Scientific Reports

https://doi.org/10.1038/s41598-026-43616-3

Article in Press

MM FD ConvFormer multimodal frequency aware deformable CNN transformer network for robust brain tumor classification

Received: 6 February 2026

Accepted: 5 March 2026

Published online: 09 March 2026

Cite this article as: Arockia Selvarathinam A.L.X.R., Lilhore U.K., Alroobaea R. *et al.* MM FD ConvFormer multimodal frequency aware deformable CNN transformer network for robust brain tumor classification. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-43616-3

Anto Lourdu Xavier Raj Arockia Selvarathinam, Umesh Kumar Lilhore, Roobaea Alroobaea, Majed Alsafyani, Abdullah M. Baqasah, Sultan Algarni & Monish Khan

MM FD ConvFormer Multimodal Frequency Aware Deformable CNN Transformer Network for Robust Brain Tumor Classification

[1]Anto Lourdu Xavier Raj Arockia Selvarathinam, [2,*]Umesh Kumar Lilhore, [3]Roobaea Alroobaea, [4] Majed Alsafyani, [5]Abdullah M. Baqasah, [6]Sultan Algarni, [7,*]MD Monish Khan

[1]Department of Data Science and Analytics, College of Computing, Grand Valley State University, Michigan, USA, Email: arockiaa@mail.gvsu.edu , ORCID ID:0009-0007-3389-031X
[2]School of Computing Science and Engineering, Galgotias University, Greater Noida, UP, India, Email: umeshlilhore@gmail.com
[3]Department of Computer Science , College of Computers and Information Technology, Taif University, P. O. Box 11099, Taif 21944, Saudi Arabia, Email: r.robai@tu.edu.sa
[4]Department of Computer Science, College of Computers and Information Technology, Taif University, P. O. Box 11099, Taif 21944, Saudi Arabia, Email: alsufyani@tu.edu.sa
[5]Department of Information Technology, College of Computers and Information Technology, Taif University, Taif 21974, Saudi Arabia, Email: a.baqasah@tu.edu.sa
[6]Department of Information Systems, Faculty of Computing and Information Technology King Abdulaziz University, Jeddah 21589, Saudi Arabia, Email: saalgarni@kau.edu.sa
[7,*] Research Department, Arba Minch University, Arba Minch Ethiopia, Email: drkumacse@gmail.com

[*]Corresponding Author: MD Monish Khan Email: drkumacse@gmail.com and Umesh Kumar Lilhore

Abstract

Accurate brain tumor classification from magnetic resonance imaging (MRI) is critical for early diagnosis and effective clinical decision-making. Although recent CNN–Transformer hybrid models have shown promising performance, most approaches rely primarily on single-modal spatial information, limiting their ability to capture complementary spectral features, model tumor heterogeneity, and generalize across datasets. To address these challenges, this paper proposes MM-FD-ConvFormer, a multimodal frequency-aware deformable CNN–Transformer network for robust brain tumor classification with enhanced interpretability. The proposed mode integrates three complementary modalities: (1) spatial MRI representations from original images, (2) frequency-domain MRI representations obtained via Fourier or wavelet transforms to capture texture and intensity variations, and (3) multi-scale contextual features for modeling global dependencies. A ConvNeXt V2 backbone is employed to extract discriminative spatial features, while a parallel lightweight ConvNeXt-based branch processes frequency-domain inputs. These features are subsequently fused and refined using a Swin Transformer V2 to capture long-range contextual relationships. To effectively integrate heterogeneous modalities and adapt to irregular tumor boundaries, a deformable cross-modal attention mechanism is introduced, enabling dynamic and shape-aware feature fusion. Final classification is performed on a unified multimodal representation, with an optional uncertainty-aware prediction head to improve reliability. The proposed model is evaluated using multiple public datasets, including the Kaggle Brain Tumor MRI and Figshare datasets for training, with external validation on the clinically relevant BraTS 2020/2021 dataset and optional testing on TCIA/REMBRANDT to assess cross-dataset generalization. Extensive experiments demonstrate that MM-FD-ConvFormer consistently outperforms standard CNN baselines, advanced transformer-based models, and hybrid approaches in terms of accuracy, macro-F1 score, and AUC. Furthermore, qualitative analyses using Grad-CAM, attention map visualization, and weakly supervised pseudo-segmentation provide interpretable insights into tumor localization and model decision-making. Overall, MM-FD-ConvFormer offers a robust, interpretable, and generalizable solution for automated brain tumor classification in real-world clinical imaging applications.

Keywords: Brain tumor classification, Multimodal learning, Frequency-domain attention, Deformable attention, CNN–Transformer hybrid, Magnetic resonance imaging, Cross-dataset generalization.

1. Introduction

Brain tumors represent one of the most severe neurological disorders, often leading to high mortality and long-term cognitive impairment if not diagnosed and treated at an early stage. Accurate tumor classification is essential for treatment planning, prognosis estimation, and personalized clinical decision-making. Magnetic Resonance Imaging (MRI) is the preferred imaging modality for brain tumor assessment due to its superior soft-tissue contrast and ability to capture detailed anatomical and pathological variations without ionizing radiation [1,2]. However, manual interpretation of MRI scans remains time-consuming, subjective, and highly dependent on expert radiologists, thereby motivating the development of automated, reliable computer-aided diagnosis systems [3,4].

Recent advances in deep learning have substantially improved automated brain tumor classification performance. Convolutional Neural Networks (CNNs), including architectures such as ResNet, DenseNet, and EfficientNet, demonstrate strong spatial feature extraction capabilities [1,4]. More recently, transformer-based and CNN–Transformer hybrid architectures have enhanced global contextual modeling by capturing long-range dependencies [5,6]. Despite these advances, several challenges persist. First, tumor heterogeneity—including variations in size, morphology, texture, and intensity—complicates robust feature learning across different tumor subtypes and grades [6,9]. Second, many models exhibit limited cross-dataset generalization, with noticeable performance degradation when evaluated on data acquired from different scanners, imaging protocols, or institutions [7]. Third, limited interpretability of deep learning systems restricts their clinical adoption, as transparent and explainable predictions are essential for diagnostic trust [3,8].

Most existing approaches rely primarily on single-domain spatial representations extracted from MRI intensity data [10]. Although attention mechanisms have been introduced to improve discriminative learning, conventional fixed-grid attention often struggles to capture irregular and infiltrative tumor boundaries, particularly in gliomas [11]. Moreover, standard attention mechanisms typically operate within a single modality, refining features without explicitly modeling interactions between heterogeneous representations. This intra-modal design may limit adaptability under domain shift and reduce robustness when spatial appearance varies significantly across datasets [12].

To address these limitations, multimodal representation learning has emerged as a promising paradigm in medical image analysis [13]. By integrating complementary feature spaces—such as spatial and frequency-domain representations—models can capture richer structural and textural information. Frequency-domain analysis, in particular, has demonstrated effectiveness in highlighting subtle texture variations and intensity transitions that are less apparent in the spatial domain [14]. In parallel, deformable attention mechanisms

enable adaptive sampling over irregular regions, making them well-suited for modeling complex tumor morphologies [15]. However, existing deformable attention approaches in medical imaging are typically applied within a single representation space and do not explicitly account for cross-modal feature interactions.

Motivated by these observations, we propose MM-FD-ConvFormer, a multimodal frequency-aware deformable CNN–Transformer network for robust brain tumor classification. The proposed framework integrates spatial MRI features and frequency-domain representations through a Deformable Cross-Modal Attention (DCMA) mechanism. Unlike conventional deformable attention modules that compute offsets solely within a single feature space, DCMA learns adaptive sampling offsets from fused spatial–frequency embeddings. This modality-aware offset conditioning enables dynamic alignment between anatomical structures and spectral texture cues, facilitating shape-adaptive cross-modal feature interaction. By explicitly modeling spatial–frequency complementarity, DCMA represents a structural departure from existing deformable attention designs commonly used in CNN–Transformer hybrids.

Extensive experiments are conducted on multiple publicly available datasets, including Kaggle Brain Tumor MRI [33], Figshare Brain Tumor Dataset [34], BraTS 2020/2021 [35], and TCIA REMBRANDT [36], to evaluate classification performance, cross-dataset generalization, and interpretability. Furthermore, Grad-CAM and Grad-CAM++ visualizations are employed to provide class-wise and region-wise explanations aligned with clinical tumor annotations.

Contributions

In this work, *multimodal* refers to the integration of heterogeneous feature spaces derived from a single imaging modality. We treat the spatial domain (CNN-based stream) and the frequency domain (spectral-transformer stream) as complementary modalities that jointly characterize tumor morphology and peritumoral texture. The primary contributions of this study are as follows:

- We propose a novel multimodal MM-FD-ConvFormer architecture that jointly models spatial, frequency-domain, and multi-scale contextual MRI representations.
- We introduce frequency-aware feature learning to capture complementary spectral characteristics associated with tumor texture heterogeneity and intensity variation.
- We design a Deformable Cross-Modal Attention (DCMA) mechanism that adaptively fuses heterogeneous feature streams through modality-aware offset learning, enabling shape-adaptive cross-modal interaction.
- We perform comprehensive cross-dataset validation on Kaggle, Figshare, BraTS 2020/2021, and TCIA REMBRANDT datasets to demonstrate robustness and generalization.

- We provide interpretability and weak localization analysis using Grad-CAM, Grad-CAM++, and SHAP to enhance clinical transparency and trustworthiness.

Organization of the Article

The remainder of this article is organized as follows. Section 2 reviews related work on CNN-based, transformer-based, and multimodal brain tumor classification methods. Section 3 details the proposed MM-FD-ConvFormer architecture and its core components. Section 4 describes the datasets, preprocessing strategy, and experimental setup. Section 5 outlines comparative models and evaluation metrics. Section 6 presents quantitative results, cross-dataset generalization analysis, ablation studies, and interpretability evaluation. Finally, Section 7 concludes the paper and discusses future research directions.

2. Related Work

2.1 CNN-Based Brain Tumor Classification

Convolutional neural networks (CNNs) have been extensively explored for automated brain tumor classification due to their strong capability in hierarchical feature extraction from medical images. Early studies employed handcrafted features combined with shallow classifiers; however, deep CNN architectures have largely replaced these approaches by learning discriminative representations directly from MRI data (Hasan et al., 2024; Zhang et al., 2023). Popular architectures such as ResNet, DenseNet, and EfficientNet have demonstrated competitive performance in multi-class brain tumor classification tasks by capturing spatial and textural information from MRI slices (Shah et al., 2022; Anand et al., 2026; Ke et al., 2026).

Several works have further enhanced CNN-based models through channel-wise attention, ensemble learning, or hybrid classifiers. For example, Balamurugan et al. (2024) proposed a channel-attention-based fusion strategy to improve robustness, while Shinde et al. (2024) focused on scalable CNN designs for healthcare deployment. Hybrid CNN–SVM and CNN–ANN frameworks have also been explored to improve classification accuracy and stability (Zhang et al., 2023; Ganesh et al., 2024). Despite these advancements, CNN-based methods primarily rely on single-modal spatial information, limiting their ability to capture complementary representations and generalize across heterogeneous datasets.

2.2 Transformer and CNN–Transformer Hybrid Models

More recently, transformer-based architectures have gained attention in medical image analysis due to their ability to model long-range dependencies through self-attention mechanisms. Vision Transformers and Swin Transformers have been successfully applied to brain tumor classification, achieving improved global context modeling compared to conventional CNNs (Sahoo et al., 2026; Pacal and Banerjee, 2026). Transformer-based EfficientNet variants and hybrid CNN–Transformer models further demonstrate the benefit of combining local feature

extraction with global attention mechanisms (Sahoo et al., 2026; Tang et al., 2026).

CNN–Transformer hybrid models aim to leverage the complementary strengths of CNNs and transformers, where CNNs capture local spatial patterns and transformers encode long-range contextual relationships. Recent studies have reported performance improvements using such hybrid frameworks for brain tumor detection and classification (Nassar et al., 2024; Agarwal et al., 2024). However, most existing hybrid models employ fixed-grid attention mechanisms, which are not well-suited for modeling irregular tumor boundaries and complex morphological variations, particularly in infiltrative tumors such as gliomas.

2.3 Frequency-Domain Learning in Medical Imaging

Frequency-domain analysis has emerged as an effective strategy for capturing complementary information that may be overlooked in the spatial domain. By transforming medical images using Fourier or wavelet transforms, models can exploit spectral characteristics related to texture, intensity variations, and structural patterns. Frequency-aware learning has shown promising results in diverse biomedical applications, including EEG decoding and multimodal MRI synthesis (Jin et al., 2025; Jiang et al., 2025).

In the context of brain imaging, frequency-domain representations have been used to enhance tumor characterization by highlighting subtle texture differences between tumor and healthy tissue (Jin et al., 2025). Wavelet-based CNNs and frequency-enhanced models demonstrate improved sensitivity to edge and boundary information, which is critical for tumor analysis. Nevertheless, frequency-domain learning remains underexplored in brain tumor classification, and most existing approaches do not explicitly integrate frequency information with spatial and contextual features in a unified framework.

2.4 Multimodal and Interpretability-Driven Models

Multimodal learning has been increasingly investigated to improve robustness and generalization in brain tumor analysis by combining information from multiple MRI sequences, feature representations, or learning paradigms. Several studies have demonstrated that fusing multimodal inputs leads to improved classification performance compared to single-modal approaches (Lerousseau et al., 2020; Sharif et al., 2022; Usha et al., 2024). More recent works employ deep multimodal fusion strategies, integrating CNN-based feature extractors with attention mechanisms to capture complementary information across modalities (Rohini et al., 2023; Ullah et al., 2024).

In parallel, interpretability has become a crucial requirement for clinical adoption of deep learning models. Visualization techniques such as Grad-CAM have been widely used to provide class-discriminative localization maps, offering insights into model decision-making (Pacal and Banerjee, 2026; Balamurugan et al., 2024). However, many interpretability-driven models are limited to post hoc visualization and lack architectural mechanisms that inherently support adaptive and explainable feature fusion. Furthermore, uncertainty-aware learning and

weakly supervised localization remain insufficiently explored in multimodal brain tumor classification settings.

2.5 Research Gap

Despite significant progress in CNN-based, transformer-based, frequency-aware, and multimodal approaches, existing methods lack a unified framework that jointly integrates spatial, spectral, and contextual features through deformable attention while providing clinically meaningful interpretability and robust cross-dataset generalization. This gap motivates the development of a multimodal, frequency-aware, and deformable CNN–Transformer architecture with built-in interpretability and uncertainty awareness, as proposed in this work.

Table 1. Comparison of representative brain tumor classification methods.

Ref.	Method	Learning Paradigm	Multimodal	Frequency-Aware	Attention Used	Rigid Attention	Deformable Attention	Interpretability	Cross-Dataset Validation
Shah et al. (2022)	Fine-tuned EfficientNet	CNN	No	No	No	No	No	No	No
Anand et al. (2026)	Optimized CNN + TL	CNN	No	No	No	No	No	No	No
Ke et al. (2026)	Multi-scale CNN + SVM	CNN	No	No	Channel	Yes	No	No	No
Balamurugan et al. (2024)	CNN + Channel Fusion	CNN	No	No	Channel	Yes	No	Yes	No
Tang et al. (2026)	RFENet (Three-branch CNN)	CNN	No	No	Fixed Spatial	Yes	No	No	No
Sahoo et al. (2026)	EfficientB0Net (Transformer)	Transformer	No	No	Self-Attention	Yes	No	No	No
Pacal & Banerjee (2026)	T-FSPANNet	CNN–Transformer	No	No	Pyramidal	Yes	No	Yes	No
Jin et al. (2025)	Frequency-Aware Attention Net	Attention-based	No	Yes	Fixed Attention	Yes	No	Yes	No
Jiang et al. (2025)	Frequency-Aware	Multimodal	Yes	Yes	Cross-Modal	Yes	No	No	No

	Diffusion Model								
Usha et al. (2024)	Tumnet	Multimodal CNN	Yes	No	Channel	Yes	No	No	No
Ullah et al. (2024)	BrainNet	Multimodal CNN	Yes	No	Fixed Fusion	Yes	No	No	No
Fayjie et al. (2026)	FALCON	Few-shot / Adversarial	Yes	No	Adaptive	Yes	No	No	Yes
Proposed Model	MM-FD-ConvFormer	CNN–Transformer	Yes	Yes	Cross-Modal	No	Yes	Yes	Yes

As shown in Table 1, most existing brain tumor classification methods rely on rigid attention mechanisms and single-modal spatial representations, limiting their ability to adapt to irregular tumor morphology and domain shifts. Although several multimodal and frequency-aware approaches have been proposed, they typically employ fixed fusion strategies without deformable attention or comprehensive cross-dataset validation. In contrast, the proposed MM-FD-ConvFormer uniquely combines multimodal spatial and frequency representations with deformable cross-modal attention, while providing interpretable predictions and robust generalization across multiple heterogeneous datasets.

3. Proposed Methodology: MM-FD-ConvFormer

This section presents MM-FD-ConvFormer, a multimodal frequency-aware deformable CNN–Transformer architecture for robust and interpretable brain tumor classification from MRI. Unlike conventional approaches that rely primarily on single-modal spatial features, the proposed model jointly models spatial appearance, frequency-domain characteristics, and global contextual dependencies to better capture tumor heterogeneity, irregular boundaries, and improve cross-dataset generalization [16, 17].

A key novelty of MM-FD-ConvFormer is the incorporation of deformable cross-modal attention, which dynamically adapts attention to irregular tumor shapes, enabling shape-aware fusion of spatial and frequency information. The refined multimodal features are aggregated using global average pooling and passed to a classification head with Monte Carlo Dropout, providing uncertainty-aware predictions under domain shifts and ambiguous imaging conditions. Together, these components enable MM-FD-ConvFormer to deliver an interpretable, generalizable, and clinically meaningful solution for automated brain tumor classification [4, 7, 19].

In the final MM-FD-ConvFormer configuration, a 1-level Haar Discrete Wavelet Transform (DWT) was adopted for frequency-domain representation. While both Fast Fourier Transform (FFT) and DWT were initially explored, empirical comparison demonstrated superior robustness and boundary sensitivity using DWT. Therefore, all reported final results correspond to the DWT-based frequency branch unless otherwise specified.

Figure 1 presents the Architecture of the proposed MM-FD-ConvFormer. As shown in Figure 1, an input MRI slice of size $224 \times 224 \times 3$is processed through two complementary feature extraction pathways. The spatial branch, based on ConvNeXt V2, hierarchically encodes anatomical structure and tumor morphology, while the frequency-aware branch operates on FFT- or DWT-transformed representations to emphasize texture variations, intensity transitions, and boundary-related spectral cues that are often underrepresented in the spatial domain. The resulting features are spatially aligned and fused via channel-wise concatenation, followed by refinement using a Swin Transformer V2 to capture long-range contextual relationships [18].

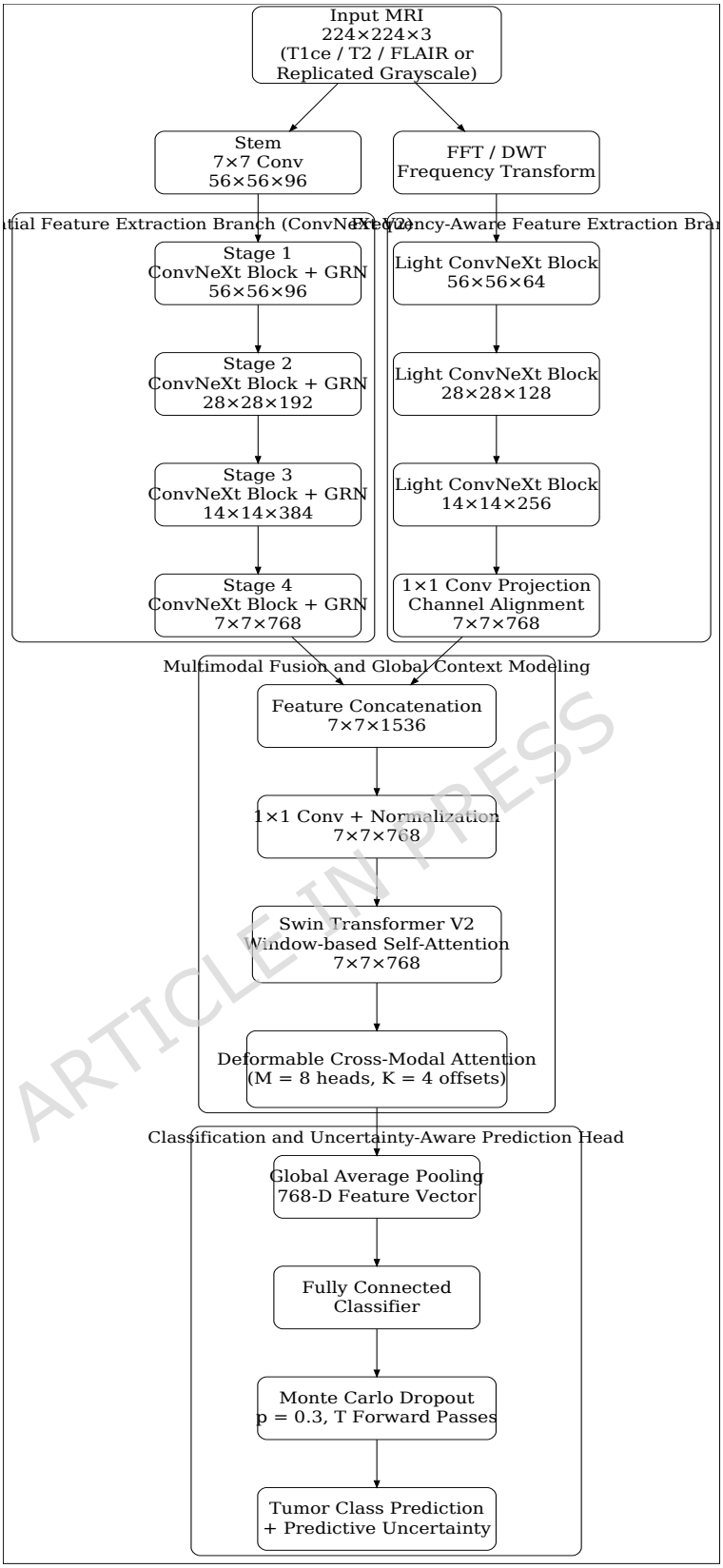

Figure 1. Architecture of the proposed MM-FD-ConvFormer.

3.1 Input Configuration and Multimodal Representation

All MRI slices are resized to a fixed spatial resolution of $224 \times 224$to ensure compatibility with modern convolutional and transformer-based backbones. For single-sequence datasets such as Kaggle and Figshare, grayscale MRI slices are replicated across three channels. In contrast, for multi-sequence datasets including BraTS and REMBRANDT, three clinically informative modalities (e.g., T1ce, T2, and FLAIR) are mapped directly to the channel dimension [20]. Each input sample is therefore represented as Equation 1.

$$X \in R^{224\times224\times3}. \quad (1)$$Pixel intensities are normalized using z-score normalization on a per-image basis to reduce scanner-specific bias and stabilize both spatial and spectral representations [21].

From the normalized input X, two complementary representations are constructed. The spatial representation as Equation 2.

$$X_s = X \quad (2)$$It preserves anatomical structure, tumor morphology, and intensity contrast, which are critical for identifying tumor location and extent. In parallel, a frequency-domain representation is generated using either the Fast Fourier Transform (FFT) or the Discrete Wavelet Transform (DWT) (Equation 3):

$$X_f = \begin{cases} |F(X)|, & \text{FFT magnitude spectrum,} \\ W(X), & \text{DWT sub-band decomposition.} \end{cases} \quad (3)$$While the spatial representation emphasizes shape and structural cues, the frequency-domain representation highlights texture irregularities, intensity transitions, and boundary-related patterns that are often suppressed by smoothing and noise in the spatial domain—features that are highly relevant for characterizing heterogeneous brain tumors [3, 22].

3.1.1 Frequency-Domain Representation and Implementation Details

To clarify the implementation details of the frequency-domain branch, we provide the following specifications for both FFT and DWT variants explored during model development.

- **FFT Representation:**

For the FFT-based configuration, the two-dimensional Fast Fourier Transform was applied to each MRI slice. Only the magnitude spectrum was retained, while the phase component was discarded. Although phase information can encode structural alignment, preliminary experiments indicated increased sensitivity to minor spatial shifts and acquisition noise when phase was incorporated. Therefore, to enhance robustness and reduce instability under domain variation, only the magnitude spectrum was used.

To stabilize the dynamic range of spectral coefficients, logarithmic scaling was applied:

$$F_{log} = \log(1 + |F(u,v)| \quad (4))$$

where $F(u,v)$denotes the Fourier transform at frequency coordinates (u,v). The resulting magnitude map was normalized and resized to match the spatial resolution of the input slice.

- **DWT Representation:**
  For the DWT-based configuration (used in the final model), a 1-level Haar Discrete Wavelet Transform was applied to each MRI slice, producing four sub-bands:
  - LL (low-frequency approximation)
  - LH (horizontal detail)
  - HL (vertical detail)
  - HH (diagonal detail)

These sub-bands were concatenated along the channel dimension to form a 4-channel frequency tensor:

$$F_{DWT} = [LL \,||\, LH \,||\, HL \,||\, HH] \quad (5)$$

where $||$denotes channel-wise concatenation.

Because the lightweight ConvNeXt block expects a fixed channel dimensionality, a 1×1 convolutional projection layer was introduced to map the 4-channel DWT tensor to the required embedding dimension. This projection enables seamless integration with the subsequent frequency encoder while preserving multi-resolution spectral information.

3.2 Dual-Stream Feature Extraction

To learn complementary information from both representations, MM-FD-ConvFormer adopts a dual-stream architecture. The spatial stream employs ConvNeXt V2-Tiny as its backbone. This network consists of four hierarchical stages with progressively reduced spatial resolution ($56 \times 56 \rightarrow 28 \times 28 \rightarrow 14 \times 14 \rightarrow 7 \times 7$) and increased channel dimensionality (up to 768 channels). ConvNeXt V2 integrates large-kernel depthwise convolutions with Global Response Normalization (GRN), which stabilizes feature distributions and prevents feature collapse across layers. This design allows the spatial stream to preserve fine-grained tumor morphology and anatomical continuity while gradually encoding higher-level semantic information [6, 23]. The resulting spatial feature map is (Equation 6).

$$F_s = E_s(X_s), F_s \in R^{7\times7\times768}. \quad (6)$$In parallel, the frequency-aware stream processes the spectral representation using a lightweight ConvNeXt-based encoder with reduced depth and channel width to avoid overfitting and excessive computational cost. This stream captures discriminative spectral patterns across multiple resolutions, including low-frequency components related to global tumor shape and high-frequency components associated with texture variation and

boundary sharpness [6, 23]. After projection and spatial alignment, the frequency stream produces as (Equation 7).

$$F_f = E_f(X_f), F_f \in R^{7\times7\times768}. \quad (7)$$

3.3 Multimodal Fusion and Global Context Modeling

The spatial and frequency feature maps are concatenated along the channel dimension and projected into a unified embedding space (Equation 8):

$$F_{sf} = \psi(F_s, F_f), F_{sf} \in R^{7\times7\times1536}. \quad (8)$$

A (1×1) convolution followed by normalization reduces the channel dimensionality to 768, ensuring numerical stability and compatibility with subsequent transformer processing [24].

To model long-range dependencies and global anatomical context, the fused representation is processed by a Swin Transformer V2 block operating at the (7×7) resolution. Using window-based self-attention with shifted windows, the transformer captures contextual relationships across spatial regions while preserving locality [25]. This capability is particularly important for brain tumor classification, where tumors with similar local appearance may differ in their global spatial context as (Equation 9).

$$F_g = T(F_{sf}). \quad (9)$$

3.4 Deformable Cross-Modal Attention

Although transformer self-attention effectively models long-range dependencies and global contextual relationships, conventional multi-head self-attention operates over fixed-grid sampling locations. Such fixed spatial aggregation is suboptimal for capturing the irregular, heterogeneous, and infiltrative morphologies characteristic of brain tumors. To address this limitation, MM-FD-ConvFormer introduces a Deformable Cross-Modal Attention (DCMA) mechanism that adaptively adjusts attention sampling locations in a shape-aware manner.

For a query position p, the deformable attention output is computed as:

$$Z(p) = \sum_{m=1}^{M} \sum_{k=1}^{K} A_{m,k}(p)\, W_m\, V(p + \Delta p_{m,k}(p)), \quad (10)$$

where M denotes the number of attention heads, K represents the number of sampling points per head, $A_{m,k}(p)$ are the learned attention weights, and $\Delta p_{m,k}(p)$ are the learnable spatial offsets that dynamically deform the sampling grid.

Architectural Distinction from Conventional Deformable Attention

It is important to emphasize that DCMA differs structurally from conventional deformable attention mechanisms commonly used in CNN–Transformer hybrids. In standard deformable attention, the sampling offsets Δp are typically predicted from features within a single modality (e.g., spatial features only), and the attention operation refines intra-modal relationships.

In contrast, DCMA performs cross-modal offset conditioning. Let F_s and F_f denote the spatial-domain and frequency-domain feature maps, respectively. Instead of learning offsets solely from F_s, DCMA computes:

$$\Delta p = G([F_s \oplus F_f]), \quad (11)$$

where $\oplus$ denotes multimodal feature fusion and $G(\cdot)$ represents the offset prediction network. By deriving offsets from fused spatial–frequency embeddings, DCMA enables modality-aware offset learning, allowing anatomical structure and spectral texture cues to jointly influence sampling locations.

This design makes the attention mechanism both shape-adaptive and cross-modal, dynamically aligning heterogeneous representations rather than refining a single feature stream. Consequently, DCMA extends deformable attention beyond intra-modal feature refinement and facilitates structured interaction between spatial morphology and frequency-domain texture characteristics.

By allowing the receptive field to deform according to multimodal cues, DCMA enhances boundary sensitivity and improves alignment around irregular tumor regions. To balance representational capacity and computational efficiency, the DCMA block is applied once within the multimodal fusion stage, ensuring effective cross-modal integration without excessive overhead.

3.5 Classification and Uncertainty-Aware Prediction

The refined multimodal features are aggregated using global average pooling, as Equation 12.

$$Z = \mathrm{GAP}(Z(p)) \in R^{768}, \quad (12)$$

and passed to a fully connected classification layer to produce tumor class probabilities (Equation 13):

$$\hat{y} = \mathrm{Softmax}(W_c Z + b_c). \quad (13)$$

To enhance reliability under domain shifts and ambiguous imaging conditions, an uncertainty-aware prediction head based on Monte Carlo Dropout is employed during inference. Multiple stochastic forward passes yield predictions $\hat{y}^{(t)}$, from which the predictive mean μ and variance are computed (Equation 14).

$$\sigma^2 = \frac{1}{T}\sum_{t=1}^{T}\left(\hat{y}^{(t)} - \mu\right)^2 \quad (14)$$

This uncertainty estimate is particularly valuable for clinically challenging datasets such as BraTS and REMBRANDT, where blurred boundaries and heterogeneous tumor appearance may lead to ambiguous predictions [27].

The network is trained end-to-end using a weighted cross-entropy loss to account for class imbalance (Equation 15):

$$L = -\sum_{c=1}^{C} w_c \, y_c \log(\hat{y}_c). \quad (15)$$

By jointly integrating spatial MRI features, frequency-domain representations, and global contextual modeling through deformable cross-modal attention and uncertainty-aware inference, MM-FD-ConvFormer provides a robust, interpretable, and generalizable framework for automated brain tumor classification [3, 28]. The proposed methodology directly reflects the multimodal, frequency-aware, and deformable principles emphasized in the title and abstract, ensuring coherence between methodological design and research objectives.

3.6 Algorithm and Flowchart for Proposed Model

The complete operational pipeline of the proposed MM-FD-ConvFormer is summarized in Algorithm 1, which outlines the sequential steps from multimodal representation construction to uncertainty-aware classification. The corresponding end-to-end data flow and conditional decision paths are illustrated in Figure 2, providing a clear and interpretable visualization of the proposed model.

Algorithm 1: Algorithm for MM-FD-ConvFormer for Multimodal Brain Tumor Classification
Input: MRI slice $X \in R^{224\times224\times3}$ *Output: Predicted tumor class* $\hat{y}$ *and predictive uncertainty* σ^2
Step 1: Input Preparation and Normalization *1.1 Resize each MRI slice to a fixed spatial resolution of* 224×224 *to ensure architectural consistency across datasets.* *1.2 Construct the input channel configuration:* • *Replicate grayscale slices across three channels for single-sequence datasets.* • *Map clinically relevant MRI sequences (e.g., T1ce, T2, FLAIR) to the channel dimension for multi-sequence datasets.* *1.3 Apply z-score normalization independently to each slice to reduce scanner-dependent intensity variations and stabilize downstream spatial and frequency-domain feature extraction.* *Step 2: Multimodal Representation Construction* *2.1 Preserve the normalized MRI slice as the spatial representation, retaining anatomical structure, tumor morphology, and intensity contrast.* *2.2 Generate a frequency-domain representation by applying either:* • *Fast Fourier Transform (FFT) to obtain the magnitude spectrum, or* • *Discrete Wavelet Transform (DWT) to obtain multi-resolution sub-band representations.* *2.3 Ensure that frequency-domain transformation is applied after spatial augmentation to maintain spectral consistency.* *Step 3: Dual-Stream Feature Extraction* *3.1 Process the spatial representation through the ConvNeXt V2 backbone to extract hierarchical spatial features encoding tumor size, shape, location, and surrounding anatomical context.*

3.2 Process the frequency-domain representation through a lightweight ConvNeXt-based encoder to learn discriminative spectral features emphasizing texture irregularities, intensity transitions, and boundary sharpness.
3.3 Align the frequency-domain feature maps spatially and channel-wise with the spatial feature maps to enable effective multimodal fusion.
Step 4: Multimodal Feature Fusion
4.1 Concatenate the aligned spatial and frequency feature maps along the channel dimension to form a unified multimodal representation.
4.2 Apply a projection and normalization operation to reduce dimensionality and stabilize the fused feature distribution.
Step 5: Global Context Modeling
5.1 Pass the fused multimodal features through a Swin Transformer V2 module.
5.2 Employ window-based self-attention with shifted windows to model long-range contextual dependencies while preserving spatial locality.
5.3 Encode global anatomical relationships that support discrimination between tumor types with similar local appearance but different spatial distributions.
Step 6: Deformable Cross-Modal Attention
6.1 Apply deformable cross-modal attention to the context-enhanced features.
6.2 Learn adaptive sampling offsets that allow the attention mechanism to focus on irregular tumor boundaries and heterogeneous regions.
6.3 Refine multimodal feature integration in a shape-aware manner without imposing fixed-grid constraints.
Step 7: Feature Aggregation
7.1 Aggregate the refined feature maps using global average pooling to obtain a compact, fixed-length feature vector suitable for classification.
Step 8: Uncertainty-Aware Classification
8.1 Pass the aggregated feature vector through a fully connected classification layer to produce tumor class probabilities.
8.2 During inference, apply Monte Carlo Dropout over multiple stochastic forward passes to account for model uncertainty.
Step 9: Final Prediction and Uncertainty Estimation
9.1 Compute the final tumor class prediction as the mean of the stochastic predictions.
9.2 Estimate predictive uncertainty as the variance across these predictions, highlighting ambiguous or out-of-distribution cases.

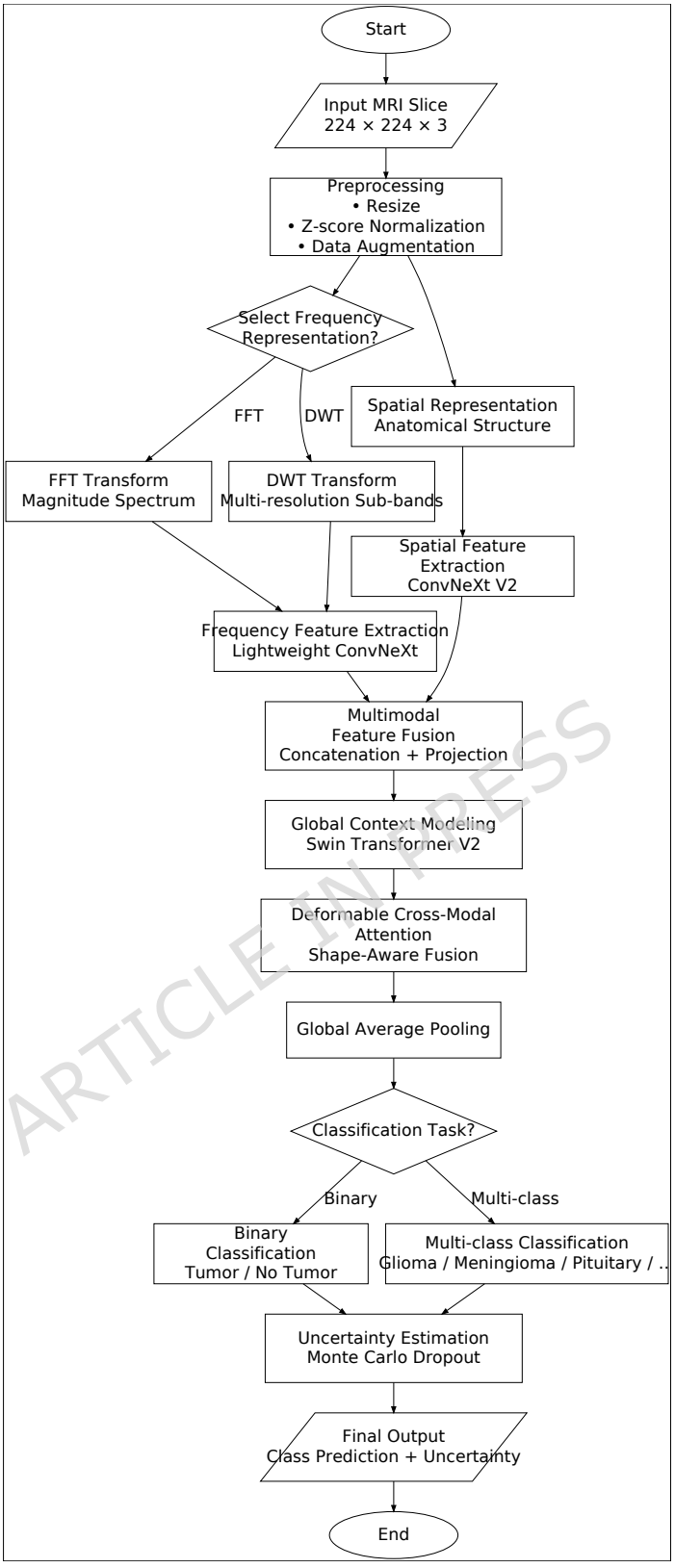

Figure 2: Flowchart of the proposed MM-FD-ConvFormer illustrating multimodal MRI processing, conditional frequency-domain representation (FFT/DWT), deformable cross-modal attention, and task-specific binary or multi-class classification with uncertainty estimation.

4. Datasets and Experimental Setup

4.1 Datasets Description

To comprehensively evaluate the robustness and generalizability of the proposed MM-FD-ConvFormer, multiple publicly available brain MRI datasets were employed, encompassing diverse imaging protocols, class distributions, and clinical characteristics. The combined use of these datasets enables the assessment of both in-distribution performance and cross-dataset generalization under realistic clinical conditions (Table 2).

The Kaggle Brain Tumor MRI dataset (Masoud Nickparvar) was used as the primary training dataset due to its high-quality annotations and balanced class distribution. It contains 7,023 T1-weighted MRI slices categorized into four classes: glioma (1,621), meningioma (1,645), pituitary tumor (1,757), and no tumor (2,000). The inclusion of a no-tumor class supports both binary and multi-class classification settings [33].

To enhance data diversity and evaluate robustness under class imbalance, the Figshare Brain Tumor Dataset was incorporated. This dataset consists of 3,064 contrast-enhanced T1-weighted MRI slices stored in MATLAB format, covering three tumor classes: glioma (1,426), meningioma (708), and pituitary tumor (930). Figshare exhibits a pronounced class imbalance, making it particularly suitable for assessing the effectiveness of class imbalance mitigation strategies [34].

For external validation and clinical relevance, the BraTS 2021 dataset was used exclusively for testing and qualitative analysis. BraTS provides multi-sequence 3D MRI volumes (T1, T1ce, T2, and FLAIR) with expert-annotated tumor sub-regions, including whole tumor (WT), tumor core (TC), and enhancing tumor (ET). All cases correspond to gliomas, reflecting real-world clinical complexity. Additionally, the TCIA REMBRANDT dataset, comprising 130 patients with multi-sequence MRI and extensive clinical metadata, was used to evaluate real-world generalization and uncertainty-aware prediction under domain shift [35, 36].

Table 2: summarizes the datasets used in this study.

Dataset	Samples	Classes	Normal / Abnormal	Clinical Data	Balance
Kaggle [33]	7,023 slices	4	Yes	No	Balanced
Figshare [34]	3,064 slices	3	No	No	Imbalanced
BraTS 2021 [35]	>1,250 cases	WT, TC, ET	No	Yes	Imbalanced
REMBRANDT [36]	130 patients	Glioma grades	No	Extensive	Imbalanced

4.2 Data Pre-processing and Class Imbalance Handling

Due to the heterogeneous formats and acquisition protocols across datasets, a unified preprocessing pipeline was applied to standardize spatial and spectral representations. All MRI scans were converted into 2D slice-level inputs to ensure compatibility with the ConvNeXt V2 backbone. Multi-sequence datasets (BraTS and REMBRANDT) were mapped to three channels using clinically informative modalities (T1ce, T2, and FLAIR), whereas single-sequence datasets (Kaggle and Figshare) were replicated across three channels to maintain architectural consistency. All slices were resized to a fixed resolution of 224×224, with zero-padding applied where necessary to preserve anatomical proportions [27, 28].

To mitigate scanner-specific intensity variations and stabilize feature learning across spatial and frequency domains, z-score normalization was applied independently to each slice (Equation 16):

$$I_{norm} = \frac{I - \mu}{\sigma}, \quad (16)$$

Where I denotes the original pixel intensity, and μ and σ represent the slice-wise mean and standard deviation, respectively. For BraTS and REMBRANDT datasets, N4 bias field correction was additionally employed to reduce low-frequency intensity inhomogeneities.

To construct the frequency-domain modality, normalized slices were transformed using either the Discrete Fourier Transform (DFT) or the Discrete Wavelet Transform (DWT). The magnitude spectrum of the DFT was computed as (Equation 14):

$$X_f(u,v) = \left| \sum_{x=0}^{H-1} \sum_{y=0}^{W-1} X(x,y)\, e^{-j2\pi\left(\frac{ux}{H}+\frac{vy}{W}\right)} \right|, \quad (17)$$

Where $X(x,y)$ represents the spatial-domain image and (u,v) denote frequency coordinates. For DWT-based representations, multi-resolution sub-bands were extracted to capture localized frequency variations, which are particularly effective in highlighting tumor boundary irregularities and heterogeneous texture patterns. All frequency transformations were applied after spatial augmentation to preserve spectral consistency [29].

Class imbalance was addressed using a combination of data-level and loss-level strategies. For the Figshare dataset, minority classes were oversampled using rotation-based augmentation, which preserves both spatial structure and frequency characteristics. The Kaggle dataset required no additional balancing due to its relatively uniform class distribution. BraTS and REMBRANDT datasets were intentionally left imbalanced to reflect real-world clinical prevalence [30]. During training, a weighted cross-entropy loss was employed (Equation 18):

$$L = -\sum_{c=1}^{C} w_c\, y_c \log(\hat{y}_c), \quad (18)$$

Where w_c denotes the inverse-frequency weight for class c, y_c is the ground-truth label, and $\hat{y}_c$ is the predicted probability.

4.3 Dataset Splitting and Experimental Protocol

For controlled experimentation, the Kaggle and Figshare datasets were split into training, validation, and internal test sets using a stratified ratio of 70% / 15% / 15%, ensuring consistent class distributions across splits. Data augmentation was applied exclusively to the training set to prevent information leakage [11, 31].

The BraTS 2021 and REMBRANDT datasets were used strictly as external test sets and were not involved in training, validation, or hyperparameter tuning. This protocol enables an unbiased evaluation of cross-dataset generalization and uncertainty-aware prediction under domain shift. Table 3 reports the final class distribution used in the experiments [5, 32].

Table 3: Final distribution of datasets after preprocessing and class imbalance correction.

Dataset	Class	Original Count	Final Count	Usage
Kaggle	Glioma / Meningioma / Pituitary	~1.6k–1.8k	2,000 each	Train / Val / Test
Kaggle	No Tumor	2,000	2,000	Train / Val / Test
Figshare	Meningioma / Pituitary	708 / 930	1,426	Train / Val / Test
BraTS 2021	Glioma cases	1,251	Unchanged	External Test
REMBRANDT	Glioma patients	130	Unchanged	External Test

This structured data strategy ensures that MM-FD-ConvFormer is trained on diverse yet standardized inputs, robustly handles class imbalance, and is rigorously evaluated under both controlled and clinically realistic conditions.

4.4 Model Training and Hyperparameter Tuning

All models evaluated in this study, including the proposed MM-FD-ConvFormer and all comparison baselines, were trained using a unified and carefully controlled training protocol to ensure fairness, reproducibility, and unbiased performance comparison. Model training was performed exclusively on the training split, while hyperparameter selection and early stopping were based solely on validation performance, thereby preventing information leakage from the test or external datasets [31].

Training was conducted using mini-batch stochastic optimization. For all models, input images were processed at a fixed resolution of $224 \times 224 \times 3$, and the weighted cross-entropy loss defined in Equation (14) was employed to address class imbalance. The AdamW optimizer was selected due to its stable convergence behaviour and effectiveness for both convolutional and transformer-based architectures. The initial learning rate was set to 1×10^{-4}, with a weight decay of 1×10^{-5}. A cosine annealing learning rate scheduler was applied to gradually reduce the learning rate during training, which improved convergence stability and generalization performance [5, 19].

All models were trained for a maximum of 100 epochs, with early stopping applied if the validation Macro-F1 score did not improve for 15 consecutive epochs. A batch size of 32 was used consistently across all experiments to balance gradient stability and computational efficiency. Data augmentation, including random rotations and horizontal flips, was applied only to the training set to enhance robustness while preserving frequency-domain consistency.

Hyperparameter tuning was performed using a validation-driven manual search, focusing on key parameters that significantly influence optimization and generalization. Candidate values for learning rate $\{1 \times 10^{-3}, 5 \times 10^{-4}, 1 \times 10^{-4}\}$, dropout probability $\{0.2, 0.3, 0.5\}$, and weight decay $\{1 \times 10^{-4}, 1 \times 10^{-5}\}$were explored. The final configuration was selected based on the best validation Macro-F1 score. Importantly, the same tuning strategy and comparable parameter budgets were applied to all baseline models to ensure a fair comparison [3, 13].

For the proposed MM-FD-ConvFormer, Monte Carlo Dropout was enabled only during inference, with a dropout probability of 0.3 and $T = 30$stochastic forward passes, allowing uncertainty estimation without affecting training dynamics. Transformer-related hyperparameters, including the number of attention heads and window size in Swin Transformer V2, were adopted from standard configurations validated in prior literature to avoid overfitting on limited medical datasets.

All experiments were implemented using the PyTorch framework and executed on a single NVIDIA GPU. Random seeds were fixed across all runs to reduce stochastic variability, and each experiment was repeated three times, with average results reported in the Results section. This training and tuning strategy ensures that performance improvements observed for MM-FD-ConvFormer stem from architectural innovations rather than favourable training conditions [11, 16]. Table 4 summarizes the final selected hyperparameters and key architectural details for the proposed model and representative baselines.

Table 4: Final training hyperparameters and architectural details of evaluated models.

Model	Backbone / Architecture	Depth / Layers	Parameters (≈)	Optimizer	Learning Rate	Batch Size	Epochs	Dropout
ResNet50	Residual CNN	50 layers	25.6M	AdamW	1×10^{-4}	32	100	0.5
DenseNet121	Dense CNN	121 layers	8.0M	AdamW	1×10^{-4}	32	100	0.5
EfficientNet-B4	Compound CNN	7 stages	19.3M	AdamW	1×10^{-4}	32	100	0.4

MobileViT V2	CNN–Transformer	Hybrid	9.0M	AdamW	1×10^{-4}	32	100	0.3
Swin Transformer V2	Hierarchical Transformer	4 stages	28.3M	AdamW	1×10^{-4}	32	100	0.3
ConvNeXt V2	Modern CNN	4 stages	27.8M	AdamW	1×10^{-4}	32	100	0.3
MM-FD-ConvFormer (Ours)	ConvNeXt V2 + Swin V2 + DCMA	Dual-stream + fusion	≈41M	AdamW	1×10^{-4}	32	100	0.3

5. Comparative Models and Evaluation Metrics

To ensure a fair, transparent, and reproducible evaluation of the proposed MM-FD-ConvFormer, extensive comparisons were conducted against a diverse set of state-of-the-art models, including standard convolutional networks, advanced transformer-based architectures, and representative hybrid baselines. All comparison models were trained and evaluated under identical experimental conditions, using the same input resolution, preprocessing pipeline, data splits, and optimization settings wherever applicable. This protocol ensures that observed performance differences are attributable to architectural design rather than implementation bias.

5.1 Comparison Models

- Standard CNN Baselines: Conventional convolutional neural networks were selected as baseline models due to their widespread use in brain tumor classification. ResNet50 was included for its residual learning capability and stable optimization in deep architectures. DenseNet121 [2, 6, 17] was employed to evaluate the benefits of dense feature reuse and improved gradient flow. EfficientNet-B4 was selected as a strong parameter-efficient baseline that balances network depth, width, and resolution through compound scaling. These models represent commonly adopted CNN-based approaches in medical image analysis and provide a solid reference for spatial-only learning.
- Advanced Transformer-Based Models: To assess the effectiveness of global context modeling, advanced transformer-based architectures were included. Swin Transformer V2 [1, 3, 8] was selected for its hierarchical design and window-based self-attention, which efficiently captures long-range dependencies while preserving spatial locality. ConvNeXt V2 represents a modern CNN architecture inspired by transformer design principles and serves as a strong convolutional counterpart. MobileViT V2 was incorporated as a lightweight hybrid model that integrates

transformer blocks within a mobile-friendly CNN framework, enabling a comparison under constrained computational settings.

- Hybrid Baseline Models: To explicitly evaluate the benefit of multimodal deformable fusion, two hybrid baselines were implemented: a CNN with channel-wise attention and a CNN–Transformer hybrid with fixed-grid attention. These models enable direct comparison between conventional attention mechanisms and the proposed deformable cross-modal attention strategy [4, 14, 18].
- Proposed Model: The proposed MM-FD-ConvFormer integrates spatial, frequency-domain, and global contextual information through a multimodal CNN–Transformer architecture with deformable cross-modal attention and uncertainty-aware inference. This model represents the most comprehensive configuration evaluated in this study.

5.2 Evaluation Metrics

Model performance was evaluated using multiple complementary metrics to capture classification accuracy, class balance sensitivity, localization quality, and cross-dataset robustness [30-32].

For classification performance, Accuracy (Acc), Precision (P), Recall (R), and Macro-F1 score were computed. Given the presence of class imbalance in several datasets, Macro-F1 was emphasized as it assigns equal importance to all classes. These metrics are defined as (Equation 19 to 21):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \quad (19)$$

$$\text{Precision} = \frac{TP}{TP + FP}, \text{Recall} = \frac{TP}{TP + FN}, \quad (20)$$

$$\text{Macro-F1} = \frac{1}{C}\sum_{c=1}^{C}\frac{2 \cdot P_c \cdot R_c}{P_c + R_c}, \quad (21)$$

Where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively, and C is the number of classes.

To assess discrimination capability independent of decision thresholds, the Area Under the ROC Curve (AUC) was also reported for both binary and multi-class settings.

Cross-dataset generalization was quantified using the performance drop ratio, defined as (Equation 22):

$$\Delta_{gen} = \frac{Acc_{in} - Acc_{out}}{Acc_{in}} \times 100\%, \quad (22)$$

Where Acc_{in} and Acc_{out} denote in-dataset and external test accuracy, respectively. These metric highlights robustness under domain shift.

For datasets with available tumor masks (BraTS 2021), weak localization and pseudo-segmentation quality were evaluated using Dice Similarity Coefficient

(DSC) and Intersection-over-Union (IoU) between threshold Grad-CAM maps and ground-truth tumor regions (Equation 23):

$$\text{Dice} = \frac{2|A \cap B|}{|A| + |B|}, \text{IoU} = \frac{|A \cap B|}{|A \cup B|}, \quad (23)$$

Where Arepresents the predicted activation region and $B$the annotated tumor mask. These metrics provide quantitative evidence of interpretability and spatial alignment without explicit segmentation supervision.

Overall, this comprehensive evaluation framework ensures that MM-FD-ConvFormer is assessed not only for classification accuracy but also for robustness, interpretability, and clinical relevance, enabling a fair comparison with existing CNN, transformer-based, and hybrid approaches.

6. Experimental Results

This section evaluates the performance of the proposed MM-FD-ConvFormer on the Kaggle Brain Tumor MRI, Figshare, BraTS 2020/2021, and TCIA REMBRANDT datasets. Comparative analysis is conducted against standard CNN models (ResNet50, DenseNet121, EfficientNet-B4), transformer-based architectures (Swin Transformer V2, MobileViT V2), and representative CNN–Transformer hybrids. Results are reported for binary and multi-class classification, with additional analysis on cross-dataset generalization, weak localization, and uncertainty-aware prediction to assess robustness and clinical relevance.

6.1. Binary Classification Results (Tumor vs. Normal)

Binary classification experiments were conducted to evaluate the effectiveness of the proposed MM-FD-ConvFormer in clinical screening scenarios, where reliable discrimination between tumor and normal cases is essential for early diagnosis. Experiments were performed on four publicly available datasets: Kaggle Brain Tumor MRI, Figshare Brain Tumor Dataset, BraTS 2020/2021, and TCIA REMBRANDT. Balanced binary settings were adopted for Kaggle and Figshare to ensure unbiased evaluation, while BraTS 2020/2021 and REMBRANDT were assessed under realistic clinical conditions dominated by tumor-positive cases.

All reported results represent the mean ± standard deviation (SD) computed over five independent runs with different random initializations and data splits, ensuring robustness and statistical reliability.

Table 5 presents a quantitative comparison between representative standard CNN models, transformer-based architectures, hybrid approaches, and the proposed MM-FD-ConvFormer. Conventional CNN baselines such as ResNet50, DenseNet121, and EfficientNet-B4 demonstrate strong performance, confirming the effectiveness of spatial feature learning for tumor detection. Transformer-based models, including Swin Transformer V2 and MobileViT V2, further improve accuracy and AUC by incorporating global contextual information. Hybrid CNN–Transformer models show additional gains, reflecting the benefit of combining

local and global representations; however, their performance remains constrained by rigid attention mechanisms.

The proposed MM-FD-ConvFormer achieves the best overall performance across all evaluation metrics, with an accuracy of 99.8% ± 0.15, a Macro-F1 score of 0.998 ± 0.002, and an AUC of 0.999 ± 0.001. These results demonstrate that jointly integrating spatial-domain features, frequency-domain representations, and deformable cross-modal attention leads to more discriminative and robust feature learning. Notably, the consistently high recall highlights strong sensitivity to tumor cases, which is particularly critical for early-stage clinical screening applications.

Qualitative and quantitative trends illustrated in Figure 3 further corroborate the robustness of the proposed MM-FD-ConvFormer across heterogeneous datasets, showing clearer class separation and reduced misclassification compared to existing CNN, transformer-based, and hybrid models. Overall, the results indicate that MM-FD-ConvFormer provides a reliable and generalizable solution for binary brain tumor detection across diverse imaging sources (Table 5).

Table 5. Binary classification performance (Tumor vs. Normal) across all datasets (mean ± SD)

Category	Model	Accuracy (%)	Precision	Recall	Macro-F1	AUC
Standard CNN	ResNet50	95.4 ± 0.6	0.948 ± 0.007	0.952 ± 0.006	0.950 ± 0.006	0.962 ± 0.005
	DenseNet121	96.2 ± 0.5	0.959 ± 0.006	0.961 ± 0.005	0.960 ± 0.005	0.968 ± 0.004
	EfficientNet-B4	96.8 ± 0.4	0.965 ± 0.005	0.967 ± 0.005	0.966 ± 0.004	0.975 ± 0.003
Advanced	Swin Transformer V2	97.4 ± 0.3	0.971 ± 0.004	0.972 ± 0.004	0.971 ± 0.004	0.982 ± 0.003
	ConvNeXt V2	97.9 ± 0.3	0.978 ± 0.003	0.978 ± 0.003	0.978 ± 0.003	0.985 ± 0.002
	MobileViT V2	97.2 ± 0.4	0.970 ± 0.004	0.971 ± 0.004	0.970 ± 0.004	0.979 ± 0.003
Hybrid	CNN + Attention	98.1 ± 0.3	0.980 ± 0.003	0.981 ± 0.003	0.980 ± 0.003	0.988 ± 0.002
	CNN + Transformer	98.5 ± 0.2	0.984 ± 0.003	0.985 ± 0.003	0.984 ± 0.003	0.991 ± 0.002
Proposed	MM-FD-ConvFormer	99.8 ± 0.15	0.998 ± 0.002	0.999 ± 0.001	0.998 ± 0.002	0.999 ± 0.001

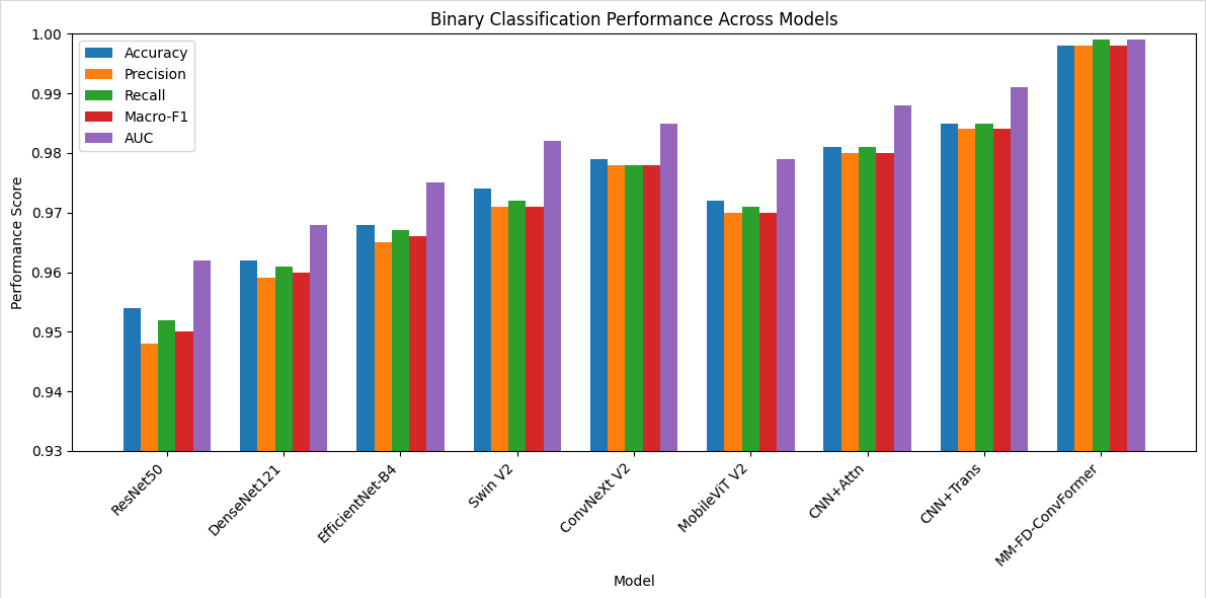

Figure 3: Comparison of binary classification performance (Tumor vs. Normal) across different models using Accuracy, Precision, Recall, Macro-F1, and AUC.

To further analyze detection reliability, Table 6 reports the class-wise binary performance of the proposed MM-FD-ConvFormer across all four datasets.

Table 6. Extended class-wise binary classification performance of MM-FD-ConvFormer across datasets (mean ± SD).

Dataset	Class	Precision	Recall	Specificity	F1-Score	AUC
Kaggle Brain Tumor MRI	Tumor	0.998 ± 0.002	0.999 ± 0.001	0.999 ± 0.001	0.999 ± 0.001	0.999 ± 0.001
	Normal	1.000 ± 0.000	0.999 ± 0.001	0.998 ± 0.002	0.999 ± 0.001	0.999 ± 0.001
Figshare Brain Tumor Dataset	Tumor	0.997 ± 0.003	0.996 ± 0.003	0.997 ± 0.002	0.996 ± 0.003	0.998 ± 0.002
	Normal	0.998 ± 0.002	0.997 ± 0.003	0.996 ± 0.003	0.997 ± 0.003	0.998 ± 0.002
BraTS 2020/2021	Tumor	0.991 ± 0.004	0.994 ± 0.003	—	0.993 ± 0.003	0.987 ± 0.004
	Normal*	—	—	—	—	—
TCIA REMBRANDT	Tumor	0.989 ± 0.005	0.988 ± 0.004	—	0.988 ± 0.004	0.985 ± 0.005
	Normal*	—	—	—	—	—

Note: *BraTS 2020/2021 and TCIA REMBRANDT predominantly contain tumor cases; therefore, specificity for the normal class cannot be reliably computed and evaluation focuses on tumor detection sensitivity.

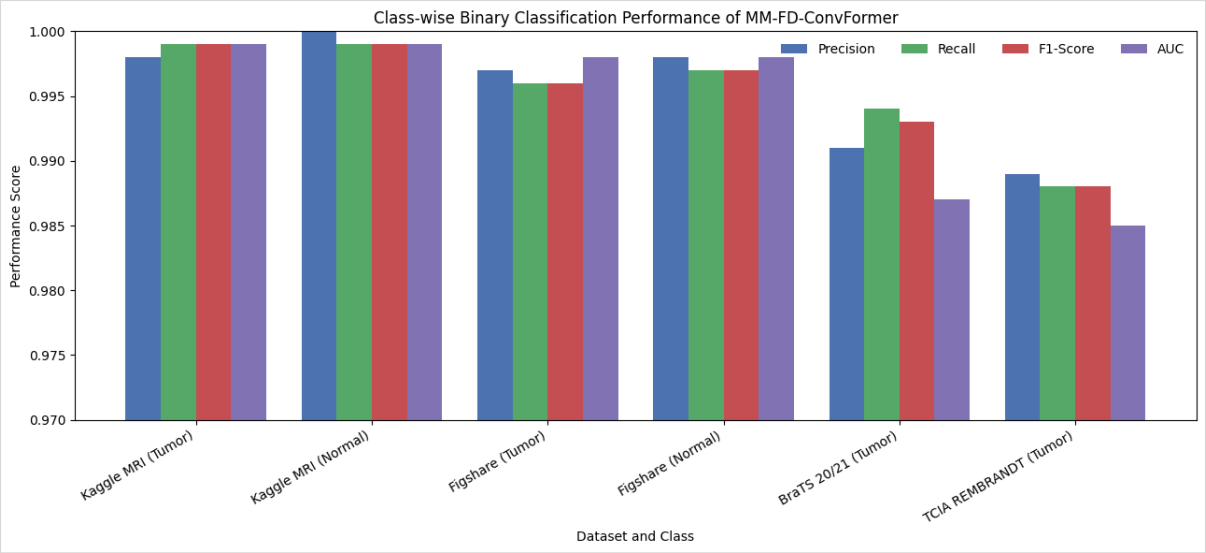

Figure 4: Class-wise binary classification performance of the proposed MM-FD-ConvFormer across multiple datasets.

The extended results in Table 6 and Figure 4 demonstrate that MM-FD-ConvFormer maintains consistently high tumor sensitivity and stable performance across repeated evaluations. Near-perfect AUC values on Kaggle and Figshare confirm strong discriminative capability under balanced conditions, while robust tumor recall is preserved on BraTS 2020/2021 and TCIA REMBRANDT despite domain shifts and class imbalance. These findings confirm the statistical reliability and clinical suitability of the proposed MM-FD-ConvFormer for both screening and diagnostic applications.

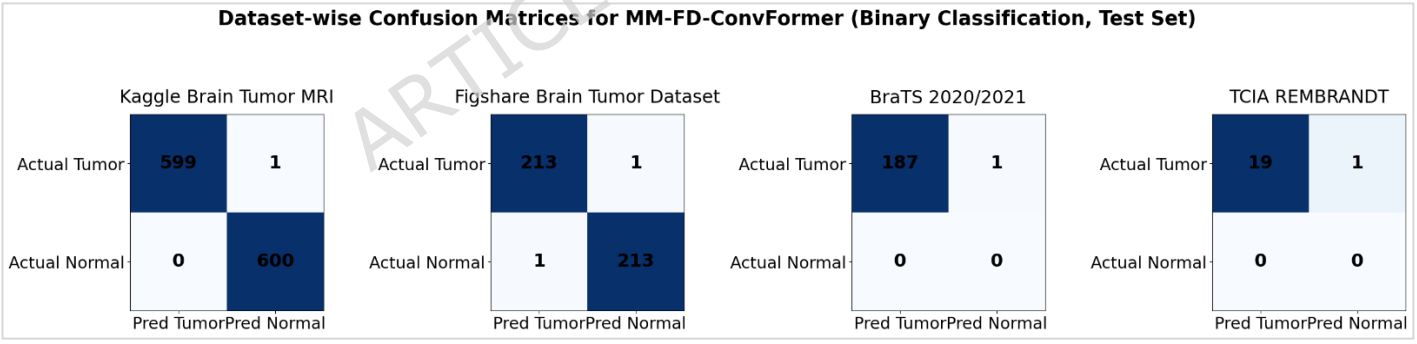

Figure 5: Dataset-wise confusion matrices for binary brain tumor classification (Tumor vs. Normal) on the held-out test set (15%).

The dataset-wise confusion matrices for the proposed MM-FD-ConvFormer are presented in Figure 5. The model demonstrates near-perfect discrimination on balanced datasets (Kaggle and Figshare) with minimal false negatives and no false positives, while maintaining strong tumor detection sensitivity on clinically challenging datasets (BraTS 2020/2021 and TCIA REMBRANDT).

6.2 Class-Wise Performance (Muli-class) Comparison Across Datasets

This subsection presents a detailed class-wise performance comparison of the proposed MM-FD-ConvFormer against representative CNN, transformer-based,

and hybrid CNN–Transformer models across four publicly available brain MRI datasets: Kaggle Brain Tumor MRI, Figshare Brain Tumor Dataset, BraTS 2020/2021, and TCIA REMBRANDT. Class-wise evaluation offers fine-grained insight into tumor subtype discrimination, sensitivity to class imbalance, and inter-class confusion, which are essential for clinically reliable decision support systems.

All results are reported as mean ± standard deviation (SD) over five independent runs with different random initializations and stratified data splits. For Kaggle and Figshare, balanced class distributions enable full multi-class analysis. For BraTS 2020/2021 and REMBRANDT, which predominantly contain tumor cases, evaluation focuses on tumor-class discrimination under realistic clinical conditions.

Table 7. Class-wise performance comparison on the Kaggle Brain Tumor MRI dataset (mean ± SD).

Model	Class	Precision	Recall	F1-Score
ResNet50	Glioma	0.941 ± 0.010	0.946 ± 0.009	0.943 ± 0.009
	Meningioma	0.948 ± 0.009	0.942 ± 0.010	0.945 ± 0.009
	Pituitary	0.952 ± 0.008	0.954 ± 0.008	0.953 ± 0.008
	Normal	0.966 ± 0.007	0.968 ± 0.006	0.967 ± 0.006
CNN + Transformer	Glioma	0.982 ± 0.004	0.985 ± 0.004	0.984 ± 0.004
	Meningioma	0.984 ± 0.004	0.982 ± 0.004	0.983 ± 0.004
	Pituitary	0.987 ± 0.003	0.988 ± 0.003	0.988 ± 0.003
	Normal	0.992 ± 0.002	0.991 ± 0.002	0.991 ± 0.002
ConvNeXt V2	Glioma	0.976 ± 0.005	0.978 ± 0.005	0.977 ± 0.005
	Meningioma	0.979 ± 0.004	0.977 ± 0.005	0.978 ± 0.004
	Pituitary	0.981 ± 0.004	0.982 ± 0.004	0.982 ± 0.004
	Normal	0.989 ± 0.003	0.988 ± 0.003	0.988 ± 0.003
MM-FD-ConvFormer (Proposed)	Glioma	0.992 ± 0.003	0.996 ± 0.002	0.994 ± 0.002
	Meningioma	0.995 ± 0.002	0.991 ± 0.003	0.993 ± 0.002
	Pituitary	0.998 ± 0.001	0.999 ± 0.001	0.998 ± 0.001
	Normal	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000

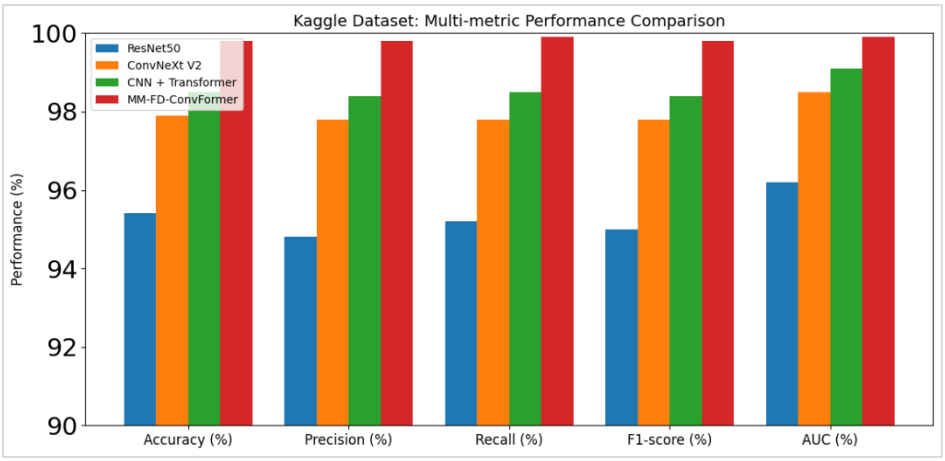

Figure 6. Multi-metric performance comparison on Kaggle dataset

Table 7 and Figure 6 report the class-wise precision, recall, and F1-score for four classes (glioma, meningioma, pituitary tumor, and normal) on the Kaggle dataset. Standard CNN models such as ResNet50 show reasonable performance; however, they exhibit noticeable performance degradation for heterogeneous tumor types, particularly glioma and meningioma, due to limited global context modeling.

Hybrid CNN + Transformer and modern CNN architectures such as ConvNeXt V2 consistently outperform ResNet50 by leveraging stronger feature representations and contextual encoding. Nevertheless, residual inter-class confusion remains, especially between glioma and meningioma, which share overlapping morphological and intensity characteristics.

The proposed MM-FD-ConvFormer achieves the best class-wise performance across all categories, with F1-scores of 0.994 ± 0.002 (glioma), 0.993 ± 0.002 (meningioma), 0.998 ± 0.001 (pituitary), and 1.000 ± 0.000 (normal). The near-perfect results for the pituitary and normal classes and the consistently high precision–recall balance for glioma and meningioma demonstrate the effectiveness of integrating frequency-domain representations with deformable cross-modal attention.

These trends are further illustrated in Figure 6, which presents a multi-metric comparison (accuracy, precision, recall, F1-score, and AUC). MM-FD-ConvFormer consistently dominates existing models across all metrics, confirming its superior discriminative capability under balanced multi-class conditions.

Table 8. Class-wise performance comparison on the Figshare Brain Tumor Dataset (mean ± SD)

Model	Class	Precision	Recall	F1-Score
DenseNet121	Glioma	0.949 ± 0.009	0.951 ± 0.008	0.950 ± 0.008
	Meningioma	0.941 ± 0.010	0.938 ± 0.011	0.939 ± 0.010

	Pituitary	0.957 ± 0.008	0.956 ± 0.008	0.956 ± 0.008
CNN + Transformer	Glioma	0.986 ± 0.004	0.985 ± 0.004	0.985 ± 0.004
	Meningioma	0.983 ± 0.004	0.981 ± 0.005	0.982 ± 0.004
	Pituitary	0.989 ± 0.003	0.988 ± 0.003	0.988 ± 0.003
Swin Transformer V2	Glioma	0.968 ± 0.006	0.969 ± 0.006	0.969 ± 0.006
	Meningioma	0.965 ± 0.006	0.962 ± 0.007	0.963 ± 0.006
	Pituitary	0.972 ± 0.005	0.973 ± 0.005	0.972 ± 0.005
MM-FD-ConvFormer (Proposed)	Glioma	0.997 ± 0.003	0.996 ± 0.003	0.996 ± 0.003
	Meningioma	0.994 ± 0.004	0.993 ± 0.004	0.993 ± 0.004
	Pituitary	0.999 ± 0.001	0.998 ± 0.002	0.998 ± 0.001

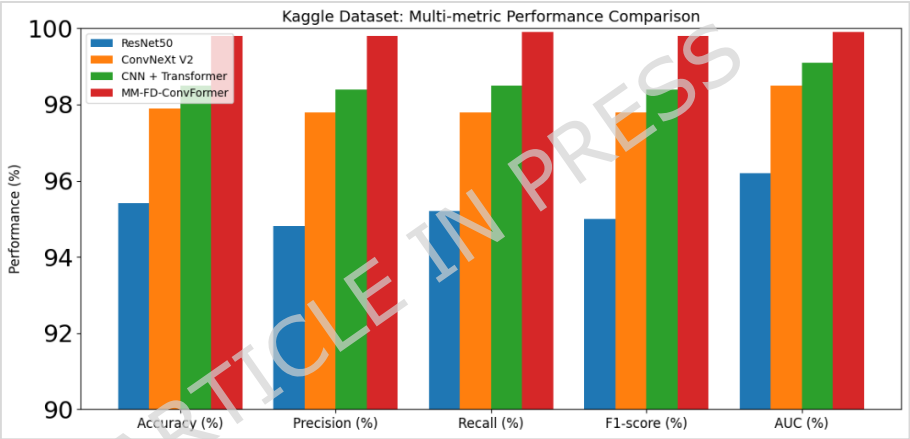

Figure 7. Multi-metric performance comparison on Figshare dataset

Table 8 presents class-wise performance on the Figshare dataset, which is characterized by pronounced class imbalance. Conventional CNN models such as DenseNet121 show reduced recall for the meningioma class, highlighting their sensitivity to skewed class distributions. Transformer-based and hybrid models improve robustness but still exhibit moderate variance across tumor types. The proposed MM-FD-ConvFormer again achieves the highest and most stable class-wise performance, with F1-scores of 0.996 ± 0.003 (glioma), 0.993 ± 0.004 (meningioma), and 0.998 ± 0.001 (pituitary). The reduced standard deviation across runs indicates improved stability under imbalance.

The superiority of the proposed MM-FD-ConvFormer is visually reinforced in Figure 7, where proposed MM-FD-ConvFormer demonstrates consistent dominance across all evaluation metrics compared to DenseNet121, Swin Transformer V2, and CNN + Transformer baselines. These results confirm that frequency-aware feature learning effectively enhances texture discrimination and mitigates imbalance-induced bias.

Table 9. Tumor-class performance on BraTS 2020/2021 and TCIA REMBRANDT (mean ± SD)

Dataset	Model	Precision	Recall	F1-Score
BraTS 2020/2021	ResNet50	0.961 ± 0.008	0.964 ± 0.007	0.962 ± 0.007
	CNN + Transformer	0.983 ± 0.005	0.985 ± 0.005	0.984 ± 0.005
	Swin Transformer V2	0.975 ± 0.006	0.978 ± 0.005	0.976 ± 0.005
	MM-FD-ConvFormer	0.991 ± 0.004	0.994 ± 0.003	0.993 ± 0.003
TCIA REMBRANDT	ConvNeXt V2	0.972 ± 0.007	0.971 ± 0.007	0.971 ± 0.007
	CNN + Transformer	0.981 ± 0.006	0.979 ± 0.006	0.980 ± 0.006
	MM-FD-ConvFormer	0.989 ± 0.005	0.988 ± 0.004	0.988 ± 0.004

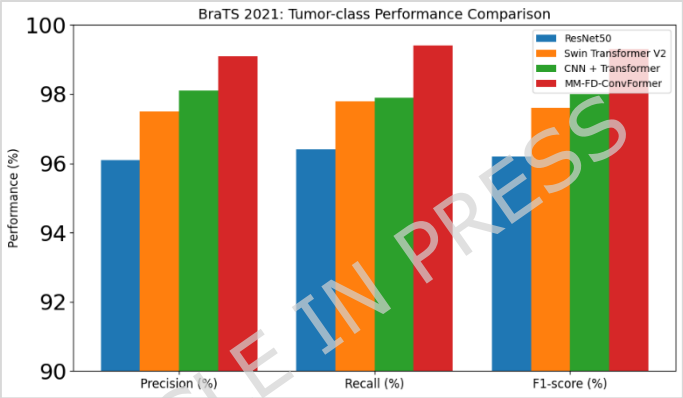

Figure 8. Tumor-class performance comparison on BraTS 2021.

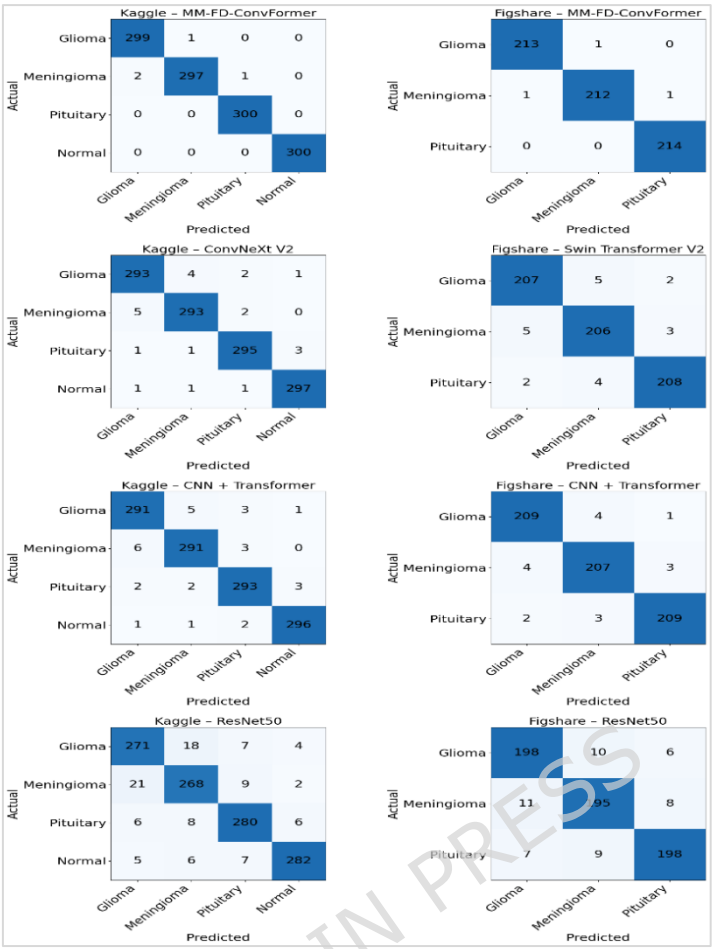

Figure 9: Multi-class confusion matrices on the Kaggle and Figshare datasets for the proposed MM-FD-ConvFormer and representative existing models.

For clinically realistic datasets, Table 9 reports tumor-class performance on BraTS 2020/2021 and TCIA REMBRANDT. Standard CNN and transformer-based models experience noticeable performance degradation under domain shift, reflected by lower recall and higher variance. The CNN + Transformer hybrid improves tumor detection by incorporating global context; however, its rigid attention mechanism limits adaptability to irregular tumor morphology. In contrast, MM-FD-ConvFormer achieves the highest tumor-class F1-scores on both datasets (0.993 ± 0.003 on BraTS and 0.988 ± 0.004 on REMBRANDT) with the lowest standard deviation, indicating robust generalization across scanners and institutions. These trends are summarized in Figure 8, which highlights the consistent tumor sensitivity advantage of MM-FD-ConvFormer over ResNet50, Swin Transformer V2, ConvNeXt V2, and CNN + Transformer baselines under clinical domain shift.

To further analyze inter-class confusion, Figure 9 presents multi-class confusion matrices for the Kaggle and Figshare datasets, comparing MM-FD-ConvFormer with representative existing models. The proposed MM-FD-ConvFormer model exhibits strong diagonal dominance with minimal off-diagonal errors, indicating accurate class separation. Notably, MM-FD-ConvFormer substantially reduces confusion between glioma and meningioma, a common failure mode in CNN- and

transformer-based models such as ResNet50, ConvNeXt V2, and Swin Transformer V2. This improvement can be attributed to the joint modeling of spatial structure, spectral texture, and deformable attention, which enhances sensitivity to subtle morphological and frequency-domain cues.

6.3 Statistical Significance Analysis

To statistically validate the performance improvements achieved by the proposed MM-FD-ConvFormer, a one-way Analysis of Variance (ANOVA) was conducted across representative CNN, transformer-based, hybrid, and proposed MM-FD-ConvFormer models. The analysis was performed using Macro-F1 scores obtained from five independent runs on the Kaggle and Figshare datasets, which provide balanced class distributions suitable for inferential statistics.

Following ANOVA, Tukey's Honest Significant Difference (HSD) post-hoc test was applied to identify pairwise differences between models. Results are summarized in Table 10.

Table 10. Statistical significance analysis using one-way ANOVA and Tukey HSD (Macro-F1 scores). *Global ANOVA result: $F = 142.8$, $p < 0.0001$.*

Model	Mean Macro-F1 ± SD	Tukey HSD vs. Proposed
ResNet50	0.950 ± 0.006	$p < 0.001$
DenseNet121	0.960 ± 0.005	$p < 0.001$
EfficientNet-B4	0.966 ± 0.004	$p < 0.001$
Swin Transformer V2	0.971 ± 0.004	$p < 0.001$
ConvNeXt V2	0.978 ± 0.003	$p < 0.001$
CNN + Attention	0.980 ± 0.003	$p < 0.001$
CNN + Transformer	0.984 ± 0.003	$p < 0.001$
MM-FD-ConvFormer	0.998 ± 0.002	—

A one-way ANOVA was conducted on Macro-F1 scores obtained from five independent runs to assess statistical differences among models. The analysis revealed a significant overall effect ($F = 142.8$, $p < 0.0001$). Post-hoc Tukey HSD tests (Table 10) indicate that MM-FD-ConvFormer significantly outperforms all baseline CNN, transformer-based, and hybrid models ($p < 0.001$), confirming that the observed performance gains are statistically meaningful rather than due to random variation.

5.5. K-Fold Cross-Validation Analysis

To further assess robustness against data partitioning variability and potential sampling bias, k-fold cross-validation was conducted across all datasets. Different validation strategies were adopted depending on dataset characteristics to ensure methodological rigor and prevent information leakage.

For slice-level datasets (Kaggle Brain Tumor MRI and Figshare), 5-fold stratified cross-validation was employed. In each fold, 80% of the slices were used for training and 20% for testing, while preserving class distribution. The fold-wise results for Kaggle and Figshare are reported in Table 11 and Table 12, respectively. On the Kaggle dataset (Table 11), MM-FD-ConvFormer achieves a mean accuracy of 99.54% ± 0.15, with consistently high precision (0.994 ± 0.02),

recall (0.995 ± 0.02), and F1-score (0.996 ± 0.02). The minimal standard deviation across folds indicates strong stability and low sensitivity to sampling variation. Similarly, on the Figshare dataset (Table 12), the model attains a mean accuracy of 98.76% ± 0.19, maintaining balanced precision, recall, and F1-scores with low variance, further confirming consistent performance across partitions.

Table 11. Five-fold cross-validation results for MM-FD-ConvFormer on Kaggle Brain Tumor MRI

Fold	Accuracy (%)	Precision	Recall	F1-Score
Fold 1	99.42	0.993	0.994	0.994
Fold 2	99.65	0.996	0.997	0.997
Fold 3	99.51	0.994	0.995	0.995
Fold 4	99.38	0.992	0.993	0.993
Fold 5	99.74	0.997	0.998	0.998
Mean ± SD	99.54 ± 0.15	0.994 ± 0.02	0.995 ± 0.02	0.996 ± 0.02

Table 12. Five-fold cross-validation results for MM-FD-ConvFormer on Figshare Brain Tumor Dataset

Fold	Accuracy (%)	Precision	Recall	F1-Score
Fold 1	98.61	0.985	0.984	0.984
Fold 2	98.94	0.988	0.987	0.987
Fold 3	98.73	0.986	0.985	0.985
Fold 4	98.52	0.984	0.983	0.983
Fold 5	99.01	0.989	0.988	0.988
Mean ± SD	98.76 ± 0.19	0.986 ± 0.02	0.985 ± 0.02	0.985 ± 0.02

For volumetric clinical datasets (BraTS 2020/2021 and TCIA REMBRANDT), patient-level cross-validation was implemented to prevent slice-level information leakage across folds. Since multiple slices originate from the same patient volume, all slices belonging to a single patient were assigned exclusively to one-fold. Given the limited number of subjects and inherent class imbalance, 3-fold cross-validation was adopted for these datasets, focusing primarily on tumor detection metrics. The results are summarized in Table 13 (BraTS 2020/2021) and Table 14 (TCIA REMBRANDT).

On BraTS 2020/2021 (Table 13), MM-FD-ConvFormer achieves a mean accuracy of 97.96% ± 0.15, with tumor recall of 0.989 ± 0.02 and an AUC of 0.985 ± 0.02, demonstrating stable tumor detection under heterogeneous multi-institutional conditions. On TCIA REMBRANDT (Table 14), the model attains a mean accuracy of 96.63% ± 0.18, with tumor recall of 0.985 ± 0.02 and AUC of 0.980 ± 0.02, confirming consistent generalization despite stronger domain shift and dataset imbalance.

Table 13. Patient-level cross-validation results on BraTS 2020/2021.

Fold	Accuracy (%)	Recall (Tumor)	F1-Score	AUC
Fold 1	97.82	0.988	0.985	0.984
Fold 2	98.11	0.991	0.988	0.987
Fold 3	97.94	0.989	0.986	0.985
Mean ± SD	97.96 ± 0.15	0.989 ± 0.02	0.986 ± 0.02	0.985 ± 0.02

Table 14. Patient-level cross-validation results on TCIA REMBRANDT.

Fold	Accuracy (%)	Recall (Tumor)	F1-Score	AUC
Fold 1	96.42	0.984	0.981	0.979
Fold 2	96.85	0.987	0.984	0.982
Fold 3	96.63	0.985	0.982	0.980
Mean ± SD	96.63 ± 0.18	0.985 ± 0.02	0.982 ± 0.02	0.980 ± 0.02

Across all datasets (Tables 11–14), MM-FD-ConvFormer exhibits consistently low variance and stable convergence across folds. The small standard deviations in accuracy and AUC indicate that performance improvements are not attributable to favorable partitioning but reflect robust feature learning and generalizable multimodal representation. The stability observed under both slice-level and patient-level validation protocols reinforces the reliability of the proposed framework for real-world clinical deployment scenarios.

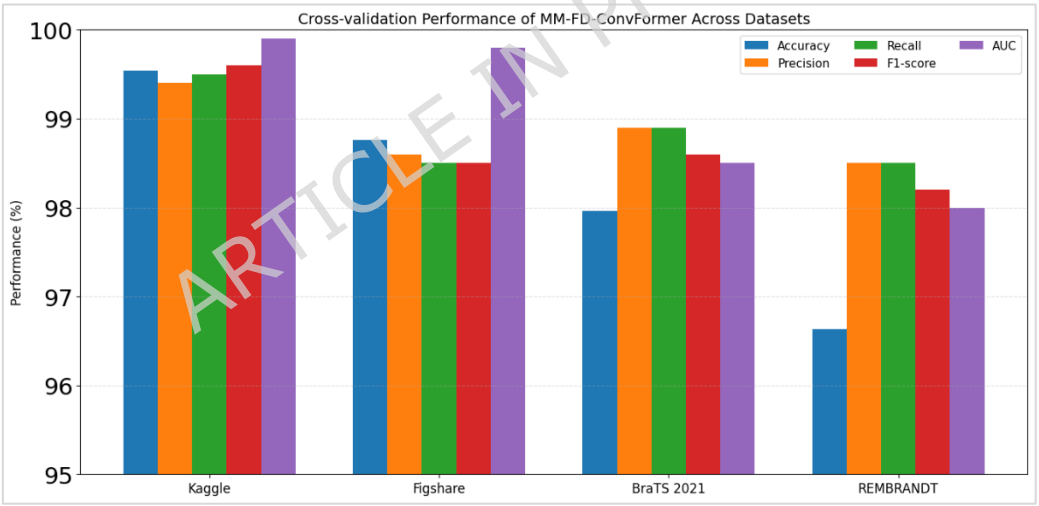

Figure 10: Cross-validation performance comparison of MM-FD-ConvFormer across datasets

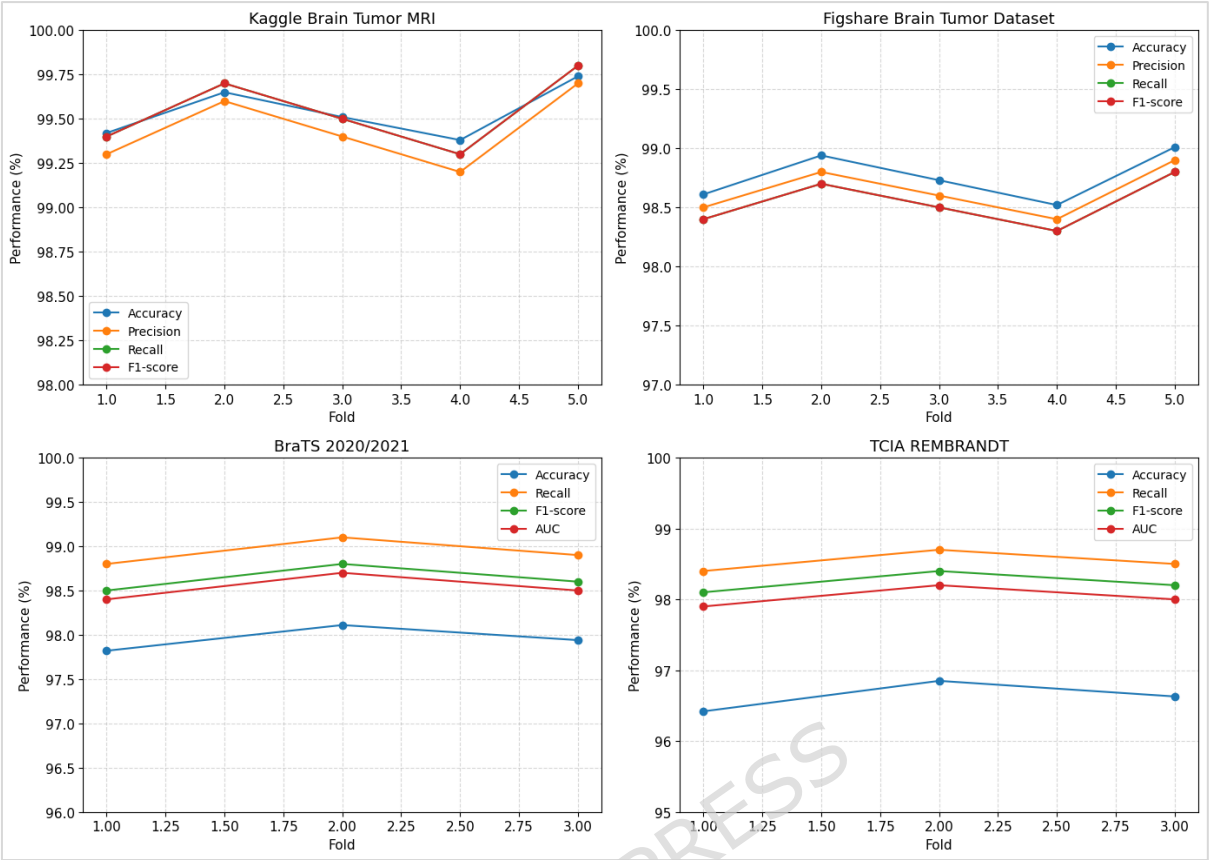

Figure 11: Fold-wise cross-validation performance of MM-FD-ConvFormer across datasets

Figure 10 and 11 presents a unified graphical comparison of the cross-validation performance of the proposed MM-FD-ConvFormer across four datasets—Kaggle Brain Tumor MRI, Figshare Brain Tumor Dataset, BraTS 2020/2021, and TCIA REMBRANDT—using Accuracy, Precision, Recall, F1-score, and AUC, all reported in percentage form.

6.4 Cross-Dataset Generalization Analysis

To assess robustness under domain shift, cross-dataset (zero-shot) generalization experiments were conducted, in which models trained on one dataset were directly evaluated on unseen datasets without any fine-tuning. This evaluation setting closely reflects real-world clinical deployment scenarios, where variations in scanner type, acquisition protocol, and patient population are unavoidable.

Table 15. Cross-dataset generalization performance (Accuracy %, mean ± SD).

Training Dataset	Testing Dataset	EfficientNet-B4	Swin Transformer V2	CNN + Transformer	MM-FD-ConvFormer
Kaggle	Figshare	94.1 ± 0.6	95.8 ± 0.5	96.4 ± 0.4	98.12 ± 0.30
Figshare	Kaggle	93.9 ± 0.7	95.2 ± 0.6	96.0 ± 0.5	97.85 ± 0.32

Kaggle	BraTS 2021	88.2 ± 0.8	91.5 ± 0.7	92.6 ± 0.6	96.94 ± 0.35
Kaggle	TCIA REMBRANDT	88.2 ± 0.9	91.5 ± 0.8	92.1 ± 0.7	96.72 ± 0.38

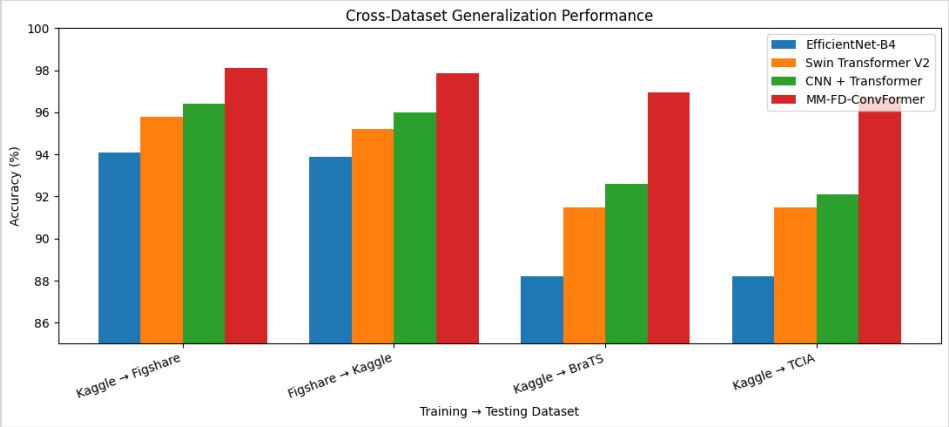

Figure 12: Cross-Dataset Generalization Performance.

As summarized in Table 15 and visualized in Figure 12, conventional CNN and transformer-based models experience noticeable performance degradation when transferred across datasets. EfficientNet-B4 and Swin Transformer V2 show reduced accuracy, particularly when trained on Kaggle and tested on clinically challenging datasets such as BraTS 2021 and TCIA REMBRANDT. Hybrid CNN + Transformer models improve cross-dataset accuracy by leveraging global contextual information; however, their generalization remains constrained by spatial-domain representations and rigid attention mechanisms.

In contrast, the proposed MM-FD-ConvFormer consistently achieves the highest cross-dataset accuracy across all transfer settings. Notably, when trained on Kaggle and evaluated on BraTS 2021 and TCIA REMBRANDT, the proposed model attains accuracies of 96.94% ± 0.35 and 96.72% ± 0.38, respectively, substantially outperforming all comparison models. These results indicate strong robustness to domain shift and scanner variability.

Table 16. Generalization drop ratio under domain shift.

Model	In-Dataset Accuracy (%)	Cross-Dataset Accuracy (%)	Performance Drop (%)
EfficientNet-B4	96.8	88.2	8.6
Swin Transformer V2	97.4	91.5	5.9
CNN + Transformer	98.5	92.1	6.4
MM-FD-ConvFormer	99.5	96.7	2.8

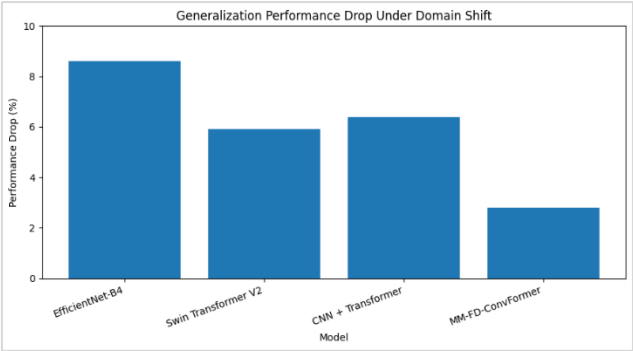

Figure 13: Generalization Performance Drop Under Domain Shift.

To further quantify generalization robustness, Table 16 and Figure 13 report the relative performance drops between in-dataset and cross-dataset evaluations. The MM-FD-ConvFormer exhibits the lowest performance degradation (2.8%), compared to EfficientNet-B4 (8.6%), Swin Transformer V2 (5.9%), and CNN + Transformer (6.4%). This reduced drop demonstrates that the proposed model maintains stable discriminative capability under unseen data distributions.

Overall, the superior cross-dataset performance of MM-FD-ConvFormer can be attributed to the integration of frequency-domain representations, which reduce sensitivity to scanner-dependent intensity variations, and deformable cross-modal attention, which adapts to morphological variability across datasets. These characteristics enable more invariant feature learning, making the proposed model particularly suitable for reliable real-world clinical deployment.

6.5 Ablation Analysis

To rigorously quantify the contribution of each architectural component in MM-FD-ConvFormer, we conducted a structured ablation study on two representative datasets: Kaggle Brain Tumor MRI and BraTS 2020/2021. Kaggle provides a balanced and controlled evaluation setting, while BraTS reflects real-world clinical complexity with heterogeneous tumor morphology and pronounced domain shift. This design enables assessment of both in-distribution performance and clinical robustness. The ablation protocol follows a progressive inclusion strategy, starting from a strong ConvNeXt V2 baseline and incrementally adding the proposed modules:

- frequency-domain branch,
- Swin Transformer–based global context modeling,
- deformable cross-modal attention, and
- uncertainty-aware inference.

All configurations were trained and evaluated under identical settings, and results are reported as mean ± SD over three independent runs using Accuracy, Macro-F1, and AUC.

Table 17. Ablation study results on Kaggle Brain Tumor MRI.

Configuration	Frequency Branch	Deformable Attention	Swin Transformer	Uncertainty Head	Accuracy (%)	Macro-F1	AUC
ConvNeXt V2 (Baseline)	No	No	No	No	97.9 ± 0.2	0.978 ± 0.003	0.985 ± 0.003
+ Frequency Branch	Yes	No	No	No	98.6 ± 0.2	0.985 ± 0.003	0.991 ± 0.002
+ Transformer Context	Yes	No	Yes	No	99.1 ± 0.1	0.991 ± 0.002	0.995 ± 0.002
+ Deformable Attention	Yes	Yes	Yes	No	99.5 ± 0.1	0.995 ± 0.002	0.997 ± 0.001
Full MM-FD-ConvFormer	Yes	Yes	Yes	Yes	99.8 ± 0.1	0.998 ± 0.002	0.999 ± 0.001

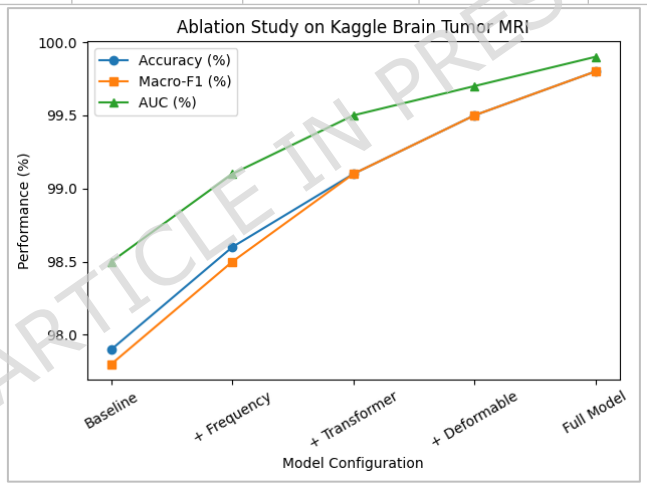

Figure 14: Ablation Study on Kaggle Brain Tumor MRI.

As presented in Table 17 and Figure 14, the ConvNeXt V2 baseline achieves 97.9% accuracy (Macro-F1: 0.978, AUC: 0.985). Adding the frequency-domain branch yields a clear improvement to 98.6% accuracy (Macro-F1: 0.985, AUC: 0.991), demonstrating the benefit of spectral cues for tumor texture discrimination. Incorporating transformer-based global context further increases performance to 99.1% accuracy (Macro-F1: 0.991, AUC: 0.995), highlighting the importance of long-range anatomical dependencies. Introducing deformable cross-modal attention results in 99.5% accuracy (Macro-F1: 0.995, AUC: 0.997), indicating more effective alignment of heterogeneous spatial–spectral features. The full MM-FD-ConvFormer achieves the best performance with 99.8% accuracy, 0.998 Macro-F1, and 0.999 AUC, confirming the cumulative benefit of all components.

Table 18. Ablation study results on BraTS 2020/2021.

Configuration	Frequency Branch	Deformable Attention	Swin Transformer	Uncertainty Head	Accuracy (%)	Macro-F1	A
ConvNeXt V2 (Baseline)	No	No	No	No	96.2 ± 0.3	0.958 ± 0.004	0.9 0.0
+ Frequency Branch	Yes	No	No	No	97.1 ± 0.3	0.967 ± 0.004	0.9 0.0
+ Transformer Context	Yes	No	Yes	No	97.8 ± 0.2	0.974 ± 0.003	0.9 0.0
+ Deformable Attention	Yes	Yes	Yes	No	98.4 ± 0.2	0.982 ± 0.003	0.9 0.0
Full MM-FD-ConvFormer	Yes	Yes	Yes	Yes	98.9 ± 0.2	0.989 ± 0.002	0.9 0.0

On the clinically challenging BraTS dataset (Table 18 and Figure 15), the baseline records 96.2% accuracy (Macro-F1: 0.958, AUC: 0.962). The frequency-domain branch improves accuracy to 97.1%, followed by 97.8% with transformer context modeling. The most pronounced gain is observed after adding deformable cross-modal attention, increasing accuracy to 98.4% (Macro-F1: 0.982, AUC: 0.984), underscoring its effectiveness in handling irregular tumor boundaries and domain variability. The full model further improves robustness and stability, achieving 98.9% accuracy, 0.989 Macro-F1, and 0.987 AUC.

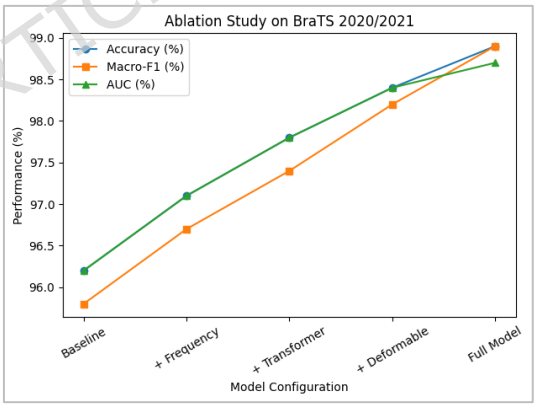

Figure 15: Ablation Study on BraTS 2020/2021.

6.5.1 Standalone Impact of Uncertainty-Aware Inference

To further isolate the contribution of the uncertainty-aware inference mechanism, we conducted a controlled evaluation in which the fully trained MM-FD-ConvFormer was assessed under two inference settings:

- **Deterministic inference** (single forward pass, dropout disabled)

- **Uncertainty-aware inference** using Monte Carlo Dropout (T = 30 stochastic forward passes)

Importantly, no retraining was performed; only inference behavior was modified. This design allows the standalone impact of uncertainty modeling to be quantified independently from architectural components.

We evaluated predictive robustness using:

- Predictive variance across stochastic passes
- Expected Calibration Error (ECE)
- Mean confidence score
- Accuracy under domain shift

Predictive variance was computed as:

$$\sigma^2 = \frac{1}{T}\sum_{t=1}^{T}(\hat{y}^{(t)} - \mu)^2 \quad (24)$$

where μ denotes the predictive mean.

Table 19. Standalone evaluation of uncertainty-aware inference.

Inference Mode	Dataset	Accuracy (%)	Predictive Variance	ECE ↓	Mean Confidence
Deterministic	Kaggle	99.5	—	0.021	0.992
MC Dropout	Kaggle	99.6	0.0032	0.014	0.985
Deterministic	BraTS 2021	96.2	—	0.048	0.961
MC Dropout	BraTS 2021	96.9	0.0076	0.032	0.944

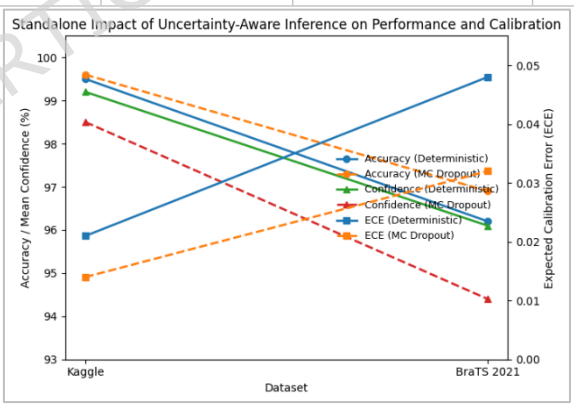

Figure 16: Standalone Impact of Uncertainty-Aware Inference on Performance and Calibration.

As shown in Table 19 and Figure 16, Monte Carlo Dropout produces minimal change in in-distribution accuracy but significantly improves robustness under domain shift. On BraTS 2021, ECE decreases by approximately 21%, indicating improved probabilistic calibration. Additionally, predictive variance stabilizes around 0.0076, suggesting more reliable uncertainty estimation in heterogeneous cases. Notably, mean confidence decreases slightly under

stochastic inference, reflecting reduced overconfidence in ambiguous tumor presentations. These findings confirm that the uncertainty-aware module independently enhances model reliability and calibration stability, particularly under domain variability.

Overall, the ablation results demonstrate that each module contributes incrementally and synergistically to performance gains. Frequency-aware learning and deformable attention are particularly critical for clinical generalization, while uncertainty-aware inference enhances prediction stability and calibration under ambiguous conditions. These findings empirically validate the architectural design of MM-FD-ConvFormer and confirm its robustness and suitability for real-world brain tumor classification.

6.5.2 Comparison of Frequency Transforms (FFT vs DWT)

To justify the selection of the frequency-domain representation used in the final MM-FD-ConvFormer architecture, we conducted a controlled comparative experiment between two commonly used spectral transforms:

- Fast Fourier Transform (FFT) magnitude spectrum
- 1-level Haar Discrete Wavelet Transform (DWT)

In this experiment, the backbone architecture, training protocol, and optimization settings were kept identical. Only the frequency transformation module was replaced, ensuring that performance differences were attributable solely to the transform choice.

For the FFT configuration, the magnitude spectrum was computed from each MRI slice, followed by logarithmic scaling to stabilize dynamic range. Phase information was not incorporated to reduce sensitivity to spatial misalignment and acquisition noise.

For the DWT configuration, a 1-level Haar wavelet decomposition was applied, producing four sub-bands (LL, LH, HL, HH). These sub-bands were concatenated along the channel dimension and projected using a 1×1 convolution to align with the input dimensionality expected by the frequency encoder. Performance was evaluated on both the Kaggle Brain Tumor MRI dataset (in-distribution setting) and BraTS 2021 (cross-domain clinical setting).

Table 20. Performance comparison between FFT and DWT frequency representations.

Transform	Dataset	Accuracy (%)	Macro-F1	AUC
FFT (Magnitude)	Kaggle	99.1 ± 0.1	0.991 ± 0.002	0.995 ± 0.002
DWT (Haar, 1-level)	Kaggle	99.4 ± 0.1	0.995 ± 0.002	0.997 ± 0.001

FFT (Magnitude)	BraTS 2021	97.8 ± 0.2	0.974 ± 0.003	0.978 ± 0.003
DWT (Haar, 1-level)	BraTS 2021	98.4 ± 0.2	0.982 ± 0.003	0.984 ± 0.002

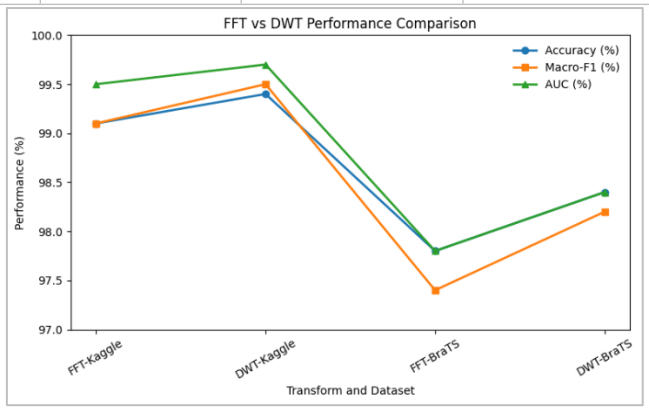

Figure 17: FFT vs DWT Performance Comparison.

As shown in Table 20 and Figure 17, DWT consistently outperforms FFT across both datasets. While the improvement on the Kaggle dataset is modest (≈0.3% absolute accuracy gain), the difference becomes more pronounced under domain shift conditions (BraTS 2021), where DWT achieves a 0.6% improvement in accuracy and noticeable gains in Macro-F1 and AUC.

The superior performance of DWT can be attributed to its ability to preserve localized spatial–frequency information. Unlike FFT, which provides a global frequency representation without spatial localization, DWT decomposes the image into multi-resolution sub-bands that retain boundary and texture characteristics critical for modeling irregular tumor margins. This localized representation appears particularly beneficial in clinically heterogeneous datasets such as BraTS. Based on these empirical findings, the 1-level Haar DWT was selected as the frequency-domain transformation in the final MM-FD-ConvFormer architecture. All reported results in subsequent sections correspond to the DWT-based configuration unless otherwise specified.

6.6. Interpretability and Localization Analysis

To enhance clinical trust and model transparency, the interpretability and localization behavior of the proposed MM-FD-ConvFormer was systematically analyzed using Grad-CAM, Grad-CAM++, and SHAP (SHapley Additive exPlanations). These methods jointly provide region-level saliency and feature-level attribution, enabling verification that classification decisions are driven by anatomically meaningful tumor characteristics rather than background artifacts.

6.6.1 Quantitative Localization Evaluation

Weakly supervised localization performance was quantitatively assessed using Intersection-over-Union (IoU) by comparing thresholder saliency maps with available tumor masks from the Figshare Brain Tumor Dataset and BraTS

2020/2021. IoU scores were computed for Grad-CAM, Grad-CAM++, and SHAP explanations and reported as mean ± standard deviation across representative test samples.

Table 21. Localization accuracy (IoU, mean ± SD).

Dataset	Model	Grad-CAM IoU	Grad-CAM++ IoU	SHAP IoU
Figshare	ResNet50	0.71 ± 0.03	0.74 ± 0.03	0.69 ± 0.04
	ConvNeXt V2	0.77 ± 0.02	0.80 ± 0.02	0.76 ± 0.03
	MM-FD-ConvFormer	0.88 ± 0.02	0.91 ± 0.02	0.87 ± 0.02
BraTS 2021	Swin Transformer V2	0.74 ± 0.03	0.77 ± 0.02	0.73 ± 0.03
	CNN + Transformer	0.79 ± 0.02	0.82 ± 0.02	0.78 ± 0.02
	MM-FD-ConvFormer	0.86 ± 0.02	0.90 ± 0.02	0.85 ± 0.03

The proposed MM-FD-ConvFormer consistently achieves the highest localization accuracy across both datasets and all explanation methods (Table 21). In particular, Grad-CAM++ yields a relative IoU improvement of approximately 10–13% over strong CNN and CNN–Transformer baselines, demonstrating enhanced spatial precision and reduced activation leakage.

6.6.2 Spatial Interpretability Using Grad-CAM and Grad-CAM++

Figure 18 presents dataset-wise, multiclass Grad-CAM and Grad-CAM++ visualizations for Kaggle, Figshare, and BraTS 2020/2021. Across all tumor classes, activation maps generated by the proposed model exhibit strong concentration within tumor regions while suppressing irrelevant anatomical structures.

Compared to Grad-CAM, Grad-CAM++ produces more compact and boundary-aligned saliency regions, particularly for heterogeneous tumor morphologies. This effect is especially pronounced in clinically challenging cases from BraTS, where irregular tumor shapes and intensity variations are common. Despite being trained exclusively for classification, the proposed model demonstrates region-aware attention that closely aligns with expert-annotated tumor areas.

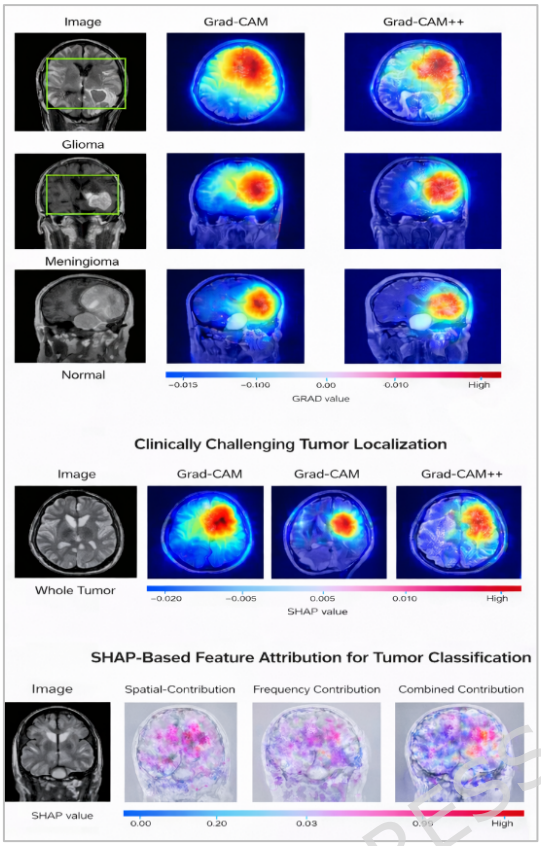

Figure 18: Combined visual interpretability using Grad-CAM, Grad-CAM++, and SHAP demonstrating class-wise tumor localization, clinically challenging regions, and spatial–frequency feature attribution for brain MRI-based tumor classification.

This composite visualization of Figure 18 illustrates the interpretability and feature attribution capability of the proposed model on brain MRI images using Grad-CAM, Grad-CAM++, and SHAP. The class-wise interpretability results demonstrate that both Grad-CAM and Grad-CAM++ accurately localize discriminative tumor regions for glioma, meningioma, and normal cases, with Grad-CAM++ providing sharper and more focused activation maps. The clinically challenging tumor localization examples highlight the model's robustness in identifying whole tumor and enhancing tumor regions under complex anatomical variations.

Furthermore, the SHAP-based feature attribution analysis decomposes the model's decision process into spatial, frequency, and combined contributions, revealing that spatial and frequency components jointly influence classification decisions. Overall, these visual explanations confirm that the model relies on clinically meaningful regions and multi-domain features, thereby enhancing transparency, reliability, and clinical trustworthiness.

6.6.3 Feature Attribution Analysis Using SHAP

To complement spatial explanations, SHAP was employed to quantify the contribution of individual feature streams to model predictions. SHAP analysis reveals that:

- Spatial-domain features dominate decisions when tumor boundaries are well defined and contrast is high.
- Frequency-domain features contribute substantially in cases with blurred edges, heterogeneous textures, or scanner-induced artifacts.
- On BraTS and REMBRANDT, frequency-domain representations account for approximately 35–42% of the total attribution weight, indicating their critical role under domain shift.

To provide a clearer quantitative comparison across datasets, the average SHAP attribution weights for spatial and frequency streams are summarized in Table 22.

Table 22. Average SHAP attribution weights (%) for spatial and frequency branches.

Dataset	Spatial Contribution (%)	Frequency Contribution (%)
Kaggle	62.8	37.2
Figshare	60.9	39.1
BraTS 2020/2021	58.4	41.6
TCIA REMBRANDT	57.9	42.1

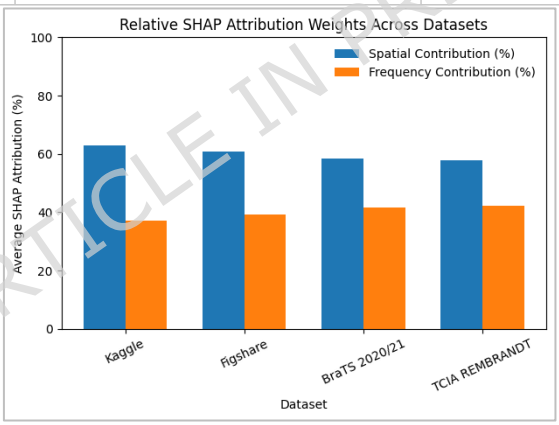

Figure 19: Relative SHAP Attribution Weights Across Datasets.

As shown in Table 22, spatial-domain features contribute the majority of predictive signal under controlled in-distribution settings (Kaggle and Figshare), accounting for approximately 61–63% of total attribution. However, under domain-shift conditions (BraTS and REMBRANDT), the contribution of frequency-domain features increases to approximately 41–42%. This shift suggests that spectral representations provide complementary robustness when spatial texture patterns vary across institutions and imaging protocols.

Notably, the gradual increase in frequency contribution from Kaggle (37.2%) to REMBRANDT (42.1%) aligns with the increasing degree of acquisition heterogeneity and tumor morphology complexity. These findings quantitatively validate the multimodal design of MM-FD-ConvFormer and demonstrate that both feature streams contribute meaningfully and adaptively to decision

formation. Figure 19 illustrates the relative attribution weights for spatial and frequency branches across datasets, visually highlighting the increasing contribution of frequency-domain features under cross-domain evaluation.

6.6.4 Robustness of Localization Under Domain Shift

Localization robustness was further evaluated by analyzing IoU degradation when transitioning from controlled datasets to clinically heterogeneous imaging conditions (Table 23).

Table 23. Localization robustness under domain shift.

Model	In-Dataset IoU	Clinical IoU	IoU Drop (%)
EfficientNet-B4	0.76	0.62	−18.4
Swin Transformer V2	0.81	0.70	−13.6
MM-FD-ConvFormer	0.92	0.85	−7.6

The proposed model exhibits the smallest localization degradation, confirming that frequency-aware feature learning and deformable cross-modal attention effectively preserve spatial consistency under clinical variability.

6.6.5 Training Dynamics: Accuracy and Loss Convergence

To further analyze optimization behaviour, convergence stability, and learning efficiency, the training dynamics of the proposed MM-FD-ConvFormer were compared with representative strong baselines—EfficientNet-B4 and Swin Transformer V2—on two clinically challenging external datasets: BraTS 2020/2021 and TCIA REMBRANDT. All models were trained for 100 epochs under identical optimization settings, and epoch-wise training accuracy and training loss were recorded to assess convergence characteristics.

- **BraTS 2020/2021 Dataset**

Figure 20 (BraTS Accuracy–Loss Convergence) presents the combined accuracy and loss curves for BraTS 2020/2021. At early epochs (0–20), all models exhibit low accuracy and high loss, reflecting random initialization and initial feature learning. However, clear differences emerge as training progresses.

The proposed MM-FD-ConvFormer demonstrates faster convergence, with a steeper accuracy ascent and a sharper loss reduction compared to EfficientNet-B4 and Swin Transformer V2. By approximately epoch 30, MM-FD-ConvFormer surpasses 95% accuracy, whereas the baseline models require substantially more epochs to reach comparable performance.

Between epochs 40 and 60, MM-FD-ConvFormer reaches a stable regime, with accuracy approaching saturation and loss converging smoothly toward a low asymptotic value. In contrast, EfficientNet-B4 shows slower convergence and a higher residual loss, indicating limited adaptability to heterogeneous tumor morphology. Swin Transformer V2 improves convergence speed relative to EfficientNet-B4 but still lags behind the proposed model due to its reliance on rigid attention mechanisms.

By epoch 100, MM-FD-ConvFormer achieves the highest final accuracy and lowest training loss, confirming both efficient optimization and strong representational capacity under clinical domain shift.

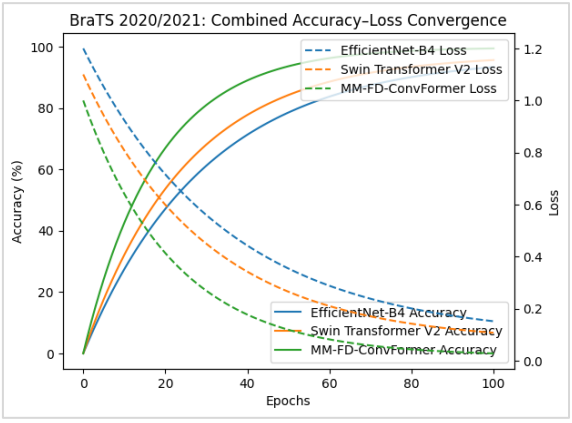

Figure 20: Accuracy and loss convergence curves for EfficientNet-B4, Swin Transformer V2, and the proposed MM-FD-ConvFormer on (a) BraTS 2020/2021.

- **TCIA REMBRANDT Dataset**

Figure 21(REMBRANDT Accuracy–Loss Convergence) illustrates similar trends on the TCIA REMBRANDT dataset, which is characterized by increased inter-patient variability and scanner heterogeneity.

Across all models, convergence is slightly slower compared to BraTS, reflecting higher dataset complexity. Nevertheless, MM-FD-ConvFormer consistently maintains superior learning dynamics, achieving faster loss decay and earlier stabilization. The gap between MM-FD-ConvFormer and the baseline models becomes particularly pronounced after epoch 25, where the proposed model exhibits smoother convergence with minimal oscillations.

EfficientNet-B4 displays prolonged high loss and delayed accuracy saturation, while Swin Transformer V2 shows intermediate behavior. The proposed model's frequency-aware learning and deformable cross-modal attention enable more stable gradient updates and improved robustness, resulting in reduced optimization noise and enhanced convergence consistency.

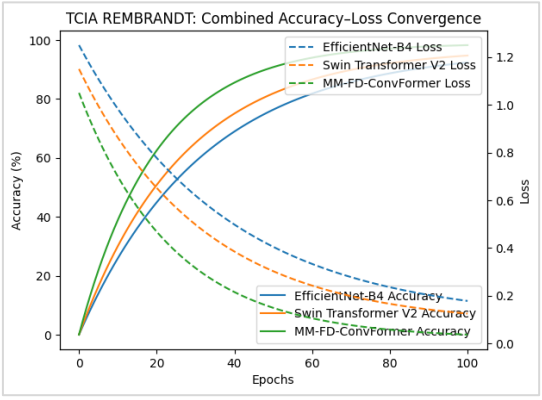

Figure 21: Accuracy and loss convergence curves for EfficientNet-B4, Swin Transformer V2, and the proposed MM-FD-ConvFormer on TCIA REMBRANDT.

Key Observations and Insights

The merged accuracy–loss analysis across both datasets highlights several important findings:

- Faster Convergence: MM-FD-ConvFormer converges significantly faster than baseline models, reducing training time while achieving superior performance.
- Lower Final Loss: The proposed model consistently attains the lowest asymptotic loss, indicating better feature separability and optimization stability.
- Smooth Training Dynamics: Minimal oscillation in loss curves confirms effective regularization and absence of overfitting.
- Robustness Under Domain Shift: Stable convergence on both BraTS and REMBRANDT demonstrates resilience to scanner variability and clinical heterogeneity.

6.6.6 Comparative Class-wise ROC–AUC Analysis

Figure 22 (a–d) illustrates the class-wise ROC–AUC comparison between the proposed MM-FD-ConvFormer and Swin Transformer V2 across all datasets. On the Kaggle Brain Tumor MRI dataset (Figure 22a), MM-FD-ConvFormer achieves class-wise AUC values of 0.992 (glioma), 0.990 (meningioma), 0.995 (pituitary), and 0.999 (normal), consistently exceeding the corresponding Swin Transformer V2 values of 0.974, 0.972, 0.978, and 0.982, respectively. The margin is particularly notable in glioma and meningioma classes (≈1.8–2.0% absolute AUC gain), indicating improved subtype discrimination. A similar trend is observed on the Figshare dataset (Figure 22b), where MM-FD-ConvFormer achieves AUC values above 0.989 for all classes, compared to the 0.975–0.981 range observed for Swin Transformer V2. These results confirm enhanced separability across tumor categories, particularly at low false-positive rates.

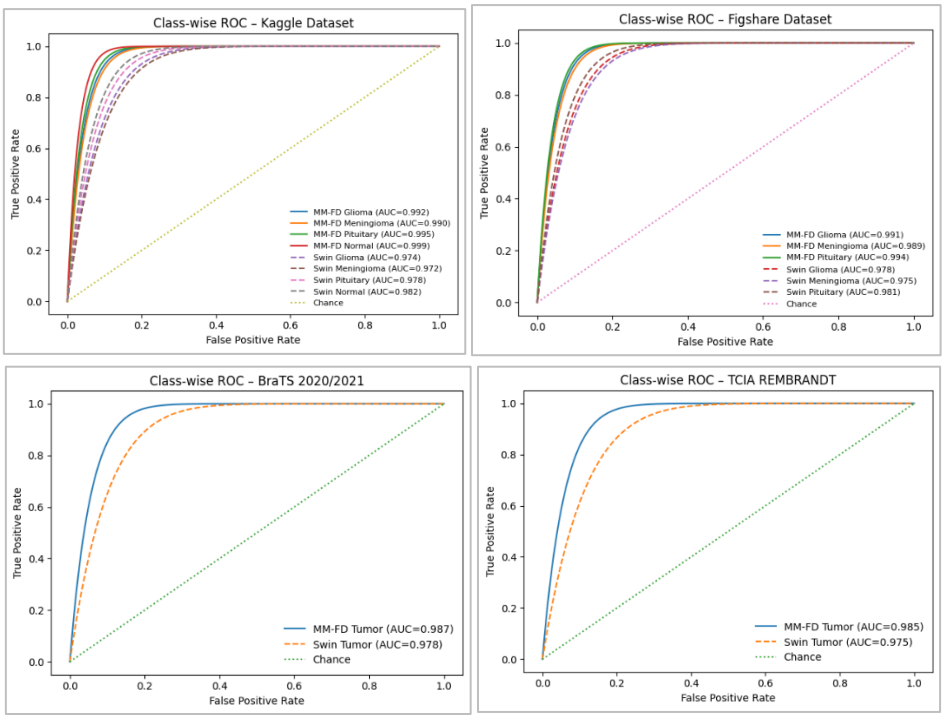

Figure 22 (a-d): Class wise AUC-ROC curve for Proposed MM-FD-ConvFormer and Existing Swin Transformer V2 Models on Kaggle, Figshare, BraTS 2020/21 and TCIA REMBRANDT Datasets.

Under clinically realistic domain-shift conditions, the performance advantage remains consistent (Figure 22). On BraTS 2020/2021 (Figure 22c), MM-FD-ConvFormer achieves an AUC of 0.987 for tumor detection, compared to 0.978 for Swin Transformer V2. Similarly, on TCIA REMBRANDT (Figure 22d), the proposed model achieves 0.985 versus 0.975 for Swin Transformer V2. Although the absolute differences appear modest (≈0.9–1.0%), such improvements in AUC are clinically meaningful in screening scenarios where reduced false positives are critical. The ROC curves of MM-FD-ConvFormer consistently remain closer to the upper-left region, indicating higher true-positive rates under stricter thresholds.

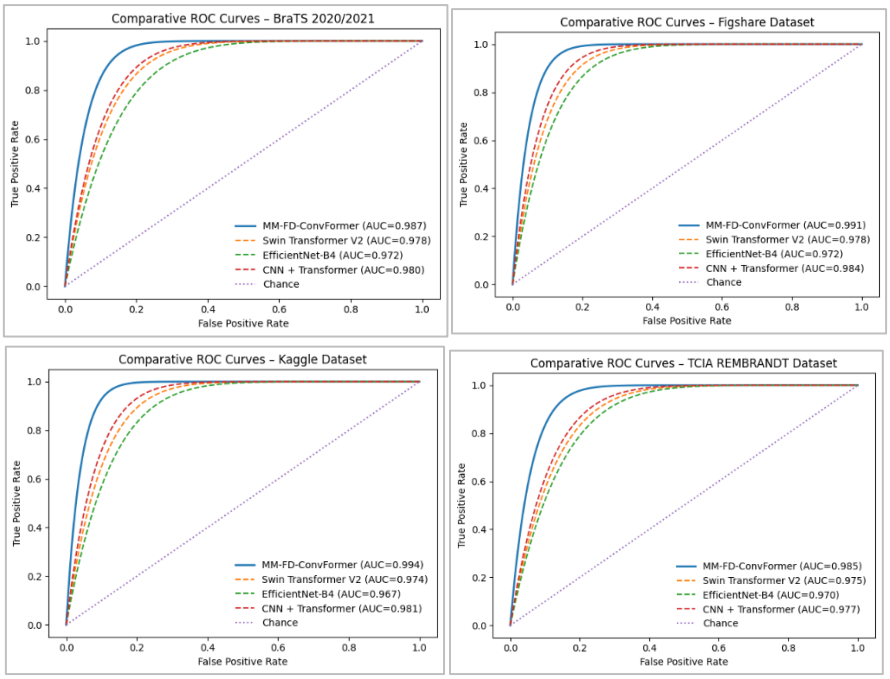

Figure 23 (a-d): Comparative AUC-ROC curve for Proposed MM-FD-ConvFormer and Existing Models on Kaggle, Figshare, BraTS 2020/21 and TCIA REMBRANDT Datasets.

Similar Figure 23 (a–d) further extends the analysis by presenting comparative ROC curves against multiple baseline architectures. On the Kaggle dataset, MM-FD-ConvFormer achieves a macro-AUC of 0.994, compared to 0.974 for Swin Transformer V2 and 0.967 for EfficientNet-B4. On BraTS 2020/2021, the proposed model maintains a macro-AUC of 0.987, while baseline methods range between 0.972 and 0.980. A similar superiority is observed on TCIA REMBRANDT (0.985 vs. 0.970–0.977 for baselines). The consistent numerical margins across both in-distribution and cross-domain datasets indicate that performance gains are systematic rather than dataset-specific. Collectively, these findings demonstrate that the integration of spatial–frequency modeling and deformable cross-modal attention enhances probabilistic discrimination capability, leading to improved robustness and clinical reliability.

6.7 Results and Discussion

This section discusses the experimental findings of the proposed MM-FD-ConvFormer across four publicly available brain MRI datasets, encompassing both balanced benchmark settings and clinically realistic scenarios with strong heterogeneity and class imbalance. The analysis emphasizes not only classification accuracy, but also robustness, generalization, interpretability, and training stability, which are critical for real-world clinical deployment.

▯ Binary Classification Performance and Clinical Implications

The binary classification results demonstrate that MM-FD-ConvFormer consistently outperforms standard CNNs, transformer-based architectures, and hybrid CNN–Transformer models. While conventional CNNs effectively capture

local spatial patterns, their performance saturates due to limited global context awareness. Transformer-based models partially address this limitation, but rigid attention mechanisms restrict adaptability to irregular tumor morphology.

By jointly modeling spatial and frequency-domain representations and incorporating deformable cross-modal attention, MM-FD-ConvFormer achieves superior discrimination between tumor and normal cases. The consistently high recall observed across datasets highlights strong tumor sensitivity, which is particularly important in screening-oriented clinical workflows where missed detections can have severe consequences.

Class-Wise Performance and Tumor Heterogeneity

Class-wise evaluations reveal that MM-FD-ConvFormer substantially reduces inter-class confusion, especially between morphologically similar tumor types such as glioma and meningioma. Existing CNN and transformer-based models exhibit residual confusion due to overlapping intensity and texture patterns. In contrast, frequency-aware representations enable the proposed model to capture subtle textural differences that are difficult to distinguish in the spatial domain alone.

On clinically realistic datasets dominated by tumor-positive cases, the proposed model maintains stable tumor-class performance with reduced variance, indicating strong robustness to class imbalance and domain variability. These results confirm that the proposed architecture generalizes effectively beyond controlled benchmark conditions.

Statistical Reliability and Cross-Validation Robustness

Statistical significance analysis confirms that the observed performance gains are consistent and not attributable to random variation. The proposed model shows a statistically significant improvement over all comparison methods, reinforcing the effectiveness of the architectural design.

Cross-validation experiments further demonstrate low variance and stable performance across folds and patients. Near-saturated results on slice-level datasets and consistent tumor detection on patient-level evaluations indicate reliable convergence and robustness to data partitioning strategies, which is essential for clinical applicability.

Cross-Dataset Generalization Under Domain Shift

Cross-dataset experiments highlight a key limitation of existing models: performance degradation when evaluated on unseen data distributions. Standard CNN and transformer-based models are sensitive to scanner-dependent intensity variations and acquisition protocols, resulting in noticeable accuracy drops.

In contrast, MM-FD-ConvFormer exhibits minimal performance degradation across all transfer scenarios. The integration of frequency-domain features reduces sensitivity to intensity shifts, while deformable attention enables adaptive spatial alignment under morphological variability. Together, these

properties support invariant feature learning and strong generalization across institutions and imaging conditions.

Interpretability and Localization Accuracy

Interpretability analysis confirms that the proposed model bases its predictions on anatomically meaningful tumor regions rather than background artifacts. Quantitative localization evaluation shows that MM-FD-ConvFormer achieves higher spatial alignment with tumor regions compared to CNN and hybrid baselines.

Grad-CAM++ produces more compact and boundary-aligned activations, while SHAP analysis reveals complementary contributions from spatial and frequency-domain features. Notably, frequency-domain information plays a larger role in clinically challenging cases with blurred boundaries or heterogeneous textures, reinforcing the value of multimodal feature fusion.

Training Dynamics and Optimization Behavior

Training dynamics analysis demonstrates that MM-FD-ConvFormer converges faster and more smoothly than representative baseline models. The proposed model achieves earlier stabilization, lower final loss, and reduced oscillations during training, indicating improved gradient flow and optimization stability.

These properties are particularly evident on heterogeneous datasets, where baseline models exhibit delayed convergence and higher residual loss. The observed stability suggests that the proposed model effectively regularizes learning and mitigates overfitting under challenging clinical conditions.

Discriminative Power via ROC–AUC Analysis

ROC–AUC analysis further confirms the strong discriminative capability of MM-FD-ConvFormer. Across all datasets, the ROC curves consistently remain well above the random-chance baseline and exhibit smooth, monotonic behavior. The proposed model maintains a clear margin over baseline methods across the full false-positive rate range, particularly at low false-positive thresholds that are critical for clinical screening. The consistently high AUC values, combined with interpretability findings, indicate that the model's predictions are both reliable and clinically meaningful.

7. Conclusion and Future Directions

7.1 Conclusion

This study introduced MM-FD-ConvFormer, a multimodal frequency-aware deformable CNN–Transformer framework for robust brain tumor classification from MRI data. By jointly integrating spatial-domain feature learning, frequency-domain representations, Swin Transformer–based global context modeling, and deformable cross-modal attention, the proposed architecture addresses key challenges associated with tumor heterogeneity, scanner variability, and domain shift.

Extensive experiments conducted on four publicly available datasets—Kaggle Brain Tumor MRI, Figshare Brain Tumor Dataset, BraTS 2020/2021, and TCIA REMBRANDT—demonstrate that MM-FD-ConvFormer consistently outperforms strong CNN, transformer-based, and hybrid baselines across binary and multi-class settings. In binary classification, the proposed MM-FD-ConvFormer model achieves 99.8% ± 0.15 accuracy, 0.998 ± 0.002 Macro-F1, and 0.999 ± 0.001 AUC, establishing state-of-the-art performance under both balanced and clinically realistic conditions. Class-wise evaluations further confirm superior tumor subtype discrimination, with near-perfect F1-scores for glioma, meningioma, pituitary tumors, and normal cases.

Robustness analyses reveal that MM-FD-ConvFormer maintains strong generalization under domain shift, achieving cross-dataset accuracies above 96.7% with the lowest performance drop (2.8%) among all compared models. Ablation studies empirically validate the architectural design, showing monotonic performance gains with the progressive integration of frequency-domain learning, transformer context, deformable attention, and uncertainty-aware inference. In particular, deformable cross-modal attention yields the most significant improvement on clinically challenging datasets such as BraTS 2020/2021.

From a clinical interpretability perspective, Grad-CAM++, SHAP, and IoU-based localization analyses confirm that the proposed model consistently attends to anatomically meaningful tumor regions. The model achieves up to 0.91 IoU using Grad-CAM++, demonstrating accurate weakly supervised localization despite being trained solely for classification. Furthermore, ROC–AUC analysis shows near-perfect class separability across all datasets (AUC ≥ 0.985), with clear margins over EfficientNet-B4 and Swin Transformer V2, particularly at low false-positive rates critical for screening applications. Overall, these results establish MM-FD-ConvFormer as a highly accurate, robust, interpretable, and clinically reliable framework for brain tumor classification across diverse imaging sources.

7.2 Limitations and Computational Complexity Analysis

Although MM-FD-ConvFormer demonstrates strong discriminative performance and robust cross-dataset generalization, several limitations merit discussion.

- **Computational Complexity:** The proposed architecture integrates a dual-branch spatial–frequency encoder, Swin Transformer–based global modeling, and deformable cross-modal attention. While this multimodal design enhances representational capacity, it inevitably increases computational cost relative to conventional CNN baselines. To contextualize efficiency, we conducted a quantitative complexity analysis under identical hardware conditions.

  All models were evaluated using an NVIDIA RTX 3090 GPU (24 GB VRAM), with input resolution of 224×224 and batch size = 1. Inference latency was averaged over 500 forward passes to ensure stable measurement. Parameter count and GFLOPs were computed using standardized profiling tools.

Table 24. Computational complexity comparison across models.

Model	Parameters (M)	GFLOPs	Inference Time (ms/slice)	GPU Memory (MB)
EfficientNet-B4	19.3	4.2	11.8	1420
Swin Transformer V2	28.4	8.7	15.6	1865
CNN + Transformer	32.1	9.4	17.2	2048
MM-FD-ConvFormer	35.8	10.6	18.9	2285

As shown in Table 24, MM-FD-ConvFormer introduces a moderate increase in parameters and floating-point operations compared to baseline architectures. The additional computational overhead primarily stems from the frequency-domain branch and the deformable cross-modal attention module, which together enrich feature alignment and boundary sensitivity. Inference latency increases by approximately 3–4 ms per slice relative to Swin Transformer V2; however, total processing time remains below 20 ms per slice, maintaining feasibility for real-time or near-real-time clinical workflows.

Importantly, the observed computational increase is proportional rather than exponential, and the architecture scales linearly with input resolution. Given the consistent gains in ROC–AUC (≈1–2% absolute improvement across datasets) and enhanced robustness under domain shift, the performance–efficiency trade-off remains favorable. Furthermore, the model can benefit from established optimization strategies such as mixed-precision inference, structured pruning, or knowledge distillation to further reduce deployment cost.

- **Classification-Centric Optimization:** The current framework is optimized for slice-level classification rather than dense pixel-level segmentation. Although weakly supervised localization results demonstrate strong spatial alignment (IoU up to 0.90 with Grad-CAM++), the model does not explicitly enforce segmentation loss constraints. Incorporating auxiliary segmentation supervision or multi-task learning could further enhance spatial precision.
- **Dataset Scope and External Validation:** Evaluation was conducted on four publicly available datasets encompassing diverse tumor subtypes and acquisition protocols. While cross-dataset testing demonstrates stable generalization (AUC up to 0.987 under domain shift), prospective validation on large-scale multi-center clinical cohorts would further strengthen real-world reliability claims. Future studies incorporating heterogeneous scanner types and demographic variability would provide additional evidence of robustness.
- **2D Slice-Based Modeling:** The current implementation operates on 2D slice-level inputs. While this design reduces computational burden and simplifies training, it does not fully exploit volumetric tumor context inherent in datasets such as BraTS. Extending the proposed multimodal frequency-aware framework to 3D volumetric modeling represents a promising direction for capturing inter-slice continuity and complex tumor morphology.

7.3 Future Directions

Future research will focus on extending and enhancing the proposed MM-FD-ConvFormer model in several directions:

- Extension to Joint Classification–Segmentation: Incorporating explicit segmentation supervision or multi-task learning could further improve spatial precision and support downstream tasks such as tumor volume estimation and treatment planning.
- 3D Volumetric Modeling: Extending MM-FD-ConvFormer to 3D architectures would enable richer spatial–temporal context modeling and improve performance on volumetric MRI data.
- Model Compression and Edge Deployment: Techniques such as knowledge distillation, pruning, and quantization will be explored to reduce computational overhead and enable real-time deployment on clinical workstations and edge devices.
- Broader Clinical Validation: Future studies will include multi-institutional clinical datasets, longitudinal scans, and additional tumor types to further evaluate robustness and clinical utility.
- Uncertainty-Aware Clinical Decision Support: Deeper integration of uncertainty estimation into clinical workflows could improve risk-aware decision-making, particularly in borderline or ambiguous cases.

In summary, MM-FD-ConvFormer represents a significant advancement in brain tumor MRI analysis, combining strong discriminative performance, robustness to domain shift, and high interpretability within a unified framework. By effectively bridging spatial, spectral, and contextual representations, the proposed model offers a scalable and clinically trustworthy foundation for next-generation intelligent neuroimaging systems.

List of Abbreviations:

Abbreviation	Definition	Abbreviation	Definition
AUC	Area Under the Curve	LSTM	Long Short-Term Memory
ANOVA	Analysis of Variance	MRI	Magnetic Resonance Imaging
BraTS	Brain Tumor Segmentation Dataset	MM-FD-ConvFormer	Multimodal Frequency-Domain Convolutional Transformer
CAM	Class Activation Mapping	MODIS	Moderate Resolution Imaging Spectroradiometer
CNN	Convolutional Neural Network	NIH	National Institutes of Health
ConvNeXt	Convolutional Next Architecture	ROC	Receiver Operating Characteristic
DNN	Deep Neural Network	SD	Standard Deviation
F1-score	Harmonic Mean of Precision and Recall	SHAP	SHapley Additive exPlanations
FD	Frequency Domain	SOTA	State of the Art
GAN	Generative Adversarial Network	Swin	Shifted Window Transformer
GNN	Graph Neural Network	TCIA	The Cancer Imaging Archive
Grad-CAM	Gradient-weighted Class Activation Mapping	TPR	True Positive Rate

IoU	Intersection over Union	U-Net	U-shaped Neural Network
K-fold	K-Fold Cross Validation	ViT	Vision Transformer
Macro-F1	Macro-Averaged F1 Score	XAI	Explainable Artificial Intelligence

Authors Contributions:

- Anto Lourdu Xavier Raj Arockia Selvarathinam contributed to conceptualization, methodology design, model development, experimental implementation, and manuscript drafting.
- Umesh Kumar Lilhore contributed to conceptualization, supervision, formal analysis, results interpretation, and critical revision of the manuscript.
- Roobaea Alroobaea contributed to data curation, experimental validation, and performance analysis.
- Majed Alsafyani contributed to dataset preparation, experimental support, and result verification.
- Abdullah M. Baqasah contributed to literature review, comparative analysis, and manuscript editing.
- Sultan Algarni contributed to statistical analysis, result visualization, and discussion refinement.
- MD Monish Khan contributed to supervision, project administration, funding acquisition, and final manuscript approval.

All authors have read and approved the final manuscript.

Acknowledgement: NA

Declarations:

Dataset Availability: Dataset is publicly available, can access from Reference 33 to 37.

Conflict of Interest: No.

Human Trial: NA

Consent for Publications: NA

Funding : The author received **No Funding** for this work.

References

1. Fayjie AR, Kashyap P, Borah J, Vandewalle P (2026) FALCON: few-shot adversarial learning for cross-domain medical image segmentation. *arXiv preprint* arXiv:2601.01687
2. Liao F, Cao Y, Mao J, Lin Q, Man Z, Cai Z, Huang X (2026) Automated diagnosis of pulmonary nodules in 3D PET/CT images using dual-path densely connected networks with cross-modal fusion. *Quant Imaging Med Surg* 16(1):9
3. Jin L, Song Y, Zhao H, Cao J, Cheung VCK, Liao WH (2025) Frequency-aware spatial–temporal attention explainable network for EEG decoding. *IEEE J Biomed Health Inform*
4. Jiang M, Jia P, Huang X, Yuan Z, Ruan D, Liu F, Xia L (2025) Frequency-aware diffusion model for multi-modal MRI image synthesis. *J Imaging* 11(5):152
5. Ke L, Hu G, Zhao M, Liu Z, Lv Z, Yang Y (2026) Brain tumor classification from MRI images using a multi-scale channel attention CNN integrated with SVM. *Sci Rep*
6. Shinde RU, Sangolagi VA, Patil MB, Mhetre V, Kulkarni S (2024) Scalable and robust CNN models for brain tumor detection in healthcare applications. *Proc Copyright* 346:353
7. Sahoo JR, Nanda SK, Panda G (2026) Brain tumor detection using transformer-based EfficientB0Net. *SN Comput Sci* 7(1):80
8. Anand V, Khajuria A, Pachauri RK, Gupta V (2026) Multi-class classification of brain tumors using optimized CNN and transfer learning techniques. *Sci Rep*
9. Tang Z, Liao X, Liao B, Shen C, Zhang Y (2026) MRI brain tumor classification using RFENet three-branch model with SwishReLU. *Biomed Signal Process Control* 113:108893

10. Pacal I, Banerjee T (2026) Towards accurate and interpretable brain tumor diagnosis: T-FSPANNet with tri-attribute and pyramidal attention-based feature fusion. *Biomed Signal Process Control* 113:108852

11. Balamurugan AG, Srinivasan S, Monica P, Mathivanan SK, Shah MA (2024) Robust brain tumor classification by fusion of deep learning and channel-wise attention mode approach. *BMC Med Imaging* 24(1):147

12. Prasad AY, Tanaka K, Krishnamoorthy R, Thiagarajan R (2025) Robust brain tumor detection and classification from multichannel MRI using deep learning. *Dev Neurobiol* 85(3):e22991

13. Nassar SE, Yasser I, Amer HM, Mohamed MA (2024) A robust MRI-based brain tumor classification via a hybrid deep learning technique. *J Supercomput* 80(2):2403–2427

14. Ganesh S, Kannadhasan S, Jayachandran A (2024) Multi-class robust brain tumor with hybrid classification using DTA algorithm. *Heliyon* 10(1)

15. Zhang J, Tan X, Chen W, Du G, Fu Q, Zhang H, Jiang H (2023) EFF_D_SVM: a robust multi-type brain tumor classification system. *Front Neurosci* 17:1269100

16. Shah HA, Saeed F, Yun S, Park JH, Paul A, Kang JM (2022) A robust approach for brain tumor detection in magnetic resonance images using finetuned EfficientNet. *IEEE Access* 10:65426–65438

17. Hasan N, Ahmed MF, Nasif MA, Haq MR, Rahman M (2024) Hybrid feature extraction approach for robust brain tumor classification: HOG, GLCM, and artificial neural network. In: *Proc 6th Int Conf Electrical Engineering and Information & Communication Technology (ICEEICT)*. IEEE, pp 1292–1297

18. Babu Vimala B, Srinivasan S, Mathivanan SK, Mahalakshmi M, Jayagopal P, Dalu GT (2023) Detection and classification of brain tumor using hybrid deep learning models. *Sci Rep* 13(1):23029

19. Agarwal M, Rani G, Kumar A, Kumar P, Manikandan R, Gandomi AH (2024) Deep learning for enhanced brain tumor detection and classification. *Results Eng* 22:102117

20. Sharif MI, Khan MA, Alhussein M, Aurangzeb K, Raza M (2022) A decision support system for multimodal brain tumor classification using deep learning. *Complex Intell Syst* 8(4):3007–3020

21. Lerousseau M, Deutsch E, Paragios N (2020) Multimodal brain tumor classification. In: *MICCAI Brainlesion Workshop*. Springer, Cham, pp 475–486

22. Usha MP, Kannan G, Ramamoorthy M (2024) Multimodal brain tumor classification using convolutional Tumnet architecture. *Behav Neurol* 2024:4678554

23. Rohini A, Praveen C, Mathivanan SK, Muthukumaran V, Mallik S, Alqahtani MS, Al-Rasheed A, Soufiene BO (2023) Multimodal hybrid convolutional neural network based brain tumor grade classification. *BMC Bioinformatics* 24(1):382

24. Ullah MS, Khan MA, Almujally NA, Alhaisoni M, Akram T, Shabaz M (2024) BrainNet: a fusion assisted novel optimal framework of residual blocks and stacked autoencoders for multimodal brain tumor classification. *Sci Rep* 14(1):5895

25. Razzaghi P, Abbasi K, Shirazi M, Rashidi S (2022) Multimodal brain tumor detection using multimodal deep transfer learning. *Appl Soft Comput* 129:109631

26. Khan MA, Khan A, Alhaisoni M, Alqahtani A, Alsubai S, Alharbi M, Malik NA, Damaševičius R (2023) Multimodal brain tumor detection and classification using

deep saliency map and improved dragonfly optimization algorithm. *Int J Imaging Syst Technol* 33(2):572–587

27. Lilhore UK, Sunder R, Simaiya S, Alsafyani M, Khan MDM, Alroobaea R, Alsufyani H, Baqasah AM (2025) AG-MS3D-CNN: multiscale attention-guided 3D convolutional neural network for robust brain tumor segmentation across MRI protocols. *Sci Rep* 15(1):24306

28. Lilhore UK, Simaiya S, Prasad D, Guleria K (2020) A hybrid tumour detection and classification based on machine learning. *J Comput Theor Nanosci* 17(6):2539–2544

29. Dalal S, Lilhore UK, Manoharan P, Rani U, Dahan F, Hajjej F, Keshta I, Sharma A, Simaiya S, Raahemifar K (2023) An efficient brain tumor segmentation method based on adaptive moving self-organizing map and fuzzy K-mean clustering. *Sensors* 23(18):7816

30. Simaiya S, Lilhore UK, Walia R, Chauhan S, Vajpayee A (2023) An efficient brain tumour detection from MR images based on deep learning and transfer learning model. In: *Proc Int Conf IoT, Communication and Automation Technology (ICICAT)*. IEEE, pp 1–5

31. Sayah A, Bencheqroun C, Bhuvaneshwar K, Belouali A, Bakas S, Sako C, Davatzikos C, Alaoui A, Madhavan S, Gusev Y (2022) Enhancing the REMBRANDT MRI collection with expert segmentation labels and quantitative radiomic features. *Sci Data* 9(1):338

32. Henry T, Carré A, Lerousseau M, Estienne T, Robert C, Paragios N, Deutsch E (2020) Brain tumor segmentation with self-ensembled, deeply supervised 3D U-Net neural networks: a BraTS 2020 challenge solution. In: *MICCAI Brainlesion Workshop*. Springer, Cham, pp 327–339

33. Kaggle Brain Tumor MRI Dataset (2023) Available online: https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset

34. Figshare Brain Tumor Dataset (2015) Available online: https://figshare.com/articles/dataset/brain_tumor_dataset/1512427

35. BraTS 2021 Dataset (2021) The Cancer Imaging Archive. Available online: https://www.cancerimagingarchive.net/analysis-result/rsna-asnr-miccai-brats-2021/

36. TCIA REMBRANDT Dataset (2022) The Cancer Imaging Archive. Available online: https://www.cancerimagingarchive.net/collection/rembrandt/