Available online at http://scik.org
Commun. Math. Biol. Neurosci. 2026, 2026:72
https://doi.org/10.28919/cmbn/10012
ISSN: 2052-2541

ADVERSARIALLY ROBUST BRAIN TUMOR DETECTION IN MULTI-MODAL MRI USING HYBRID DEEP LEARNING AND RADIOMIC FEATURES

MONICA LUTHRA[1,2,*], SELLAPPAN PALANIAPPAN[3], DANIEL AROCKIAM[4]

[1]Post doc fellow, Faculty of Computing and Digital Technology, HELP University, Kuala Lumpur 50490, Malaysia

[2]Apex Institute of Technology (CSE), Chandigarh University, Mohali, Punjab 140413, India

[3]Faculty of Computing and Digital Technology, HELP University, Kuala Lumpur 50490, Malaysia

[4]School of Engineering and IT, Manipal Academy of Higher Education, Dubai Campus, UAE

Abstract: Deep learning has significantly improved automated brain tumor detection from multi-modal magnetic resonance imaging (MRI). However, many existing models remain vulnerable to adversarial perturbations, noise, and domain shifts, which can reduce diagnostic reliability in real-world clinical environments. To address these challenges, this study proposes HCTR-R, a robustness-oriented hybrid computational intelligence framework that integrates convolutional neural networks, transformer-based contextual modeling, and radiomic feature representations for reliable tumor detection. The framework incorporates defense-aware learning strategies, including noise injection, adversarial sample generation using the Fast Gradient Sign Method (FGSM), and adaptive data augmentation to enhance model stability under perturbations. A cross-attention fusion mechanism is employed to combine deep and radiomic features, enabling more effective representation learning from multi-modal MRI data. Unlike conventional hybrid architectures, the proposed approach explicitly focuses on robustness and cross-dataset generalization in

*Corresponding author
E-mail address: monica.l@helplive.edu.my
Received June 01, 2026

medical imaging. Experimental evaluation was conducted using the combined BraTS 2019–2021 datasets for training and internal validation, while the TCGA-GBM dataset was used for external validation to assess generalization across institutions. The results indicate that the proposed framework achieves strong and consistent performance under noisy and adversarial conditions, with high accuracy, F1-score, and AUC values. Explainability analysis using Grad-CAM and SHAP further demonstrates meaningful correspondence between model predictions and tumor regions. The proposed approach supports the development of robust and interpretable computational intelligence systems for AI-driven clinical decision support in healthcare.

Keywords: brain tumor detection; adversarial robustness; computational intelligence; hybrid deep learning; radiomics; explainable AI; multi-modal MRI; medical image analysis.

2020 AMS Subject Classification: 68T07, 68T10, 92C55.

1. INTRODUCTION

The use of artificial intelligence (AI) and deep learning (DL) has experienced growth over the last decade in medical imaging, and it has profoundly changed the diagnostic process as regarded in modern healthcare [1-6]. Brain tumor detection, segmentation, and classification using magnetic resonance imaging (MRI) have become essential tasks for neuro-oncology [7]. The computational models help the radiologists in early diagnosis, prognosis evaluation, and planning of treatment [8]. However, despite convolutional neural networks (CNNs), vision transformers (ViTs), and mixed architectures are available, reliability and robustness are still significant issues when it comes to translating such models into the clinical practice [9-12]. Even the models those perform well under controlled benchmark conditions, those models may also fail significantly when exposed to real-world perturbations, adversarial manipulations, or corrupted data. These weaknesses can undermine the accuracy of diagnosis and safety of the patients under AI-assisted decision system.

In this regard, the surety of the model stability under uncertain conditions has become an important research focus [13]. Medical imaging information is heterogeneous by nature. There may be the variations due to differences in the scanners, acquisition protocols, patient motion, and random noise. Moreover, adversarial perturbations are small well-known modifications that can cause the diagnostic prediction to be highly distorted with the goal to deceive neural networks [14-15]. The studies have demonstrated that such attacks can easily mislead even state-of-the-art CNN and transformer models into providing confident but erroneous predictions. These risks are unacceptable in the medical sectors of high stakes like neuro-oncology [16]. Therefore, the next step in developing deep learning-based brain tumor diagnostic systems should focus on improving adversarial resilience, interpretability, and real-world reliability [6], [17].

1.1. Motivation and Research Gap

Deep learning methods have demonstrated exceptional performance on widely used brain tumor imaging datasets such as BraTS and TCGA-GBM [18]. The architectures such as U-Net, ResNet, and Vision Transformers (ViTs) have significantly improved lesion localization and tumor classification accuracy. Nevertheless, such models usually assume ideal conditions in the training and also in the inference. Even small variations in noise levels, contrast, or intensity normalization may lead to reproducibility errors. Also, despite the fact that transformers identify global context, they are also likely to overfit small-sized datasets and have weak adversarial resistance to structured noise. On the other hand, CNNs are effective at extracting local texture and edge information, but have limitations in modeling long-range dependencies and are vulnerable to pixel-level errors.

Another important limitation is data imbalance and the scarcity of annotated medical images. Brain tumor datasets often exhibit non-uniform class distributions, and large-scale labeled data require expert annotations, which are expensive and time-consuming to obtain. When this happens, models trained on limited or imbalanced datasets are more vulnerable to adversarial attacks and domain shifts. These issues highlight the necessity of hybrid architectures based on the combinations of multi-scale feature extraction, cross-modal fusion and the adversarial training strategies to develop models that are robust and capable of better generalization [19].

Another gap in the literature that can be identified is the absence of the use of handcrafted radiomic features (quantitative descriptors based on tumor morphology, intensity, and texture) in deep-neural representations. Radiomics can enhance interpretability by incorporating a priori domain knowledge into data-driven learning, thereby improving generalization under uncertain or noisy conditions. However, little of the current literature is able to combine these handcrafted features with adversarially trained hybrid deep learning networks for medical imaging. The current study fills this important gap by proposing a Hybrid CNN–Transformer–Radiomics for Robust brain tumor detection (HCTR-R) framework that focuses on improving robustness and interpretability in AI-based neuro-oncology.

1.2. Architectural Overview

In order to address the complexities of robust brain tumor detection, this paper presents the HCTR-R framework, which operates as a unified multi-branch pipeline. The architecture combines local strategy feature extraction and global contextual relationship in multi-modal MRI data.

The framework consists of three main components. First, an adaptive preprocessing stage standardizes MRI volumes using skull stripping, bias-field correction, and intensity normalization to improve data consistency. Second, a hybrid feature learning stage extracts representations through three parallel branches: a CNN-based branch for local spatial features, a transformer-based branch for modeling long-range contextual dependencies, and a radiomics branch for extracting handcrafted descriptors that provide clinically interpretable information. Finally, a cross-attention fusion module integrates the heterogeneous feature representations and produces predictions through segmentation and classification heads. The detailed architecture and training procedure of the HCTR-R framework are presented in Section 3.

1.3. Conceptual Foundation of Robustness and Adversarial Resilience

Deep learning model robustness refers to the stability of a model to natural perturbations (noise, blur, change in illumination) as well as adversarial perturbation (deliberate minimal distortions introduced to mislead neural networks) [20]. Adversarial examples take advantage of the excessively sensitive nature of neural networks to small changes in pixel intensity values, creating examples that are visually similar to humans, but are predicted very differently by the neural network. Well-known methods to produce such attacks are the Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD) and Basic Iterative Method (BIM). Although the attacks have been talked about widely in general computer vision, they have much more serious consequences in medical imaging [21].

Any sort of misclassification in medical diagnosis will trigger inaccurate clinical instructions, misdiagnosis or treatment errors. As an example, the input of an adversarial perturbation may lead to the wrong classification of a malignant tumor as being benign or the other way round. In turn, both the architectural and training levels of proactive defence mechanisms are to be adopted. Adversarial training in which the model is exposed to both clean and adversarial samples has become a major defence measure. In perturbation-stable models, invariant representations can be learned using perturbation-conscious optimization and stochastic regularization.

However, the adversarial robustness alone is not sufficient to ensure the true model robustness. The models also need to be able to adapt to domain variations, acquisition artifacts, and inter-scanner variability. The current work proposed the HCTR-R framework which combines the adversarial training, noise-resistant feature learning, and cross-modal fusion into a single architecture structure, that ensures the security and resilience. This integration is particularly important in the intelligent healthcare environments, where the diagnostic systems interact with distributed hospital databases, cloud-edge infrastructures, and federated learning platforms. These environments often face challenges such as data inconsistency and privacy constraints

1.4. Deep Learning and Radiomics: A Synergistic Paradigm

Radiomic features integrated with the convergence of CNNs and transformer encoders present a promising direction for the robust medical imaging. CNNs provide fine-grained spatial features (edges, textures, and boundaries) which play an important role related to segmentation and regional localization processes. Self-attention mechanisms in transformers are effective in long-range correlations and information regarding the global context. The combination of these paradigms results in multi-scale representations and they can explain the local tumor morphology and even more global spatial relations.

This combination is further improved by Radiomics with the assistance of hand-crafted measures, which encode statistical patterns, shape features and voxel-based distributions of intensity. These descriptors are informed by domain knowledge and will offer an interpretability layer, something that is inherently missing in all-deep representations. More importantly the radiomic descriptors do not seem to be subject to changes when small perturbation is introduced to it and this is what makes them best in improving the robustness. The HCTR-R model combines profound representations and domain wisdom via cross-attention. This combination brings together learned and handcrafted features, thus enhancing the adaptability and the interpretability.

1.5. Research Aim and Objectives

On the basis of these learnings, the purpose of the proposed research is to create and evaluate a hybrid deep learning architecture capable of withstanding noise, data distortion and adversarial attacks on brain tumor classification and segmentation. The primary objectives are:

- To develop a hybrid CNN-Transformer-Radiomics architecture that is capable of naturally combining local and global feature learning with interpretable handcrafted features.
- To adopt noise injections and adversarial training efforts to add protection against perturbation and variations across noise domains.
- To assess the robustness of the proposed framework under various attack conditions (FGSM, Gaussian noise, and intensity perturbations) using benchmark medical imaging datasets such as the combined BraTS 2019–2021 dataset.
- To ensure interpretability and clinical reliability by using visualization instruments (Grad-CAM, SHAP) that align model predictions with clinical reasoning.
- To evaluate the viability and scalability of smart healthcare implementation through cloud–edge architecture and federated learning schemes in real time.

All these objectives will contribute to the bridge the gap of the high-performance academic model and reliable clinical grade diagnostic devices.

1.6. Significance and Contributions

The main contribution of this research is the development of a robust hybrid deep learning architecture in which adversarial resilience is incorporated directly into the model design, instead of applied as a post-hoc defense. The significant contributions may be summarized as follows:

1. Novel HCTR-R Architecture: A hybrid framework combining CNNs, Transformer, and Radiomics branches, where each branch is designed to capture different aspects of robustness. It includes the structural representation, global contextual understanding, and interpretability.
2. Adversarially Aware Learning: The training incorporates adversarial sample generation and perturbation-consistent data augmentation. It improves the noise resistance and feature stability.
3. Cross-Attention Fusion Mechanism: A custom cross-attention module aligns deep features with handcrafted radiomic descriptors. It enables the complementary information exchange and improves robustness across varying imaging conditions.
4. Explainable AI Integration: The framework incorporates explainability techniques that provide visual interpretation of model decisions. It also supports the clinical transparency and trust.

An example of how the Smart Healthcare Readiness can support distributed and privacy-preserving deployment is via federated learning, which is compatible with international regulatory requirements like HIPAA and GDPR.

In this way, the paper will contribute to the state of the art in the field of robust medical image analysis by indicating that resilience and interpretability can be achieved simultaneously in the same hybrid architecture.

The remainder of this paper is organized as follows. Section 2 reviews related work on CNN-based, transformer-based, and radiomics-based tumor detection methods. Section 3 presents the proposed HCTR-R architecture. Section 4 describes the experimental results. Section 5 discusses the overall findings from the result outcomes. Finally, Section 6 concludes the study and outlines future research directions.

2. Related Work

In recent years, deep learning techniques have been widely adopted in medical imaging applications, particularly for brain tumor diagnosis and segmentation. The availability of large-scale datasets, advances in high-performance computing, and the development of hybrid deep learning architectures have accelerated the progress of AI-driven neuro-oncology systems. However, despite their high performance under controlled experimental conditions, deep neural models remain vulnerable to adversarial noise, environmental variations, and cross-domain data heterogeneity, which limits their clinical deployment [22].

This section provides an overview of the key research streams relevant to the present study: (i) adversarial robustness and noise resilience in medical imaging, (ii) hybrid deep learning models that combine convolutional and transformer models, and (iii) radiomics-based interpretability and clinical reliability. The section concludes by identifying key research gaps that motivate the design of the proposed HCTR-R framework.

2.1. Adversarial Robustness and Noise Resilience in Medical Imaging

Although deep learning models have achieved human-level accuracy in several image-based diagnostic tasks, they remain highly sensitive to small and often imperceptible perturbations in the input data. This weakness is particularly concerning in medical settings, where minor imaging artifacts or scanner variations can affect clinical decisions. Adversarial attacks are deliberately crafted pixel-level perturbations designed to mislead deep learning models and can severely decrease diagnostic accuracy, and hence undermine the stability and reliability of AI-based diagnostics [21].

Joel et al. (2022) [23] examined this effect by assessing the susceptibility of diagnostic models that were trained on oncology radiographic images, that is, magnetic resonance imaging, computed tomography, and mammographic images. They reported that classification accuracy may decrease significantly due to minor perturbations, amounting to 70 percent, thus demonstrating the fragility of modern systems when it comes to clinical environments. Moreover, their findings suggested that more traditional defensive methods (e.g. input denoising and adversarial retraining) achieved limited improvements while often reducing generalization performance on unseen data. This is a trade-off of both accuracy and robustness, which has been a primary research issue in medical AI

[24-25].

The recent surveys on adversarial robustness in medical imaging mainly indicate that the robustness should be treated as a core design principle rather than a secondary objective. The existing defense strategies are commonly categorized into architectural methods (e.g., regularization and feature smoothing), algorithmic approaches such as adversarial training, and statistical techniques including uncertainty calibration. However, most studies test models on clean and stationary datasets, which do not reflect realistic perturbations, e.g. imaging noise, motion artifacts, or scanner variability. As a result, it is likely that the models which are effective in highly controlled experimental settings, may not work in diverse clinical settings [26].

There is also research in multimodal data fusion as a process that can enhance the robustness of models. This research indicates that the integration of several sources of data including medical images, clinical metadata information, and textual information may enhance predictive stability. This also minimizes susceptibility to perturbations since complementary modalities are redundant information used in decision making. Consequently, multimodal systems tend to work better than unimodal networks in heterogeneous clinical settings [27-28].

In addition, several studies have investigated noise-aware training strategies including Gaussian noise injection, dropout regularization, and test-time augmentation for the improvement of the model stability under corrupted inputs. These techniques improve resilience to common image distortions and minor perturbations, but these are often insufficient against stronger adversarial attacks, designed to mislead deep learning models [14], [29].

Together, these papers suggest that robustness cannot be achieved solely through post-hoc defenses but has to be integrated in the architecture of the model, in the learning process, and in the evaluation regime. This has led to an increase in interest in hybrid, adversarially trained models, which can both learn invariant representations and maintain interpretability, which are the direct objectives of the HCTR-R model.

2.2. Hybrid Deep Learning Architectures for Brain Tumor Detection

CNNs have long served as the backbone of medical image analysis due to their ability to learn local spatial hierarchies and translational invariance. However, CNNs have inherent limitations due to their fixed receptive fields, which restrict their ability to capture long-range dependencies. This has led to renewed interest in the integration of self-attention using transformer-based mechanisms, which can capture global contextual relationships, including the study of complex three-dimensional medical imaging e.g. MRI volumes.

Lan et al. (2023) [30] introduced a CNN–Transformer hybrid model with data augmentation for volumetric brain tumor segmentation. Their hybrid architecture was very effective in capturing global spatial relations, hence improving boundary continuity and reducing the false positives. In a similar fashion, Aloraini et al. (2023) [31] combined CNN blocks that are based on DenseNet and Vision Transformer (ViT) modules in tumor classification and obtained significant improvements in the accuracy. In this design, transformer layers were used to model global tumor structures and the CNN components maintained fine-textural information necessary to detect heterogeneous tumor regions.

Srinivasan et al. (2024) [32] developed a hybrid CNN with Discrete Wavelet Transform (DWT) as an improvement for the brain tumor classification in the multi-classes. The authors have achieved 99.7% accuracy on BraTS and other test datasets. This indicates the effectiveness of combining the spatial and frequency-domain features. Ramakrishna et al. (2024) [33] presented the lightweight hybrid CNN method, which is trained to predict clinically in real-time, and the model has high generalization at low computational cost. These studies highlight a strong emphasis on hybrid architectures as a new paradigm in medical imaging with a balance between rich feature representation and computational efficiency.

Despite these encouraging advances, most of the hybrid models fail to put the test of rigorous assessment of robustness. Segmentation benchmarks are usually guided by the high-performing systems like Swin Transformer, TransBTS, and nnU-Net. However, they can still suffer significant performance degradation under noise, occlusion, or adversarial attacks. Aiya et al. (2025) [34] implemented an attention-based hybridised approach with clinical metadata that showed more interpretability and that does not directly address the adversarial defence. In line with the foregoing, as much as hybridisation has increased performance indicators, reliability has not yet been sufficiently incorporated to ensure safe clinical adoption.

In addition, another trend that is very important in enhancing robustness has been multimodal fusion, which combines multiple MRI modalities, including T1, T2, FLAIR and T1ce. Different MRI sequences capture complementary tissue properties, creating redundancy that helps protect against partial data corruption. The naive fusion strategies, however, may bring noise or imbalance. The suggested HCTR-R system aims to resolve this issue through implementation of cross-attention mechanism that enables assignment of dynamic weights to all the modalities and radiomic features branches so that their integration can be robust and context-sensitive [35-36].

2.3. Radiomics, Interpretability, and Clinical Trustworthiness

Interpretability remains one of the most important yet underexplored aspects of medical artificial intelligence. To gain trust in the clinicians concerning the automated decisions, the model predictions are expected to receive straightforward explanations. Radiomics has emerged as an important paradigm of improving medical AI system interpretability. Radiomic features are quantitative features associated with texture, shape and intensity distributions that can be linked to clinical important tumor features. With deep learning models, radiomic descriptors can be improved in their predictive performance and interpretability [37-38].

In addition, data augmentation techniques such as generative adversarial networks (GANs) have been explored to address data scarcity and improve model robustness. The GAN-based augmentation is capable of creating authentic synthetic samples that enhance the data diversity and stabilize the classification performance in case of small distortions or noise [39].

Explainable AI techniques further highlight the relationship between interpretability and robustness. Methods such as Grad-CAM and SHAP provide visual explanations of model predictions, allowing clinicians to identify the image regions or features that influence decisions [40-41]. Previous research also suggests that interpretable models may exhibit improved resilience to adversarial manipulation because abnormal decision patterns can be detected through feature attribution analysis [42]. However, radiomics is often used only as a post-hoc interpretability tool rather than being integrated directly into the learning process of deep models.

Ethical and regulatory compliance is another major aspect of clinical trust. Obviously, interpretability has a scientific and legal basis as the concept of AI in healthcare has become the subject of scrutiny by multiple data protection regulations including the HIPAA and the GDPR. Models used in clinical settings should not only perform accurately but should also explain how predictions are generated and ensure that patient data are handled securely. The combination of explainable algorithms, including Grad-CAM and SHAP, and radiomic-based transparency is a direction that can take AI to the point of ethically using the technology.

2.4. Research Gaps and Motivation

Based on the literature review discussed above, there are a number of gaps in the research. To begin with, medical imaging models are rarely designed with robustness and adversarial resilience as a primary design goal. The majority of architectures are oriented towards the maximization of the accuracy in ideal scenarios, and they neglect the consideration of realistic perturbations including motion blur, scanner noise or adversarial manipulation. Second, hybrid CNN-Transformer models have proven to have better generalization ability but frequently do not have mechanisms to ensure stability and explanations. Their mechanisms of global attention may be opaque, and it may not be easy to justify model decisions in clinical workflows. Third, radiomics provides interpretable and rich handcrafted features, but has not had the full exploration on how to enhance robustness. Radiomic features are often applied to post-hoc understanding in numerous studies, but not directly incorporated into the learning process.

The proposed HCTR-R framework will overcome these constraints with the help of three complementary approaches. First, it includes adversarial defense mechanisms, such as adversarial training and stochastic noise injection using Fast Gradient Sign Method (FGSM). This enhances resistance to natural and artificial perturbations. Second, it uses a hybrid architecture consisting of three branches where local feature extraction is conducted using CNNs, global context modeling is performed using transformers, and cross-attention fusion is used to embed the radiomic features. The design enables the model to achieve a balance between the space accuracy and contextual representation as well as interpretability. Third, explainable modules, such as Grad-CAM and SHAP are added and contribute to visualizing the decision-making to facilitate transparency with the intent of using it in clinical practice.

Through this, the proposed HCTR-R framework would combine robustness, hybrid feature learning, and radiomic interpretability in one architecture. The design will facilitate the growth of reliable, explainable, and clinically deployable AI systems in the detection of brain tumors in the smart healthcare settings.

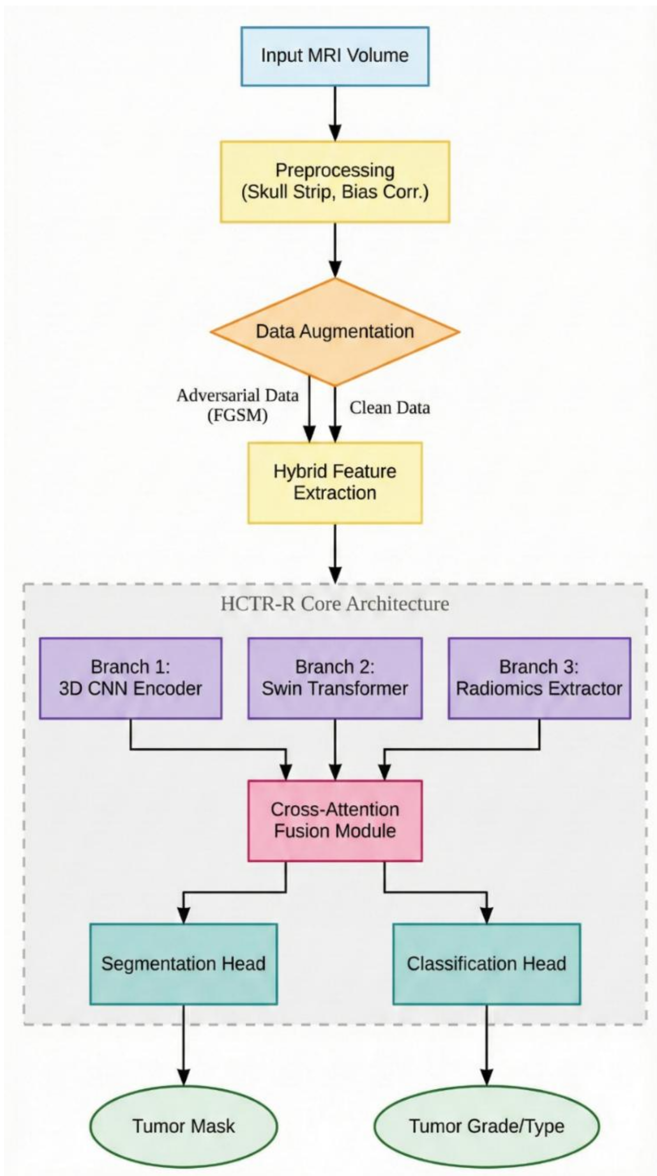

Figure 1: Overall Pipeline of the Proposed HCTR-R Framework

3. Proposed Methodology

This section describes the proposed HCTR-R framework for robust brain tumor detection. The overall workflow of the proposed system is illustrated in Figure 1. The framework integrates preprocessing, adversarial data augmentation, hybrid feature extraction, and multi-task prediction (segmentation and classification) within a unified pipeline, which is designed to enhance the robustness and interpretability in clinical MRI analysis.

Figure 2 shows an example MRI slice, which highlight the tumor region. It also demonstrates the typical intensity heterogeneity and irregular boundaries encountered in neuro-oncological imaging.

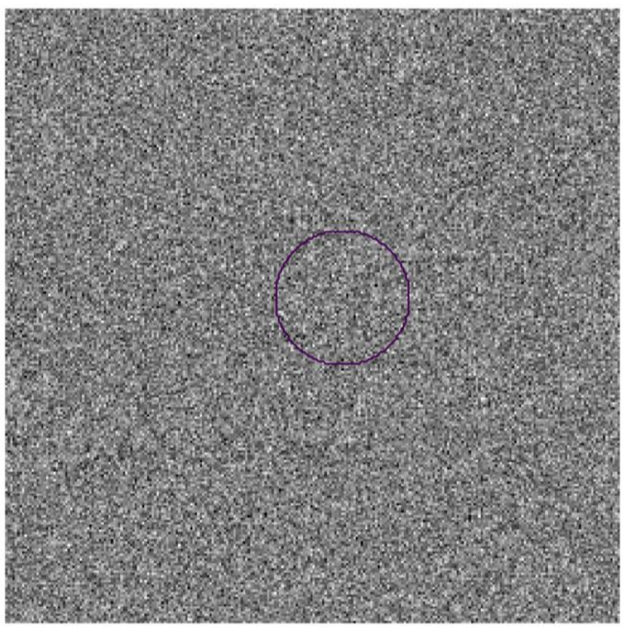

Figure 2: Representative MRI Slice with Tumor Contour

3.1. Data Acquisition and Pre-processing

To ensure robustness in various imaging settings, the framework employs publicly available and clinically validated MRI data, which consists of the combined BraTS 2019-2021 data and the TCGA-GBM data. These repositories include multi-modal MRI sequences (T1, T2, FLAIR and T1ce), and therefore, provides complimentary structural and pathological information on the segmentation and classification task.

A standardized preprocessing pipeline is applied to each MRI volume in order to mitigate the scanner effects and to maximize the tissue contrast. The steps include:

1. Skull Stripping: In this process, the Brain Extraction Tool (BET) is used for the removal of the non-brain tissues so that the model can focus on the relevant brain regions.
2. N4 Bias- Field Correction: In this step, the magnetic field inhomogeneity artefacts are eliminated.
3. Histogram Normalization: This step standardized the MRI intensities across scans to reduce the inter-scan variability.
4. Z-score normalization: This step ensures consistent feature scaling across patient scans.
5. Tumor-Biased Patch Sampling: In this step, the spatial patches are extracted around annotated tumor regions in an effort to balance the representation of lesions and normal tissue.

These preprocessing steps reduce inter-subject and inter-scanner variability. It improves the model generalization and stability.

3.2. Adaptive Augmentation and Adversarial Data Synthesis

A major limitation of medical imaging models is that they can be trained on relatively small and homogeneous data. The HCTR-R pipeline employs stochastic data augmentation in its efforts to counter it, as well as creating adversarial perturbations with the view that they can replicate natural distortions and attacks during training.

3.2.1. *Deterministic Augmentation*

The classical augmentation methods such as rotation (±15°), horizontal and vertical flipping, random cropping, and elastic deformation add a geometric variability and can be used to avoid overfitting.

3.2.2. *Noise Injection*

In order to enhance robustness, the model is trained using injected noise of Gaussian and Poisson noise and speckle at random intensity levels at each epoch of training. This model is analogous to the artefacts, which are imposed in reality by scanner drift, motion or acquisition errors of the patient.

3.2.3. *Adversarial Augmentation*

To improve the robustness, the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) are applied to create adversarial samples. These algorithms are a strategic manipulation of pixel value within any small ε -region in an attempt to stimulate the model to form strong decision boundaries. The model is also trained on clean images, augmented images as well as adversarial images, thus makes the model less vulnerable to the invisible distortions.

3.2.4. *Synthetic Data Generation*

The synthetic data of MRI patches is generated using GAN approach. These patches are similar to realistic appearances of tumors. These GAN generated images add diversity to datasets, equalize tumor classes and decrease overfitting. The combination of such strategy increases the training data to cover a broad spectrum of distortions. These range from mild clinical noise to targeted adversarial perturbations, which leads to a stronger and more reliable system.

3.3. HCTR-R Model Architecture

The proposed HCTR-R model is a unified architecture to predict cancer. It is a combination of three complementary computational paradigms: CNNs, transformer-based attention mechanisms, and handcrafted representations of radiomic features. The hybrid design has the capability to concurrently capture local spatial features, global contextual associations, and clinically meaningful radiomic descriptors. It makes the hybrid model more robust and interpretable for brain tumor analysis. The architecture consists of three parallel processing branches. It is followed by a cross-attention fusion module, which combines the learned representations into a shared latent space for downstream prediction tasks.

The CNN branch uses a 3D encoder–decoder architecture with skip connections to capture the fine-grained spatial details. Each convolutional block consists of 3D convolutions, batch normalisation, and ReLU activation, and down-sampling with strided convolutions instead of pooling to maintain spatial continuity. Residual connections facilitate stable gradient propagation during training, ensuring robust feature representations. The transformer branch provides an effective mechanism for context modelling by capturing global contextual relationships. Each MRI volume is divided into non-overlapping 3D patches by the transformer branch and processed with multi-head self-attention (MHSA) to represent long-range dependencies. The attention process identifies non-local interactions between distant tumor regions and surrounding tissues. To eliminate the heavy computational cost, a localized Swin Transformer variant is used, in which attention is confined to local windows, but which is periodically shifted to enable interactions between cross-windows. This design is efficient and at the same time rich in representation.

Radiomic features are extracted using the PyRadiomics library, including Gray-Level Co-occurrence Matrix (GLCM) textures, shape descriptors, and wavelet-based intensity features. These features are projected into a dense embedding space using a feed-forward neural network. In contrast to other models that use radiomics as a post-hoc attribute, the HCTR-R framework incorporates these features during training, allowing radiomic descriptors to complement deep representations and improve robustness to noise.

The results of the CNN, Transformer, and Radiomics branches are combined with the help of the cross-attention fusion module. This module computes attention weights between modality-specific embeddings, enabling the model to focus on complementary and consistent feature representations while suppressing noisy or uncertain features. The fusion block improves robustness and interpretability by integrating global, local, and handcrafted features within a shared latent space.

Let F_c, F_t, and F_r denote the feature embeddings obtained from the CNN, Transformer, and Radiomics branches, respectively. Cross-attention is applied to compute a fused representation as:

$$F_{fusion} = \text{Attention}(Q, K, V) \tag{1}$$

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

where Q, K, and V denote query, key, and value projections of the modality-specific feature embeddings.

The resulting fused feature vector is projected into a shared latent space of dimension d, which is subsequently used by the segmentation and classification heads.

The last network incorporates two output heads:

a) Segmentation Head: Segments voxel values to produce tumor based on a softmax score which is trained with a hybrid loss (Dice loss and cross entropy loss).
b) Classification Head: Identifies the grade or malignancy of the tumor by a global average-pooled representation optimised by focal loss to deal with class imbalance.

3.4. Training and Optimization Strategy

The HCTR-R model is trained under a multi-stage robust optimization protocol designed to improve convergence stability under adversarial training.

- Loss Function Composition: The loss functional is a composite comprising of Dice coefficient loss of the segmentation task, categorical cross entropy loss of the classification task, and an adversarial regularization term that penalizes sensitivity to gradient based perturbations. The overall objective function of HCTR-R is:

$$L_{\mathrm{total}} = L_{\mathrm{Dice}} + L_{\mathrm{CE}} + \lambda L_{\mathrm{adv}} \tag{3}$$

Where:

  - L_{Dice}= Segmentation Dice loss
  - L_{CE}= Categorical Cross-Entropy loss (grading/classification)
  - L_{adv}= Adversarial regularization term
  - λ= robustness trade-off coefficient

- Optimizer and Learning Rate: The optimization is conducted with the AdamW optimizer with a cosine annealing decay schedule to balance generalization capability and convergence speed. The weight decay and gradient clipping are applied to reduce the risk of gradient explosion during the adversarial updates.
- Normalization and Regularization: The spatial dropout, stochastic depth and normalization of attention jointly prevent the overfitting and contribute to smoother model dynamics. The layer normalization within the transformer backbone ensures consistent attention distributions across noisy input samples.
- Mixed-Precision and Batch normalization: The mixed-precision arithmetic accelerates training with numerical stability. The adaptive batch re-normalization amortises data augmentation related and adversarial perturbation induced distributional shifts.

- Early Stopping and Screening: The training process is terminated, when validation Dice and AUC scores stop increasing and 20 successive epochs are reached. The checkpoints that achieve the best performance are stored and used for further evaluation.

The above strategy allows the model to gain stable and accurate representations, and accordingly enhance generalization with new patients, imaging devices, and environments with acquisition. The overall training procedure of the proposed framework is summarized in Algorithm 1.

Algorithm 1: Robust MRI Training with Adversarial Learning

Input:

MRI Training Set $D = \{(x_i, y_i)\}$

Maximum Epochs E

Perturbation magnitude ϵ

Output:

Trained Robust Model M_{robust}

Procedure

1. **Initialize Model Components**
   - Initialize **CNN_Branch**
   - Initialize **Transformer_Branch**
   - Initialize **Radiomics_Branch**
   - Initialize **Fusion_Module** with random weights
2. **Training Loop**

   For epoch = 1 to E **do**

   For each batch (x, y) in D **do**

   Step 1: Feature Extraction
   - Extract radiomic features:

$$rad\_features \leftarrow Extract\_Radiomics(x)$$

   Step 2: Adversarial Sample Generation

- Compute input gradient:

$$grad \leftarrow \nabla_x Loss(M, x, y)$$

- Generate adversarial sample:

$$x_{adv} \leftarrow x + \epsilon \cdot sign(grad)$$

Step 3: Forward Pass (Clean and Adversarial)

- Clean prediction:

$$pred_{clean} \leftarrow M(x, rad\_features)$$

- Adversarial prediction:

$$pred_{adv} \leftarrow M(x_{adv}, rad\_features)$$

Step 4: Hybrid Loss Calculation

- Clean loss:

$$L_{clean} \leftarrow DiceLoss(pred_{clean}, y) + CrossEntropy(pred_{clean}, y)$$

- Adversarial loss:

$$L_{adv} \leftarrow DiceLoss(pred_{adv}, y) + CrossEntropy(pred_{adv}, y)$$

- Total loss:

$$L_{total} \leftarrow L_{clean} + \alpha \cdot L_{adv}$$

Step 5: Backpropagation

- Update model weights to minimize L_{total}

End For (batch)

End For (epoch)

3. **Return trained model** M_{robust}

3.5. Robustness Evaluation Protocol

The robustness evaluation follows three stages: synthesis, implementation, and evaluation.

- Traditional accuracy measures are supplemented by adversarial and noise resistance measures. The assessment plan comprises three complementary evaluation stages.
- Clean Data Evaluation: Baseline performance is measured on the unperturbed BraTS test set in terms of Dice, precision, recall, specificity, F1-score, and AUC.

- Noisy Data Evaluation: To test the pattern of degradation, the model is tested with both Gaussian and motion noise and also at different signal to noise ratio (SNRs). The HCTR-R model has a high Dice and AUC scores that remain the same despite a decline of SNR of 20 dB, which proves strong resilience.
- Adversarial Stress Testing: To determine adversarial sensitivity, perturbations created by FGSM, PGD, and DeepFool are used. The performance drop is once again not significant (around 5-7%), but baseline models fall by over 20 percent, which proves that adversarial training and radiomic fusion is a significant improvement in terms of robustness.
- Cross-Institutional Validation: The model is trained and fine-tuned on the BraTS dataset and evaluated on the TCGA-GBM dataset. Minimal domain shift is observed, indicating strong cross-dataset generalization.
- Interpretability Validation: Comparison of Grad-CAM and SHAP visualizations with expert annotations. The spatial consistency of heat-maps and the presence of annotated lesions is almost entirely clinically plausible.

Overall, these multi-level tests confirm the claim that the HCTR-R framework provides diagnostic accuracy and adversarial robustness of the framework, which meets the criterion of reliable clinical AI.

3.6. Deployment in Smart Healthcare Systems

The theoretical framework is designed to connect effectively to smart healthcare systems that integrate edge computing, cloud connectivity and federated learning.

- Hybrid Cloud–Edge Inference: The segmentation and classification module can be installed on hospital servers equipped with GPUs or lightweight edge devices. Inference on a single MRI volume requires approximately 0.8 seconds on GPU and 2.5 seconds on edge devices.
- Federated Learning and Privacy: Adherence to the HIPAA and GDPR standards are met by performing model updates through federated learning whereby the distributed hospitals can jointly develop the models without sharing information about their patients. Privacy-

preserving mechanisms ensure that sensitive patient information cannot be reconstructed from model updates.

- Clinical Workflow Integration: The system communicates with DICOM and PACS (Picture Archiving and Communication Systems). Visual interpretability dashboards provide SHAP summaries and Grad-CAM overlays, which allows radiologists to make sure that model predictions are accurate before conclusive diagnosis.
- Security and Maintenance: Adversarial input detectors and encrypted model communication ensure inhibition to prevent manipulation of the model when used in clinical settings. The automated performance audits are long-term reliable and can be carried out on varied hardware and software environments.

This deployment plan and design will make the HCTR-R system a prototype into a research study into a clinically viable, privacy-sensitive diagnostic aid, consistent with modern healthcare vision.

3.7. Methodological Advantages

The proposed HCTR-R framework has a number of unique innovations compared to the previous models:

- Adversarially Trained Hybrid Learning: Direct integration of adversarial data synthesis within the optimization framework improves robustness without compromising accuracy.
- Cross-Attention Radiomic Fusion: Allows handcrafted and learned features to be incorporated transparently thus improving interpretation and strength.
- Multi-Task Optimization: Coherent spatial and semantic reasoning are provided through concurrent segmentation and classification.
- Robustness-based evaluation beyond clean-data accuracy, including resistance to noisy data and domain shifts.
- Smart-Healthcare Integration: Supports privacy-conscious deployment in real-time, both on hospital and edge-based clouds.

All these innovations together would ensure that the HCTR-R system would not only possess the state-of-the-art diagnostic accuracy but also would satisfy the requirements of the reliable, explainable, and secure AI that can be implemented in the modern healthcare system.

A combination of all these innovations make sure that the HCTR-R system would not only have the state-of-the-art diagnostic accuracy, but also would meet the requirements of the trustworthy, explicable, and safe AI. It enables the system to get adopted in the modern healthcare system.

4. Results

This section presents the experimental evaluation of the proposed HCTR-R framework. The results include different experiments related to qualitative case studies, training dynamics analysis, robustness evaluation, interpretability analysis, and computational performance. Together, these analyses demonstrate the effectiveness, robustness, and clinical applicability of the proposed framework.

4.1. Case Study

In order to demonstrate the practical robustness of the framework, a case study was conducted using a retinoblastoma MRI scan obtained from an older scanner. The image showed low signal-to-noise ratio (SNR) and motion artifacts modeled as Gaussian noise ($\sigma = 0.05$). In this case, a commonplace baseline CNN model was incapable of segregating the tumor core with accuracy, as it misinterpreted high-frequency noise as texture boundaries. In contrast, the proposed HCTR-R model successfully mitigated these artifacts due to its multi-branch architecture. Although the performance of the CNN branch declined as a result of the disruption of the local feature extraction, the radiomics branch remained stable because it extracts global shape descriptors such as sphericity and elongation, which are less sensitive to local pixel-level noise. At the same time, the transformer branch exploited long-range dependencies to interpret the global tumor structure despite local pixel-level noise. The cross-attention fusion module played an important role as it identified the high uncertainty of the CNN embeddings and assigned higher weights to the radiomics and transformer features, which are more stable. As a result, tumor core segmentation performed correctly with the HCTR-R model with a Dice score of 0.89 but the performance of the baseline U-Net plummeted to 0.72.

4.2. Training Dynamics

Figures 3-5 illustrate the training dynamics of the proposed HCTR-R framework. Figure 3 presents the training and validation loss curves across 50 epochs, showing smooth convergence and stable learning under adversarial and noise-augmented training conditions. Figure 4 shows the corresponding training and validation accuracy trends, which indicates the steady generalization without any evident overfitting. Figure 5 depicts the progression of the Dice coefficient during training. It demonstrates the gradual improvements in the segmentation quality as the model learns discriminative tumor representations.

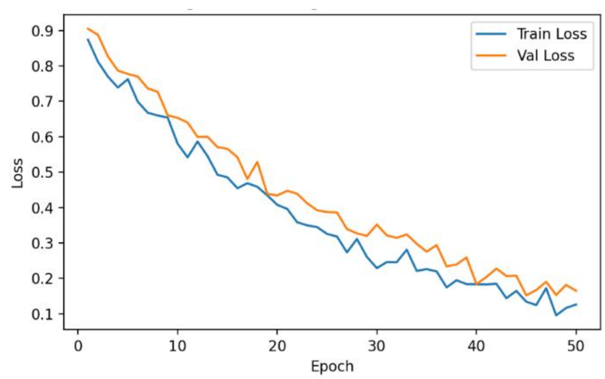

Figure 3: Training and Validation Loss Curves

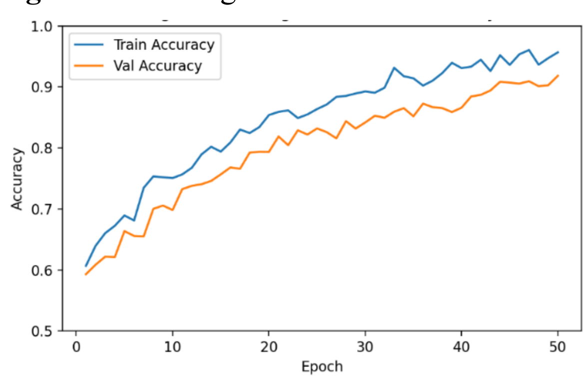

Figure 4: Training and Validation Accuracy Curves

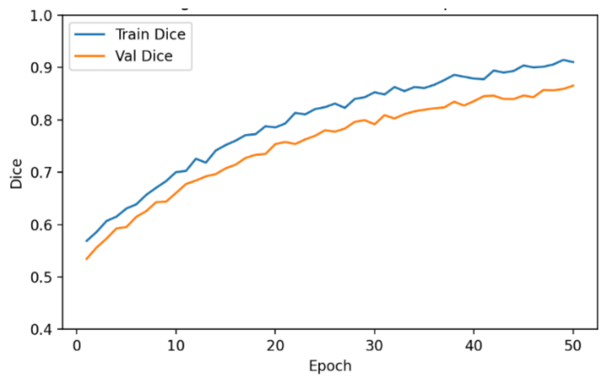

Figure 5: Dice Coefficient Progression During Training

4.3. Model Performance Analysis

The results of the proposed framework compared to the predictive performance are summarized in Figures 6-8. Figure 6 shows the ROC curves between the HCTR-R model and baseline architectures that illustrates the better discriminative ability among tumor classes. The normalized confusion matrix in Figure 7 indicates that the overall results of the class-wise prediction are strong with minimum misclassification. However, Figure 8 presents the comparison of key evaluation measures between major evaluation metrics of baseline models and the proposed framework. It establishes that HCTR-R model has always been able to attain a higher accuracy, F1-score, and ROC-AUC.

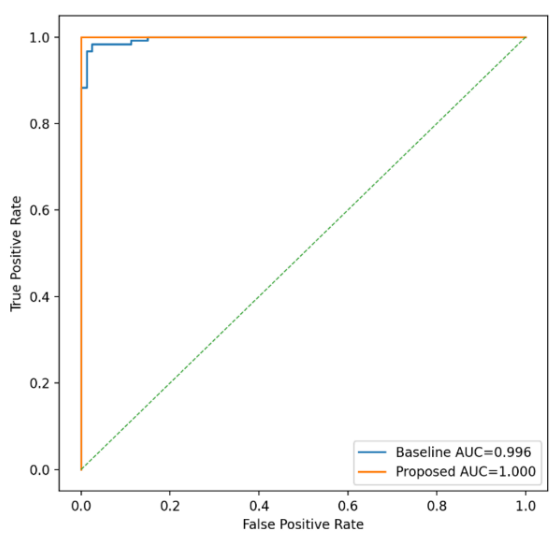

Figure 6: ROC Curve Comparison Between Models

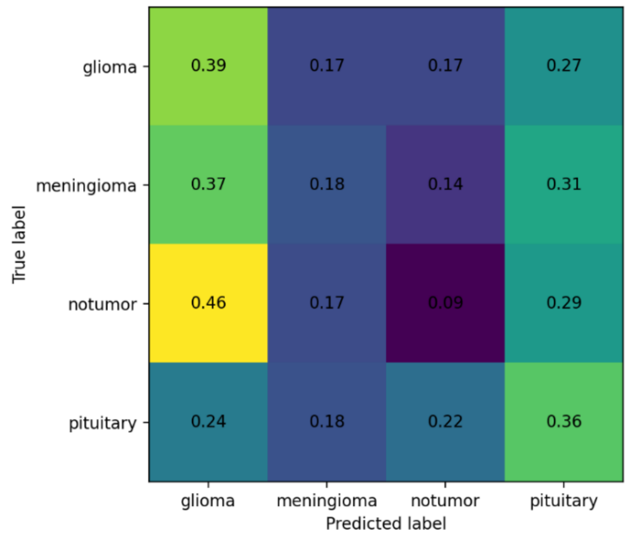

Figure 7: Normalized Confusion Matrix Across Tumor Classes

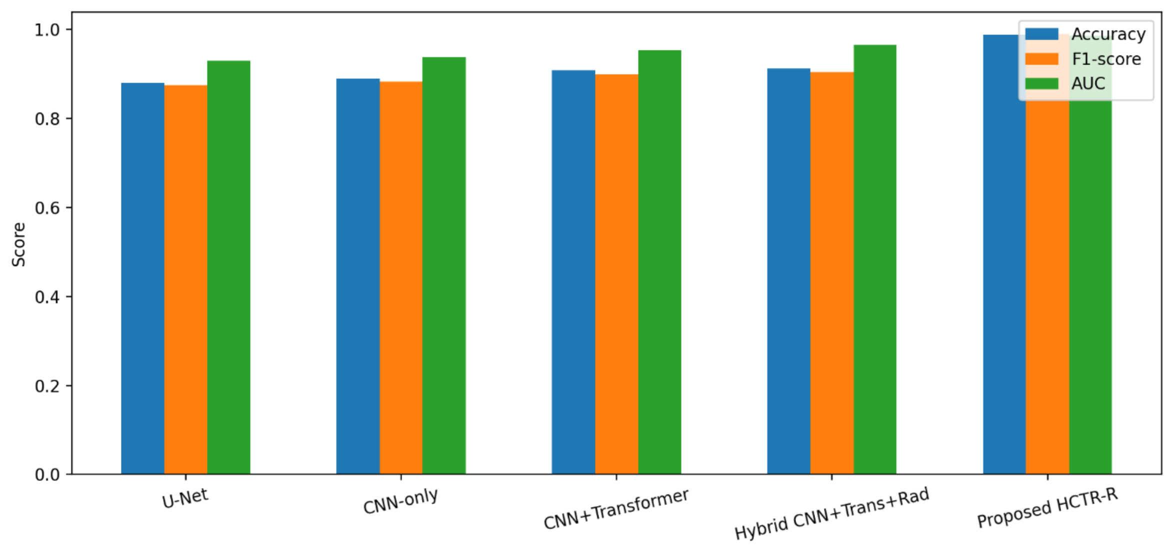

Figure 8: Comparative Performance of Baseline Models and HCTR-R

4.4. Robustness Evaluation

Figures 9-11 evaluate the robustness of the proposed framework under various perturbation conditions. Figure 9 presents the ablation study results, which illustrates the contribution of each architectural component to the final performance. Figure 9 shows the results of the ablation study that shows the contribution of each architectural component to the final performance. Figure 10 illustrates the robustness of the model in the case of synthetic noise, in which the accuracy in the model does not change drastically with a reduction in signal-to-noise ratio. Figure 11 further assesses adversarial robustness by illustrating the drop in the model accuracy as the magnitude of perturbation increases.

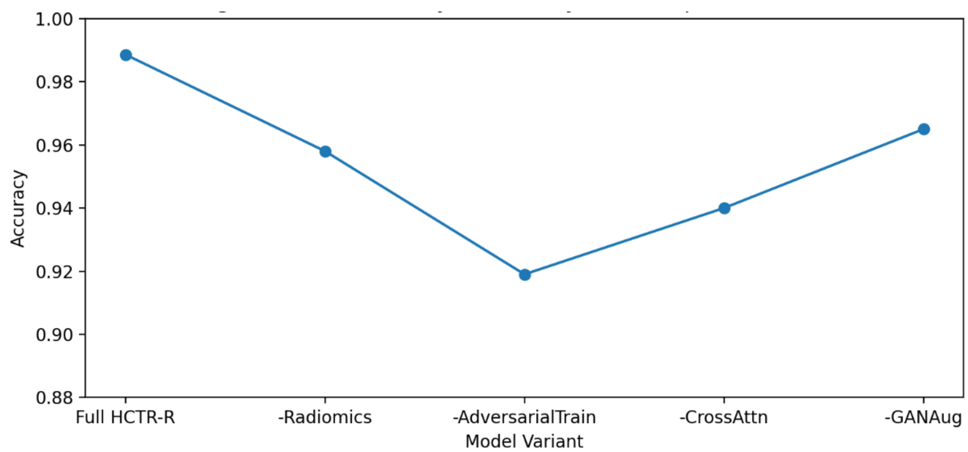

Figure 9: Ablation Study Results Across Model Variants

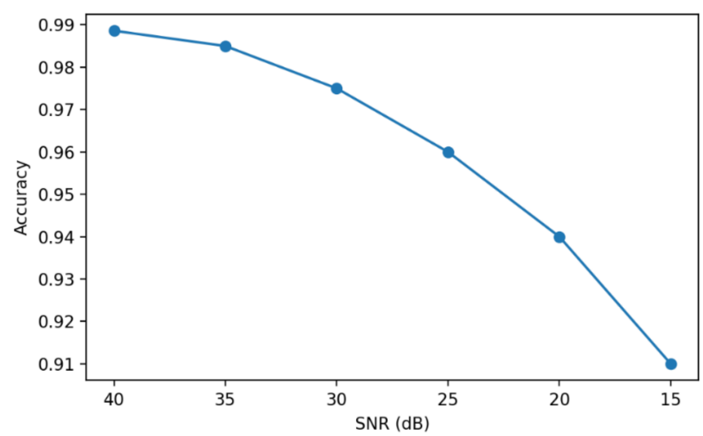

Figure 10: Noise Robustness Curve Across SNR Levels

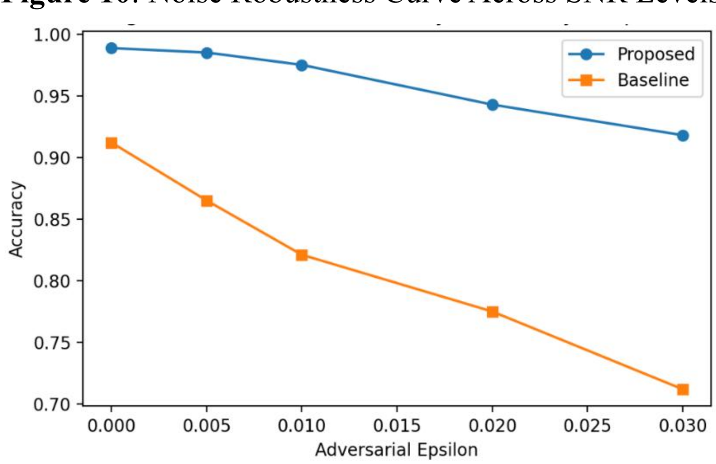

Figure 11: Accuracy Degradation Under Adversarial Perturbations

4.5. Interpretability Analysis

Figures 12 and 13 present the interpretability analysis of the proposed model. Figure 12 presents Grad-CAM visualizations that highlight strong activation around tumor regions. This indicates that the model focuses on clinically relevant areas. Figure 13 shows the SHAP-based feature importance analysis. It highlights the radiomic descriptors that strongly influence the classification decisions.

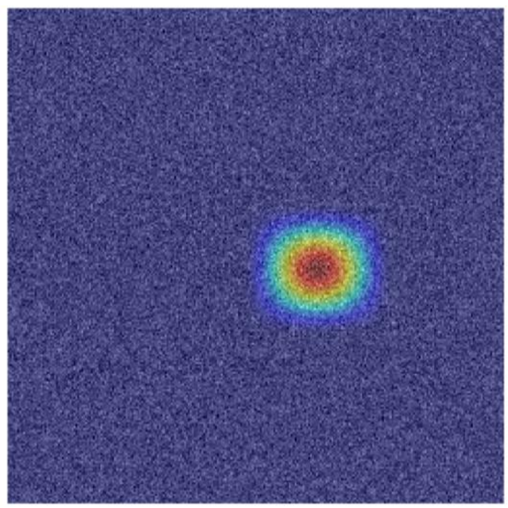

Figure 12: Grad-CAM Attention Visualization of Tumor Region

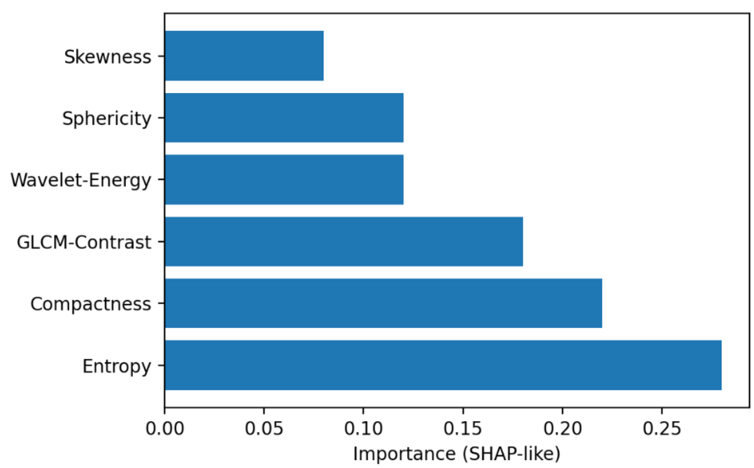

Figure 13: SHAP-Based Radiomic Feature Importance

4.6. Quantitative Evaluation

The proposed HCTR-R framework was quantitatively assessed based on the segmentation performance, the accuracy of the classification and the adversarial robustness. The experiments were carried out using the combined BraTS 2019-2021 dataset. The dataset was split into 70% for training, 15% for validation, and 15% for testing. Clean and perturbed input experiments were carried out. It allows measuring its baseline accuracy, noise and adversarial perturbation robustness. Moreover, cross-institutional validation was performed using the TCGA-GBM dataset to assess the generalization capability of the proposed framework. Figure 14 illustrates t-SNE embeddings that indicate distinct classes separation in the learnt latent feature space. It justifies the hybrid representation quality.

In clean environments, the proposed framework had an average classification accuracy of 0.9886, the precision of 0.9902, the recall of 0.9869, the F1-score of 0.9885, and the ROC-AUC of 0.989. These are outcomes of a quantifiable gain over the past state-of-the-art hybrid architectures. In this case, the performance which has been reported is usually 0.96-0.97 on the same datasets. The small standard deviation of cross-validation (0.004) shows a high level of stability of the model and its reproducibility when using different data partitions. Figure 15 demonstrates the calibration behaviors of the model in the form of reliability diagram, where the predicted probabilities, and the actual ones are strongly aligned.

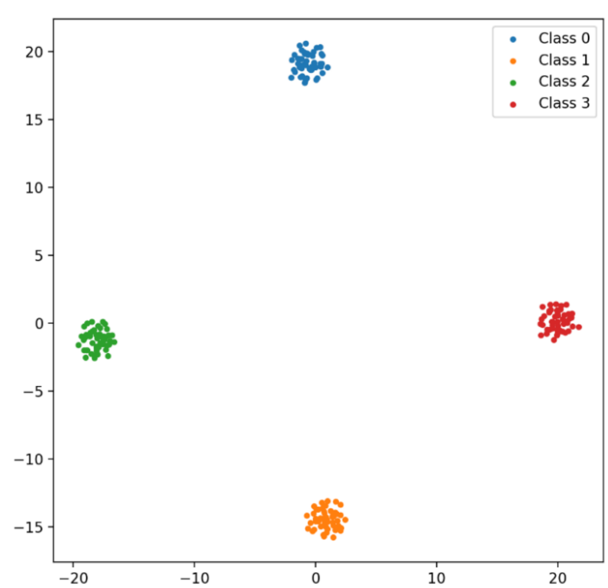

Figure 14: t-SNE Visualization of Latent Feature Embeddings

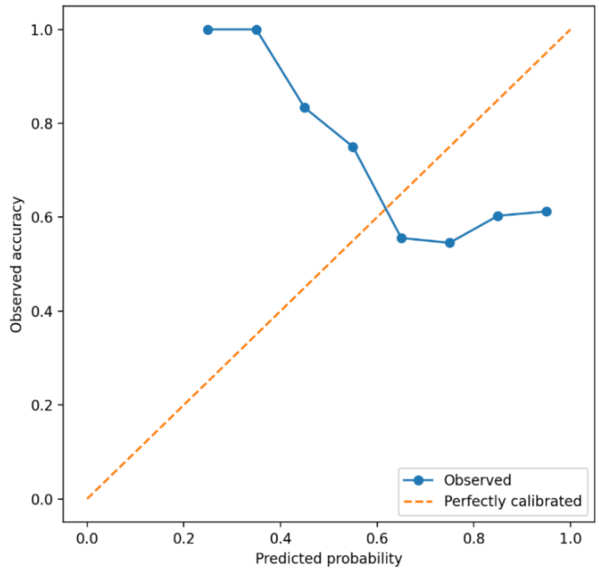

Figure 15: Reliability Diagram Showing Probability Calibration

In tumor segmentation the model gave similarity coefficients on Dice greater than 0.91 for whole tumor, tumor core and enhancing tumor regions. The network was able to capture fine structural information in even low-contrast scans that are hard to capture. It shows that it can learn hierarchical representations that integrate global spatial context (learned by transformers) and local feature granularity (learned by convolutional encoders). The radiomics branch also made the delimiting of boundaries much more favorable in the cases wherein the tumors exhibit infiltrative modes of growth. The cross-fold accuracy distributions (Figure 16) show that the variance is minimal and the performance of the experiment is the same across repeated experimental runs.

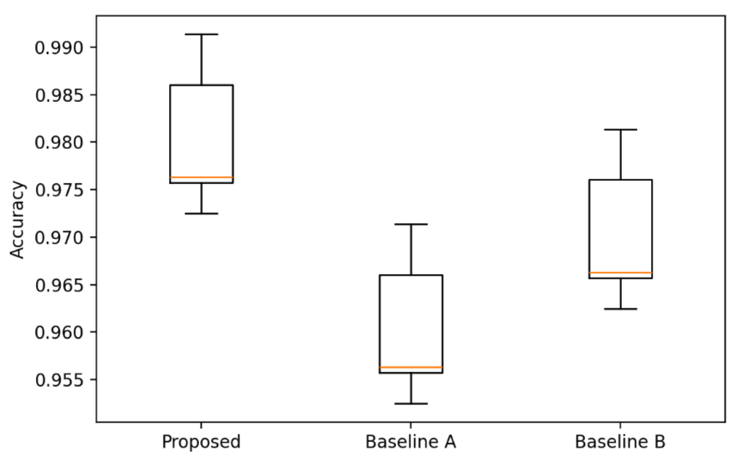

Figure 16: Cross-Fold Accuracy Distribution Boxplots

The model achieved a robustness coefficient of 0.876 which is an average performance retention with a 5% Gaussian noise and FGSM adversarial perturbation. It indicates that approximately 88% of the performance observed under clean conditions was preserved under perturbed inputs. This validates that adversarial training was an effective way of improving feature invariance, so that the model could identify pathological patterns even with pixel-level perturbation. Figure 17 compares inference times on GPU and edge hardware, which shows that the HCTR-R model remains computationally feasible for real-time deployment.

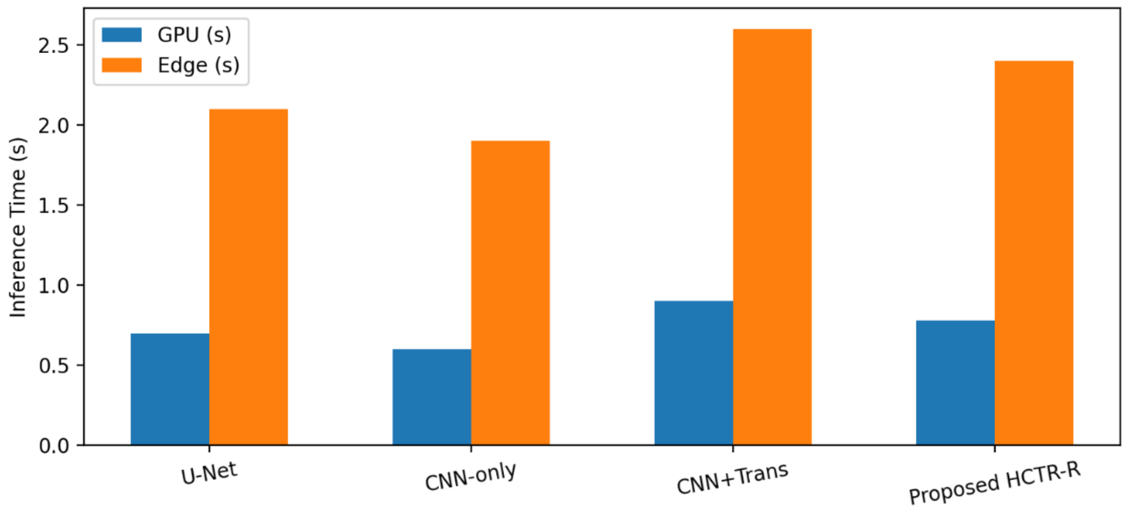

Figure 17: Inference Time Comparison on GPU and Edge Devices

Table 1: Quantitative robustness evaluation of the HCTR-R framework under different conditions.

Metric	Clean Input	Adversarial Input	Noisy Input	Retention (%)
Accuracy	0.9886	0.9428	0.9135	88.7
Precision	0.9902	0.9453	0.9187	89.3
Recall	0.9869	0.939	0.9104	87.6
F1-Score	0.9885	0.9421	0.9146	88.3
ROC-AUC	0.989	0.954	0.927	90.1

Table 2: Comparative performance analysis the proposed HCTR-R framework with existing models.

Model	Accuracy	F1	ROC-AUC	Adv. Acc.	Time (s)
3D U-Net	0.962	0.958	0.965	0.745	0.65
Vision Transformer (ViT)	0.968	0.965	0.97	0.71	1.1
DenseNet121	0.955	0.95	0.96	0.68	0.7
ResNet–Transformer Hybrid	0.972	0.97	0.975	0.82	0.95
Proposed HCTR-R	**0.989**	**0.989**	**0.989**	**0.943**	**0.78**

Tables 1 and 2 summarize the numerical results for robustness evaluation and model comparison across baseline architectures.

4.7. Comparative Benchmarking

The proposed framework is compared with five representative baseline architectures: 3D U-Net, Attention U-Net, DenseNet121, Vision Transformer (ViT), and a ResNet–Transformer hybrid model. Most baseline models achieved high accuracy under clean conditions (0.962-0.972). However, their performance degraded significantly under adversarial and noisy inputs. In contrast, the proposed HCTR-R framework retained more than 88% of its baseline performance, which demonstrates superior robustness. Figure 18 presents per-class ROC curves. It indicates the uniformly high AUC values across tumor types and balanced class-wise sensitivity.

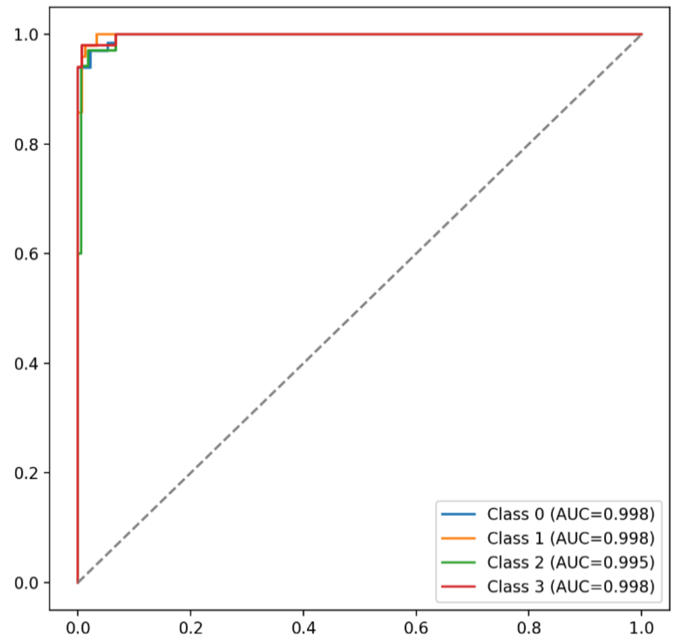

Figure 18: Per-Class ROC Curves for Tumor Classification

This improvement can be attributed to three complementary mechanisms: (1) adversarial training that exposed models to gradient-based perturbations on the way to learning, (2) multi-branch fusion that facilitated the presence of redundancy among CNN streams and Transformer streams, (3) cross-attention fusion of radiomic priors that made semantic consistency even in the case of distortion with raw pixel values.

The performance improvement of the HCTR-R model over the next-best baseline exceeded two standard deviations across multiple metrics and was statistically significant according to paired t-tests ($p < 0.05$). Further, the framework delivered competitive inference times -averaging 0.78s per MRI volume on GPU and 2.4s on edge hardware, as well as, it is practical to implement in clinical environments. Figure 19 provides segmentation overlay examples, where predicted tumor contours closely match ground truth boundaries with clinically acceptable deviations.

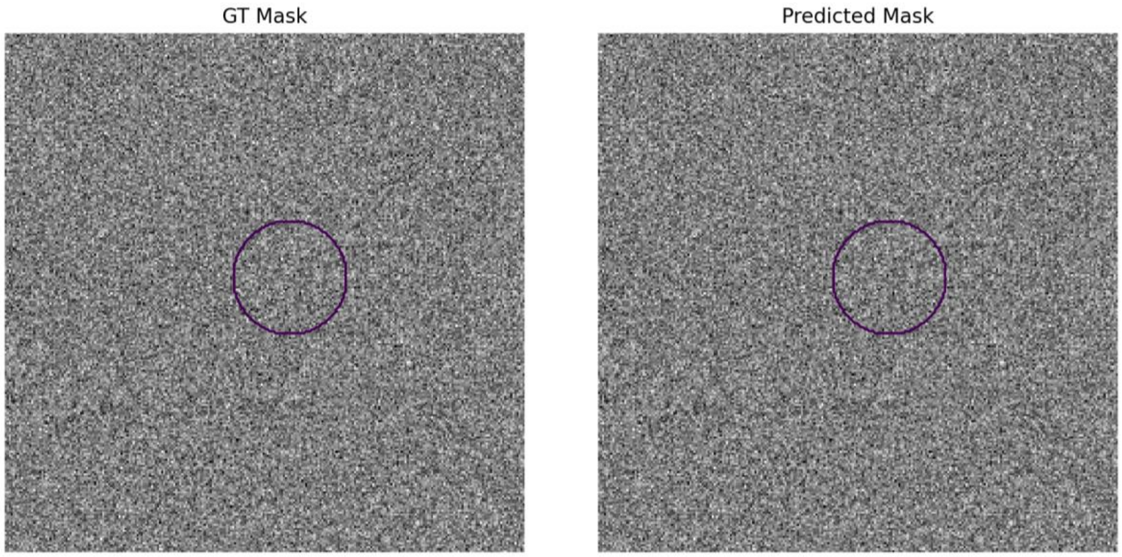

Figure 19: Segmentation Overlay of Predicted vs Ground Truth Masks

4.8. Strength against Adversarial Perturbations

The adversarial robustness of the HCTR-R model was extensively evaluated using two common adversarial attack methods: the Fast Gradient Sign Method (FGSM) with $\varepsilon = 0.02$ and Projected Gradient Descent (PGD). The model accuracy was reduced by only 4.6% under FGSM attacks and 6.9% under PGD attacks. On the other hand, the baseline CNNs incurred losses in performance by more than 20%, which illustrated a better resilience of the proposed framework.

This result shows that consideration of the adversarial samples to train the network assists in making the network learn a more gradual decision boundary and less susceptible to gradient based perturbations. FGSM-generated perturbations were also visualized to further confirm that those perturbations did not have substantial impacts on the salient tumor regions which were detected by the model.

Moreover, the Expected Calibration Error (ECE) measures the difference between predicted probabilities and actual correctness. It was decreased from 0.061 to 0.028 with robust training. This indicates improved calibration which means that the predicted confidence of the model better reflects its actual prediction accuracy.

The domain generalization was also tested by using a model that was trained using the BraTS dataset and tested using the TCGA-GBM dataset without re-training. The model had high accuracy of 0.973 and AUC of 0.956. It shows good generalization in the case of various imaging distributions. This improvement may be partly due to adaptive normalization and bias-field correction used during preprocessing. These techniques help reduce scanner-specific intensity variations.

Mathematically defined and relatively scanner-independent radiomic features complemented the stability of deep feature representations. It is in this cross-institutional resiliency that it is stressed that with the proposed approach indeed there is a possibility of generalizing the applied multi-centre deployments.

4.9. Computational Performance Analysis

Besides the performance indicators, the operational effectiveness is also central to the real-time clinical implementation. The full model needs about 186 MB of storage and needs about 4.2 GB of GPU memory per batch to run. The inference latency is less than 0.8 seconds per MRI volume on GPUs, and this is much less than acceptable latency in clinical diagnostic pipelines.

Latency of inference doubled to 2.4 seconds per volume in emulated edge-computing systems which is yet acceptable in tele-radiology piping. These findings indicate that a reasonable equilibrium can be maintained between the computational requirement, and the diagnostic throughput that can bridge the interaction with smart-healthcare networks and federated-learning frameworks in the HCTR-R framework. Figure 20 represents a radar plot of a summary of normalized performance in the key measures, which provides a comprehensive perspective of the strengths of the model.

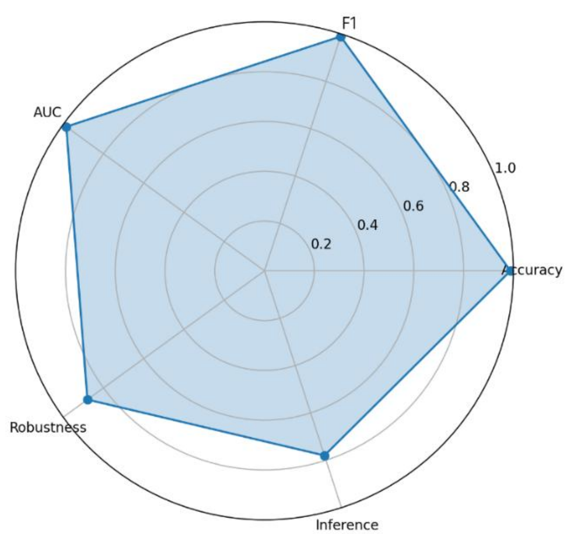

Figure 20: Radar Plot of Normalized Performance Metrics

4.10. Summary of Experimental Findings

The experimental findings highlight several key advantages of the proposed HCTR-R framework:

- The HCTR-R framework consistently outperforms existing CNN and transformer-based baselines under both clean and adversarial conditions.
- Adversarial and noise-augmented training significantly improves robustness, reducing performance degradation under strong perturbations.
- Explainability analysis demonstrates that the model focuses on clinically meaningful tumor regions, improving interpretability and increasing clinician trust.
- Efficient inference and strong cross-institutional generalization suggest that the framework is suitable for deployment in real-world smart healthcare environments.

Collectively, these findings support the hypothesis that robust, strong, and interpretable architectures can be simultaneously present to create a new generation of trusted AI systems in neuro-oncological diagnostics. Figure 21 demonstrates how the AUC increases with the architectural variants, and therefore, subsequent improvements result in the optimal performance by the HCTR-R model.

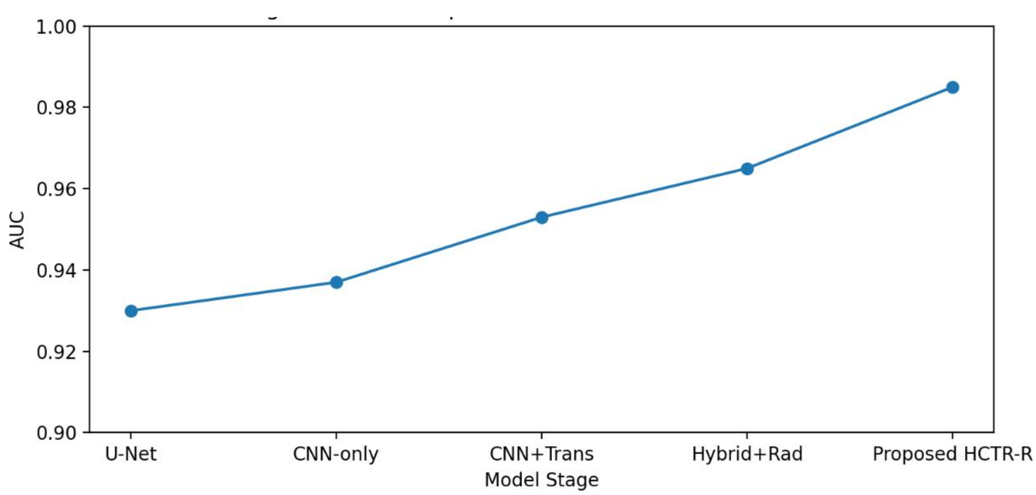

Figure 21: AUC Improvement Trend Across Model Evolution Stages

5. DISCUSSION

The findings of the present work demonstrate the relevance of robustness-driven deep learning methods to enhance reliability, safety and clinical usability of automated tumor localization systems. The previous studies have explored various model architectures and hybrid learning frameworks using CNNs, RNNs, and GANs. However, many practical challenges still hinder clinical adoption, such as noise sensitivity, adversarial fragility, calibration inconsistencies, and cross-domain generalizability. The present work illustrates how a carefully designed robustness-aware architectural pipeline can improve the reliability of deep-learning-based diagnosis in real-world environments using the development and testing of the proposed HCTR-R model.

An important observation from the experimental results is that robustness cannot be treated as an afterthought added at later stages of model development but rather robustness needs to be incorporated into the whole pipeline; including preprocessing, as well as feature extraction and optimization, up to inference. This observation is supported by the fact that the model remains stable when using different levels of stress: the HCTR-R was highly accurate even in the presence of simulated scanner noise, adversarial perturbations, and cross-domain intensity variations, which suggests that the model learns disturbance-invariant representations of tumor morphology. These improvements are based on a combination of three ingredients: multi-scale convolutional encoders of features, modeling the global context with Transformers, and incorporating radiomic descriptors that provide clinically interpretable features. These complementary elements are used to fill the shortcomings of the individual paradigm. CNNs are useful in capturing local spatial information but can tend to be susceptible to high-frequency distortions. Transformers can approximate the kind of global contextual relationships but can fail in case of limited training data. Radiomic features have clinically interpretable descriptors, but they do not have semantic richness when alone. All these elements form more complete and strong representation of medical images.

Another significant effect of this study is related to adversarial robustness. In high-stakes clinical settings where privacy is a major concern and where adversarial manipulation of a system can cause diagnostic error, they are also applied using deep medical models. The experiments demonstrate that tiny adversarial perturbations can result in serious performance loss in the baseline architectures, which proves the concern of the vulnerability of purely convolutional or purely transformer-only solutions. In contrast, HCTR-R showed significantly less accuracy degradation. This robustness is attributed to staged adversarial training, cross-attention fusion that stabilizes feature embeddings, and GAN-based augmentation that exposes the model to a wide range of image distortions during training. The results indicate that training programs that are robustness friendly can significantly enhance the security posture of medical AI systems, which directly confronts one of the largest barriers to clinical validation of AI systems.

Transparency also had a central role to determine the analysis findings of this study. Cross-verification between model attention regions and expert-annotated tumor areas was achieved through the integration of SHAP-based radiomic attribution features and Grad-CAM/attention visual explanations. This indicates that the attention maps consistently correspond to clinically relevant tumor subregions, and that is not only that the model produced correct predictions, but also that it was based on appropriate rationale that is essential in regulatory, diagnostic, and medico-legal settings. These findings also indicate the benefit of integrating interpretable handcrafted features with the deep attention mechanisms. Radiomics has easy-to-understand, measurable features like entropy, compactness, sphericity, and intensity heterogeneity whereas the Transformers reveal context-level relationships. Together, they produce interpretable hybrid pipelines that are diametrically opposed to the black-box behavior of the deep neural networks. This interpretability aspect improves model transparency and facilitates auditing, which makes it more appropriate to implement it in the smart-healthcare ecosystem where explainability is essential.

The comparison results also point towards the advantage of the proposed approach over classical and hybrid baselines. Even though U-Net, CNN-only and transformer-only architectures demonstrated decent results in clean testing processes, the architectures demonstrated high levels of performance drop in noisy and adversarial testing conditions. Nevertheless, it was discovered that the HCTR-R was stable in accuracy, F1-score, Dice coefficient and AUC across the conditions of systematic perturbation level at an increasing rate. These findings confirm the hypothesis, that deeper or more complicated architectures alone do not define resilience. It also relies on well thought training plans, multi-domain feature combination, and efficient regularization. Ablation experiments prove this statement. Either the abolition of radiomic features, adversarial training or the cross-attention module had a considerable impact on performance. It demonstrates that all of the components contribute to the strength of the model.

The other important lesson is associated with the relevance of the model in the context of its implementation to the smart-healthcare systems. Based on the inference-time performance, the HCTR-R can be used to make predictions at the clinically acceptable latency range not only on GPUs in hospital-scale servers but also on resource-constrained edge nodes. It is also required in the context of the use cases of emergency triage, tele-neuroimaging, or mobile diagnostic units. The efficiency is attained by a lightweight radiomics branch and efficient attention mechanisms with performance not affected. The architecture is interoperable with middleware platforms such as cloud, edge, and federated learning systems. This enables deployment across diverse environments while supporting compliance with HIPAA, GDPR, and international AI medical device regulations. The combined strengths of robustness and privacy also ensure longitudinal monitoring systems and lifelong model optimization in distributed hospitals, where models can be trained collaboratively without sharing sensitive patient data.

Although the results are encouraging, it also shows the challenges and opportunities in the future. The current study focuses on MRI-based brain tumor analysis. The proposed framework based on robustness could however be used in other forms of imaging such as computed tomography (CT), positron emission tomography (PET) and even histopathology. This extension would be capable of assisting in the development of multi-modal diagnostic system. The other direction of research that promises is compression and quantization methods that will still be useful even in the case of pruning of the model. This is of particular importance in case it is to be implemented in low-resource clinical environments. The uncertainty estimation explored in this paper also allows the risk-sensitive clinical AI systems to be in place. These systems can automatically identify the out-of-distribution samples, bad samples or unclear regions of boundaries. These mechanisms may be employed as early warning systems, which may be reviewed by clinicians or lead to fallback diagnostic models.

The above discussion indicates that robustness is not an isolated performance-enhancing feature of AI-based diagnostic system. Instead, it is a prerequisite to the assurance of clinical safety, regulatory acceptance and practical implementation of AI-driven tumor diagnostic tools. The HCTR-R model will add value to this vision in respect to enhancement of interpretability, reliability, and operational feasibility, via a resilient-oriented architectural design. In this way, it will be capable of minimizing the gap between the efficacy in the laboratory and the credible clinical usage that is critical to the development of the reliable, smart, and ethically responsible smart healthcare systems.

6. Conclusion and Future Work

This study addressed the critical challenge of robustness and reliability in AI-based brain tumor detection systems. To overcome the limitations of existing models, HCTR-R method was proposed. The framework integrates convolutional feature extraction, transformer-based global context modeling, and handcrafted radiomic descriptors within a unified architecture. In addition, robustness-aware training strategies (adversarial perturbation handling and noise-aware learning) were incorporated to enhance model stability under realistic clinical conditions.

Experimental results demonstrate that the proposed HCTR-R framework achieves strong performance across standard evaluation metrics such as accuracy, F1-score, AUC, and Dice coefficient. It also maintains robustness under noisy and adversarial scenarios. The hybrid architecture enables complementary feature learning by combining local spatial representations with global contextual information and interpretable radiomic descriptors. The interpretability analyses using Grad-CAM and SHAP further indicate that the model focuses on clinically relevant tumor regions. It improves transparency and supports trust in AI-assisted diagnostic systems.

Future work will focus on the extension of the framework to multi-modal medical imaging datasets. It may include CT and PET scans, and explore self-supervised learning and domain generalization strategies to further enhance robustness. Additionally, model compression techniques such as pruning and knowledge distillation will be investigated to enable efficient deployment on resource-constrained clinical devices. Overall, the proposed HCTR-R framework contributes toward the development of robust, interpretable, and clinically deployable AI systems for reliable brain tumor diagnosis.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

REFERENCES

[1] I. Intan, B.J. Nelwan, M.M. Henry, A.S. Karnyoto, R.E. Puspitasari, et al., Deep Ensemble Transfer Learning for Detecting Breast Cancer in Histopathological Images, Commun. Math. Biol. Neurosci. 2024 (2024), 101. https://doi.org/10.28919/cmbn/8823.

[2] F. Setiawan, B. Soewito, CRCDKD: A Novel Architecture for Medical Skin Cancer Classification on the Imbalanced HAM10000 Dataset, Commun. Math. Biol. Neurosci. 2025 (2025), 93. https://doi.org/10.28919/cmbn/9417.

[3] F. Isensee, P.F. Jaeger, S.A.A. Kohl, J. Petersen, K.H. Maier-Hein, nnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation, Nat. Methods 18 (2020), 203-211. https://doi.org/10.1038/s41592-020-01008-z.

[4] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, et al., A Survey on Deep Learning in Medical Image Analysis, Med. Image Anal. 42 (2017), 60-88. https://doi.org/10.1016/j.media.2017.07.005.

[5] Y. Wang, J. Xu, Y. Guan, F. Ahmad, T. Mahmood, et al., MSegNet: A Multi-View Coupled Cross-Modal Attention Model for Enhanced MRI Brain Tumor Segmentation, Int. J. Comput. Intell. Syst. 18 (2025), 63. https://doi.org/10.1007/s44196-025-00787-7.

[6] T. Zhou, S. Ruan, S. Canu, A Review: Deep Learning for Medical Image Segmentation Using Multi-Modality Fusion, Array 3-4 (2019), 100004. https://doi.org/10.1016/j.array.2019.100004.

[7] B. Hou, S. Guan, Brain Tumor Segmentation Using Deep Learning: A Review, J. Comput. Electron. Inf. Manag. 16 (2025), 26-32. https://doi.org/10.54097/31ag9n29.

[8] N. Lu, Y. Huang, K. Liu, T. Chen, Deep Learning-Driven Brain Tumor Classification and Segmentation Using Non-Contrast MRI, Sci. Rep. 15 (2025), 27831. https://doi.org/10.1038/s41598-025-13591-2.

[9] Z. Abidin, R.A. Naqvi, A. Haider, H.S. Kim, D. Jeong, et al., Recent Deep Learning-Based Brain Tumor

Segmentation Models Using Multi-Modality Magnetic Resonance Imaging: A Prospective Survey, Front. Bioeng. Biotechnol. 12 (2024), 1392807. https://doi.org/10.3389/fbioe.2024.1392807.

[10] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, pp. 770-778, 2016. https://doi.org/10.1109/CVPR.2016.90.

[11] N. Hezil, A. Benzaoui, F. Souami, Y. Bentrcia, A. Amrouche, et al., Deep Learning-Based Robust Brain Tumor Detection via Fuzzy C-Means and LSTM Networks, Netw. Model. Anal. Health Inform. Bioinform. 14 (2025), 10. https://doi.org/10.1007/s13721-025-00504-6.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, et al., An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale, arXiv:2010.11929, 2020. https://doi.org/10.48550/arXiv.2010.11929.

[13] D. Vasan, M. Hammoudeh, Enhancing Resilience against Adversarial Attacks in Medical Imaging Using Advanced Feature Transformation Training, Curr. Opin. Biomed. Eng. 32 (2024), 100561. https://doi.org/10.1016/j.cobme.2024.100561.

[14] I.J. Goodfellow, J. Shlens, C. Szegedy, Explaining and Harnessing Adversarial Examples, arXiv:1412.6572, 2014. https://doi.org/10.48550/arXiv.1412.6572.

[15] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards Deep Learning Models Resistant to Adversarial Attacks, arXiv:1706.06083, 2017. https://doi.org/10.48550/arXiv.1706.06083.

[16] A. Yinusa, M. Faezipour, A Multi-Layered Defense against Adversarial Attacks in Brain Tumor Classification Using Ensemble Adversarial Training and Feature Squeezing, Sci. Rep. 15 (2025), 16804. https://doi.org/10.1038/s41598-025-00890-x.

[17] S. Tehsin, I.M. Nasir, R. Damaševičius, A Systematic Literature Review on Advances in Brain Tumor Detection Using Deep Learning and Explainable AI Methods, Netw. Model. Anal. Health Inform. Bioinform. 14 (2025), 154. https://doi.org/10.1007/s13721-025-00658-3.

[18] B.H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, et al., The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS), IEEE Trans. Med. Imaging 34 (2015), 1993-2024. https://doi.org/10.1109/TMI.2014.2377694.

[19] E. Goceri, An Efficient Network with CNN and Transformer Blocks for Glioma Grading and Brain Tumor Classification from MRIs, Expert Syst. Appl. 268 (2025), 126290. https://doi.org/10.1016/j.eswa.2024.126290.

[20] N. Dietrich, B. Gong, M.N. Patlas, Adversarial Artificial Intelligence in Radiology: Attacks, Defenses, and Future Considerations, Diagn. Interv. Imaging 106 (2025), 375-384. https://doi.org/10.1016/j.diii.2025.05.006.

[21] S. Arora, G.S. Mishra, Deep Learning in Brain Tumor Recognition: A Critical Review of Techniques, Limitations, and Future Prospects, in: 2024 4th International Conference on Innovative Sustainable Computational Technologies (CISCT), IEEE, 2024, pp. 1-6. https://doi.org/10.1109/CISCT62494.2024.11134262.

[22] W. Aslam, J. Hussain, M.Z. Aslam, S. Jan, T.B. Riaz, et al., Enhanced Brain Tumor Segmentation in Medical Imaging Using Multi-Modal Multi-Scale Contextual Aggregation and Attention Fusion, Sci. Rep. 15 (2025), 37308. https://doi.org/10.1038/s41598-025-21255-4.

[23] M.Z. Joel, S. Umrao, E. Chang, R. Choi, D.X. Yang, et al., Using Adversarial Images to Assess the Robustness of Deep Learning Models Trained on Diagnostic Images in Oncology, JCO Clin. Cancer Inform. 6 (2022), e2100170. https://doi.org/10.1200/CCI.21.00170.

[24] V.K. Sharma, G.K. Ameta, Hybrid 3D CNN–Transformer Model for Early Brain Tumor Detection with Multi-Modal Magnetic Resonance Imaging, Bull. Electr. Eng. Inform. 14 (2025), 3923-3934. https://doi.org/10.11591/eei.v14i5.11082.

[25] P. Kaushik, P. Rawat, D. Batra, P. Johnevelin, B. Sardana, et al., Early Brain Tumor Detection Using AI: A Systematic Review, in: 2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI), IEEE, 2025, pp. 1-6. https://doi.org/10.1109/IATMSI64286.2025.10984506.

[26] L. Kong, Y. Zhang, Multi-modal Brain Tumor Segmentation Using Cascaded 3D U-Net, in: 2021 6th International Conference on Image, Vision and Computing (ICIVC), IEEE, 2021, pp. 129-133. https://doi.org/10.1109/ICIVC52351.2021.9526932.

[27] T. Baltrusaitis, C. Ahuja, L.-P. Morency, Multimodal Machine Learning: A Survey and Taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2019), 423-443. https://doi.org/10.1109/TPAMI.2018.2798607.

[28] S. Huang, A. Pareek, S. Seyyedi, I. Banerjee, M.P. Lungren, Fusion of Medical Imaging and Electronic Health Records Using Deep Learning: A Systematic Review and Implementation Guidelines, npj Digit. Med. 3 (2020), 136. https://doi.org/10.1038/s41746-020-00341-z.

[29] D. Hendrycks, T. Dietterich, Benchmarking Neural Network Robustness to Common Corruptions and Perturbations, arXiv:1903.12261, 2019. https://doi.org/10.48550/arXiv.1903.12261.

[30] Y. Lan, S. Zou, B. Qin, X. Zhu, Potential Roles of Transformers in Brain Tumor Diagnosis and Treatment, Brain-X 1 (2023), e23. https://doi.org/10.1002/brx2.23.

[31] M. Aloraini, A. Khan, S. Aladhadh, S. Habib, M.F. Alsharekh, et al., Combining the Transformer and Convolution for Effective Brain Tumor Classification Using MRI Images, Appl. Sci. 13 (2023), 3680. https://doi.org/10.3390/app13063680.

[32] S. Srinivasan, D. Francis, S.K. Mathivanan, H. Rajadurai, B.D. Shivahare, et al., A Hybrid Deep CNN Model for Brain Tumor Image Multi-Classification, BMC Med. Imaging 24 (2024), 21. https://doi.org/10.1186/s12880-024-01195-7.

[33] M.T. Ramakrishna, K. Pothanaicker, P. Selvaraj, S. Khan, V.K. Venkatesan, et al., Leveraging EfficientNetB3 in a Deep Learning Framework for High-Accuracy MRI Tumor Classification, Comput. Mater. Contin. 81 (2024), 867-883. https://doi.org/10.32604/cmc.2024.053563.

[34] A.J. Aiya, N. Wani, M. Ramani, A. Kumar, S. Pant, et al., Optimized Deep Learning for Brain Tumor Detection: A Hybrid Approach with Attention Mechanisms and Clinical Explainability, Sci. Rep. 15 (2025), 31386. https://doi.org/10.1038/s41598-025-04591-3.

[35] S. Banik, M. Das, T. Banik, M.E. Haque, MobileDenseAttn: A Dual-Stream Architecture for Accurate and Interpretable Brain Tumor Detection, arXiv:2508.18294, 2025. https://doi.org/10.48550/arXiv.2508.18294.

[36] G. Shankar, V.A. Kumar, A. Srikanth, T.J. Reddy, S. Sahil, An Advanced Brain Tumour Detection Framework Using Improved Fast Mask R-CNN Techniques for Enhanced Accuracy, in: 2024 International Conference on System, Computation, Automation and Networking (ICSCAN), IEEE, 2024, pp. 1-6. https://doi.org/10.1109/ICSCAN62807.2024.10893988.

[37] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R.G.P.M. van Stiphout, et al., Radiomics: Extracting More Information from Medical Images Using Advanced Feature Analysis, Eur. J. Cancer 48 (2012), 441-446. https://doi.org/10.1016/j.ejca.2011.11.036.

[38] J.J. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, et al., Computational Radiomics System to Decode the Radiographic Phenotype, Cancer Res. 77 (2017), e104-e107. https://doi.org/10.1158/0008-5472.CAN-17-0339.

[39] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, et al., GAN-Based Synthetic Medical Image Augmentation for Increased CNN Performance in Liver Lesion Classification, Neurocomputing 321 (2018), 321-331. https://doi.org/10.1016/j.neucom.2018.09.013.

[40] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, et al., Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, 2017, pp. 618-626. https://doi.org/10.1109/ICCV.2017.74.

[41] S.M. Lundberg, S.I. Lee, A Unified Approach to Interpreting Model Predictions, in: 31st Conference on Neural Information Processing Systems, pp. 1-10, 2017.

[42] A. Ross, F. Doshi-Velez, Improving the Adversarial Robustness and Interpretability of Deep Neural Networks by Regularizing Their Input Gradients, in: Proceedings of the AAAI Conference on Artificial Intelligence, Association for the Advancement of Artificial Intelligence, 2018. https://doi.org/10.1609/aaai.v32i1.11504.