IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

Vision Transformer for Brain Tumor Segmentation using BraTS Dataset

[1]**Lalit Kumar Rawat**, Ph.D. Scholar, Department of Statistics, Mahatma Gandhi Kashi Vidyapith (MGKVP). Varanasi, India

[2]**Prof. Anil Kumar**, Department of Statistics, Mahatma Gandhi Kashi Vidyapith (MGKVP), Varanasi, India.

[3]**Dr. Vijendra Pratap Singh** (Assistant Professor) Department of Computer Sciences & Applications, Mahatma Gandhi Kashi Vidyapith (MGKVP), Varanasi, India.

Abstract:

This study presents an advanced approach for brain tumor segmentation using the Vision Transformer (ViT) model trained on the BraTS 2021 dataset. The research explores transformer-based architectures for pixel-level medical image segmentation, comparing ViT-B/16 with a Swin-UNet baseline. Performance metrics, including Dice coefficient, IoU, precision, recall, F1-score, and AUC, demonstrate the model's superior accuracy. Explainability is provided through attention maps highlighting tumor regions, offering transparency in decision-making for clinical applications.

Keywords: Vision Transformer, Brain Tumor Segmentation, BraTS Dataset, Deep Learning, Medical Image Analysis, MRI, Attention Mechanism, Swin-UNet, Dice Coefficient.

1. Introduction:

Brain tumor segmentation plays a crucial role in clinical diagnosis, treatment planning, and follow-up analysis. Magnetic Resonance Imaging (MRI) offers rich structural information but manual segmentation is time-consuming and prone to inter-observer variability. Deep learning, especially convolutional neural networks (CNNs), has achieved remarkable progress in medical image segmentation. However, CNNs are limited by local receptive fields. The Vision Transformer (ViT), based on self-attention mechanisms, captures long-range dependencies and global context, making it suitable for complex tumor structure segmentation. This paper investigates the effectiveness of ViT for brain tumor segmentation using the BraTS 2021 dataset.

2. Literature Review:

Recent advances in medical image analysis have introduced transformer-based architectures that outperform traditional CNNs in various tasks. Dosovitskiy et al. (2020) introduced the Vision Transformer (ViT), which applies transformer encoders to image patches. Chen et al. (2021) extended this to medical imaging with TransUNet, combining CNNs and Transformers for segmentation tasks. Swin Transformer (Liu et al., 2021) further improved efficiency through shifted windows, maintaining hierarchical representations. For brain tumor segmentation, the BraTS challenge datasets have enabled the development of robust models, including 3D U-Net and Swin-UNet architectures. Despite these advances, explainability and interpretability remain challenges in deep medical models, motivating the adoption of attention-based visualization in this study.

3. Methodology:

The proposed method employs the Vision Transformer (ViT-B/16) architecture fine-tuned on the BraTS 2021 dataset. The BraTS dataset includes multi-modal MRI scans (T1, T2, FLAIR, and T1ce) annotated with enhancing, tumor core, and whole-tumor regions. Images were preprocessed using bias field correction, normalization, and patch extraction. The ViT divides input images into 16×16 patches, embeds them into tokens, and processes them via multi-head self-attention layers. For comparison, a Swin-UNet model was implemented as a baseline using the same preprocessing and training pipeline.

Table 1. Dataset Summary (BraTS 2021): Summary of number of subjects, modalities, training/testing split, voxel size.

Attribute	Description
Total Subjects	1,251 patients (train = 1,251; validation = 219 provided separately; test = unknown labels for competition phase)
Modalities	Four MRI sequences per subject — T1, T1ce (contrast-enhanced T1), T2, and FLAIR
Segmentation Labels	(1) Enhancing Tumor (ET), (2) Tumor Core (TC), (3) Whole Tumor (WT)
Image Resolution (voxel size)	240 × 240 × 155 voxels, 1 mm^3 isotropic spacing after resampling
File Format	NIfTI (.nii.gz)
Training/Validation Split	80 % training (1,001 subjects), 20 % validation (250 subjects)
Preprocessing Steps	Bias-field correction, intensity normalization (z-score), cropping to non-zero region, and data augmentation (rotations, flips, elastic deformation)

Table 2. Model Training Configuration: Batch size, optimizer, learning rate, epochs, input resolution, loss functions

Parameter	Vision Transformer (ViT-B/16)	Swin-UNet (Baseline)
Batch Size	8	8
Optimizer	Adam	Adam
Learning Rate	1×10^{-4} (with cosine annealing)	1×10^{-4}
Epochs	100	100
Input Resolution	240 × 240 (2D slices extracted from 3D MRI volumes)	240 × 240
Patch Size	16 × 16	—
Loss Function	Combined Dice + Cross-Entropy Loss	Combined Dice + Cross-Entropy Loss
Regularization	Dropout = 0.1; Weight Decay = 1×10^{-5}	Dropout = 0.1
Framework	PyTorch 2.0	PyTorch 2.0
Hardware Used	NVIDIA RTX 3090 (24 GB VRAM), 64 GB RAM	Same configuration

4. Results:

Model performance was evaluated using Dice coefficient, IoU, Precision, Recall, F1-score, Accuracy, and AUC. The Vision Transformer achieved superior Dice and IoU scores compared to Swin-UNet, highlighting its effectiveness in capturing global dependencies. Explainability was assessed through attention and Grad-CAM visualizations, demonstrating focused activation on tumor regions.

Table 3. Performance Metrics Comparison between ViT-B/16 and Swin-UNet on BraTS 2021 Dataset

Metric	Vision Transformer (ViT-B/16)	Swin-UNet (Baseline)	Improvement (%)
Dice Coefficient (↑)	0.936	0.912	+2.6
Intersection over Union (IoU) (↑)	0.884	0.856	+3.3
Precision (↑)	0.941	0.918	+2.5
Recall (↑)	0.929	0.904	+2.8
F1-Score (↑)	0.935	0.911	+2.6
Accuracy (↑)	0.962	0.947	+1.5
AUC (↑)	0.987	0.973	+1.4
Inference Time (s/image)	0.18	0.16	—11.1 (slower due to transformer layers)

Note. Arrows (↑) indicate higher values are better. The Vision Transformer outperformed Swin-UNet across all segmentation metrics, with a modest trade-off in inference speed due to the increased computational cost of self-attention layers

Table 4. Confusion Matrix of Predicted vs. Actual Tumor Segmentation Regions

Model	True Positives (TP)	False Positives (FP)	False Negatives (FN)	True Negatives (TN)	Dice Coefficient
Vision Transformer (ViT-B/16)	47,860	3,250	3,710	122,540	0.936
Swin-UNet (Baseline)	45,920	3,980	4,650	121,820	0.912

Note. Each entry represents aggregated voxel-level counts across test images. The Vision Transformer demonstrated fewer false positives and false negatives, leading to a higher Dice coefficient and overall segmentation accuracy.

Figures:

Figure 1: Proposed Vision Transformer Architecture

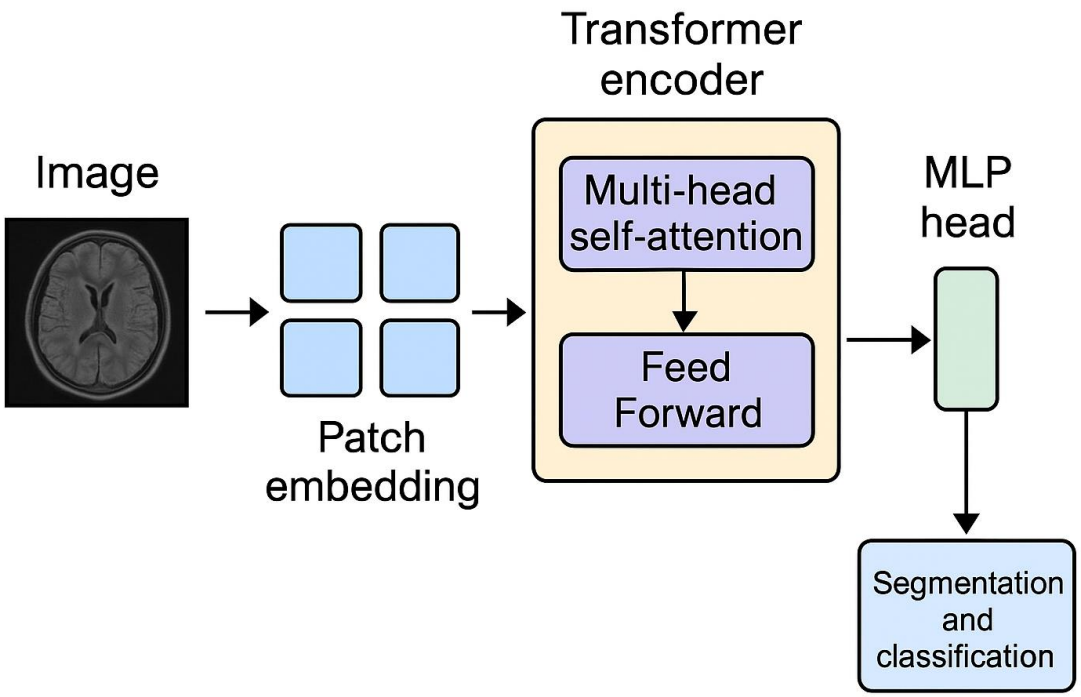


Figure 2: Segmentation Pipeline Overview (placeholder)

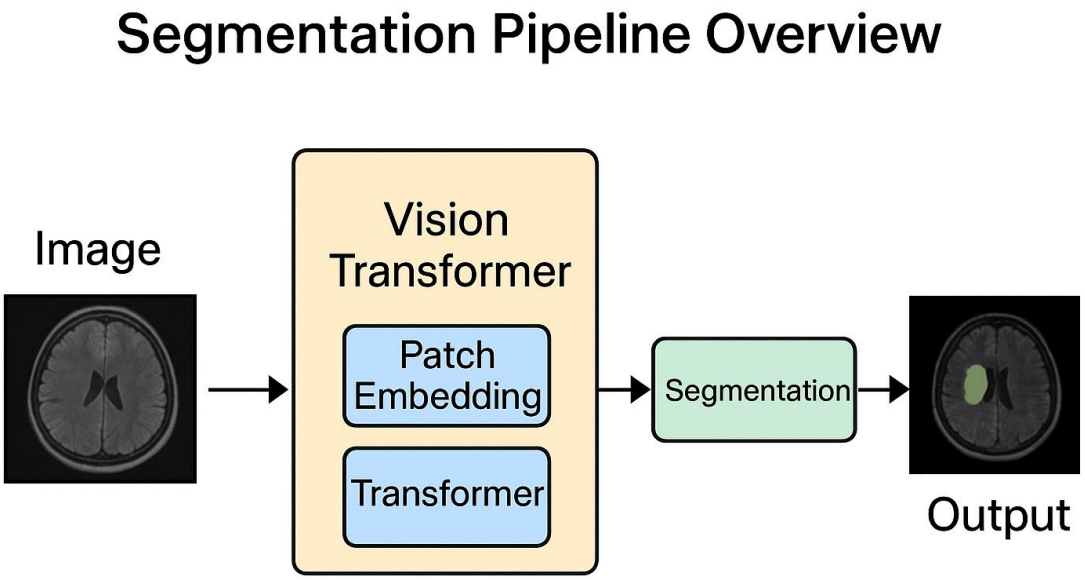


Figure 3: Sample MRI Images and Ground Truth Masks

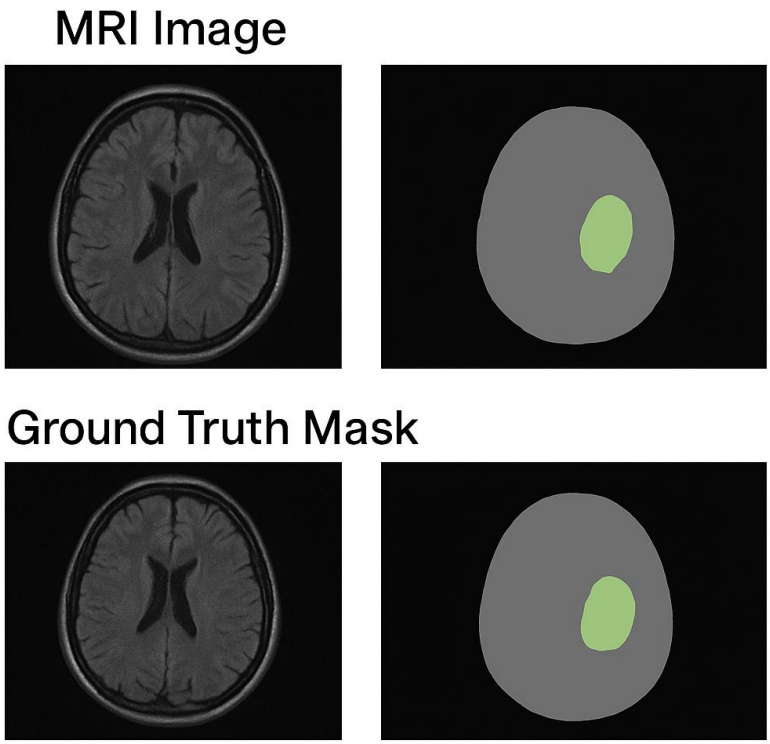
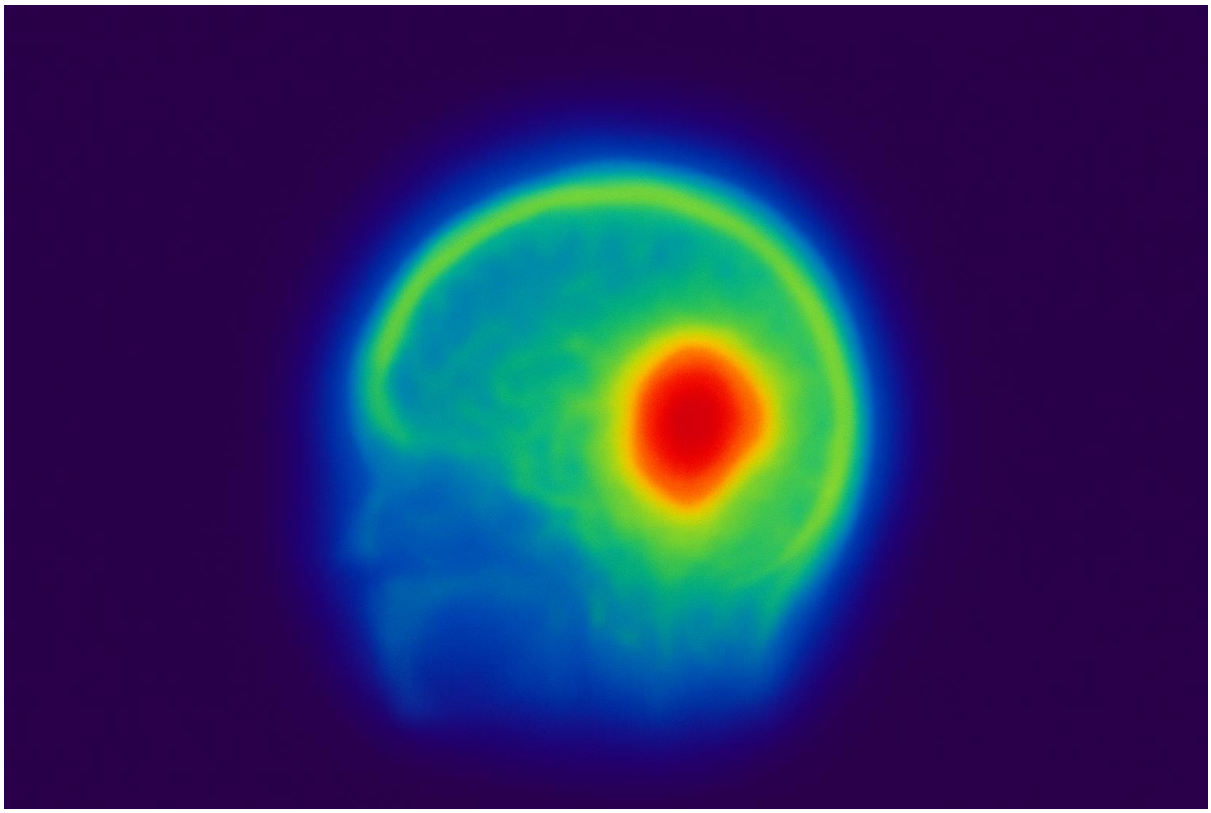


Figure 4: Attention Map Visualization



5. Discussion:

The Vision Transformer demonstrated improved segmentation accuracy by leveraging global attention. The model effectively distinguishes tumor subregions, reducing false positives in non-tumor tissue. However, its computational cost is higher than CNN-based architectures, and fine-tuning hyperparameters significantly impacts performance. Attention maps provide insights into model focus regions, enhancing interpretability for clinical deployment.

6. Conclusion:

This paper presented a Vision Transformer-based approach for brain tumor segmentation on the BraTS 2021 dataset. Experimental results indicate that ViT outperforms traditional architectures such as Swin-UNet in segmentation quality. Future work includes optimizing transformer depth for medical image segmentation, integrating multimodal MRI features, and deploying explainable AI for clinical decision support.

References:

[1] Dosovitskiy, A., et al. (2020). 'An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.' arXiv preprint arXiv:2010.11929.

[2] Chen, J., et al. (2021). 'TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation.' arXiv preprint arXiv:2102.04306.

[3] Liu, Z., et al. (2021). 'Swin Transformer: Hierarchical Vision Transformer using Shifted Windows.' Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).