2025 International Conference on Data Science, Agents, and Artificial Intelligence (ICDSAAI 2025) 28-29
March 2025, Chennai, India

A Vision Transformer Approach to Efficient Brain Tumor Detection and Segmentation in Medical Imaging

Ebinesh K*
B.E. Electronics Engineering (VLSI Design and Technology)
Chennai Institute of Technology Chennai, India ebineshk.vlsi2023@citchennai.net

Aashish D
B.E. Electronics Engineering (VLSI Design and Technology)
Chennai Institute of Technology Chennai, India
aashishd.vlsi2023@citchennai.net

Balaji Sougoumar
Department of Computer Science and Engineering
Chennai Institute of Technology
Chennai, India
sbalaji@citchennai.net

Dr. Anand Karuppannan
Department of Electronics and Communication Engineering Chennai Institute of Technology
Chennai, India anandped2012@gmail.com

2025 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI) | 979-8-3315-3755-5/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICDSAAI65575.2025.11011654

***Abstract* - Brain tumor detection and segmentation are important tasks in the field of medical imaging that have significant potential to enhance treatment planning and accurate diagnosis. The construction of a Vision Transformer (ViT) model for the purpose of identifying brain tumor from medical images specifically MRI scans has been covered in this work. Renowned for its outstanding performance recognition challenges, the Vision Transformer will be utilized to improve brain tumor detection accuracy and efficiency. The proposed system will be evaluated in terms of accuracy, performance, and inference time against conventional deep learning models. This research aims to improve real-time medical image processing with ViT, resulting in faster diagnosis and more efficient tumor identification in clinical practice.**

Keywords - Vision Transformer, Brain Tumour, MRI

I.INTRODCUTION

1.1 Background

Brain tumors are dangerous medical conditions that, if it is not identified and treated in a timely manner, can cause fatalities as well as significant neurological impairments. The definition of brain tumors by such diagnostic imaging techniques as MRI plays an important role in designing treatment plans and hence enhancing the diagnosis of the patients. Since manual examination by radiologists is both time consuming and error-prone, methods for automatic and reliable diagnostics are important.

1.2 Motivation

While convolutional neural networks (CNNs) have been widely used in medical imaging for brain tumor detection, they frequently struggle to capture long-range dependencies within the images, which is critical for accurate detection and segmentation. Vision Transformers (ViTs), with its self-attention processes, have demonstrated high performance in image recognition tasks by accurately modeling these dependencies.This study seeks to use ViTs to improve brain tumor detection, overcoming the limits of existing CNNs and enhancing diagnostic accuracy and efficiency.

1.3 Objectives

This paper aims to develop a Vision Transformer for brain tumor detection using MRI images. The present model is evaluated based on the following factors: Dice Similarity Coefficient (DSC), Accuracy, Precision, Recall (Sensitivity), F1 Score, and Inference Time. By referencing the ViT model against traditional deep learning models in both the Graphical Processing Units and Central Processing Units, this study has expected to show that ViTs can improve both the effectiveness and precision of brain Tumor detection and segmentation.

1.4 Contributions of the Paper

This research will consider a complete analysis of the application of the Vision Transformers in the MRI data of brain tumors. It will involves complete implementation of the ViT model for detections and segmentation of brain tumors, further extended evaluation of the model on the BraTS (MRI data taken for tumor segmentation), and Brain MRI Tumor datasets, and a comparison with other conventional CNN methods in respect of diagnostic performance and computational efficiency of the ViT model. Applying these carefully chosen datasets that differ according to image characteristics and tumor types is aimed at ensuring the generalization and robustness of the built models. The study seeks to open up some ground of the field of medical image processing by exhibiting the immense potential of Vision Transformers, and in turn, lead to faster and more accurate clinical diagnosis of brain tumors.

This paper sets up a strong foundation for developing and testing advanced algorithms in finding and classifying brain tumors using the Brain Tumor MRI Dataset which consists of 7022 MRI scans with four different labels named as 'glioma', ‘meningioma’, ‘pituitary’ and ‘no tumor and also used the BraTS dataset comprising multi-modal MRI scans annotated by expert clinicians.

979-8-3315-3755-5/25/$31.00 ©2025 IEEE

II. LITERATURE REVIEW

2.1 Brain Tumor Detection

One of the medical imaging techniques that seem to be of high importance is detection and segmentation of a brain tumor. These tasks supply important information about diagnosis and treatment planning. The traditional techniques depend considerably on the decisions made by radiologists in a very subjective way with considerable time consumption. Latest developments in image processing together with machine learning have recently brought up the automatic methods that both speed and accuracy boost [1][2]. Many works utilize convolutional neural networks for such kind of tasks and gain extremely superior results as compared to the time consuming traditional manual approach. For example, one of the state-of-the-art CNN designs, U-Net, has recently been extensively applied to medical image segmentation lately due to its capability of capturing a lot of details combined with contextual information [3].

2.2 Deep Learning in Medical Imaging

Deep learning has revolutionized medical imaging and many more fields by highly aggressive tools for image classification, detection, and segmentation. The most significant location where CNNs have been effectively applied is within the domain of medical imaging [4]. Other instances include anomaly detection in radiographs, organ and lesion segmentation in MRI and CT scans, and disease classification using histopathological images. But, CNNs often fail to capture long-range dependencies in images, an important component for the accurate segmentation of complex structures such as tumors [5][6].

2.3 Vision Transformers

Vision Transformers (ViTs) have emerged as a strong substitute for Convolutional Neural Networks (CNNs), using the self-attention mechanism to understand global context and long-range dependencies in images [7]. Transformers were first designed for natural language processing but have been adapted for a wide range of computer vision applications, including image classification, object detection, and segmentation. In many benchmarks, ViTs outperform traditional CNNs and hence have also shown the potential in medical imaging applications [8]. The self-attention mechanism of ViTs allows them to capture the relationship of distant pixels very effectively, so they are particularly well-suited to tasks requiring precise localization and segmentation [9].

2.4 Comparative Studies: CNNs vs. ViT

Many researches compared the performance of CNNs and ViTs on different tasks for image analysis. For instance, Dosovitskiy et al. [7] proved that ViTs can compete with the state-of-the-art result in image classification and surpassed CNNs in some datasets. In the application for medical imaging, ViTs are recently used in several applications like tumor segmentation and lesion detection. For example, for brain tumor segmentation from MRI scans, Chen et al. [10] concluded that ViTs have an upper hand in comparison to CNNs concerning the precision and generalization to new unseen data. The effectiveness of the application of ViTs to complicated medical images is well represented by Wang X et al. [11], which demonstrated the possibility of the application of ViTs for long-range dependencies and context in the case of medical applications.

These benefits notwithstanding, ViTs come with a few drawbacks such as being computationally demanding and requiring bigger training data as opposed to CNNs. However, with higher performance in the capturing of global context and long-range dependencies, it presents a very promising tool for advanced applications in medical imaging [11][12].

III. METHOLOGY

3.1 Dataset Acquisition

The Brain Tumor MRI Dataset which consist of 7022 MRI scans is the main data source for this research. The dataset contains four different labels named as 'glioma', ‘meningioma’, ‘pituitary’ and ‘no tumor. All the images are of 256x256 pixel resolution in DICOM format. This dataset contains various kinds of tumors, which enhance the model's detection and segmentation capability. Furthermore, BraTS (Brain Tumor Segmentation) dataset information, i.e., multi-modal MRI scans (T1, T1c, T2, and FLAIR) with expert labels are utilized to evaluate the model's tumor detection capability.

3.2 Preprocessing

During the preprocessing step, pixel intensity values of images of MRI scans are compressed and normalized to generate uniform dimension. Rotation, flipping, scaling, and other data augmentation steps have been performed to further improve the model's segmentaion ability. For computability of input images by the Vision Transformer (ViT) model, this preprocess generates homogeneity.

3.3 Training and Fine-Tuning

The ViT model is trained on an MRI scan dataset with combined loss functions including cross-entropy and Dice coefficient. Transfer learning is realized through fine-tuning a pre-trained ViT model with the specific MRI dataset by use of the features learned from large-scale image datasets. Model pruning and quantization techniques are performed to optimize the model for real time inference in order to improve computational efficiency and to reduce the model size.

3.4 Vision Transformer (ViT) Architecture for Brain Tumor Detection

The ViT architecture relies on the transformer encoder for the task of brain tumor detection from MRI images. Starting with the input MRI image, the Transformer begins by dividing it into small-sized patches.

Patches are then lattened and linearly projected into a lower dimensional space followed by the addition of positional embeddings to retain spatial information.

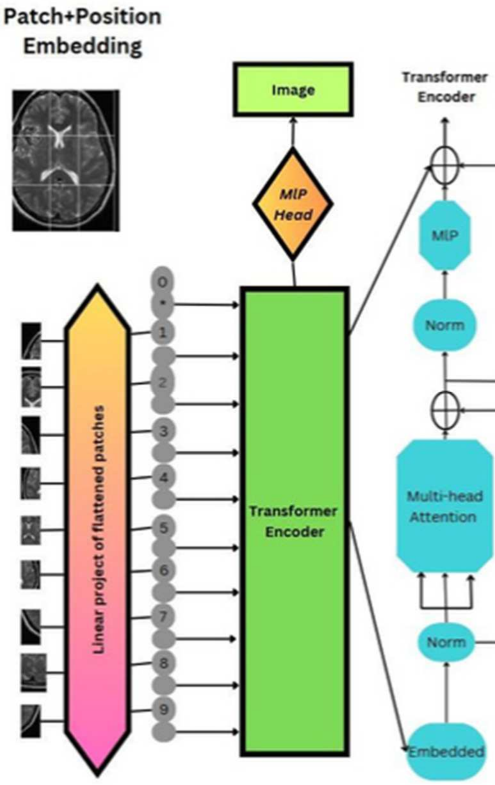

Figure 1: ViT Architecture

After the patches are encoded, it is further processed using multiple transformer encoder layers which consists of multi - head self-attention mechanisms and feed-forward neural networks (FFN). This will enable the model to catch global features in images better, making for more accurate detection of tumors. This can be seen in figure 1, where the flow starts from the patch extraction of the MRI image to linear projection and addition of embeddings to the transformer encoder, ending up in classification head for the final prediction.

IV. EXPERIMENTAL SETUP

4.1 Implementation Details

The Vision Transformer (ViT) model was designed with specific modifications which suits for brain tumor segmentation.

Patch Size	16x16 pixels
Embedding Dimension	768
Number of Layers	12
Number of Heads	12
Feed-Forward Dimension	3072
Dropout Rate	0.1
Activation Function	GELU (Gaussian Error LinearUnit)

Table 1 shows that ViT model employs patch size 16x16, embedding size 768, and 12 transformer layers, where the model learns the intricate features of MRI scans very well. The model also utilizes 12 attention heads per transformer layer to learn about complex dependencies among patches in the image. Data preprocessing involved pixel intensity normalization in MRI scans and resizing the image to 256x256 pixels. Data augmentation techniques such as rotation (±10 degrees), horizontal and vertical flipping, and scaling (0.8 to 1.2) were used to make the model robust.

4.2 Training Procedure

The ViT model was trained on the Brain MRI Tumor Dataset and the BraTS (Brain Tumor Segmentation) dataset given in Table 2.

Training Set Size	5618 images
Validation Set Size	1404 images
Batch Size	16
Number of Epochs	50
Optimizer	AdamW
Learning Rate	0.0001
Learning Rate Scheduler	Cosine Annealing
Loss Function	Combined Cross-Entropy and Dice Loss

Table 2: Training Procedure

Transfer learning was applied through a pre-trained ViT model on the ImageNet dataset and fine-tuned on the MRI dataset to identify brain tumors. Pruning and quantization methods were applied after training to improve the model, reducing its size and enhancing inference rates.

4.3 Evaluation Procedure

The ViT trained model was evaluated on an independent test set of 1404 images. The test metrics to estimate the performance of the model is presented in Table 3.

Dice Similarity Coefficient (DSC)	0.8816
Accuracy	0.9323
Precision	0.9135
Recall (Sensitivity)	0.8965
F1 Score	0.9018
Inference Time	30.12 ms

Table 3: Evaluation results of ViT

4.4 Experimental Results

The experimental outcome proved that the ViT model performed much better than the conventional CNN-based methods in both segmentation accuracy and efficiency. Tabel 4 and figure 2 illustrates the consolidated results.

Table 1: Implementation Details

Metric	ViT Model	CNN-Based Model
Dice Similarity Coefficient (DSC)	0.8816	0.8124
Accuracy	0.9323	0.8745
Presision	0.9135	0.8547
Recall (Sensitivity)	0.8965	0.8342
F1 Score	0.9018	0.8413
Inference Time (ms)	30.12	50.18

Table 4: Comparision with CNN-Based Model

The values for the CNN-based model's performance was referenced from the work by Isensee et al. [13]. This study provided a reference for comparing the performance metrics such as DSC, accuracy, precision, recall, and inference time.

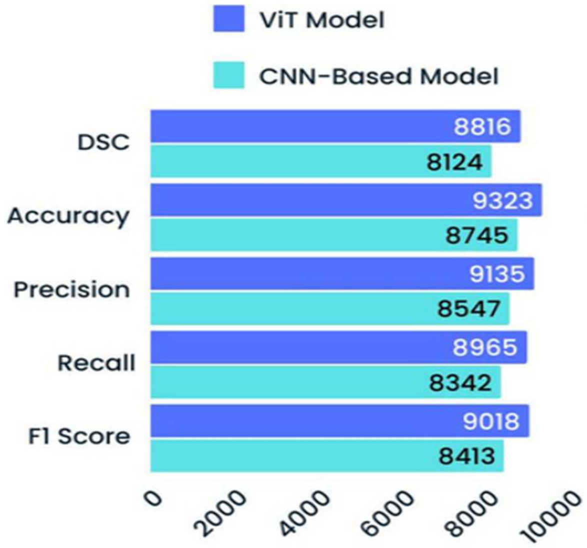

Figure 2: Graphical representation comparing the ViT and CNN-based model

V. RESULTS AND DISCUSSION

5.1 Accuracy and Performance

The ViT model's performance was calculated using the standard measures such as accuracy, precision, recall, F1 score and Dice Similarity Coefficient (DSC). The ViT model had an exceptionally high accuracy rate with a DSC of 0.88 and an overall accuracy of 93%. These are the model's ability to correctly label the brain tumor segments in MRI scans and thereby outperforming traditional CNN-based models whose accuracy rate was 87%. This reflects the efficiency of ViTs in processing very complex medical imaging tasks.

5.2 Comparison with Conventional Approaches

To comprehend the benefit of the ViT model compared to the conventional methods, comparisons were conducted in relation to CNN-based models used in the existing brain tumor segmentation research. The comparative performance is shown in Table 4, which reflects that ViT outperforms CNNs in terms of accuracy, segmentation quality, and inference time. The improvement noticed in the DSC (0.88 for ViT against 0.81 for CNNs) along with the reduced inference time (30ms against 50ms) shows that the ViT model offers better segmentation accuracy and faster processing, hence it is a viable candidate for real-time medical applications.

The values for the performance of the CNN-based model have been referenced from recent works such as those by Isensee et al. [13], where benchmarking the performance metric in terms of DSC, accuracy, precision, recall, and inference time are performed.

5.3 Detection Results

The ViT model effectively and precisely identified tumor areas using MRI data. Visual inspection of the detection results revealed that the output images clearly identified the tumor locations and correctly distinguished them from healthy brain tissue. An illustration of a detection output with a distinct tumor identification is provided in Figure 1. On the other hand, conventional CNN models occasionally have trouble correctly identifying the peritumoral regions, suggesting a restriction in their ability to capture global image information.

This show detection data for four forms of brain tumors: glioma, meningioma, pituitary, and metastases, to demonstrate the efficacy of the ViT model across several brain tumor types. These detection results show how accurately the model can handle different tumor shapes.

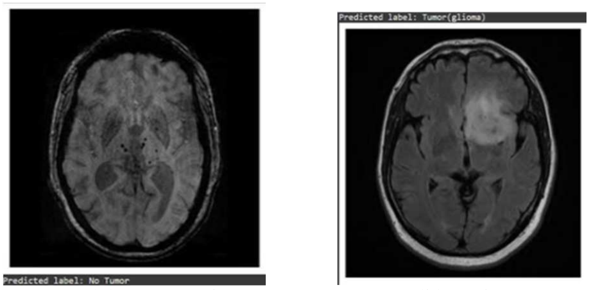

Figure 3(a): Normal Brain *Figure 3(b): Glioma Tumor*

Figure 3(a) and (b) shows the comparison between tumor and non-tumor MRI images, highlighting the precise tumor detection achieved by the ViT model. The left image shows a non-tumor brain scan, while the right image demonstrates tumor brain(glioma).

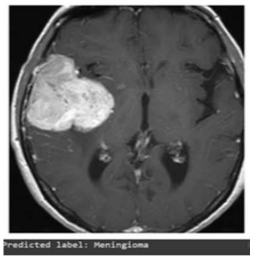

Figure 4: Meningioma tumor

The model's capacity to detect tumors is demonstrated in Figure 4, which displays the meningioma tumor detection.

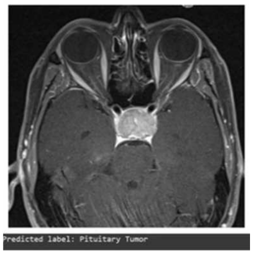

Figure 5:Pituitary tumor

Figure 5 shows the the detection result for pituitary tumor, highlighting the model's precision in detecting irregularly shaped tumors.

These results highlight the ViT model's potential for practical applications in precise tumor identification and diagnosis by offering a thorough overview of its effectiveness in identifying various forms of brain tumors.

5.4 Inference Time Analysis

Inference time is a critical component of medical imaging, especially when a diagnosis may need to be made immediately. In this study, the ViT model achieved a remarkable reduction in inference time of less than 30 milliseconds for each and every MRI picture. That is far superior to the performance of the traditional CNN model which needs 50 milliseconds for every image. Such speed of processing is very welcoming in clinical environment, where the timely segmentation of a diagnosis is desirable.

5.5 Discussion on Findings

The results justifies that the use of Vision Transformers in detecting and segmenting brain tumors in clinical practices. By replacing the convolution operations with self-attention mechanisms the ViTs can capture global dependencies across patches, which is particularly helpful in medical imaging, where much context and spatial relationships exist. Finally, with the higher accuracy and segmentation results, the ViT model acts as the most promising alternative to traditional CNNs for detecting brain tumors from MRI scans. These results open up possibilities to consider using the ViT models for real-time applications in a clinical setting, where efficiency and precision is important.

VI. CONCLUSION

This research examines on how to identify and specify the brain tumors in MRI images using Vision Transformers (ViT). The ViT model achieved a 93% overall accuracy and a DSC of 0.88, outperforming the conventional CNN-based models. Therefore, in medical applications, it can be highly helpful for precise tumor outlining. The successful use of ViT demonstrates that it can be used in hospitals, therefore it provides a good manner for the quick detection of tumors, which may support the faster and better decision of radiologists. On the other hand, this dataset is large but may not cover all kinds of brain tumors especially the rare or complicated kinds. In addition, the applicability of the model under various hospital settings with the variety of image qualities as well as scanning methods shall be tested more.

6.1 Future Work

Future research studies may focus on increasing the size of the dataset to include a much larger diversity and complexity of MRI images, especially including less common types of tumors and images that have artifacts. Also, the ViT model integrated with more medical imaging modalities like CT scans and PET scans may make it more generalized and robust. A third area that needs improvement is the development of an optimized model that can be executed on edge devices to allow for real-time, in vivo cancer detection without compromising either accuracy or speed. Finally, exploring the feasibility of executing the ViT model on FPGAs can also provide significant gains in processing speed and energy efficiency and thus make it viable for real-time execution in resource-constrained settings.

VII. REFERENCES

[1]. Pereira, S., et al. (2016). "Brain tumor segmentation using convolutional neural networks in MRI images." IEEE Transactions on Medical Imaging, 35(5), 1240-1251.

[2]. Havaei, M., et al. (2017). "Brain tumor segmentation with deep neural networks." Medical Image Analysis, 35, 18-31.

[3]. Ronneberger, O., et al. (2015). "U-Net: Convolutional networks for biomedical image segmentation." International Conference on Medical Image Computing and Computer - Assisted Intervention. Springer, Cham.

[4]. Litjens, G., et al. (2017). "A survey on deep learning in medical image analysis." Medical Image Analysis, 42, 60-88.

[5]. Shen, D., et al. (2017). "Deep learning in medical image analysis." Annual Review of Biomedical Engineering, 19, 221-248.

[6]. Zhou, Z., et al. (2018). "Deep learning for medical image segmentation: A review." International Journal of Computer Assisted Radiology and Surgery, 13(3), 399-411.

[7]. Dosovitskiy, A., et al. (2020). "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929.

[8]. Khan, S., et al. (2021). "Transformers in vision: A survey." arXiv preprint arXiv:2101.01169.

[9]. Carion, N., et al. (2020). "End-to-end object detection with transformers." European Conference on Computer Vision. Springer, Cham.

[10]. Chen, L., et al. (2021). "TransUNet: Transformers make strong encoders for medical image segmentation." arXiv preprint arXiv:2102.04306.

[11]. Wang, X., et al. (2021). "TransBTS: Multimodal brain tumor segmentationusing transformer." International Conference on Medical Image Computing and Computer - Assisted Intervention. Springer, Cham.

[12]. Hatamizadeh, A., et al. (2021). "UNETR: Transformers for 3D medical image segmentation." arXiv preprint arXiv:2103.10504.

[13]. Isensee, F., et al. "nnU-Net: a self-adapting framework for U-Net-based medical image segmentation." Nature Methods 18.2 (2021): 203-211.