Hybrid-CNNViT: A Deep Learning Framework for Multi-Class Brain Tumor Classification and Computer-Assisted Intervention Support

Md Fahimul Kabir Chowdhury[1] MdFahimulKabirChowdhury@my.unt.edu
Jannatul Ferdous[2] jannatulferdous0624@gmail.com
Wael Korani[1] wael.korani@unt.edu
[1] *University of North Texas*
[2] *International Islamic University Chittagong*

Editors: Under Review for MIDL 2026

Abstract

Accurate and early detection of brain tumors is critical for effective treatment planning and surgical decision-making. Manual evaluation of Magnetic Resonance Imaging (MRI) scans is time-consuming and subject to inter-observer variability. This study aims to develop a robust deep learning framework that automates tumor classification while supporting computer-assisted intervention (CAI) workflows. We propose a practical Hybrid-CNNViT architecture that integrates the local feature extraction strength of Convolutional Neural Networks (CNNs) with the global contextual understanding of the Vision Transformer (ViT). The model was trained and evaluated on two benchmark brain tumor MRI datasets using ten-fold cross-validation. Performance was assessed through accuracy, precision, recall, F1-score, and ROC metrics. The proposed framework achieved superior performance, obtaining accuracies of 99.29% (95% CI: 99.12-99.46%) and 98.46% (95% CI: 98.19-98.73%) on the two datasets, with consistently high precision, recall, and F1-scores across all tumor classes. Comparative experiments confirmed that the Hybrid-CNNViT outperformed several state-of-the-art pre-trained CNN and transformer-based models, demonstrating strong generalization and stability. The Hybrid-CNNViT framework delivers accurate and interpretable brain tumor classification, positioning it as a potential component of CAI-driven diagnostic systems. By combining efficiency, precision, and scalability, this approach advances automated neuroimaging analysis and offers meaningful support for clinical decision-making in neurosurgical contexts. Code available at: `https://github.com/fahimulkabir/Hybrid-CNNViT`

Keywords: Brain tumor, Deep learning, MRI, Vision Transformer, Machine Learning.

1. Introduction

Brain tumors represent one of the most critical neurological disorders, contributing substantially to global morbidity and mortality across all age groups (Ranjbarzadeh et al., 2024). Malignant tumors are particularly aggressive, characterized by rapid proliferation and the potential to invade adjacent or distant tissues (Satyanarayana et al., 2023). Brain cancers are typically classified as either Primary, arising within the brain itself, or Metastatic, originating from malignancies in other organs and spreading to the brain. Primary tumors may be benign or malignant and include major subtypes such as gliomas, pituitary adenomas, and meningiomas. Among these, gliomas (Pedada et al., 2023) are clinically significant and are histologically graded from I to IV, with higher-grade variants exhibiting accelerated growth, treatment resistance, and unfavorable prognoses.

MRI is a widely utilized non-invasive diagnostic modality known for its exceptional ability to generate high-resolution images of soft tissues, making it particularly effective in the detection, localization, and assessment of brain tumors (Abu Mhanna et al., 2024). Despite its diagnostic value, manual interpretation of MRI scans by radiologists is inherently time-consuming, reliant on specialized expertise, and subject to human error (Chatterjee, 2021; Li et al., 2025). To address these limitations, automated computer-aided diagnostic systems based on Artificial Intelligence (AI) and Deep Learning (DL) have emerged as powerful alternatives, offering improved accuracy and consistency in image evaluation (Athamnah et al., 2024; Mamun et al., 2022).

Hybrid CNN architectures have emerged as a powerful paradigm in medical image analysis, combining the spatial feature extraction strength of CNNs with complementary mechanisms. In (Semwal et al., 2025), a hybrid CNN-SVM model optimized through particle swarm optimization (PSO) achieved an accuracy of 84.77%. In (Islam et al., 2024), using three merged datasets, 2D CNN, CNN-LSTM, and ensemble models achieved up to 98.82% accuracy. In (Hashemzehi et al., 2020), a hybrid CNN-NADE model achieving 95% accuracy under 6-fold cross-validation. Applied Gabor filters and ResNet50 for feature extraction, followed by SVM classification used in (Ullah et al., 2023). The hybrid approach combining both feature sets achieved 95.73% accuracy. (Shanjida et al., 2024) introduces a lightweight CNN-SVM framework that achieved 96.7% accuracy, demonstrating efficient and accurate tumor detection with fewer parameters.

ViT models have gained prominence in medical imaging by leveraging self-attention mechanisms to capture long-range spatial dependencies. In (Karuppanan et al., 2025), a ViT model was implemented for automatic brain tumor classification. With 12 encoder layers yielding optimal results of 97.4%. In (Mzoughi et al., 2025), several explainable AI (XAI) methods-Grad-CAM, LIME, and SHAP were integrated with ViT and achieved a test accuracy of 91.61%, outperforming the CNN's 83.37%. In (Rahman et al., 2025), a hybrid ViT-Gated Recurrent Unit (ViT–GRU) framework achieved 92% accuracy on the binary dataset and 87.73% on the multiclass. In (Tariq et al., 2025), a combined EfficientNetV2 and ViT architecture was proposed, attaining a peak accuracy of 96%.

Automated brain tumour classification from MRI not only enhances diagnostic precision but also serves as a critical enabler for CAI systems. By delivering rapid, reliable tumour categorization and spatial delineation, such tools support neurosurgeons in preoperative planning, intraoperative navigation, and postoperative monitoring - thereby integrating seamlessly into the CAI workflow and improving patient outcomes (Vercauteren et al., 2019).

This research introduces a practical Hybrid-CNNViT framework for brain tumor classification, designed to integrate the local feature extraction strength of CNNs with the global contextual awareness of ViTs. The model aims to achieve a balanced representation that enhances classification accuracy while maintaining computational efficiency and interpretability. The major contributions of this work are summarized as follows:

- A unified deep learning framework combining CNN and ViT components is developed to capture both fine-grained local textures and long-range dependencies within MRI images.
- The transformer module is intentionally kept compact, ensuring that CNN remains the dominant contributor while the ViT branch supplements global spatial context without increasing overfitting risk.
- The proposed approach is systematically benchmarked against several leading pre-trained architectures, including EfficientNetB4, Xception, MobileNetV2, InceptionV3, ResNet50, VGG19, and DenseNet121, under identical experimental conditions.
- Experimental results are evaluated using an unbiased cross-validation (CV) strategy, and assess performance using metrics including F1-score, recall, accuracy, precision, and ROC Curve to identify the best-performing configuration and validate its effectiveness for real-world diagnostic applications.

2. Methodology

Our experimental setup was executed on a high-performance workstation equipped with dual NVIDIA H100 NVL GPUs, each offering 95,830 MB of dedicated memory and operating under CUDA version 12.2. The system's extensive computational resources facilitated efficient training, fine-tuning, and evaluation of all deep learning models. A schematic overview of the proposed architecture is illustrated in Figure 1.

2.1. Data Description and Splitting

Two publicly accessible brain tumor datasets are used in this study to assess the suggested Hybrid-CNNViT. Glioma, meningioma, pituitary tumor, and no tumor are the four categories into which 7,023 brain MRI images are divided in the first dataset, which was acquired via Kaggle (Nickparvar, 2021), showing in Fig. 2. The second dataset, which included 3,064 T1-weighted contrast-enhanced pictures taken from 233 individuals, was obtained from Figshare (Cheng, 2017). This dataset was obtained from Nanfang Hospital and General Hospital in China and includes 708 slices of meningioma, 1,426 slices of glioma, and 930 slices of pituitary tumor.

To enhance generalization and mitigate overfitting, data augmentation was applied exclusively to the training set using Keras `ImageDataGenerator`. Each image was rescaled to the range $[0, 1]$ and subjected to light. Augmentations included random horizontal flips, small in-plane rotations, mild brightness adjustments, and limited zoom operations.

2.2. Hybrid CNN-ViT Architecture

The proposed classification framework integrates a convolutional feature extractor with a lightweight ViT to balance local texture sensitivity and global spatial reasoning. The input brain MRI slices, resized to $168 \times 168 \times 3$, pass through a five-stage convolutional backbone composed of 3×3 kernels with $LeakyReLU$ activations. Channel depth increases

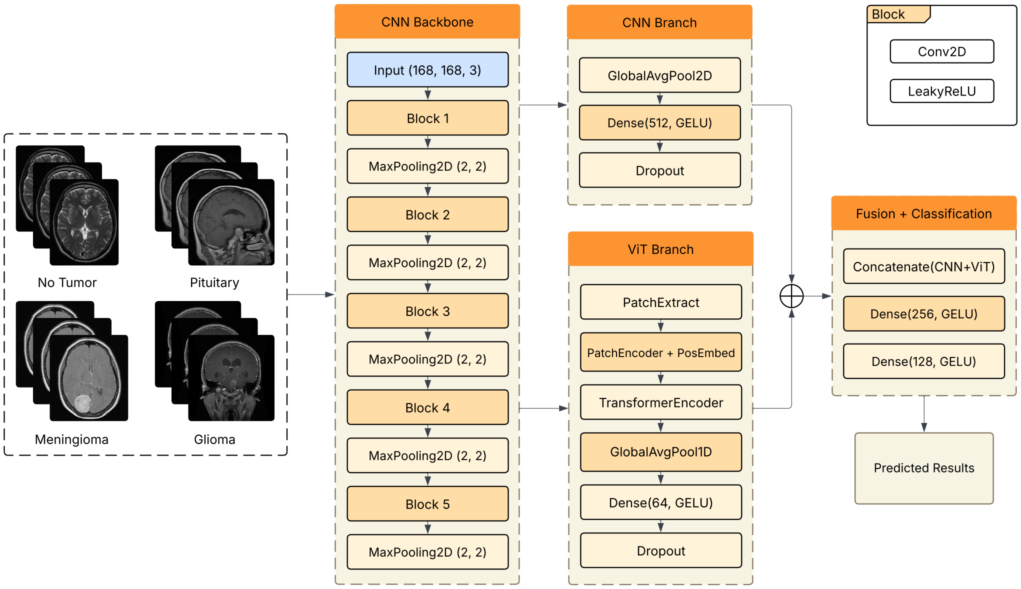

Figure 1: Architecture of the Hybrid-CNNViT: the CNN backbone serves as a feature extractor; the CNN branch captures local features, and the ViT branch captures global context.

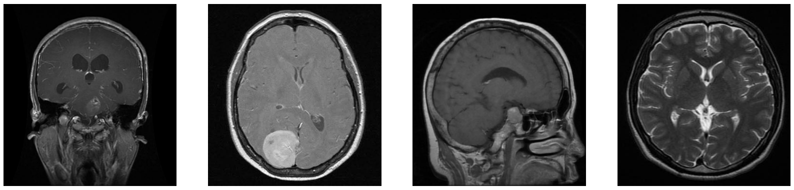
Glioma Meningioma Pituitary No tumor

Figure 2: Representative MRI images from the brain tumor datasets

progressively (32, 64, 128, 256, 256), and each block is followed by 2×2 max-pooling to condense spatial information while enriching semantic abstraction.

From the final convolutional feature map, two parallel branches are formed. The first, referred to as the CNN branch, aggregates local representations through global average pooling, followed by dense layers of 512 units with *GELU* activation and a dropout rate of 0.3 to reduce overfitting. This branch acts as the dominant feature pathway, emphasizing spatial locality and fine-grained morphological cues. The second, the ViT branch, extracts non-overlapping 5×5 patches from the CNN feature map instead of the raw image to reduce computational load. Each patch is linearly projected into a 64-dimensional embedding and enriched with learnable positional encodings to maintain spatial coherence. These patch embeddings are processed with four attention layers, four attention heads, and a multi-layer

Table 1: Training hyperparameters used for pre-trained and proposed Hybrid-CNNViT models.

Hyperparameter	**Pre-trained Models**	**Proposed Hybrid-CNNViT**
Input shape	(168, 168, 3)	(168, 168, 3)
Batch size	32	32
Dropout rate	0.5	(0.2, 0.3, 0.5)
Learning rate	0.001	0.001
Optimizer	AdamW	AdamW
Activation function	LeakyReLU	GELU
Cross-validation	10-fold	10-fold
Patch size	–	5
Projection dimension	–	64
Transformer layers	–	4
Number of heads	–	4
Total parameters	19,513,955	1,873,539
Trainable parameters	19,388,748	1,873,539

perceptron (MLP) of width 64. This configuration enables the model to capture long-range dependencies and inter-region relationships in a computationally efficient manner.

The two embeddings, one from the CNN branch and one from the ViT, are concatenated to form a fused latent representation. This combined feature vector is refined by a two-layer dense network (256 and 128 neurons, both with $GELU$ activation) to integrate local and global information before classification. The architectural design prioritizes stability and generalization by maintaining the CNN as the primary contributor while leveraging the ViT's contextual modeling to capture broader spatial dependencies.

2.3. Implementation of Pre-Trained CNN Models

To benchmark the proposed hybrid model, several state-of-the-art CNN architectures pre-trained on ImageNet were evaluated under identical experimental conditions. The comparative models included *EfficientNetB4*, *Xception*, *MobileNetV2*, *InceptionV3*, *ResNet50*, *VGG19*, and *DenseNet121*. For consistency, all models received the same input image resolution of 168×168 and were trained using the identical data preprocessing, augmentation strategy, loss function, optimizer, learning schedule, and ten-fold cross-validation framework as the proposed model. Each pre-trained network was employed as a fixed convolutional feature extractor by removing its top classification layer (`include_top=False`). A uniform classification head, consisting of a global average pooling layer, a fully connected layer of 1024 units. The list of hyperparameters used for pre-trained (EfficientNetB4) and proposed Hybrid-CNNViT models is shown in Table 1.

2.4. Cross Validation

To ensure a rigorous and unbiased performance evaluation, a ten-fold cross-validation scheme was adopted. In this approach, the dataset was divided into ten mutually exclusive subsets, or folds, where each fold served once as a testing partition while the remaining nine folds were used for training. This process was repeated iteratively across all folds so that every sample contributed to both training and validation phases. The fold-wise accuracies were then aggregated to compute the final mean performance, thereby reducing the influence of data partition randomness.

2.5. Model evaluation and validation

Among the various measures available for assessing classification performance, accuracy remains the most widely adopted and intuitive indicator. In this study, accuracy serves as the primary evaluation metric, representing the ratio of correctly classified instances to the total number of samples evaluated. It reflects the overall reliability and predictive effectiveness of the model.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

where TP denotes true positives, TN represents true negatives, FP indicates false positives, and FN corresponds to false negatives.

However, for datasets with class imbalance, a confusion matrix was employed to summarize the correct and incorrect predictions across all categories. Based on this, additional metrics were derived as follows:

- **Precision:** measures the proportion of correctly identified positive cases among all predicted positives.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

- **Recall (Sensitivity):** quantifies the proportion of actual positive cases that were correctly detected by the model.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

To achieve a balanced perspective between precision and recall, the F-score was utilized as their harmonic mean.

$$\text{F-score} = \frac{2}{3} \sum_{c=1}^{3} \frac{\text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{4}$$

Furthermore, the Receiver Operating Characteristic (ROC) curve was used to visualize the trade-off between true positive and false positive rates, while the Area Under the Curve (AUC) provided a quantitative measure of class separability. An ideal classifier yields an AUC value of 1.0, indicating perfect discrimination between categories.

3. Experiments and results

In our first experiment, we evaluate our proposed model with dataset 1. Table 2 presents a detailed performance comparison of the proposed Hybrid-CNNViT model against several state-of-the-art pre-trained architectures, including *VGG19*, *EfficientNetB4*, *MobileNetV2*, *DenseNet121*, *ResNet50*, *InceptionV3*, and *Xception*. The evaluation metrics, Precision, Recall, F1-score, Accuracy, and ROC are computed for four tumor classes: glioma (GL), meningioma (ME), no-tumor (NT), and pituitary (PI).

Among all evaluated models, the proposed Hybrid-CNNViT achieved the highest mean cross-validation accuracy of 99.29% ± 0.28% (95% CI: [99.12%, 99.46%]) with an ROC of 99.91%, consistently reaching near-perfect precision, recall, and F1-scores across all tumor categories. In contrast, conventional CNN architectures such as *VGG19* and *EfficientNetB4* showed markedly lower precision and recall in several classes, reflecting weaker feature generalization. Figure 3 presents the confusion matrices of representative pretrained models compared with our approach, while Table 3 reports the aggregated statistics. Hybrid-CNNViT achieves both the highest accuracy (0.9929) and the lowest variance (std = 0.0028), resulting in an exceptionally narrow confidence interval [0.9912, 0.9946], confirming its robustness across folds. Furthermore, Wilcoxon signed-rank tests show statistically significant improvements ($p < 0.05$) over all pretrained baselines, demonstrating that the performance gains are consistent rather than due to random variation.

Table 2: Class-wise performance metrics of pre-trained models and the proposed Hybrid-CNNViT (Dataset 1).

Models	**Precision**				**Recall**				**F1-score**				**Acc (%)**	**ROC (%)**
	GL	ME	PI	NT	GL	ME	PI	NT	GL	ME	PI	NT		
VGG19	49	99	99	32	28	10	9	96	28	18	17	48	36.55	61.91
EfficientNetB4	82	95	71	50	68	42	61	96	60	58	66	66	63.56	85.53
MobileNetV2	85	62	93	96	82	90	75	79	84	74	83	86	81.43	94.36
DenseNet121	97	95	94	98	95	92	99	98	96	93	96	98	95.97	99.66
ResNet50	96	96	96	98	95	93	99	99	96	95	97	98	96.65	99.78
InceptionV3	98	95	97	99	95	96	99	98	97	96	98	99	97.29	99.79
Xception	98	97	99	99	98	97	99	100	98	97	99	99	98.31	99.82
Hybrid-CNNViT	**100**	**99**	**100**	**99**	**99**	**99**	**100**	**99**	**99**	**99**	**100**	**99**	**99.29**	**99.91**

Table 3: Statistical summary of cross-validation results on Dataset 1. CI = 95% confidence interval. Wilcoxon test compares each model against Hybrid-CNNViT

Models	**Mean**	**Std**	**95% CI**	**Wilcoxon *p*-value**
VGG19	0.3656	0.2108	[0.2275, 0.5037]	0.0020
EfficientNetB4	0.6357	0.2309	[0.4862, 0.7852]	0.0020
MobileNetV2	0.8143	0.1059	[0.7492, 0.8794]	0.0020
DenseNet121	0.9597	0.0308	[0.9405, 0.9789]	0.0020
ResNet50	0.9665	0.0092	[0.9607, 0.9723]	0.0020
InceptionV3	0.9729	0.0143	[0.9641, 0.9817]	0.0137
Xception	0.9831	0.0084	[0.9779, 0.9883]	0.0195
Hybrid-CNNViT	**0.9929**	**0.0028**	**[0.9912, 0.9946]**	–

In our second experiment, we have evaluated our proposed architecture with dataset 2. Table 4 illustrates the comparative performance of the proposed CNN architecture against the same state-of-the-art pretrained models. As observed, traditional CNN-based models such as *ResNet50* and *EfficientNetB4* demonstrated moderate accuracy, reflecting limitations in capturing complex spatial and structural variations in MRI scans. Advanced architectures like *Xception* and *DenseNet121* exhibited notable improvements, achieving accuracies of 97.49% and 95.01%, respectively. However, the Hybrid-CNNViT architecture

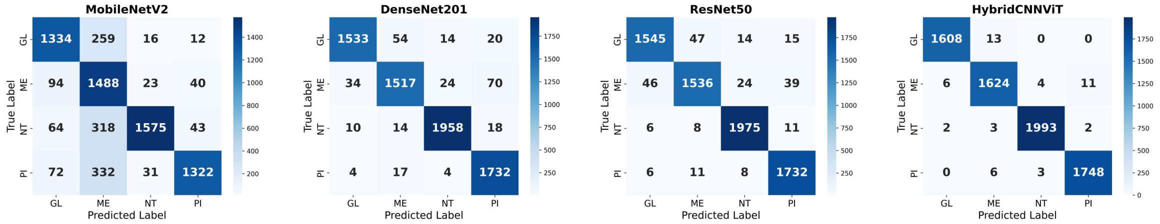

Figure 3: Confusion matrices of Pre-trained models and the proposed Hybrid-CNNViT.

surpassed all benchmarks, achieving the highest performance with an accuracy of 98.46% and ROC of 99.84%. In Fig. 4, we show the ROC-Curve of randomly picked two pre-trained models and our proposed model.

Table 4: Class-wise performance metrics of pre-trained models and the proposed Hybrid-CNNViT (Dataset 2).

Models	**Precision**			**Recall**			**F1-score**			**Acc (%)**	**ROC (%)**
	GL	ME	PI	GL	ME	PI	GL	ME	PI		
ResNet50	49	100	87	100	6	17	66	12	9	50.78	62.15
EfficientNetB4	60	80	47	70	8	67	65	14	55	54.93	70.77
MobileNetV2	59	79	92	98	41	55	74	54	38	66.32	78.69
VGG19	60	84	92	96	35	62	74	49	47	67.00	86.08
InceptionV3	94	96	95	99	85	95	96	90	96	94.75	99.25
DenseNet121	95	93	98	98	89	96	96	91	97	95.01	99.33
Xception	98	96	98	99	95	98	98	96	97	97.49	99.75
Hybrid-CNNViT	**98**	**99**	**99**	**98**	**97**	**100**	**98**	**98**	**99**	**98.46**	**99.84**

Table 5 presents aggregated statistics derived from the 10-fold evaluation. Hybrid-CNNViT achieves the highest mean accuracy (0.9846) with the lowest standard deviation (0.0043), resulting in an exceptionally narrow 95% confidence interval of [0.9819, 0.9873]. These metrics confirm that the model's performance is both accurate and stable across folds. The Wilcoxon p-values (compares each model against Hybrid-CNNViT) differ between baselines because Hybrid-CNNViT does not dominate all models in the same way. For most baselines (e.g., VGG19, ResNet50, MobileNetV2, EfficientNetB4), Hybrid-CNNViT outperforms them in every fold, yielding a perfect signed-rank pattern and the minimum possible two-tailed p-value (0.0019). However, Xception performs competitively in several folds, producing a mixture of positive and negative paired differences. This weaker sign consistency results in a slightly larger but still highly significant p-value (0.0059). Thus, the difference in p-values reflects the degree of fold-wise dominance rather than the average accuracy gap.

The outstanding performance of the proposed approach across both datasets can be attributed to its hybrid architecture, which effectively integrates the local feature extraction strength of CNNs with the global contextual learning capability of the ViT.

Table 5: Statistical summary of cross-validation performance on Dataset 2.

Models	**Mean**	**Std**	**95% CI**	**Wilcoxon p-value**
ResNet50	0.5078	0.1496	[0.4159, 0.5997]	0.0019
EfficientNetB4	0.5493	0.1852	[0.4349, 0.6637]	0.0019
MobileNetV2	0.6632	0.1715	[0.5559, 0.7705]	0.0019
VGG19	0.6702	0.2335	[0.5242, 0.8162]	0.0019
InceptionV3	0.9474	0.0231	[0.9330, 0.9618]	0.0019
DenseNet121	0.9501	0.0205	[0.9374, 0.9628]	0.0019
Xception	0.9749	0.0236	[0.9603, 0.9895]]	0.0059
Hybrid-CNNViT	**0.9846**	**0.0043**	**[0.9819, 0.9873]**	–

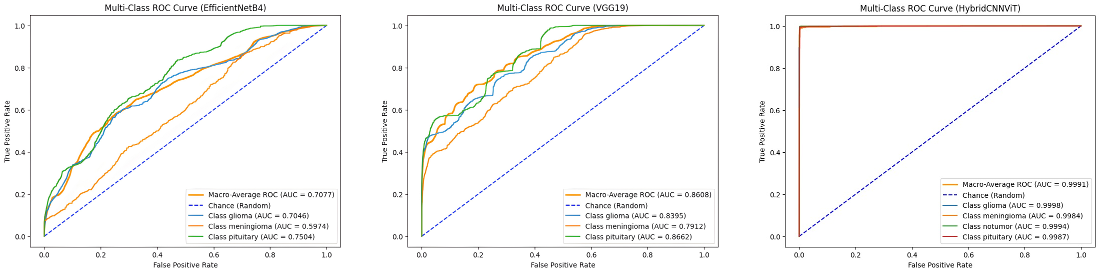

Figure 4: ROC curves of Pre-trained models and the proposed Hybrid-CNNViT model.

4. Discussion

Across both experimental datasets, the proposed Hybrid-CNNViT framework consistently achieved superior classification performance, surpassing all pre-trained state-of-the-art models. On Dataset 1, the model attained an accuracy of 99.29% and an ROC of 99.91%, while on Dataset 2, it achieved a closely aligned accuracy of 98.46% and ROC of 99.84%. These results indicate remarkable generalization capability and robustness across datasets with differing sample distributions. In contrast, traditional convolutional backbones such as ResNet50, VGG19, and EfficientNetB4 exhibited notable variability in precision and recall, particularly for underrepresented tumor classes.

The exceptional performance of the Hybrid-CNNViT can be attributed to the complementary strengths of its dual architecture. The CNN component excels in extracting localized spatial and texture-based features crucial for identifying subtle variations in tumor morphology, whereas the ViT branch captures global spatial dependencies and long-range contextual relationships. By fusing these feature representations, the model benefits from both detailed local sensitivity and comprehensive global awareness.

Compared with existing methods in Table 6, the proposed Hybrid-CNNViT shows the top performance on both datasets, and consistently outperforming prior CNN, transformer, and hybrid baselines; earlier works report strong but lower results. For example, ViT-only (Karuppanan et al., 2025) at 97.40% accuracy on Dataset 1, EfficientNetV2+ViT (Tariq et al., 2025) at 96%, and classical hybrids such as PSO-optimized CNN–SVM (Semwal et al., 2025) trailing at 84.77% on Dataset 2, the overall trend underscores that combining CNN's local feature sensitivity with a lightweight ViT for global context yields superior and more stable multi-class tumor discrimination across distinct datasets.

Table 6: Comparison with previous studies on Dataset 1 and Dataset 2.

Study	Technique	Precision	Recall	F1-Score	Acc (%)
Dataset 1 Used					
(Ullah et al., 2023), 2023	ResNet50 + Gabor + SVM	95.9	–	95.72	95.73
(Islam et al., 2024), 2024	2D CNN + LSTM	99	99	98.5	98.82
(Karuppanan et al., 2025), 2025	ViT	96.33	97.16	96.33	97.40
(Mzoughi et al., 2025), 2025	D-CNN, ViT	–	–	–	83.37, 91.61
(Tariq et al., 2025), 2025	EfficientNetV2 + ViT	96	96	96	96
Proposed	**Hybrid-CNNViT**	**99.5**	**99.25**	**99.25**	**99.29**
Dataset 2 Used					
(Hashemzehi et al., 2020), 2020	CNN-NADE	94.49	–	94.56	95
(Shanjida et al., 2024), 2024	CNN-SVM	95.31	–	–	98.31
(Semwal et al., 2025), 2025	PSO-optimized CNN-SVM	91	–	80	84.77
(Rahman et al., 2025), 2025	ViTGRU	87.11	87.33	87.43	87.73
Proposed	**Hybrid-CNNViT**	**98.33**	**98.33**	**98.33**	**98.46**

The proposed Hybrid-CNNViT framework holds strong potential for clinical adoption, offering radiologists an automated and reliable decision-support tool for accurate brain tumor identification from MRI scans within the scope of computer-assisted interventions (CAI). Its ability to capture both local and global spatial information enhances diagnostic precision and reduces reliance on manual interpretation.

5. Conclusion

This study introduced a practical Hybrid-CNNViT framework for automated brain tumor classification using MRI scans. The proposed model effectively combines the local feature extraction power of CNNs with the global contextual learning ability of the ViT, enabling comprehensive spatial feature representation. Extensive experiments conducted on two benchmark MRI datasets demonstrated the model's superior performance, achieving accuracies of 99.29% and 98.46%, which significantly outperform several state-of-the-art CNN and transformer-based approaches. Its strong discriminative capability allows for precise tumor differentiation across multiple categories, providing a dependable and interpretable framework for clinical use. In particular, the model's architecture and performance suggest strong potential for integration into computer-CAI systems, offering radiologists and neurosurgeons a reliable diagnostic aid for preoperative planning, intraoperative decision support, and postoperative monitoring. Future work will focus on expanding validation using multi-center datasets, incorporating advanced explainable AI techniques for improved interpretability, and optimizing model inference for real-time deployment within CAI-driven diagnostic environments.

References

Hamad Yahia Abu Mhanna, Ahmad Fairuz Omar, Yasmin Md Radzi, Ammar A Oglat, Hanan Fawaz Akhdar, Haytham Al Ewaidat, Abdallah Almahmoud, Laith Al Badarneh, Amer Ali Malkawi, and Ahmed Malkawi. Systematic review between resting-state fmri and task fmri in planning for brain tumour surgery. *Journal of Multidisciplinary Healthcare*, pages 2409–2424, 2024.

Sema Athamnah, Enas Abdulhay, Firas Fohely, Ammar A Oglat, and Mohammed Ibbini. Unraveling gender-specific structural brain differences in drug-resistant epilepsy using advanced deep learning techniques. *Informatics in Medicine Unlocked*, 51:101592, 2024.

Ishita Chatterjee. Artificial intelligence and patentability: review and discussions. *International Journal of Modern Research*, 1(1):15–21, 2021.

Jun Cheng. Brain tumor dataset. https://doi.org/10.6084/m9.figshare.1512427.v8, 2017. Dataset.

Raheleh Hashemzehi, Seyyed Javad Seyyed Mahdavi, Maryam Kheirabadi, and Seyed Reza Kamel. Detection of brain tumors from mri images base on deep learning using hybrid model cnn and nade. *biocybernetics and biomedical engineering*, 40(3):1225–1232, 2020.

Md Naim Islam, Md Shafiul Azam, Md Samiul Islam, Muntasir Hasan Kanchan, AHM Shahariar Parvez, and Md Monirul Islam. An improved deep learning-based hybrid model with ensemble techniques for brain tumor detection from mri image. *Informatics in Medicine Unlocked*, 47:101483, 2024.

Sakthisudhan Karuppanan et al. Optimizing brain tumor classification with vit: A metaheuristic approach. *Journal of Electrical Engineering & Technology*, pages 1–22, 2025.

Mengqi Li, Jingchao Fang, Haonan Hou, Li Yuan, Jin Guo, and Zhenlong Liu. Multi-model segmentation algorithm for rotator cuff injury based on mri images. *Bioengineering*, 12 (3):218, 2025.

Muntasir Mamun, Siam Bin Shawkat, Md Salim Ahammed, Md Milon Uddin, Md Ishtyaq Mahmud, and Asm Mohaimenul Islam. Deep learning based model for alzheimer's disease detection using brain mri images. In *2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pages 0510–0516. IEEE, 2022.

Hiba Mzoughi, Ines Njeh, Mohamed BenSlima, Nouha Farhat, and Chokri Mhiri. Vision transformers (vit) and deep convolutional neural network (d-cnn)-based models for mri brain primary tumors images multi-classification supported by explainable artificial intelligence (xai). *The Visual Computer*, 41(4):2123–2142, 2025.

Masoud Nickparvar. Brain tumor mri dataset. https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset, 2021. Dataset.

Kameswara Rao Pedada, Bhujanga Rao A., Kiran Kumar Patro, Jaya Prakash Allam, Mona M. Jamjoom, and Nagwan Abdel Samee. A novel approach for brain tumour detection using deep learning based technique. *Biomedical Signal Processing and Control*, 82:104549, 2023. ISSN 1746-8094.

Fahad Ibne Rahman, Nusrat Islam, Md Emamul Hossen, Md khairul Islam, et al. A deep learning approach based on xai and vit-gru hybrid model for brain tumor classification using mri images. In *2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)*, pages 1–6. IEEE, 2025.

Ramin Ranjbarzadeh, Payam Zarbakhsh, Annalina Caputo, Erfan Babaee Tirkolaee, and Malika Bendechache. Brain tumor segmentation based on optimized convolutional neural network and improved chimp optimization algorithm. *Computers in Biology and Medicine*, 168:107723, 2024. ISSN 0010-4825.

Gandi Satyanarayana, P. Appala Naidu, Venkata Subbaiah Desanamukula, Kadupukotla Satish kumar, and B. Chinna Rao. A mass correlation based deep learning approach using deep convolutional neural network to classify the brain tumor. *Biomedical Signal Processing and Control*, 81:104395, 2023. ISSN 1746-8094.

Tanay Semwal, Sania Jain, Agradeep Mohanta, and Ankur Jain. A hybrid cnn-svm model optimized with pso for accurate and non-invasive brain tumor classification. *Neural Computing and Applications*, pages 1–30, 2025.

Shaila Shanjida, Md Saiful Islam, and Mohammad Mohiuddin. Hybrid model-based brain tumor detection and classification using deep cnn-svm. In *2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)*, pages 1467–1472. IEEE, 2024.

Anees Tariq, Muhammad Munwar Iqbal, Muhammad Javed Iqbal, and Iftikhar Ahmad. Transforming brain tumor detection empowering multi-class classification with vision transformers and efficientnetv2. *IEEE Access*, 2025.

Sajeed Ullah, Mehran Ahmad, Shahzad Anwar, and Muhammad Irfan Khattak. An intelligent hybrid approach for brain tumor detection. *Pakistan Journal of Engineering and Technology*, 6(1):42–50, 2023.

Tom Vercauteren, Mathias Unberath, Nicolas Padoy, and Nassir Navab. Cai4cai: the rise of contextual artificial intelligence in computer-assisted interventions. *Proceedings of the IEEE*, 108(1):198–214, 2019.

Appendix A. Statistical Analysis of Cross-Validation Results

To complement the quantitative results reported in the main manuscript, we provide additional statistical measures that characterize the stability and significance of model performance across cross-validation folds. These include the mean accuracy, standard deviation, 95% confidence interval, and a Wilcoxon signed-rank test comparing the proposed model to baseline architectures.

A.1. Mean and Standard Deviation

Given fold-wise accuracies $\{a_1, a_2, \ldots, a_n\}$, the mean accuracy is computed as

$$\bar{a} = \frac{1}{n} \sum_{i=1}^{n} a_i. \tag{5}$$

The sample standard deviation, which reflects the variability of performance across folds, is defined as

$$\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (a_i - \bar{a})^2}. \tag{6}$$

A.2. 95% Confidence Interval

To estimate the range in which the true mean accuracy lies, we compute a 95% confidence interval (CI) using the standard error of the mean:

$$\text{CI}_{95} = \bar{a} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}}. \tag{7}$$

A narrower CI indicates higher consistency and reduced variance across folds.

A.3. Wilcoxon Signed-Rank Test

To assess whether the improvement of the proposed Hybrid-CNNViT model over baseline networks is statistically significant, we employ the Wilcoxon signed-rank test. This non-parametric test evaluates paired differences between two models' fold accuracies and does not assume normality.

Given paired accuracies $X = \{x_i\}$ and $Y = \{y_i\}$, the test examines the median of the differences $d_i = x_i - y_i$. The resulting p-value indicates statistical significance:

- $p < 0.05$: the proposed model significantly outperforms the baseline,
- $p \geq 0.05$: no statistically significant difference is detected.

The Wilcoxon test is well suited for cross-validation analysis because it accounts for the paired nature of fold-wise comparisons.

These statistical measures demonstrate the stability of the Hybrid-CNNViT model across folds and confirm that its improvements over baseline architectures are statistically meaningful, providing a reproducible and rigorous evaluation framework.

Appendix B. Fold wise result analysis

Across both Dataset 1 and Dataset 2, the fold-wise accuracy tables reveal clear and consistent trends in model robustness under 10-fold cross-validation shown in Table 7 and 8. In Dataset 1, while strong baselines such as DenseNet121 and InceptionV3 fluctuate between 0.900–0.986 and 0.952–0.993, Hybrid-CNNViT remains tightly bounded between 0.987-0.997, achieving the highest accuracy in every single fold. Similarly, in Dataset 2,

Table 7: Fold-wise accuracy of pre-trained models and Hybrid-CNNViT (Dataset 1).

Models	**F0**	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**F9**
VGG19	0.2731	0.2930	0.2646	0.2778	0.4288	0.9829	0.2721	0.2863	0.3034	0.2735
EfficientNetB4	0.4282	0.3272	0.8307	0.3632	0.9687	0.9516	0.8077	0.5470	0.4544	0.6781
MobileNetV2	0.8578	0.7767	0.8606	0.7507	0.9929	0.9473	0.6624	0.7308	0.8846	0.6795
DenseNet121	0.9772	0.9431	0.9004	0.9715	0.9858	0.9060	0.9929	0.9715	0.9758	0.9729
ResNet50	0.9659	0.9587	0.9502	0.9672	0.9715	0.9701	0.9744	0.9744	0.9530	0.9801
InceptionV3	0.9587	0.9915	0.9687	09929	0.9729	0.9786	0.9715	0.9886	0.9516	0.9544
Xception	0.9858	0.9701	0.9829	0.9843	0.9886	0.9801	0.9929	0.9701	0.9972	0.9786
Hybrid-CNNViT	**0.9943**	**0.9943**	**0.9872**	**0.9915**	**0.9929**	**0.9900**	**0.9915**	**0.9957**	**0.9972**	**0.9943**

Table 8: Fold-wise accuracy of pre-trained models and Hybrid-CNNViT (Dataset 2).

Models	**F0**	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**F9**
VGG19	0.4169	0.4821	0.4300	0.9414	0.4477	0.8562	0.8562	0.9477	0.4183	0.9052
ResNet50	0.4169	0.4821	0.4300	0.9479	0.4477	0.4706	0.4837	0.4673	0.4183	0.5131
MobileNetV2	0.9511	0.4821	0.4463	0.6254	0.6634	0.9020	0.7745	0.6307	0.4281	0.7288
EfficientNetB4	0.4202	0.6515	0.5570	0.5244	0.3105	0.4608	0.6503	0.9706	0.3137	0.6340
DenseNet121	0.9544	0.9544	0.9055	0.9707	0.9248	0.9542	0.9804	0.9575	0.9412	0.9575
InceptionV3	0.9609	0.9609	0.9414	0.9642	0.9314	0.9118	0.9477	0.9085	0.9641	0.9837
Xception	0.9837	0.9935	0.9805	0.9772	0.9510	0.9902	0.9967	0.9150	0.9706	0.9902
Hybrid-CNNViT	**0.9883**	**0.9883**	**0.9883**	**0.9813**	**0.9813**	**0.9883**	**0.9836**	**0.9860**	**0.9742**	**0.9859**

pretrained models show large instability-for example, VGG19 ranges from 0.416–0.947, EfficientNetB4 from 0.310–0.970, and even Xception drops to 0.915, whereas Hybrid-CNNViT stays extremely stable between 0.974–0.988 across all folds. This narrow performance band, combined with hybrid's consistently superior values, demonstrates numerically that Hybrid-CNNViT is the only model that performs well everywhere, with no failure folds and the smallest variance across datasets, establishing it as the most reliable and generalizable architecture among all evaluated models.

Appendix C. Architectural Rationale, Statistical Rigor, and Study Limitations

C.1. Architectural Motivation and Conceptual Distinction from Existing Hybrids

The Hybrid-CNNViT architecture is motivated by practical considerations in biomedical imaging rather than architectural novelty. Traditional hybrids typically apply transformer layers directly to raw image patches, which remain highly sensitive to noise and imaging variability. This behavior is reflected in the pretrained baselines: for example, in Dataset 2, VGG19 fluctuates from 0.4169 to 0.9477 and EfficientNetB4 from 0.3105 to 0.9706 across folds. By contrast, Hybrid-CNNViT extracts patches from intermediate CNN feature maps, producing semantically enriched and more stable tokens; consequently, its accuracies remain tightly bounded between 0.9742 and 0.9883. These results indicate that transforming high-level CNN features into tokens yields more discriminative and robust representations than using raw pixel patches.

The lightweight ViT branch further aligns with the dataset scale and numerical behavior. Because the CNN encoder already captures strong local structure, only shallow global modeling is needed. This is evident in performance stability: in Dataset 1, Hybrid-

CNNViT achieves a mean accuracy of 0.9929 with a standard deviation of 0.0028, whereas stronger baselines such as DenseNet121 and InceptionV3 exhibit higher variability (0.0308 and 0.0143, respectively). These results show that a light transformer is sufficient when the CNN backbone has already distilled the key spatial information.

Finally, the simplicity of the fusion mechanism, direct concatenation followed by projection also contributes to stability. Unlike complex cross-attention fusion modules, which often overfit small datasets, this streamlined integration consistently outperforms all baselines. In Dataset 2, Hybrid-CNNViT reaches a mean accuracy of 0.9846, surpassing VGG19 (0.6702), ResNet50 (0.5078), MobileNetV2 (0.6632), and even high-performing models such as DenseNet121 (0.9501) and InceptionV3 (0.9474). These results collectively demonstrate that combining CNN-based local representations with lightweight global reasoning provides a practical and empirically superior hybridization strategy for biomedical imaging.

C.2. Statistical Indicators of Reliability and Variability

The reliability of Hybrid-CNNViT is further supported by statistical analyses. Across both datasets, it consistently displays much lower variability than the pretrained baselines. In Dataset 2, its standard deviation (0.0043) is substantially smaller than that of MobileNetV2 (0.1715), EfficientNetB4 (0.1852), or VGG19 (0.2335), confirming stability across folds. Similarly, its 95% confidence intervals remain narrow, for example, [0.9912, 0.9946] in Dataset 1, whereas those of weaker baselines such as VGG19 ([0.2275, 0.5037]) or EfficientNetB4 ([0.4862, 0.7852]) are much wider, reflecting greater uncertainty. Wilcoxon signed-rank tests also provide strong evidence of superiority: for Dataset 2, the hybrid model achieves a p-value of 0.001953 against all baselines except Xception, meaning it outperforms them in all 10 folds. Even for Xception, the p-value of 0.0059 remains highly significant. These statistical patterns demonstrate that the hybrid model's performance gains are both consistent and unlikely to be attributable to random variation.

C.3. Study Limitations

Despite these strengths, several limitations must be acknowledged. Both datasets are relatively small, with limited subject diversity, which inherently constrains the generalizability of the results. Although cross-validation and statistical testing mitigate some concerns, larger and more varied cohorts, including multi-center clinical MRI scans would be needed to validate the model's robustness across acquisition conditions and patient populations. Additionally, the study relies on 2D slice-level representations rather than full volumetric MRI sequences, which reduces spatial context and may overlook multi-slice dependencies. Volumetric extensions of the hybrid model may further enhance performance. Finally, although the architecture demonstrates strong numerical stability, the lack of interpretability analyses such as Grad-CAM or transformer attention visualizations limits the clinical transparency of the model's decision process.

In summary, the numerical evidence from both datasets strongly supports the practical value of the Hybrid-CNNViT architecture. Its stability across folds, low variance, narrow confidence intervals, and highly significant p-values collectively validate the effectiveness of combining CNN-based feature patches with a lightweight transformer branch. While limitations related to dataset size, imaging diversity, and model interpretability remain, the results indicate a strong foundation for future methodological and clinical extensions.