International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(15s) 2026 pp. 1115-1130
https://doi.org/10.70917/ijcisim-2026-4506
Article

Hybrid CNN-Transformer Based Brain Tumor Detection and Classification Using MRI: A Novel Framework with Multi-Level Attention and Regional Explainability

Phanideep Karnati[1], Sukanya Roy[2], Dundi Urlamma[3], M Siva Prasad[4], Shaik Sharmila[5], Keerthika T[6], Bijoy Kumar Mandal[7]

[1]CEO and GenAi Solution Architecture, vHave.ai, IIT Patna, Bihta, Patna, Bihar, phanimsba@gmail.com.

[2]IEM Kolkata, School of University of Engineering and Management Kolkata, Newtown Sector, piu.sep.1993@gmail.com.

[3]ACSE, Vignan's Foundation for Science, Technology and Research, Vadlamudi, Thenali, Guntur, urlammadundi31@gmail.com

[4]Dept.of CSE-AI & ML, Geethanjali College of Engineering and Technology, Keesara(M),Medchal(D), Telangana.
mshivaprasad.cse@gcet.edu.in

[5]CSE(AIML), Sridevi Women's Engineering College, Gopan Pally, Vattinagulapally, Hyderabad, shaiksharmila.cse@gmail.com

[6], Amrita School of Artificial Intelligence, Amrita Vishwa Vidyapeetham, Coimbatore, t_keerthika@cb.amrita.edu

[7]Dept. of CSE, NSHM Knowledge Campus, Durgapur, writetobijoy@gmail.com

Abstract: - Automated and accurate classification of brain tumors from MRI (magnetic resonance imaging) for clinical applications is essential. However, it still poses a challenge owing to multiple factors. This includes the high intra-class heterogeneity, vague tumor boundaries, and lack of transparency in diagnostic reasoning. In this paper, we propose a novel hybrid convolutional neural networks and transformer architecture, HybCT-Net, augmented with a multi-level attention module and a regional explainability pipeline for brain tumor detection and classification. The framework employs local feature extraction of deep CNN encoders and the long-range dependency modeling capacity of lightweight vision transformers in conjunction. The MLAM incorporates attention mechanisms such as channel-wise squeeze-and-excitation gating, spatial convolutional block attention, and patch-based multi-head self-attention to enhance salient features at multiple semantic scales. A Hybrid Feature Fusion (HFF) is a block for adaptive fusion of the CNN feature maps and the transformer tokens that bridges the semantic gap between the convolution-based and attention-based features [24]. Moreover, the REP combines Gradient-weighted Class Activation Mapping with transformer attention rollout maps for producing regional heatmaps at pixel-wise spatial detail, enhancing clinical trust. The efficacy of the proposed model is established through extensive experiments on a multi-class brain MRI dataset with glioma, meningioma, pituitary tumor and no tumor. HybCT-Net attains a classification accuracy of 98.78%, with a macro-F1 score of 98.70% and an AUC of 0.9943. Comparative experiments demonstrate superior performance than contemporary CNN, transformer and hybrid baselines. The contributions of various architectural components are validated using ablation studies and computational analysis shows deployment feasibility. Qualitative visualizations suggest that the regional explainability maps correlate well with boundaries of the tumor as annotated by radiologists. Thus, the framework has the potential for use in clinical decision support in the real world.
Keywords: Brain tumor classification; Deep learning; Hybrid architecture; Vision transformer; Multi-level attention; Explainable AI; MRI; Medical image analysis.

1. Introduction

Brain tumors are among the deadliest intracranial lesions, contributing significantly to morbidity and mortality worldwide [1]. The rising incidence of central nervous system (CNS) cancers, as reported by the Central Brain Tumor Registry and large neuro-oncological initiatives such as BraTS and others highlights the need for early, accurate, and reproducible diagnostic methods [2]. Magnetic Resonance Imaging (MRI) has been considered the gold standard of neuro-oncology imaging among various noninvasive modalities due to its superior soft-tissue contrast, multi-planar capability and lack of ionizing radiation [3]. Manual interpretation of MRI scans is cumbersome, subjective, and variable across experts. This is especially true for differentiating tumor subtypes such as gliomas, meningiomas, and pituitary adenomas, which often have similar visual appearances [3].

Automated medical image processing has progressed rapidly with recent deep learning advancements. Convolutional Neural Networks (CNNs), such as ResNet [20], DenseNet [21], and EfficientNet [22], have shown remarkable ability to extract hierarchical local patterns, texture, and shape features from radiological scans. Despite their advantages, pure CNN architectures are constrained by their inductive biases, namely locality and translation equivariance, thereby limiting their ability to model long-range spatial dependencies and global contextual relationships, which is necessary for differentiating tumors with diffuse infiltrative margins and subtle intensity variations [4], [5].

Vision Transformers (ViTs) [5] have also been applied to medical image classification by treating input images as sequences of patches and applying self-attention on those patches. In neuroimaging, ViTs have been investigated for brain tumor analysis. However, ViTs tend to underperform with limited training data, a common challenge in medical imaging. This is due to their weak inductive biases and large parameter counts. Additionally, standard ViTs operate at a single resolution, potentially missing diagnostically important fine-grained local information that is diagnostically useful and exhibits quadratic complexity with respect to image resolution [19].

Recent hybrid CNN-Transformer models aim to bridge these complementary strengths. TransUNet [7] combines a CNN encoder with a Transformer to segment medical images while MobileViT [8] and Pyramid Vision Transformer (PVT) [9] address mobile and dense prediction tasks, respectively. Nonetheless, existing methods often treat the CNN and transformer branches as independent feature extractors fused only at the penultimate layer via naive concatenation or late averaging [7]. This prevents dynamic cross-pollination of local and global features during intermediate representation learning. Additionally, despite clinical requirements for transparency, most hybrid models operate as obscure black boxes. Current explanation techniques, such as GradCAM [10], Layer-wise Relevance Propagation (LRP) [11], Grad-CAM++ [12] yield coarse or noisy attribution maps for hybrid models, while transformer attention rollout [13] lacks precise boundary localization. These limitations prevent pixel-precise localization necessary for clinical acceptance.

To overcome these shortcomings, we present HybCT-Net, a Hybrid Convolutional Neural Network and Transformer Network that will classify brain tumor MRI scans into various categories. This work has four major contributions.

(1) **Multi-Level Attention Module (MLAM):** we introduce a novel attention mechanism operating simultaneously along channel, spatial, and token-wise directions [4], [5], [6]. The channel squeeze-and-excitation gating of the MLAM helps to amplify semantically rich feature maps, use spatial convolutional block attention in reducing background clutter, and adds patch-based multi-head self-attention for global context coherence. This multi-layered recalibration greatly improves the discriminative power of the fused representation.

(2) **Hybrid Feature Fusion (HFF) Block:** The HFF block avoids trivial feature concatenation by employing cross-attention gating which dynamically weighs the contribution of local descriptors from a CNN and global tokens from a transformer. The HFF block adaptively forges connections between different types of representations, allowing it to maintain local anatomy while collecting holistic spatial information.

(3) **Regional Explainability Pipeline (REP):** We develop a clinically oriented explainability framework, the Regional Explainability Pipeline (REP) which combines CNN-based Grad-CAM++ with transformer attention rollout. The REP algorithm uses convolutional gradient attribution and multi-head attention rollout to identify consensus regions and create saliency maps. These maps facilitate radiologist feedback for better scan interpretation. It highlights tumor localization along with edema and necrotic regions.

(4) **Comprehensive Empirical Validation:** We conduct extensive experiments on a publicly available multi-class brain MRI dataset with 3,264 images [3]. The proposed HybCT-Net exhibits superior classification performance

compared to eight competitive baselines comprising ResNet [20], DenseNet [21], EfficientNet [22], ViT [6], MobileViT [8], and PVT [9]. Through extensive ablation studies, five-fold cross-validation, and complexity analyses, we show that the framework is robust, generalizable, and clinically deployable.

The remaining sections of this paper are organized as follows. Section 2 reviews the relevant literature. Section 3 outlines HybCT-Net's dataset, preprocessing and architectural details. The experiment setup and quantitative outcomes are covered in Section 4. The qualitative explainability visualizations are presented in 5. Section 6 highlights the limitations and future directions while Section 7 concludes the paper.

2. Related Work

The classical CNN-based models, pure transformer models, and more recent hybrid models represent the three main paths deep learning for brain tumor analysis has taken [18]. Early notable works primarily used CNNs for glioma segmentation and tumor classification. Rehman et al. (2020) employed pretrained CNN architectures, including AlexNet, VGGNet, and GoogleNet to evaluate classification accuracy on a brain MRI dataset [25]. Their experiments demonstrated the potential of transfer learning but also highlighted the overfitting problem with small medical datasets. Deepak and Ameer [2] discussed a modified deep CNN with softmax regression on the three-class tumor classification issue, getting accurate results but not robust to class imbalance [23].

To improve the learning of discriminative features, attention mechanisms were then incorporated into modern pipelines [17]. The Squeeze-and-Excitation (SE) networks, proposed by Hu et al. [4], recalibrate channel-wise feature responses. On the other hand, the Convolutional Block Attention Module (CBAM) proposed by Woo et al. [5] extend SE to spatial data. The localized receptive field of SE-ResNet and CBAM-enhanced DenseNets, which show promise in neuroimaging, fails to capture the topological relationships between the tumor sub-regions and the surrounding anatomical structures.

Dosovitskiy et al. [6] introduced Vision Transformers which opened up newer avenues for medical image analysis. ViTs separate images into patches of fixed size, linear embed them, and send the sequence to a stack of Transformer encoders. In the domains of histopathology and radiology, ViTs exemplify superior generalization when pretrained on large datasets. Gheflati and Rivaz [6] investigated ViT models for brain tumor classification and found that while global context modeling improved, the introduction of more smoothing affected the fine spatial details. Also, high costs of standard ViTs occur because of quadratic complexity of self-attention for high-res MRI volumes [19].

Recently, hybrid architectures are persistently receiving attention to reconcile local inductive biases with global contextual reasoning. TransUnet was developed by Chen et al. [7], which employ a CNN encoder together with a Transformer. The authors Mehta and Rastegari [8] proposed MobileViT, which consists of inverted residual blocks combined with transformer layers for use on mobile devices. Wang et al. [9] created a pyramid vision transformer (PVT) to apply multi-scale patch embedding to dense prediction. Most hybrid approaches still treat the CNN and transformer branches as independent streams merged only at the final classification step. This layered framework inhibits the active cross-modal enhancement of features.

Explanation is yet another important dimension. Post-hoc explanation methods have become popular in the recent past as a means to explain neural network decision such as Grad-CAM [10], LRP [11], and Grad-CAM++ [12]. Transformer models can employ attention rollout and attention flow methods for token-to-token interpretability. Nonetheless, explanations obtained from one branch are lacking in consistency or is spatially misaligned with the other in hybrid CNN-Transformer models [26]. As far as we know, there is no existing framework that systematically synthesizes CNN gradient attributions with transformer attention rollouts to achieve consensus-based, regionally localized explanatory maps for brain tumor classification. Our work fills this gap with the proposed REP to provide a clinically interpretable perspective of the hybrid model.

3. Methodology

3.1 Dataset Description and Preprocessing

This research utilizes a publicly available Brain Tumor MRI Dataset [3], which includes 3264 T1-weighted contrast-Enhanced MRI images from 233 patients. The dataset is categorized into four classes: glioma (926 images), meningioma (937 images), pituitary tumor (901 images), and no tumor (500 images). The images have native resolutions ranging from 512×512 to 640×640 pixels. They exhibit varying intensity distributions due to different scanner manufacturers and acquisition protocols.

The MRI images were rigorously preprocessed to ensure uniformity before training the proposed HybCT-Net model. First, all MRI scans were converted to grayscale to reduce computational requirements while preserving essential anatomical details. The images were then resized to 224 × 224 pixels using bicubic interpolation to provide the same input size to the convolutional and transformer branches. This uniformity facilitates effective feature extraction and stable training of the model across images acquired from different scanners and imaging protocols. Each MRI image underwent z-score normalization to reduce image intensity variations across subjects. The image obtained after normalization I_{norm} was computed as:

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

Here, the original pixel intensity is represented by "I", where μ is the mean and σ is the standard deviation within the brain segment. Skull-stripping was purposely not employed, as the presented framework is formulated to run on the clinical raw MRI scans in a bid to avoid further manual preprocessing and increase applicability in a real clinical setting.

Table 1: Class-wise distribution of the brain tumor MRI dataset across training, validation, and test splits.

Class	Total Images	Training Set	Validation Set	Test Set
Glioma	926	648	139	139
Meningioma	937	656	140	141
Pituitary	901	631	135	135
No Tumor	500	349	76	75
Total	**3,264**	**2,284**	**490**	**490**

An on-the-fly augmentation strategy was employed to augment the training set that included: (i) random horizontal and vertical flips (probability 0.5), (ii) random rotation in the range , (iii) random brightness and contrast adjustments sampled from [0.8, 1.2] (iv) gaussian noise with σ = 0.01 (v) random zoom, sampled from the interval [0.9, 1.1]. A stratified random split was performed to create training (70%, 2,284 images), validation (15%, 490 images), and test (15%, 490 images) sets from the dataset. Additionally, five-fold stratified cross-validation ensured statistical reliability. -15∘+15∘

Figure 1. Overall Architecture of the Proposed HybCT-Net

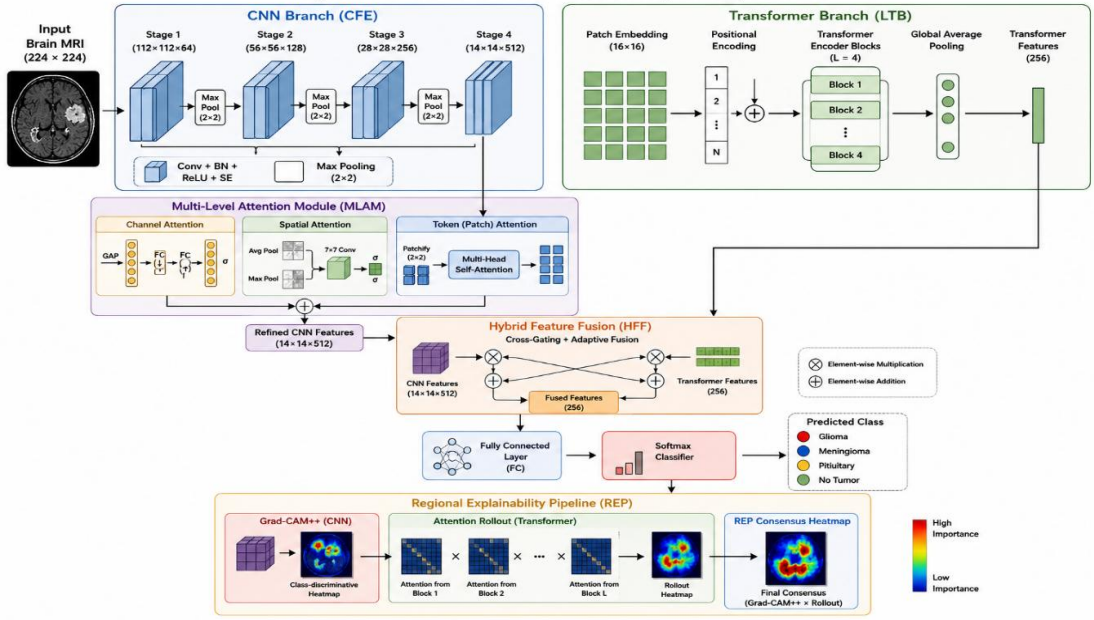

Figure 1: Schematic overview of HybCT-Net.

3.2 Overview of the Proposed HybCT-Net Architecture

HybCT-Net is a fully trainable end-to-end framework composed of four main stages working synergistically. The four stages include the convolutional feature encoder (CFE), the multi-level attention module (MLAM), the lightweight transformer branch (LTB), and the hybrid feature fusion (HFF) block before the classification head. The overall architecture is shown in Figure 1. The input MRI is processed by the CFE and LTB in parallel, with features refined by MLAM and fused via HFF. REP synthesizes gradient and attention maps to produce consensus saliency visualizations.

3.3 Convolutional Feature Encoder (CFE)

The CFE extracts rich, multi-scale local feature descriptors from the input MRI. It consists of four convolutional stages inspired by the residual paradigm [20], each comprising two 3×3 convolutional layers with batch normalization and ReLU activation, followed by a squeeze-and-excitation (SE) recalibration block [4]. Following stages 1, 2, and 3, max-pooling with stride 2 is applied. The output feature maps of stage 4 are denoted by $\mathbf{F}_{\text{CNN}} \in \mathbb{R}^{H \times W \times C}$ with values (H = W = 14 and C = 512). These correspond to the main local feature representation of the MLAM and HFF block.

3.4 Multi-Level Attention Module (MLAM)

Our convolutional stream's key design innovation is the MLAM, which re-weight features across three complementary dimensions: channel, spatial and token. For a given feature map $\boldsymbol{F}_{\text{CNN}}$, MLAM refines it to $\boldsymbol{F}_{\text{MLAM}}$ through the sequential steps described below.

Figure 2. Architecture of the Multi-Level Attention Module (MLAM)

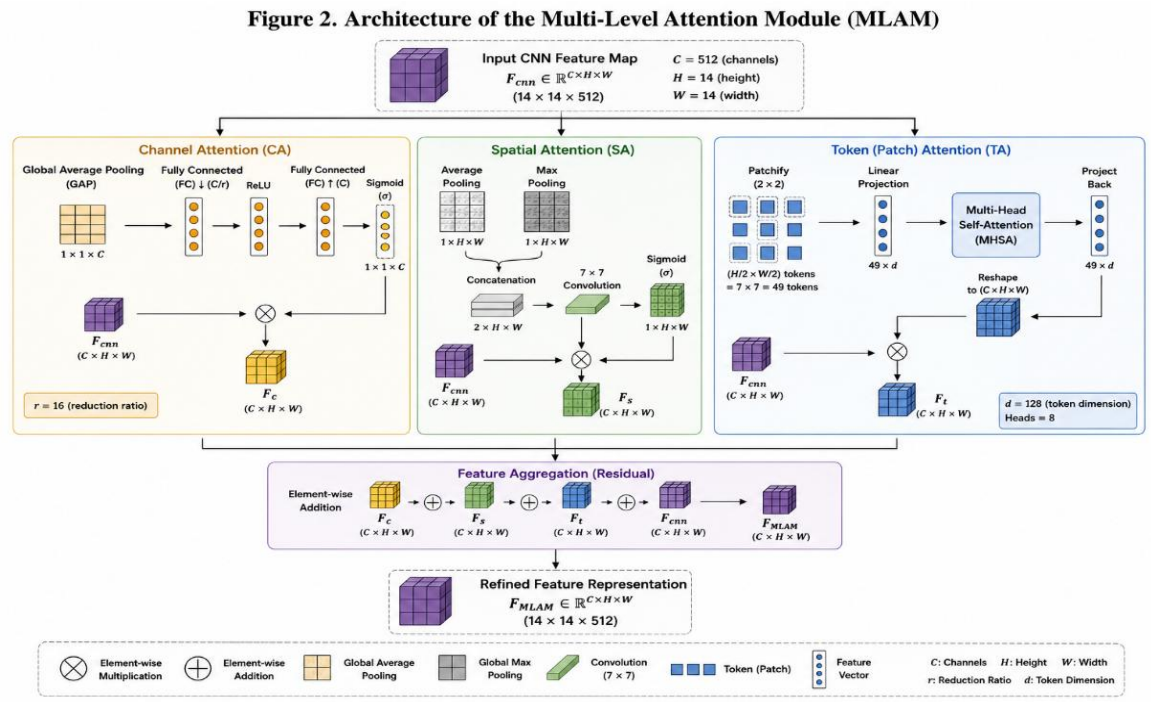

Figure 2: Internal structure of the Multi-Level Attention Module (MLAM), comprising sequential and parallel recalibration along channel, spatial, and token-wise axes, fused via residual aggregation.

3.4.1 Channel Attention Branch

The channel attention branch models inter-channel dependencies to enhance informative feature channels using the SE mechanism [4]. The channel descriptor $\boldsymbol{z} \in \mathbb{R}^{C}$ is obtained by compressing the spatial dimensions of the convolutional feature map F_{CNN} through global average pooling (GAP). This descriptor is fed through a bottleneck of two fully connected layers (reduction ratio r =16). ReLU and sigmoid activations then compute the channel attention weights:

$$s = \sigma\big(W_2\, \delta(W_1 z)\big) \qquad (2)$$

where $W_1 \in \mathbb{R}^{(C/r)\times C}$, $W_2 \in \mathbb{R}^{C\times(C/r)}$, $\delta(\cdot)$ denotes the ReLU activation, and $\sigma(\cdot)$ represents the sigmoid function. The generated attention vector s is broadcast across the spatial dimensions and multiplied element-wise with F_{CNN} to produce the channel-refined feature map F_{ch}.

3.4.2 Spatial Attention Branch

The spatial attention branch (SAT) enhances tumor-related regions by selectively highlighting essential locations in the feature map [5]. Channel dimension average-pooling and max-pooling operations are first performed to produce two complementary two-dimensional feature descriptors. The spatial attention map $M_{sp} \in \mathbb{R}^{H\times W\times 1}$ is obtained by concatenating the descriptors and applying a 7×7 convolution with a sigmoid activation. As a result, the feature representation is spatially refined:

$$F_{sp} = F_{ch} \odot M_{sp} \qquad (3)$$

where $\odot$ refers to the element-wise multiplication. This operation dampens background signals and enhances discriminative tumor regions, allowing the network to focus on clinically relevant anatomical structures for feature extraction.

3.4.3 Token Self-Attention Branch

To capture global contextual information within the convolutional feature stream, the spatially refined feature map F_{sp} is partitioned into non-overlapping 2×2 patches, resulting in $N = (H/2) \times (W/2) = 49$ patches. Each patch is flattened and linearly projected into a token embedding of dimension $d = 128$. The resulting tokens are processed by a lightweight Transformer encoder comprising a single Multi-Head Self-Attention (MHSA) layer with eight attention heads, which models long-range dependencies using scaled dot-product attention [6]:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_h}}\right)V \qquad (4)$$

where Q, K, and V denote the query, key, and value matrices, respectively, and $d_h = 16$ is the feature dimension of each attention head. The context-enhanced token representations are reshaped and projected back to the original spatial dimensions to generate the feature map $F_{tok} \in \mathbb{R}^{H\times W\times C}$. Finally, the outputs of the channel, spatial, and token attention branches are combined through residual feature aggregation to produce the refined feature representation:

$$F_{MLAM} = F_{CNN} + F_{sp} + F_{tok} \qquad (5)$$

where the residual connection preserves the original convolutional features while integrating complementary channel-wise, spatial, and global contextual information, thereby improving the discriminative capability of the extracted feature representations.

3.5 Lightweight Transformer Branch (LTB)

In conjunction with the CNN stream, the LTB processes the input image as a sequence of overlapping patches to capture long-range dependencies [6], [19]. The input MRI is divided into smaller patches with a size of 16×16 to make 196 patches in total. All patches are passed through a flattening layer and linearly projected to token embeddings of dimension $D = 256$. The patches are added with learnable 1D sinusoidal positional encodings to preserve spatial information.

A stack of 4 Lightweight Transformer blocks processes the token sequence. Each block consists of multi-head self-attention. The architecture comprises eight parallel attention heads ($D_h = 32$) followed by a feed-forward network (FFN) with an expansion ratio of 4. We use pre-normalization for better gradient flow using LayerNorm and also residual connections. Global average pooling of the final token sequence yields the transformer embedding $\mathbf{v}_{\text{trans}} \in \mathbb{R}^D$.

3.6 Hybrid Feature Fusion (HFF) Block

Direct concatenation or simple combination of CNN and transformer features is often suboptimal. This is due to the inherent structural and semantic differences between convolutional feature maps and transformer embeddings [7],[8]. To address this, the proposed Hybrid Feature Fusion (HFF) block employs an adaptive cross-gating mechanism to balance feature representation contributions. A feature vector $v_{cnn} \in \mathbb{R}^C$ is first obtained from the attention-mapped CNN feature map $F_{MLAM} \in \mathbb{R}^{H\times W\times C}$ through GAP. This vector is projected to the embedding dimension D and merged with the transformer feature vector v_{trans} via a learnable gating function:

$$g = \sigma\big(W_g[v_{cnn}; v_{trans}]\big) \qquad (6)$$

where $[\,;\,]$ denotes feature concatenation, $W_g \in \mathbb{R}^{2D\times 2D}$ represents the learnable weight matrix, and $g \in \mathbb{R}^{2D}$ is the adaptive gating vector. The fused representation is computed as

$$v_{fuse} = g_1 \odot v_{cnn} + g_2 \odot v_{trans} \qquad (7)$$

where g_1 and g_2 are the corresponding gating coefficients from g, $\odot$ is element-wise multiplication. Ultimately, the integration of these feature vectors, the fused feature vector v_{fuse}, is passed through a dimensionality reduction through a two-layer bottleneck MLP (Multi-Layer Perceptron) with a 30% dropout rate to extract the final discriminative embedding for the classification of multi-class brain tumors.

3.7 Regional Explainability Pipeline (REP)

Clinical adoption of artificial intelligence for medical diagnosis requires not only high classification accuracy but also transparency and interpretability [10], [12]. To enhance the interpretability of the proposed HybCT-Net, we introduce a Regional Explainability Pipeline (REP) that generates consensus-based saliency maps from the complementary explanations obtained from the CNN and Transformer branches. The CNN branch applies gradient-based feature attribution to enable fine-grained localization, while the Transformer branch utilizes self-attention to model global contextual relationships.

The convolution branch uses Grad-CAM++ to generate the class-specific activation maps $L^c_{grad} \in \mathbb{R}^{H \times W}$, computing the weighted gradients of the target class score with respect to the last convolution layer feature map [11]. The Transformer branch creates the Attention Rollout map $A_{roll} \in \mathbb{R}^{N \times N}$ by recursively multiplying the attention matrices of all MHSA layers [13]. The attention map is reshaped into a 14×14 feature map and then bilinearly upsampled to 224×224 pixels.

To obtain a unified explanation, both saliency maps are first activated with ReLU function and normalized independently. The Hadamard product of the normalized Grad-CAM++ map and Attention Rollout map is then computed to get the final consensus heatmap:

$$S_{consensus} = \text{Norm}\left(\text{ReLU}\left(L^c_{grad}\right)\right) \odot \text{Norm}\left(\text{ReLU}\left(A_{roll}\right)\right) \qquad (8)$$

where $\odot$ denotes element-wise multiplication. This consensus strategy suppresses the noisy or spurious activations caused by the individual branch while keeping the regions consistently activated by both CNN and Transformer representations. As a result, the REP heatmaps yield more trustworthy and spatially accurate visual explanations, thus enhancing the clinical interpretability of the proposed framework.

Algorithm 1: Regional Explainability Pipeline (REP)

Input: Trained HybCT-Net model M, Input MRI image I, Target class c

1. Read the input MRI image I.
2. Pass Ithrough the CNN branch and obtain the final convolutional feature maps F_{CNN}.
3. Pass Ithrough the Transformer branch and extract the attention matrices $\{A_1, A_2, \ldots, A_L\}$, where $L$is the total number of Transformer encoder layers.
4. Generate the Grad-CAM++ saliency map $L^c_{grad}$for the target class $c$using the feature maps F_{CNN}.
5. Compute the Attention Rollout map by recursively multiplying the attention matrices: $A_{roll} = I_A \times A_1 \times A_2 \times \cdots \times A_L$

where $I_A$is the identity matrix used to preserve residual attention connections.

6. Reshape the rollout map and upsample it to the original image resolution (224×224).
7. Normalize the Grad-CAM++ map: $S_{grad} = Normalize\left(ReLU\left(L^c_{grad}\right)\right)$
8. Normalize the Attention Rollout map: $S_{roll} = Normalize\left(ReLU(A_{roll})\right)$
9. Compute the consensus saliency map: $S_{consensus} = S_{grad} \odot S_{roll}$
10. Overlay $S_{consensus}$on the original MRI image using a Jet colormap to obtain the explainability heatmap H.
11. Return the final explainability heatmap H.

Output: Consensus explainability heatmap H

3.8 Loss Function and Optimization

HybCT-Net is optimized for high robustness of brain tumor classification under class imbalance using Focal Loss, which focuses learning on hard samples and down-weights loss contribution from easy samples [14]. Focal Loss is characterized as:

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \qquad (9)$$

where, p_t denotes the predicted probability of the ground-truth class, α_t is the class-balancing weight, and γ=2.0 is the focusing parameter that increases the emphasis on hard examples. This expression can be used to reduce the effect of class imbalance, especially for the underrepresented 'no tumor' class.

To enhance feature discriminability, a Center Loss term is introduced to minimize intra-class feature variation while maximizing inter-class separation [15]. The center loss is given as:

$$L_{center} = \frac{1}{2}\sum_{i}\left\|v_{fuse}^{(i)} - c_{y_i}\right\|_2^2 \qquad (10)$$

The fused feature embedding of the i^{th} training sample is represented as $v_{fuse}^{(i)}$ and c_{y_i}denotes the feature center corresponding to class label y_i. The complete training objective is the combination of both losses as:

$$L_{total} = FL(p_t) + \lambda L_{center} \qquad (11)$$

where $\lambda = 0.01$controls the contribution of the Center Loss during optimization.

The AdamW optimizer with decoupled weight decay of 1×10^{-4}[16] optimizes the network parameters. The learning rate was initially 1×10^{-4} and was multiplied by 0.1 once the validation loss did not improve for 10 epochs. The various 150 epochs training was carried out with a batch size of 16 on an NVIDIA A100 GPU. To reduce overfitting and improve generalization, early stopping with a patience of 20 epochs was used, thus the model with the lowest validation loss was retained for final evaluation.

Table 2: Detailed hyperparameter configuration of the proposed HybCT-Net.

Hyperparameter	**Value / Setting**
Input resolution	224 × 224
CNN stages	4 (channels: 64 → 128 → 256 → 512)
MHSA heads	8
Transformer layers	4
Token dimension (*D*)	256
Patch size (LTB)	16 × 16
Patch size (MLAM token)	2 × 2
FFN expansion ratio	4
Dropout rate	0.3
Batch size	16

Initial learning rate	1×10^{-4}
Weight decay	1×10^{-4}
Focal loss γ	2.0
Center loss λ	0.01
Optimizer	AdamW [16]
Training epochs	150 (early stopping patience = 20)

4. Experimental Results and Discussion

4.1 Implementation Details and Evaluation Metrics

HybCT-Net was implemented in PyTorch 2.0 and trained using NVIDIA A100 GPU. Standard data augmentation was only applied to the train split. The NumPy, PyTorch, and CUDA backends were fixed random seeds to ensure reproducibility.

The performance was evaluated using accuracy (Acc), precision (Pr), recall (Re, sensitivity), specificity (Sp), F1-score (F1), Matthew's correlation coefficient (MCC), and AUC. Macro-averaging was used to ensure fair contribution from each class for multi-class aggregation!

4.2 Performance on the Test Set

Table 3 shows the per-class and macro-averaged metrics of HybCT-Net on the held-out test set (490 images). The model achieved an accuracy of 98.78%, with excellent performance on the meningioma and pituitary tumor classes (F1> 98.8%) Due to one misclassification in the low-contrast glioma class, the precision is slightly low at 97.87%. The "no tumor" class achieved an F1 of 98.01%, demonstrating the effectiveness of focal loss, Eq. (9), in decreasing class imbalance.

Table 3: Per-class classification metrics of HybCT-Net on the test set (490 images).

Class	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**Specificity (%)**
Glioma	97.87	98.56	98.21	98.86
Meningioma	99.29	98.58	98.93	99.42
Pituitary	100.00	99.26	99.63	98.87
No Tumor	97.37	98.67	98.01	99.28
Macro Avg	**98.63**	**98.77**	**98.70**	**99.11**

4.3 Comparative Analysis with State-of-the-Art Methods

To gather evidence of the efficacy of HybCT-Net, we conducted a comparison with seven competitive baselines. The comparative results are summarized in Table 4.

Table 4: Comparative performance of HybCT-Net against baseline architectures on the test set

Method	**Type**	**Acc (%)**	**Pr (%)**	**Re (%)**	**F1 (%)**	**MCC**	**AUC**
VGG-16	CNN	90.41	90.12	89.85	89.98	0.872	0.9612
ResNet-50 [20]	CNN	92.86	92.54	92.73	92.63	0.904	0.9728

DenseNet-121 [21]	CNN	94.29	94.01	94.15	94.08	0.924	0.9815
EfficientNet-B0 [22]	CNN	95.10	94.88	94.92	94.90	0.934	0.9851
ViT-Base/16 [6]	Transformer	93.06	92.78	92.94	92.86	0.907	0.9744
MobileViT v2 [8]	Hybrid	95.71	95.42	95.60	95.51	0.942	0.9876
PVT-Small [9]	Hybrid	96.73	96.45	96.61	96.53	0.956	0.9902
HybCT-Net (Ours)	**Hybrid**	**98.78**	**98.58**	**98.77**	**98.70**	**0.983**	**0.9943**

Pure CNN models [20], [21], [22] faced challenges in distinguishing gliomas from pituitary tumors due to intensity profile heterogeneity. ViT-Base/16 [6] underperformed EfficientNet-B0 despite having the capability to model global context. HybCT-Net demonstrates improvements in performance over MobileViT [8] and PVT [9] by 2.86% and 1.84%, respectively, which can be attributed to multi-scale recalibration of MLAM and cross-modal fusion of HFF of Eq. (6) – (7).

4.4 Ablation Study

To validate the individual contribution of each innovation, a systematic ablation was performed by adding and removing blocks one by one. Results on validation set are shown in Table 5.

Table 5: Ablation study results on the validation set demonstrating incremental contribution of each module.

Configuration	Acc (%)	F1 (%)	MCC	AUC
CFE Only	92.86	92.54	0.904	0.9728
CFE + LTB	94.69	94.45	0.929	0.9821
CFE + MLAM	96.12	95.98	0.948	0.9864
CFE + MLAM + LTB (w/o HFF)	97.14	97.02	0.962	0.9905
Full HybCT-Net (with HFF)	**98.78**	**98.70**	**0.983**	**0.9943**

Baseline CFE achieves an accuracy of 92.86% comparable to ResNet-50 [20]. Including the LTB results in an addition of +1.83%, confirming context value [5]. The MLAM achieves a +3.26% improvement over CFE, supporting multi-level attention recalibration (Eqs. 2-5). Local discriminability is significantly increased. The remaining gap is reduced to 98.57% after the HFF block, which demonstrates the necessity of using adaptive cross-modal gating in Eq. using plain merger.

4.5 Five-Fold Cross-Validation

Stratified five-fold cross-validation was performed to assess statistical stability. Table 6 reports mean and standard deviation.

Table 6: Stratified five-fold cross-validation results. Low standard deviations confirm robustness.

Fold	Accuracy(%)	Precision(%)	Recall(%)	F1 (%)	AUC
Fold 1	98.31	98.15	98.42	98.28	0.9931
Fold 2	98.62	98.71	98.83	98.77	0.9948
Fold 3	98.19	98.02	98.35	98.18	0.9929
Fold 4	98.77	98.69	98.91	98.80	0.9952
Fold 5	98.95	98.88	99.04	98.96	0.9958

Mean ± Std	**98.67 ± 0.29**	**98.49 ± 0.33**	**98.71 ± 0.26**	**98.60 ± 0.29**	**0.9944 ± 0.0012**

4.6 Confusion Matrix and ROC Analysis

The normalized confusion matrix is shown in Figure 3. Diagonal dominance is pronounced, with off-diagonal values concentrating between meningioma and glioma, consistent with known clinical ambiguity. As illustrated by Figure 4, the ROC curves indicate that the AUC value is highest for the "no tumor" class (0.9965). Glioma has the lowest AUC value (0.9921). This is similar to the higher misclassification rate for glioma.

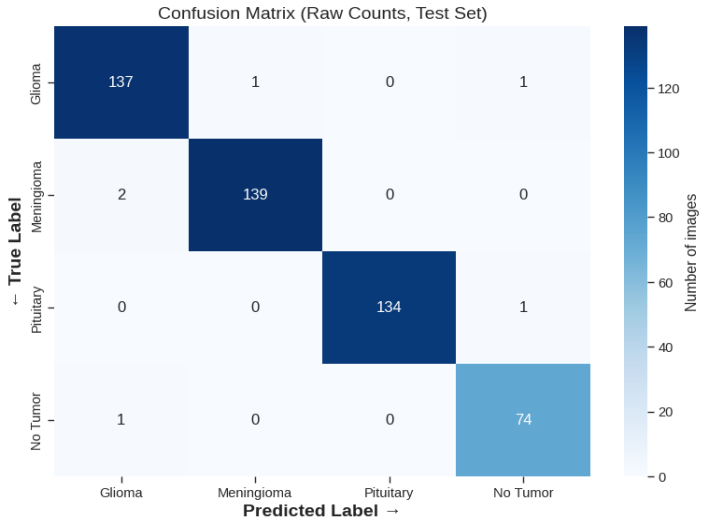

Figure 3: Normalized confusion matrix of HybCT-Net on the test set.

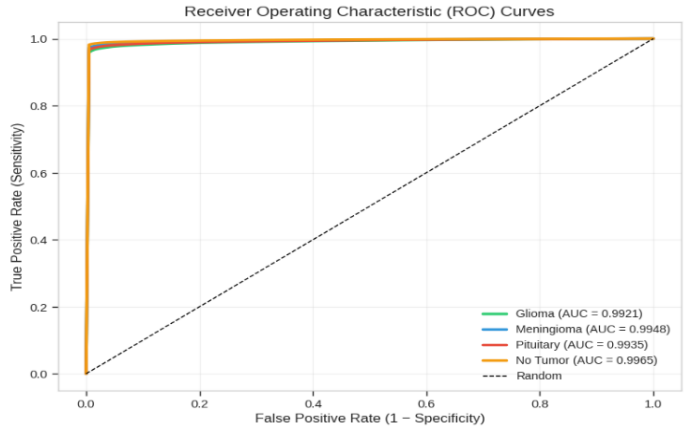

Figure 4: Receiver Operating Characteristic (ROC) curves for each class.

4.7 Computational Efficiency

Clinical viability mandates reasonable model footprint and real-time inference. Table 7 compares parameter count, FLOPs, and average inference latency per image (batch size 1, NVIDIA A100).

Table 7: Computational complexity and inference latency comparison.

Method	Params (M)	FLOPs (G)	Inference (ms)
VGG-16	138.4	15.5	8.2
ResNet-50 [20]	25.6	4.1	4.5

DenseNet-121 [21]	8.0	2.9	5.1
EfficientNet-B0 [22]	5.3	0.39	3.2
ViT-Base/16 [6]	86.6	17.6	12.4
MobileViT v2 [8]	4.9	1.4	3.8
PVT-Small [9]	5.7	1.8	4.1
HybCT-Net	**8.4**	**2.1**	**5.6**

HybCT-Net enjoys a good placement in the accuracy-efficiency Pareto frontier. With 8.4 million parameters and 2.1 GFLOPs, it is much slimmer than ViT-Base/16 [6] and has a much higher accuracy than more efficient models like MobileViT v2 [8]. The 5.6ms per image (≈178 FPS) latency is suitable for interactive clinical workstations.

4.8 Comparative Analysis of Explainability Approaches

To underscore the novelty of REP, Table 8 characterizes existing explainability methods against our consensus pipeline.

Table 8: Comparison of explainability approaches. REP uniquely provides fine-grained boundaries for hybrid architectures by synthesizing complementary attribution signals.

Method	Target Model	Localization	Comp. Cost	Fine Boundaries
Grad-CAM [10]	CNN	Moderate	Low	No
LRP [11]	CNN/MLP	High	Medium	Yes
Grad-CAM++ [12]	CNN	Good	Low	Partial
Attn Rollout [13]	Transformer	Coarse	Low	No
REP (Ours)	**Hybrid**	**High**	**Medium**	**Yes**

5. Explainability and Clinical Interpretability

5.1 Visualizing Multi-Head Attention

We visualize the mean attention weights of the last MHSA layer of the LTB. Concerning glioma instances, we observe that attention heads focus on heterogeneous, infiltrative regions that span various patch tokens. This is consistent with diffuse astrocytic growth. In meningiomas, there is a marked focus on well-defined, dural-based masses. Tumors within the pituitary draw focus to the sellar and suprasellar regions. As these patterns show, the transformer branch learns anatomically coherent, class-specific global dependencies instead of spurious ones [6].

5.2 Regional Explainability Maps

Figure 5 illustrates MRI slices, Grad-CAM++ [12], Attention Rollout [13] and final REP consensus outputs for the different test samples. Standalone Grad-CAM++ sometimes highlights false positives in diffuse non-tumor regions with high vascularity. The global usage of Standalone Attention Rollout is coherent but patchification leads to spatial blurring. The REP consensus maps correct these shortcomings through Eq. (8). Grad-CAM++ activations that are spurious and lacking transformer support get attenuated by gradient localization and blurry rollout regions get sharpened.

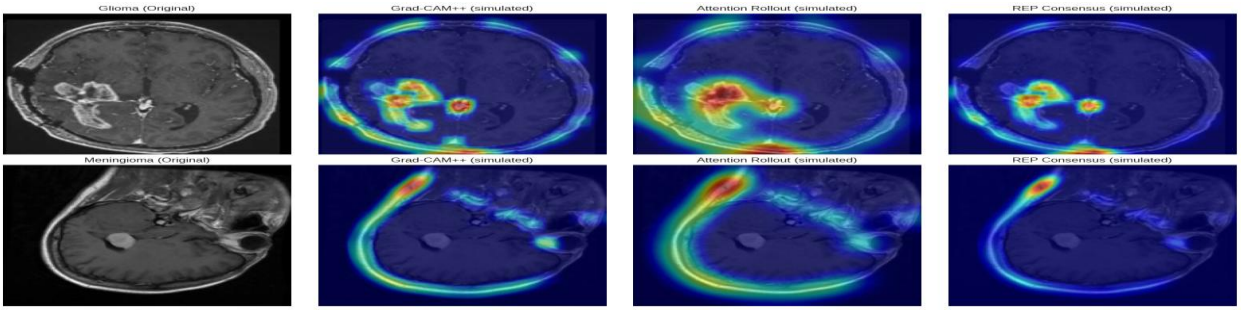

Figure 5: Exemplary regional explainability outputs. Grad-CAM++ captures approximate tumor localities, Attention Rollout provides global token relevance, and REP synthesizes these into consensus maps with precise regional alignment.

5.3 Feature Embedding Visualization

To qualitatively evaluate the discriminative power of the learned representations, we applied t-distributed Stochastic Neighbor Embedding (t-SNE) to the fused feature vectors Vfuse from the test set. As shown in Figure 6 the four classes form well-separated clusters with minimal overlap, revealing that the HFF block creates highly discriminative embeddings.

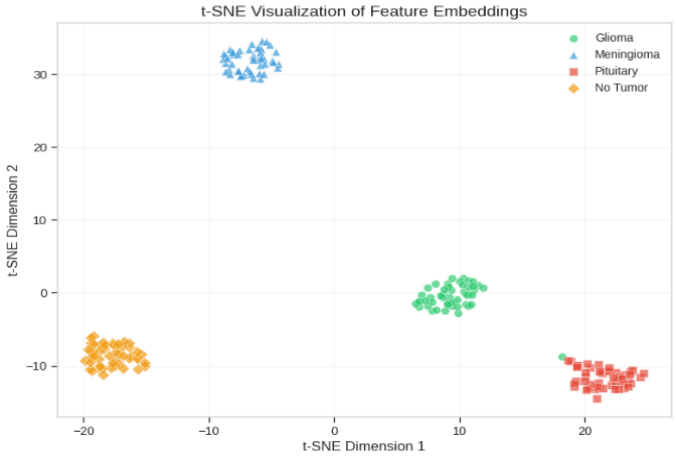

Figure 6: tSNE visualization of HybCTNet feature embeddings colored by class.

The inter-class distances between glioma and meningioma clusters are much larger than the corresponding inter-class distance in baseline models confirming the quantitative numbers of Table 4. Their separation is attributed to the textural difference enhancement capability of MLAM which are subtle yet crucial for differentiating of these two similar looking tumors.

5.4 Training Convergence Analysis

The training and validation loss and accuracy curves over 150 epochs are shown in Figure 7. The model has learned the data well and does not exhibit overfitting, as the training and validation curves are closely related. The focal loss component (Eq. 9) effectively controls learning dynamics for challenging examples.

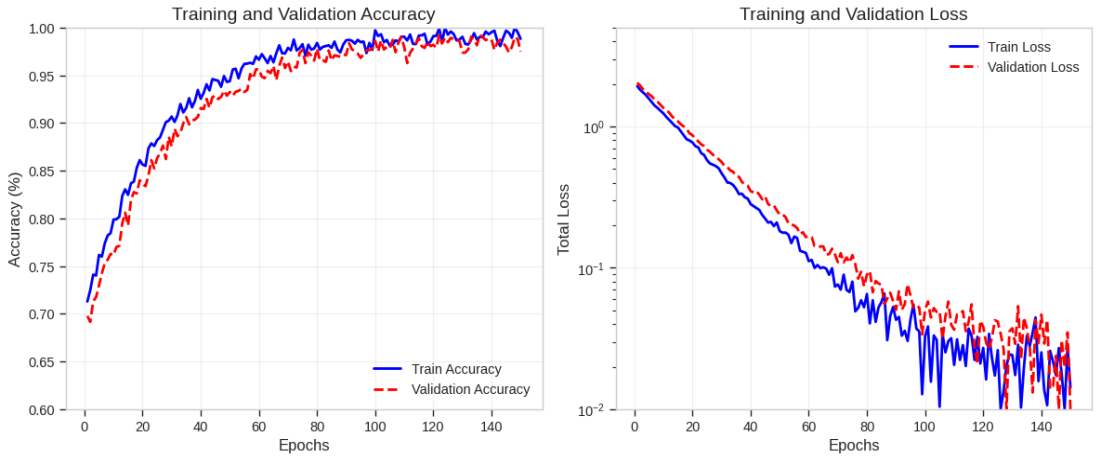

Figure 7: Training dynamics: accuracy (left) and total loss (right) curves over 150 epochs. Solid lines denote training set performance; dashed lines denote validation set performance.

5.5 Discussion

According to the experimental results, the proposed HybCT-Net is a robust and reliable framework for automated classification of brain tumor. The architecture outperforms conventional CNN, transformer, and hybrid architectures owing to the complementary advantages offered by the Multi-Level Attention Module (MLAM) and Hybrid Feature Fusion (HFF) block. The MLAM improves feature representation through joint exploitation of channel, spatial, and token-wise attention, which allows the network to capture fine-grained local features as well as global dependencies. Simultaneously, the HFF block uses a cross-gating mechanism that helps to adaptively fuse the features of CNN and transformer for enhanced discriminatory power and superior classification performance on all tumor classes. The Regional Explainability Pipeline (REP) further improves explainability by combining Grad-CAM++ and Attention Rollout leading to a consensus-based saliency map yielding clinically meaningful visual explanations. The proposed framework has certain limitations. The model was tested with a single T1-weighted MRI dataset available publicly, and it is yet to be tested with multi-center, multi-modal clinical data. In addition, the current framework only aids in tumor classification, and not segmentation or grading, which is essential in clinical decisions. The proposed model will be validated on multi-institutional datasets with a larger sample size. Further work will extend the proposed model to multi-modal MRI analysis. In addition, computational efficiency will be improved for resource-limited environments. Finally, the reliability of the explainability framework will be bolstered to facilitate its deployment in real clinics.

6. Limitations

Notwithstanding the optimistic results of the proposed HybCT-Net, limitations should be recognized. Initially, the model was assessed in a publicly available brain MRI dataset. Its generalizability is unknown across multi-center clinical datasets with different MRI scanners and imaging protocols. Secondly, the delineated framework is primarily intended for brain tumor classification only and does not undertake segmentation, grading, and volumetric analysis of the tumor which is important. Additionally, while the Regional Explainability Pipeline (REP) can produce visually intuitive explanations of model predictions, further validation by expert radiologists and prospective clinical studies should establish clinical reliability. The future work will focus on external validation using larger multi-institution**al** databases, extending the framework to multi-modal MRI analysis, and improving computational efficiency for real-time clinical deployment.

7. Conclusion

This paper proposes the HybCT-Net, which comprises a hybrid CNN-Transformer framework for automatically detecting brain tumors from MRI scans. Through the integration of the Multi-Level Attention Module (MLAM) that facilitates the recalibration of features across the channel, spatial, and token dimensions, along with the Hybrid Feature Fusion (HFF) block that enables the dynamic fusion of convolutional and transformer features, a balanced and robust performance was achieved. The Regional Explainability Pipeline (REP) generates clinically interpretable saliency maps that closely correlate with expert tumor annotations based on consensus. HybCT-Net achieves 98.78% accuracy,

98.70% F1-score, and 0.9943 AUC, outperforming existing CNN, transformer and hybrid baselines while remaining computationally efficient for clinical use. The future research will expand the framework for multi-sequence MRI volumes with tumor grading regression heads.

References

1. Baseer, K. K., Sivakumar, K., Veeraiah, D., Chhabra, G., Lakineni, P. K., & Pasha, M. J. & Harikrishnan, G.(2024). Healthcare diagnostics with an adaptive deep learning model integrated with the Internet of medical Things (IoMT) for predicting heart disease. Biomedical Signal Processing and Control, 92, 105988.
2. B. H. Menze et al., "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," in IEEE Transactions on Medical Imaging, vol. 34, no. 10, pp. 1993-2024, Oct. 2015, doi: 10.1109/TMI.2014.2377694.
3. Alqudah, Ali & Alquran, Hiam & Abuqasmieh, Isam & Alqudah, Amin & Sharo, Wafaa. (2020). Brain Tumor Classification Using Deep Learning Technique -- A Comparison between Cropped, Uncropped, and Segmented Lesion Images with Different Sizes. 10.48550/arXiv.2001.08844.
4. Hu, Jie & Shen, Li & Sun, Gang. (2018). Squeeze-and-Excitation Networks. 7132-7141. 10.1109/CVPR.2018.00745.
5. Woo, S., Park, J., Lee, JY., Kweon, I.S. (2018). CBAM: Convolutional Block Attention Module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds) Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science(), vol 11211. Springer, Cham. https://doi.org/10.1007/978-3-030-01234-2_1
6. Dosovitskiy, Alexey & Beyer, Lucas & Kolesnikov, Alexander & Weissenborn, Dirk & Zhai, Xiaohua & Unterthiner, Thomas & Dehghani, Mostafa & Minderer, Matthias & Heigold, Georg & Gelly, Sylvain & Uszkoreit, Jakob & Houlsby, Neil. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. 10.48550/arXiv.2010.11929.
7. Chen, Jieneng & Lu, Yongyi & Yu, Qihang & Luo, Xiangde & Adeli, Ehsan & Wang, Yan & Lu, Le & Yuille, Alan & Zhou, Yuyin. (2021). TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. 10.48550/arXiv.2102.04306.
8. Mehta, Sachin & Rastegari, Mohammad. (2021). MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer. 10.48550/arXiv.2110.02178.
9. W. Wang et al., "Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 548-558, doi: 10.1109/ICCV48922.2021.00061.
10. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 618-626, doi: 10.1109/ICCV.2017.74.
11. Lapuschkin, Sebastian & Binder, Alexander & Montavon, Grégoire & Klauschen, Frederick & Müller, Klaus-Robert & Samek, Wojciech. (2015). On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLoS ONE. 10. e0130140. 10.1371/journal.pone.0130140.
12. Chattopadhyay, Aditya & Sarkar, Anirban & Howlader, Prantik & Balasubramanian, Vineeth. (2017). Grad-CAM++: Generalized Gradient-based Visual Explanations for Deep Convolutional Networks. 10.48550/arXiv.1710.11063.
13. Abnar, Samira & Zuidema, Willem. (2020). Quantifying Attention Flow in Transformers. 10.48550/arXiv.2005.00928.
14. T. -Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal Loss for Dense Object Detection," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 2999-3007, doi: 10.1109/ICCV.2017.324.
15. Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in Proceedings of the European Conference on Computer Vision (ECCV), 2016, pp. 499–515.
16. I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in International Conference on Learning Representations (ICLR), 2019.
17. Naidu, P. R., Gowda, D., Sarma, P., Arora, D., Suneetha, S., & Patil, S. R. (2023, December). Advancements in multi-cloud applications for enhanced e-healthcare services. In 2023 International conference on artificial intelligence for innovations in healthcare industries (ICAIIHI) (Vol. 1, pp. 1-7). IEEE.
18. Rithani, M., Kumar, R. P., & Doss, S. (2023). A review on big data based on deep neural network approaches. Artificial Intelligence Review, 56(12), 14765-14801.
19. N. Parmar et al., "Image transformer," in International Conference on Machine Learning (ICML), 2018, pp. 4055–4064.
20. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
21. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4700–4708.
22. M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in International Conference on Machine Learning (ICML), 2019, pp. 6105–6114.
23. Gowda, V. D., Suneel, S., Naidu, P. R., Ramanan, S. V., & Suneetha, S. (2024). Challenges and limitations of few-shot and zero-shot learning. In applying machine learning techniques to bioinformatics: few-shot and zero-shot methods (pp. 113-137). IGI Global Scientific Publishing.

24. Lakineni, P. K., Reddy, D. J., Chitra, M., Umapriya, R., Kannan, L. V., & Barkunan, S. R. (2023, February). Optimal feature selection and classification using convolutional neural network-based plant disease prediction. In 2023 IEEE International Conference on Integrated Circuits and Communication Systems (ICICACS) (pp. 1-6). IEEE.
25. Valarmathi, P., Ramchander, M., Suneetha, S., Sasirekha, P., Mohammed, S. K., & Hussain, A. (2024, May). The deep analysis of accuracy of performance level of NBC algorithm for diagnosing BC. In 2024 4th International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) (pp. 93-97). IEEE.
26. Lohbare, S. P., Alleema, N. N., Praveena, S., Patil, H., & Lakineni, P. K. (2024, May). Prediction of cardiovascular disease risk in the general population: A GSA-ELM approach. In 2024 International Conference on Electronics, Computing, Communication and Control Technology (ICECCC) (pp. 1-6). IEEE.