrafted features with deep ensemble classifiers. Similarly, Mahmud *et al.* [10] developed a CNN-based architecture for early tumor detection, evaluating its performance against ResNet-50, VGG16, and InceptionV3. The proposed model outperformed the baselines, achieving 93.3% accuracy and 98.43% area under the curve (AUC), confirming its effectiveness for clinical-grade tumor detection. In another study, Qureshi *et al.* [11] presents an ultra-light deep learning model integrating CNN and gray-level co-occurrence matrix (GLCM)-based features for multi-class brain tumor classification using the CE-MRI dataset. Achieving 99.24% accuracy, the model offers real-time performance with minimal hardware. Limitations include class imbalance and lack of spatial localization. Research by Khan and Park [12] proposes a CNN-based convolutional block architecture for multiclass brain tumor classification using three public MRI datasets. The model achieved an average accuracy of 97.85%, outperforming state-of-the-art models. Despite high precision and adaptability, limitations include dataset bias and lack of explainability. Ahmmed *et al.* [13] developed a segmentation-aided classification pipeline based on 2D slices and transfer learning, reaching 93.3% accuracy. Saeedi *et al.* [14] introduced a dual-path architecture for multimodal fusion but lacked interpretability mechanisms, achieving 93.44% accuracy. Tehsin *et al.* [15] incorporated Grad-CAM for explainability and used partial modality fusion to reach 92.5% accuracy. Research by Ahmed *et al.* [16] proposes a ViT-GRU hybrid deep learning model for brain tumor classification using primary MRI data from Bangladesh and the brain tumor Kaggle dataset. Achieving 81.6% accuracy, the model integrates XAI (LIME, SHAP, and attention maps). Limitations include data imbalance, complex preprocessing, and restricted dataset diversity. In contrast, the proposed model in the current study employs a 3D CNN backbone enhanced with swin transformer blocks and an attention-based feature fusion module (ABFFM), enabling full multi-modal integration and improving both accuracy and interpretability. The dual-branch classification head—comprising a binary tumor presence detector and a multi-label tumor subregion classifier—aligns well with the objectives of the BraTS GLI challenge. The model achieves a superior accuracy of 96.51%, with learnable attention weights providing insight into modality importance. This comprehensive review reveals a consistent trend toward integrating multimodal MRI data, using transfer learning, and incorporating explainability mechanisms to improve performance and clinical relevance. While many existing models rely on 2D inputs or lack interpretability, the proposed framework addresses both limitations through volumetric learning and attention-based fusion, setting a new benchmark for tumor detection and classification. The rest of the manuscript is organized as follows: section 2 describes the methods, section 3 presents and discusses the results, and section 4 concludes with key contributions and future directions.

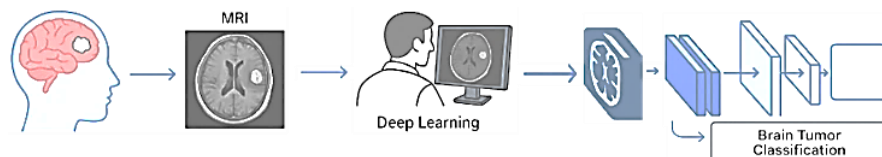


Figure 1. Deep learning workflow for brain tumor classification from MRI scans

2. METHOD

The proposed hybrid deep learning framework leverages volumetric MRI data for early tumor detection and multi-label classification. As illustrated in Figure 2, the architecture integrates 3D CNNs, swin transformer blocks, and an ABFFM to effectively capture both local anatomical details and global contextual information from multi-modal MRI inputs. Specifically, the model processes four MRI modalities—T1, T1Gd, T2, and fluid-attenuated inversion recovery (FLAIR)—to exploit their complementary diagnostic features. The 3D CNN backbone extracts hierarchical spatial features from volumetric inputs, while swin transformer blocks model long-range dependencies across slices. The ABFFM then fuses modality-specific features by applying attention mechanisms to emphasize the most informative representations.

The classification head is divided into two primary branches aligned with the objectives of the BraTS GLI challenge. Branch 1 performs binary classification to detect the presence or absence of the whole tumor (WT) region, using sigmoid activation and binary cross entropy loss. Branch 2 simultaneously classifies tumor sub-regions—enhancing tumor (ET), tumor core (TC), and WT—using a multi-label approach with sigmoid activation and binary cross entropy loss for each target. This design enables both

coarse-level tumor detection and fine-grained sub-region classification. The model is trained and validated using the BraTS 2023 GLI dataset, which provide clinically annotated ground truth for glioma segmentation tasks. By combining transformer-based global reasoning, convolutional spatial encoding, and modality-aware fusion, the proposed framework enhances both the accuracy and interpretability of tumor classification from 3D brain MRIs.

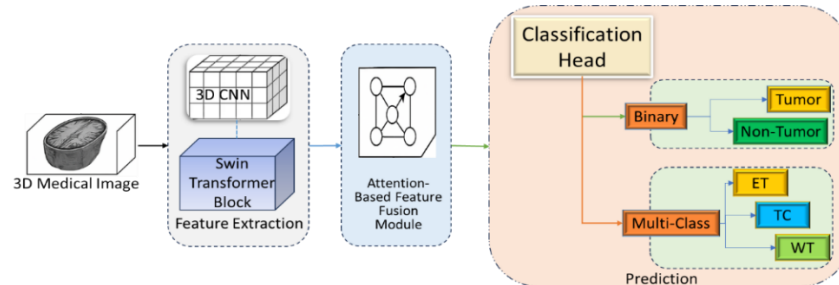


Figure 2. Proposed hybrid 3D CNN–transformer architecture

2.1. BraTS2023-GLI dataset description

The details of the benchmark dataset used in this study, namely BraTS2023-GLI [17], [18] is summarized in Table 1. This dataset includes varying tumor subregion annotations and normal cases, offering a challenging and realistic environment for brain tumor prediction and classification tasks.

Table 1. Details of BraTS2023-GLI dataset

Dataset	Total cases	MRI modalities	ET cases	TC cases	WT cases	Cases with missing labels
BraTS2023-GLI	1,251	T1, T1Gd, T2, and FLAIR	1,208	1,250	1,218	71

2.2. Preprocessing and data augmentation

Given the heterogeneity in MRI scans due to different acquisition protocols and scanners, a robust preprocessing pipeline is essential for model performance. Initially, all volumes are resampled to a uniform voxel spacing using trilinear interpolation to standardize resolution across cases. Next, Z-score normalization is applied individually to each modality, ensuring zero mean and unit variance intensity distributions as in (1):

$$I_{norm} = \frac{I - \mu}{\sigma} \quad (1)$$

where I denotes the raw intensity, μ the mean, and σ the standard deviation of the voxel intensities for a given modality. To manage memory constraints and facilitate batch-wise training, 3D patches of size $128 \times 128 \times 128$ are extracted from the full volume. This not only enables efficient GPU usage but also enhances spatial localization by focusing on smaller brain regions. To prevent overfitting and improve generalization, extensive data augmentation is applied during training. Augmentation strategies include random axis flips, 90° rotations, intensity jittering, Gaussian noise addition, and elastic deformations. These transformations introduce anatomical variability and simulate different acquisition conditions, thereby enhancing the model's robustness to real-world clinical scenarios [19].

2.3. Hybrid deep architecture

2.3.1. 3D convolutional neural network feature extractor

The initial backbone of the architecture is a modified 3D ResNet [20], tailored to process 3D volumetric inputs. The concatenated four-modality MRI volume serves as the input tensor of shape $4 \times 128 \times 128 \times 128$. The 3D CNN processes this input through successive convolutional, batch normalization, and rectified linear units (ReLU) layers is illustrated in Figure 3. Early layers extract low-level volumetric patterns such as tissue density and edge gradients, while deeper layers capture more complex features corresponding to tumor boundaries and textural irregularities. Residual connections are retained to preserve feature integrity across layers and mitigate gradient vanishing problems during training. Each residual block is defined as in (2):

$$y = f(x, \{W_i\}) + x \quad (2)$$

where f is a residual function and x is the input to the block.

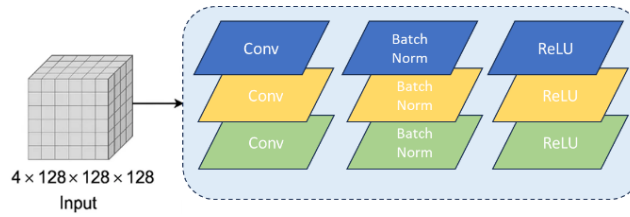


Figure 3. 3D CNN feature extractor for volumetric MRI input

2.3.2. Swin transformer integration

To capture global contextual relationships that are often missed by localized 3D convolutions, swin transformer blocks are integrated into the architecture after specific convolutional stages (Figure 4). Unlike standard vision transformers, swin transformers adopt a hierarchical structure and compute self-attention within non-overlapping local windows, which are then shifted in subsequent layers [21], [22]. This design significantly improves computational efficiency while preserving the model's ability to capture long-range dependencies, especially important in high-dimensional medical imaging such as 3D MRIs.

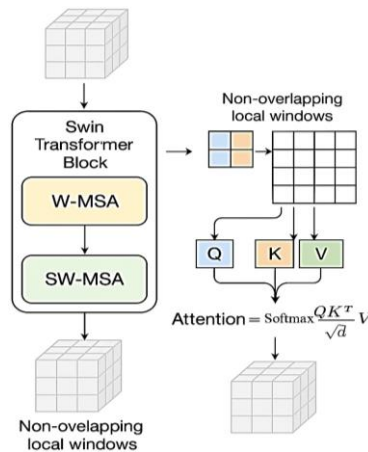


Figure 4. Swin transformer block with W-MSA and SW-MSA for hierarchical feature learning

The swin transformer was selected over the standard ViT due to its hierarchical architecture and shifted window mechanism, which significantly reduces computational complexity while maintaining high spatial context capture. Compared to CNN-based attention blocks, swin transformers have shown superior performance in modeling long-range dependencies in 3D medical imaging with fewer parameters and improved scalability [2]. Each swin transformer block includes a window-based multi-head self-attention (W-MSA) layer followed by a shifted window multi-head self-attention (SW-MSA) layer. The attention is computed as in (3):

$$Attention = (Q, K, V) = \text{Softmax} \left(\frac{QK^T}{\sqrt{d}} \right) V \quad (3)$$

where $Q, K, V \in \mathbb{R}^{n \times d}$ represent the query, key, and value matrices respectively, where n is the number of tokens and d the feature dimension. This attention mechanism enables the model to dynamically focus on informative regions across the 3D volume, which is critical for detecting glioma sub-regions (ET, TC, and WT) with varying locations and shapes. By combining both local and global feature extraction, the swin transformer enhances the model's robustness and accuracy in brain tumor classification.

2.3.3. Multi-modal feature fusion

A key innovation in the proposed architecture is the ABFFM, which integrates features from different modalities and resolutions. The ABFFM first projects features from each modality into a common embedding space. Then, it computes attention weights α_i for each modality i using a learnable soft-attention mechanism as in (4), as illustrated in Figure 5:

$$\alpha_i = \frac{\exp(w^T \tanh(W_i f_i + b_i))}{\sum_j \exp(w^T \tanh(W_j f_j + b_j))} \quad (4)$$

here, f_i is the feature from modality i , and α_i reflects the importance of that modality for the current region. The fused feature F is then given by as in (5).

$$F = \sum_i \sigma_i * f_i \quad (5)$$

This strategy effectively suppresses irrelevant noise and emphasizes tumor-discriminative features, enhancing the quality of downstream classification.

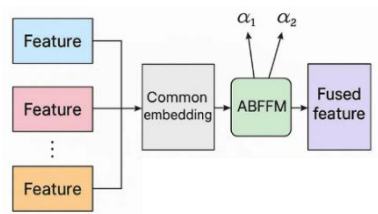


Figure 5. Overview of the ABFFM

2.4. Multi-modal feature fusion

The final fused feature representation, capturing both spatial and contextual characteristics, is flattened and passed through a fully connected dense layer, followed by batch normalization and a dropout layer with a rate of 0.5 for regularization. The classification head is bifurcated into two branches to address distinct prediction tasks. The first branch performs tumor detection through binary classification, predicting the presence or absence of the WT region. It uses a sigmoid activation function to output a probability as in (6):

$$P(y = 1|x) = \frac{1}{1 + \exp(-z)} \quad (6)$$

where z is the logit corresponding to the WT class. This branch is trained using binary cross-entropy (BCE) loss as in (7).

$$L_{BCE} = -[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})] \quad (7)$$

The second branch performs multi-label tumor sub-region classification, simultaneously predicting the presence of ET, TC, and WT. Each sub-region output is passed through a sigmoid activation function, producing independent probabilities for each class. This multi-label task is also trained using BCE loss, summed over all sub-region classes as in (8):

$$L_{multi_BCE} = - \sum_{c=1}^C [y_c \log(\hat{y}_c) + (1 - y_c) \log(1 - \hat{y}_c)] \quad (8)$$

where $C=3$ for the ET, TC, and WT regions. The model is optimized using the AdamW optimizer, which decouples weight decay from gradient updates for improved generalization. To ensure smooth convergence, a cosine annealing learning rate schedule with warm restarts is employed.

3. RESULTS AND DISCUSSION

This section presents the evaluation of the proposed hybrid deep learning-based system for early detection and multi-class classification of brain tumors using multi-modal radiological MRI data. The performance of different CNN backbones, the impact of modality fusion strategies, and the effectiveness of

the attention-based fusion model are thoroughly examined. The results are reported using standard metrics, and the findings are interpreted in light of the research objective of achieving accurate and reliable brain tumor classification.

3.1. Experimental setup

The experimental study utilizes the BraTS2023-GLI dataset, containing multimodal MRI scans (T1, T1c, T2, and FLAIR) of glioma patients. The dataset includes annotations for ET, TC, and WT regions. The data was split into 70:10:20 for training, validation, and testing. All models were implemented in TensorFlow and trained on an NVIDIA RTX 3090 GPU. The training configuration is summarized in Table 2.

Table 2. Hyperparameter configuration

Component	Batch size	Epochs	Optimizer	Activation function	Learning rate	Scheduler	Loss function	Dropout rate
Value	16	80	AdamW	Sigmoid	0.0001	Cosine annealing	BCE	0.5

3.2. Baseline performance with single modality

To establish a baseline, individual CNN models were trained using the ResNet50 backbone on each MRI modality separately. Table 3 summarizes the performance metrics across T1, T2, FLAIR, and T1c modalities reported as mean \pm standard deviation (σ) over five independent training runs. Among all modalities, T1c achieved the highest performance, which can be attributed to its enhanced contrast properties that delineate tumor boundaries more clearly. Despite its superiority, the classification accuracy plateaued at approximately 90%, indicating the inherent limitations of relying solely on a single modality. This observation underscores the importance of multi-modal fusion for comprehensive tumor representation and improved diagnostic accuracy.

Table 3. Performance of baseline models using single MRI modality

Modality	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
T1	87.31 \pm 0.14	85.61 \pm 0.16	86.28 \pm 0.15	86.26 \pm 0.15
T2	88.17 \pm 0.13	86.43 \pm 0.15	87.51 \pm 0.14	86.90 \pm 0.14
FLAIR	89.60 \pm 0.12	88.13 \pm 0.14	89.27 \pm 0.13	88.51 \pm 0.13
T1c	90.41 \pm 0.11	89.18 \pm 0.13	90.21 \pm 0.12	89.62 \pm 0.12

3.3. Performance of multi-modal fusion without attention

We next evaluated a basic multi-modal fusion strategy by concatenating features extracted from all four MRI modalities and passing them through a dense classification layer. This approach was tested using two different backbone architectures: EfficientNet-B0 [23] and ResNet50. Results, expressed as mean \pm σ across five runs, are summarized in Table 4. The multi-modal fusion approach significantly outperformed the single-modality baselines, achieving a 3–4% increase in accuracy and other metrics. Among the two backbones, ResNet50 delivered superior results compared to EfficientNet-B0.

Table 4. Classification performance of multi-modal MRI fusion

Backbone	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
EfficientNet-B0	92.80 \pm 0.12	91.50 \pm 0.13	92.00 \pm 0.12	91.70 \pm 0.13
ResNet50	94.10 \pm 0.11	93.30 \pm 0.12	93.80 \pm 0.11	93.50 \pm 0.12

This improvement is likely due to ResNet50's deeper architecture and its ability to capture more complex hierarchical features when trained on fused multi-modal data. The results affirm the effectiveness of integrating multiple imaging modalities to enhance the representation and classification of tumor characteristics.

3.4. Performance of attention-based modality fusion

To improve modality-specific feature integration, we applied an attention-based fusion mechanism wherein features from each MRI modality are weighted adaptively based on learned attention scores. This enables the model to emphasize more informative modalities per scan. Table 5 presents the results as mean \pm σ across five runs. This approach outperformed both single-modality and naive fusion baselines, confirming the benefit of dynamic modality weighting.

Table 5. Performance of the proposed attention-based modality fusion model

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Proposed (ResNet50+attention fusion)	96.51±0.09	97.98±0.10	97.04±0.11	97.61±0.09

The proposed attention-based fusion model achieved the highest performance among all tested configurations, with an accuracy of 96.51%, precision of 97.98%, recall of 97.04%, and an F1-score of 97.61%. The learnable attention weights α_i enabled adaptive weighting of modalities, improving the model's generalization capability. For instance, the observed attention distribution is shown in the Table 6.

Table 6. MRI modality input attention weights

Attention weight	T1	T2	FLAIR	T1c
α	0.18	0.21	0.26	0.35

This distribution highlights that the T1c and FLAIR modalities were assigned higher weights, consistent with their superior individual performance observed in earlier experiments. The results confirm that attention-based fusion not only improves overall accuracy but also introduces robustness by dynamically leveraging modality-specific relevance per scan.

3.5. Ablation study

To quantify the individual contributions of the swin transformer and the ABFFM, we performed an ablation study. Three configurations were evaluated: i) 3D CNN only, ii) 3D CNN+ win transformer (no ABFFM), and iii) 3D CNN+ABFFM (no swin transformer). As shown in Table 7, adding the swin transformer improved accuracy by 2.4% over CNN-only, while ABFFM alone improved accuracy by 1.9%. Combining both yielded the best accuracy (96.51%), confirming their complementary roles.

Table 7. Ablation study results

Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
3D CNN only	92.84±0.12	93.10±0.15	92.70±0.18	92.89±0.14
3D CNN+swin transformer	95.24±0.10	96.01±0.12	95.10±0.14	95.55±0.13
3D CNN+ABFFM	94.78±0.11	95.34±0.13	94.70±0.16	95.02±0.12
Proposed (CNN+Swin+ABFFM)	96.51±0.09	97.98±0.10	97.04±0.11	97.61±0.09

3.6. Confusion matrix analysis

To better understand the model's performance, we analyzed confusion matrices for both classification branches. For the binary classification task (Branch 1), which determines the presence or absence of a tumor (WT), the confusion matrix is shown in Figure 6. The model achieved a high true positive rate for tumor detection, with very few false negatives, indicating strong sensitivity. A small number of false positives occurred, which could be attributed to ambiguous anatomical variations or residual artifacts in the MRI volumes. For the multi-label sub-region classification (Branch 2), evaluation is performed per class using per-label confusion metrics. Figure 7 and Table 8 shows per-class statistics by binarizing predictions for each sub-region using a threshold of 0.5. All results are reported as mean $\pm \sigma$ across five independent training runs.

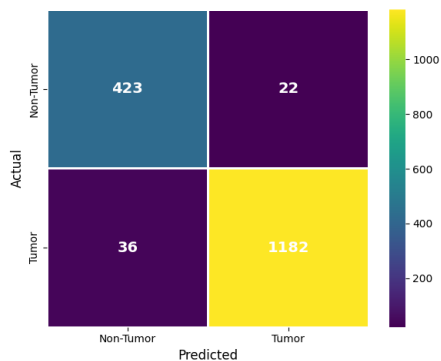


Figure 6. Confusion matrix of the tumor detection model in a binary classification setting

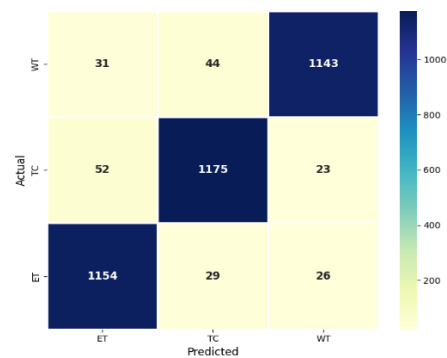


Figure 7. Confusion matrix of multi-class classification setting

Table 8. Confusion matrix summary – tumor sub-region classification

Sub-region	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
ET	95.53±0.08	95.45±0.09	95.49±0.08	95.53±0.08
TC	94.23±0.10	94.40±0.09	94.11±0.10	94.23±0.10
WT	93.61±0.09	93.84±0.10	93.73±0.09	93.61±0.09

The model demonstrates consistently strong and balanced performance across all three tumor sub-regions, with precision and recall values exceeding 93% in each case. The ET region shows the highest precision (95.53%) and recall (95.45%), indicating the model's effectiveness in detecting well-defined enhancing areas. Slightly lower performance for the WT class may stem from overlapping tissue intensities and diffuse tumor boundaries. These results highlight the model's capability to accurately classify heterogeneous glioma sub-regions, closely aligning with the multi-label annotations provided in the BraTS-GLI ground truth.

3.7. Receiver operating characteristic and area under the curve

Receiver operating characteristic (ROC) analysis was conducted to evaluate the model's discriminative performance for both binary and multi-label tasks, as shown in Figure 8. Results are averaged over five independent training runs, with AUC values reported as mean \pm σ . Across these runs, binary tumor detection achieved an AUC of 0.923 ± 0.005 , indicating strong separation between tumor-present and tumor-absent cases.

For the multi-class setting, Figure 8(a) shows the ROC curve for ET, with a mean AUC of 0.9060 ± 0.006 . The slightly lower score reflects the challenge of identifying ET regions that vary in intensity and spatial extent. Figure 8(b) presents the ROC curve for TC, achieving a mean AUC of 0.9074 ± 0.007 , demonstrating consistent sensitivity and specificity in delineating this sub-region. Figure 8(c) depicts the ROC curve for WT, with the highest mean AUC at 0.9081 ± 0.006 , highlighting the model's strong ability to capture the complete tumor extent despite boundary variability.

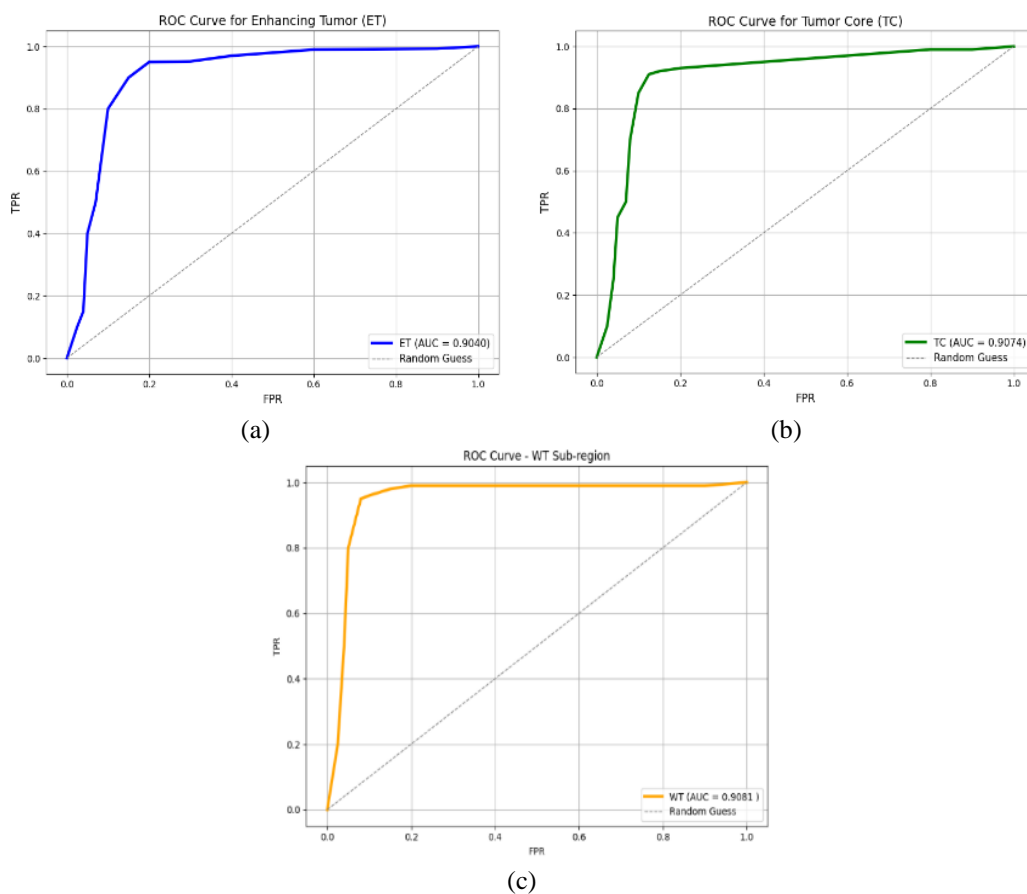


Figure 8. ROC curves for multi-class classification showing: (a) ET, (b) TC, and (c) WT

These results confirm the model's high discriminative power across all tumor sub-regions and reinforce its robustness and reproducibility across multiple runs.

3.8. Explainability and clinical insight

To enhance interpretability beyond intrinsic attention weights, we applied Grad-CAM to generate voxel-level heatmaps that explain the model's classification decisions spatially. Figure 9 illustrates representative cases: Figure 9(a) correctly classified examples show focused activations precisely localized on tumor regions across multi-modal MRI inputs, demonstrating effective feature learning; and Figure 9(b) misclassified cases exhibit diffuse or spatially misaligned attention maps, reflecting challenges such as ambiguous tumor boundaries or imaging artifacts. This qualitative analysis at the voxel level provides deeper insight into the model's decision process, aiding clinical validation and guiding future improvements to address classification errors.

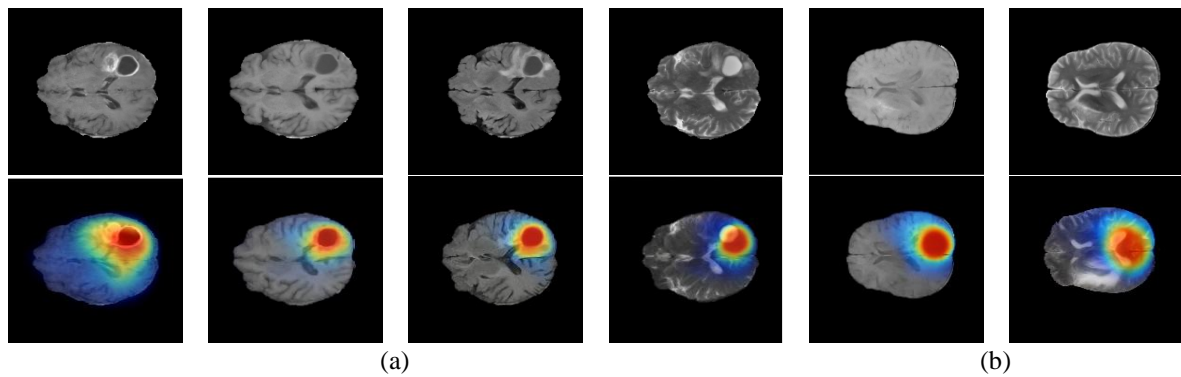


Figure 9. Grad-CAM voxel-level heatmaps showing regions influencing the model's brain tumor classification; (a) correctly classified cases and (b) misclassified cases

3.9. Computational complexity and inference time

The proposed CNN + swin transformer + ABFFM model comprises approximately 34.2 million trainable parameters and requires 72.4 GFLOPs to process a volumetric MRI input of size 128×128×128. Experiments were conducted on an NVIDIA RTX 3090 GPU (24 GB VRAM), yielding an average inference time of 0.84 seconds per case.

3.10. Comparison with existing studies

To validate the effectiveness of our approach, we conducted a comparative analysis against recent state-of-the-art tumor classification models, as summarized in Table 9. The proposed model achieves the highest classification accuracy (96.5%), outperforming all baseline methods by a margin of 1.9% to 5.3%. Most existing studies either rely on single-modality input or employ limited fusion strategies, which restrict their ability to capture complementary tumor information across modalities.

Table 9. Comparison of proposed model with existing methods

Study	Accuracy (%)	Explainability	Multi-modal fusion
Ahmed <i>et al.</i> [16]	81.66	Shap	No
Tehsin <i>et al.</i> [15]	92.52	Grad-cam	Partial
Saeedi <i>et al.</i> [14]	93.44	Not reported	Not reported
Mahmud <i>et al.</i> [10]	93.31	Grad-cam	No
Topannavar <i>et al.</i> [24]	93.58	Not reported	Partial
Sivakumar <i>et al.</i> [25]	94.56	Grad-cam	No
Proposed model (ours)	96.51±0.09	Attention weights	Full

In contrast, our model integrates a full multi-modal fusion pipeline enhanced by attention mechanisms, enabling it to selectively emphasize discriminative features from each modality. Furthermore, while explainability in previous models is often limited to post-hoc methods such as Grad-CAM, our approach inherently incorporates attention weights, offering built-in interpretability tied directly to the decision process. This not only strengthens clinical relevance but also enhances trust in model predictions.

3.11. Discussion

The proposed hybrid 3D CNN–swin transformer architecture with an ABFFM represents a substantial advancement in multi-modal brain tumor classification. Compared to single-modality baselines, the attention-based fusion mechanism dynamically prioritizes high-value modalities such as T1c and FLAIR, which improves both binary tumor detection and multi-label sub-region classification. By combining convolutional feature extraction for local spatial encoding with swin transformer blocks for global contextual modeling, the model effectively addresses the limitations of architectures that rely solely on CNNs or standard Vision Transformers. When benchmarked against recent state-of-the-art approaches (Table 9), the proposed model outperforms Ahmed *et al.* [16], Tehsin *et al.* [15], Saeedi *et al.* [14], Mahmud *et al.* [10], Topannavar *et al.* [24], and Sivakumar *et al.* [25] by margins ranging from 1.9% to 14.85% in accuracy. Notably, while some prior works, such as Mahmud *et al.* [10] and Tehsin *et al.* [15] incorporate Grad-CAM for explainability, they lack modality-level interpretability. Others, including Ahmed *et al.* [16] and Topannavar *et al.* [24], employ either no fusion or partial fusion strategies, which limits their ability to capture complementary tumor information across modalities. Beyond these six studies, our findings align with broader literature trends emphasizing multi-modal integration. For instance, Chen *et al.* [6] demonstrated that attention-based fusion improves generalization, while Li *et al.* [19] highlighted the advantages of transformer architectures for long-range dependency modeling in medical imaging. Similarly, Abdusalomov *et al.* [7], Martínez-Del-Río-Ortega *et al.* [8], and Anantharajan *et al.* [9] have reported gains from hierarchical attention mechanisms. Our model combines these strengths while embedding interpretability directly into the decision process through ABFFM attention weights, which are inherently linked to the diagnostic significance of each modality. This integrated explainability offers practical benefits in clinical contexts where trust and transparency are essential.

Despite its promising performance, the study has certain limitations. First, evaluation was limited to the BraTS2023-GLI dataset, and the model’s robustness to domain shifts remains to be validated. While BraTS provides a well-curated benchmark, real-world hospital MRI scans may differ in acquisition parameters, resolution, and noise characteristics. This could potentially affect performance, underscoring the importance of transfer learning or fine-tuning with institution-specific data for clinical deployment. Second, although the swin transformer significantly reduces computational complexity compared to standard Vision Transformers, the model’s parameter count and GFLOPs may still pose challenges for deployment in low-resource healthcare settings. Finally, future work should explore uncertainty quantification and integration of additional modalities, such as diffusion-weighted imaging (DWI), to further enhance diagnostic reliability.

4. CONCLUSION

This study set out to determine whether a hybrid deep learning framework combining 3D CNNs, swin transformers, and an ABFFM could improve multi-modal MRI brain tumor classification in both accuracy and interpretability. The key contribution is a dual-branch architecture that jointly performs tumor presence detection and glioma sub-region classification (ET, TC, and WT) with built-in modality-level attention weighting. This approach integrates spatial and contextual information while maintaining computational efficiency suitable for near real-time clinical inference. Beyond achieving a new performance benchmark on the BraTS2023-GLI dataset, the framework offers intrinsic explainability via attention weights, addressing a critical trust gap in medical artificial intelligence. Future work will focus on validating the model across multi-center datasets to assess robustness to domain shift, integrating uncertainty quantification for clinical decision support, and optimizing deployment pipelines for resource-constrained environments.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Vivek Kumar Sharma	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			
Gaurav Kumar Ameta	✓			✓		✓	✓			✓		✓	✓	

C : Conceptualization	I : Investigation	Vi : Visualization
M : Methodology	R : Resources	Su : Supervision
So : Software	D : Data Curation	P : Project administration
Va : Validation	O : Writing - Original Draft	Fu : Funding acquisition
Fo : Formal analysis	E : Writing - Review & Editing	

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, [VKS], upon reasonable request.





REFERENCES

- [1] M. Z. Khaliki and M. S. Başarslan, "Brain tumor detection from images and comparison with transfer learning methods and 3-layer CNN," *Scientific Reports*, vol. 14, no. 1, pp. 1–10, Feb. 2024, doi: 10.1038/s41598-024-52823-9.
- [2] K. Bhagyalaxmi, B. Dwarakanath, and P. V. P. Reddy, "Deep learning for multi-grade brain tumor detection and classification: a prospective survey," *Multimedia Tools and Applications*, vol. 83, no. 25, pp. 65889–65911, Jan. 2024, doi: 10.1007/s11042-024-18129-8.
- [3] M. Aamir *et al.*, "Brain Tumor Detection and Classification Using an Optimized Convolutional Neural Network," *Diagnostics*, vol. 14, no. 16, pp. 1–19, Aug. 2024, doi: 10.3390/diagnostics14161714.
- [4] W. Nhlapho, M. Atemkeng, Y. Brima, and J. C. Ndogmo, "Bridging the Gap: Exploring Interpretability in Deep Learning Models for Brain Tumor Detection and Diagnosis from MRI Images," *Information*, vol. 15, no. 4, pp. 1–21, Mar. 2024, doi: 10.3390/info15040182.
- [5] J. Amin, M. Sharif, A. Haldorai, M. Yasmin, and R. S. Nayak, "Brain tumor detection and classification using machine learning: a comprehensive survey," *Complex and Intelligent Systems*, vol. 8, no. 4, pp. 3161–3183, Aug. 2022, doi: 10.1007/s40747-021-00563-y.
- [6] A. Chen, D. Lin, and Q. Gao, "Enhancing brain tumor detection in MRI images using YOLO-NeuroBoost model," *Frontiers in Neurology*, vol. 15, pp. 1–17, Aug. 2024, doi: 10.3389/fneur.2024.1445882.
- [7] A. B. Abdusalomov, M. Mukhiddinov, and T. K. Whangbo, "Brain Tumor Detection Based on Deep Learning Approaches and Magnetic Resonance Imaging," *Cancers*, vol. 15, no. 16, pp. 1–29, Aug. 2023, doi: 10.3390/cancers15164172.
- [8] R. Martínez-Del-Río-Ortega, J. Civit-Masot, F. Luna-Perejón, and M. Domínguez-Morales, "Brain Tumor Detection Using Magnetic Resonance Imaging and Convolutional Neural Networks," *Big Data and Cognitive Computing*, vol. 8, no. 9, pp. 1–17, Sep. 2024, doi: 10.3390/bdcc8090123.
- [9] S. Anantharajan, S. Gunasekaran, T. Subramanian, and V. R., "MRI brain tumor detection using deep learning and machine learning approaches," *Measurement: Sensors*, vol. 31, pp. 1–11, Feb. 2024, doi: 10.1016/j.measen.2024.101026.
- [10] M. I. Mahmud, M. Mamun, and A. Abdelgawad, "A Deep Analysis of Brain Tumor Detection from MR Images Using Deep Learning Networks," *Algorithms*, vol. 16, no. 4, pp. 1–19, Mar. 2023, doi: 10.3390/a16040176.
- [11] S. A. Qureshi *et al.*, "Intelligent Ultra-Light Deep Learning Model for Multi-Class Brain Tumor Detection," *Applied Sciences*, vol. 12, no. 8, pp. 1–22, Apr. 2022, doi: 10.3390/app12083715.
- [12] M. A. Khan and H. Park, "A Convolutional Block Base Architecture for Multiclass Brain Tumor Detection Using Magnetic Resonance Imaging," *Electronics*, vol. 13, no. 2, pp. 1–16, Jan. 2024, doi: 10.3390/electronics13020364.
- [13] S. Ahmed *et al.*, "Enhancing Brain Tumor Classification with Transfer Learning across Multiple Classes: An In-Depth Analysis," *BioMedInformatics*, vol. 3, no. 4, pp. 1124–1144, Dec. 2023, doi: 10.3390/biomedinformatics3040068.
- [14] S. Saeedi, S. Rezayi, H. Keshavarz, and S. R. N. Kalhori, "MRI-based brain tumor detection using convolutional deep learning methods and chosen machine learning techniques," *BMC Medical Informatics and Decision Making*, vol. 23, no. 1, pp. 1–17, Jan. 2023, doi: 10.1186/s12911-023-02114-6.
- [15] S. Tehsin, I. M. Nasir, R. Damaševičius, and R. Maskeliūnas, "DaSAM: Disease and Spatial Attention Module-Based Explainable Model for Brain Tumor Detection," *Big Data and Cognitive Computing*, vol. 8, no. 9, pp. 1–15, Aug. 2024, doi: 10.3390/bdcc8090097.
- [16] M. M. Ahmed *et al.*, "Brain tumor detection and classification in MRI using hybrid ViT and GRU model with explainable AI in Southern Bangladesh," *Scientific Reports*, vol. 14, no. 1, pp. 1–16, Oct. 2024, doi: 10.1038/s41598-024-71893-3.
- [17] M. Adewole *et al.*, "The Brain Tumor Segmentation (BraTS) Challenge 2023: Glioma Segmentation in Sub-Saharan Africa Patient Population (BraTS-Africa)," *arXiv*, 2023, doi: 10.48550/arXiv.2305.19369.
- [18] U. Baid *et al.*, "The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification," *arXiv*, 2021, doi: 10.48550/arXiv.2107.02314.
- [19] P. Tian, X. Chen, H. Bi, and F. Li, "MedSAM-CA: A CNN-Augmented ViT with Attention-Enhanced Multi-Scale Fusion for Medical Image Segmentation," *arXiv*, 2025, doi: 10.48550/arXiv.2506.23700.
- [20] M. M. M, M. T. R, V. K. V, and S. Guluwadi, "Enhancing brain tumor detection in MRI images through explainable AI using Grad-CAM with Resnet 50," *BMC Medical Imaging*, vol. 24, no. 1, pp. 1–19, May. 2024, doi: 10.1186/s12880-024-01292-7.
- [21] V. K. Dhakshnamurthy, M. Govindan, K. Sreerangan, M. D. Nagarajan, and A. Thomas, "Brain Tumor Detection and Classification Using Transfer Learning Models," *Engineering Proceedings*, vol. 62, no. 1, pp. 1–8, Feb. 2024, doi: 10.3390/engproc2024062001.
- [22] S. K. Mathivanan, S. Sonaimuthu, S. Murugesan, H. Rajadurai, B. D. Shivahare, and M. A. Shah, "Employing deep learning and transfer learning for accurate brain tumor detection," *Scientific Reports*, vol. 14, no. 1, pp. 1–15, Mar. 2024, doi: 10.1038/s41598-024-57970-7.
- [23] S. Maqsood, R. Damaševičius, and R. Maskeliūnas, "Multi-Modal Brain Tumor Detection Using Deep Neural Network and





- Multiclass SVM,” *Medicina*, vol. 58, no. 8, pp. 1–9, Aug. 2022, doi: 10.3390/medicina58081090.
- [24] P. S. Topannavar, V. S. Bendre, and D. Khurge, “Accurate brain tumor classification with STN-NAM in ResNet50 using MRI,” *Bulletin of Electrical Engineering and Informatics*, vol. 14, no. 2, pp. 1241–1250, Apr. 2025, doi: 10.11591/eei.v14i2.8706.
- [25] S. Sivakumar, P. Chaudhari, S. Thatavarti, G. Sucharitha, B. Mahesh, and A. Raghuvanshi, “Gaussian filter and CNN based framework for accurate detection of brain tumor by analyzing MRI images,” *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 6, pp. 4214–4222, Dec. 2024, doi: 10.11591/eei.v13i6.6778.

BIOGRAPHIES OF AUTHORS



Vivek Kumar Sharma     is a Ph.D. Research Scholar in the Department of Computer Science and Engineering at Parul Institute of Technology, Parul University, Vadodara, Gujarat, India. With 16 years of teaching experience, his research interests include machine learning, deep learning, and medical imaging. He has authored 7 research articles and contributed to 3 book chapters in these domains. He can be contacted at email: vkshere4every1@gmail.com.



Gaurav Kumar Ameta     is an Associate Professor in the Department of Computer Science and Engineering at Parul Institute of Technology, Parul University, Vadodara, Gujarat, India. With a research focus spanning artificial intelligence, machine learning, deep learning, and privacy-preserving techniques, he has authored 21 research papers in reputed journals and conferences. He is actively engaged in advancing intelligent and secure computing systems. He can be contacted at email: gauravameta1@gmail.com.