Position Papers of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 41–48
DOI: 10.15439/2025F7399
ISSN 2300-5963 ACSIS, Vol. 44

RAG^4-Unet: An Approach for Recognition and Segmentation of Brain Tumor in MRI Scans

Ameer Hamza
Centre of Real Time Computer Systems
Kaunas University of Technology
Kaunas, Lithuania
ameer.hamza@ktu.edu

Robertas Damaševičius
Centre of Real Time Computer Systems
Kaunas University of Technology
Kaunas, Lithuania
robertas.damasevicius@ktu.lt

***Abstract*—We propose a novel U-net architecture, RAG^4-Unet, based on residual attention gated for brain tumor segmentation, Swin transformer for classification task, and Yolo11 for tumor detection. For the experiments, the Figshare dataset is employed and the proposed architecture achieved 91.37% Dice for tumor segmentation task, and Swin transformer achieved 91.74% classification accuracy. The Yolo11 gained 89.6% of detection precision. Comparative evaluation with the SOTA techniques reveals that the proposed architecture outperformed the existing methods and Yolo11. The proposed architecture improved the tumor boundary detection, making it a promising solution for brain tumor recognition and segmentation.**

***Index Terms*—Tumor Segmentation, Residual Attention Gated, Unet, Yolo11, Attention Maps.**

I. Introduction

BRAIN tumors are a major health challenge, characterized by abnormal cell growth in the brain, which can affect its vital functions [1]. These tumors, from benign to malignant, are often associated with persistent headaches, seizures, cognitive impairments, and neurological diseases and have a negative impact on the quality of life of patients [2]. Manual diagnostic methods such as visual inspection of histological slides and radiological imaging have traditionally been used, but they take time, are subjective, and can cause human error [3]. Radiologists are imaging modalities more frequently because they tend to be more accurate and put patients at far lower risk. Medical imaging data can be recorded using a variety of techniques, such as tomography [4], magnetic resonance imaging (MRI) [5], radiography [6], and echocardiography [7].

The introduction of artificial intelligence (AI) has transformed the detection of brain tumors through automation and improved diagnostic accuracy [8]. Machine Learning ML) approaches depend on methods for gathering features, selecting features, and classification [9]. Deep Learning (DL) models learn by extracting features from images. Particularly, Convolutional Neural Networks (CNNs) are widely used in medical imaging analysis and show vital achievements in the identification of brain tumors, enabling advances in classification and segmentation. Several studies have proposed innovative methods for the segmentation and classification of brain tumors from MRI images. Zhang et al. [10] introduced a modified U-net method with an attention mechanism for improving segmentation accuracy. Their approach focused on addressing limitations of traditional U-net models, such as difficulties in handling small tumor regions and blurry tumor boundaries. By incorporating multi-scale feature fusion and attention mechanisms, their method demonstrated enhanced efficiency and achieved Dice coefficients of 0.876, 0.868, and 0.814 for tumor subregions.

Ahsan et al. [11] compared object detection algorithms (YOLOv5, Faster R-CNN, SSD) for brain tumor. They used Figshare dataset and paired YOLOv5 with 2D U-Net for segmentation. Yolov5 gained the highest mAP of 89.5%, and Yolov5+2D U-Net achieved 88.1% DSC. However, the dual-model framework increased learning complexity.

Arumaiththurai et al. [12] proposed two methods for classifying brain tumors using ML and DL algorithms. The first method used decision trees and SVM, while the second used pre-trained VGG19 and ResNet152 models. Figshare brain tumor dataset assessed the effectiveness of these approaches. The CNN-based method performed better in classification and attained an accuracy rate of 94.67%.

Alyami et al. [13] employed AlexNet and VGG19 models for feature extraction and the slap swarm algorithm for feature selection. They used Kaggle brain tumor dataset and achieved an accuracy of 99.1% with a cubic SVM using 4111 best selected features out of 8192.

Asiri et al. [14] introduced a customized CNN model for classification brain tumor, focusing on hyperparameter tuning of kernel size, strides, activation, and learning rates. The model was evaluated on two MRI datasets: a four-class dataset with 7,023 images and a binary dataset with 253 images. This method achieved 88% accuracy.

These studies demonstrate the growing use of deep learning models, particularly U-net, transfer learning, and attention mechanisms, to enhance the accuracy and efficiency of brain tumor segmentation and classification. The incorporation of explainable AI such as LIME, attention maps, also adds a layer of transparency, which is crucial for the deployment of these models in clinical settings.

However, challenges such as variability in growth patterns, textures, and irregularity in tumors across patients, and different tumors have overlapping visual features and irregular

Topical area: Advanced Artificial Intelligence in Applications

boundaries, especially when the tumors are in early stages. Addressing these challenges remains a critical focus of ongoing research on brain tumor analysis.

To address these challenges, we introduce an innovative architecture, RAG^4-Unet, for the segmentation process, and this framework incorporates Swin transformer and Yolo11 for precision detection of the tumor region. The key contributions of this work is summarized as follows:

- We introduce a novel Residual Attention Gated (RAG) module to focus on significant spatial and contextual features to enhance detection of brain tumor boundaries.
- We employ a Swin Transformer to leverage shifting window sizes, utilizing its attention mechanism to learn features hierarchically.
- We integrate YOLO11 for detection of growth regions in brain tumors, enhancing accuracy of tumor detection.

II. Methodology

A. Data Collection and Augmentation

The FigShare Brain tumor dataset is utilized for experiments. The dataset is available at https://Figshare.com/articles/dataset/brain_tumor_dataset/1512427. This dataset includes 233 patients with three types of tumors: glioma, meningioma, and pituitary tumor. The glioma category contains 1426 slices, meningioma has 708 slices, and the pituitary tumor has 930 slices. Each image has a dimension of 512×512 with a depth resolution of 96 dpi. The dataset is imbalanced and that there were not enough samples for the efficient learning of the deep learning model. Therefore, we performed an augmentation process to increase the diversity in the dataset. For the augmentation process, four basic transformations are utilized: horizontal flip, rotation by 10°, vertical flip, and solarization. After augmentation process, the samples in each class are 4120. The augmentation process is visually presented in Fig. 1.

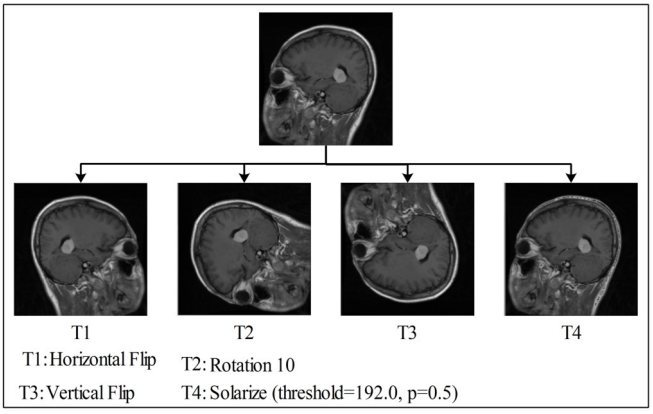

Fig. 1. Sample of augmentation operation on brain tumor dataset.

B. Overview of Swin Transformer

Swin Transformer is an enhanced version of the transformer that boosts computational effectiveness and capacity for high-resolution images. Similar to conventional CNNs, the Swin Transformer gradually reduces the image size by introducing a hierarchical structure that reflects images at various sizes. It limits attention to small windows and shifts these windows at every level. The input RGB image is separated into non-overlapping patches using a patch-splitting module like ViT. Each patch is handled as a "token" and its feature is configured as a concatenation of raw pixel RGB values. After that, many Swin Transformer blocks are applied to these patch tokens.

a) Swin Transformer Stages:

The Swin Transformer block, known as "Stage 1," maintains a token count of $\frac{\phi_h}{4} \times \frac{\phi_w}{4}$ when used with linear embedding. Hierarchical representation is achieved by reducing the number of tokens using a patch merging technique as the depth of the neural network grows. The initial patch merging layer concatenates the features of neighboring 2×2 patches and then applies a linear layer to the $4C$-dimensional features produced. This procedure reduces the token count by a factor of $2 \times 2 = 4$, while changing the output dimension to $2C$. Swin Transformer blocks are added to transform features while keeping a resolution of $\frac{\phi_h}{8} \times \frac{\phi_w}{8}$. Stage 2 begins with patch merging and feature transition. The procedure is performed twice, resulting in "Stage 3" and "Stage 4", with output resolutions of $\frac{\phi_h}{16} \times \frac{\phi_w}{16}$ and $\frac{\phi_h}{32} \times \frac{\phi_w}{32}$ correspondingly. The four stages work together to provide a hierarchical representation with feature map resolutions equivalent to those of typical CNNs. Swin Transformer replaces multi-head self-attention (MSA) module in a transformer block with a module based on the shifted window, while the other layers remain unchanged. The Swin Transformer block consists of a shifted window-based MSA module, a 2-layer MLP with GELU activation in between. The layer normalization is placed before each MSA and MLP module, followed by a residual connection.

b) Hierarchical Feature Learning:

The self-attention within localized windows enables effective modeling. The windows are positioned such that they do not overlap and divide the image equally. The computational complexity of a global MSA module and a window-based one, based on an image of $\phi_h \times \phi_w$ patches, assuming each window has $k \times k$ patches:

$$\Omega(MSA) = 4\phi_h\phi_w C^2 + 2(\phi_h\phi_w)^2 C \tag{1}$$

$$\Omega(MSA)_w = 4\phi_h\phi_w C^2 + 2k^2\phi_h\phi_w C \tag{2}$$

The computation of the Swin Transformer has linear complexity when fixed, but the computational cost of traditional ViT increases quadratically with the number of patches. Although the W-MSA of the Swin Transformer decreases the computational cost from quadratic to linear, its modeling capability may be limited by the absence of links and communication between many windows. To overcome this restriction, the Swin Transformer adds a shifted window divider that makes it easier for nearby non-overlapping windows to share information. In two successive Swin Transformer blocks, this method alternates between using W-MSA and a modified SW-MSA. By connecting adjacent non-overlapping windows, the shifted window partitioning greatly expands the receptive field. After

employing the shifted window divider, the computation within two consecutive Swin Transformers is followed as:

$$\hat{\Phi}^b = (MSA)_w(\text{LN}(\Phi^{b-1})) + \Phi^{b-1} \tag{3}$$

$$\Phi^b = \text{MLP}(\text{LN}(\hat{\Phi}^b)) + \hat{\Phi}^b \tag{4}$$

$$\hat{\Phi}^{b+1} = (MSA)_{sw}(\text{LN}(\Phi^b)) + \Phi^b \tag{5}$$

$$\Phi^{b+1} = \text{MLP}(\text{LN}(\hat{\Phi}^{b+1})) + \hat{\Phi}^{b+1} \tag{6}$$

Where $(MSA)_w$ and $(MSA)_{sw}$ represent window-based multi-head self-attention and shifted window divider, respectively. $\hat{\Phi}^b$ and Φ^b denote resultant features of the $(MSA)_{sw}$ and MLP module for block b.

Swin Transformer introduces the relative position biases for every head during the similarity calculation, which is formulated as:

$$Att(Q, K, V) = \text{Soft}\left(\frac{QK^T}{\sqrt{d}} + \psi_b\right)V \tag{7}$$

Where Q, K, V are the query, key, and value vectors, and d denotes the dimension of Q, K, V, and ψ_b is the bias vector.

The motivation behind choosing the Swin Transformer for brain tumor analysis is its ability to process high-resolution images and its window-based attention mechanism, which can learn fine-grained details about the tumor region, such as tiny tumor boundaries and growth patterns in the local context.

C. Proposed RAG^4-Unet Architecture

U-Net is a deep learning architecture proposed for image segmentation. It consists of three steps: encoder, bridge, and decoder. In this work, we proposed a Residual Attention-Gated U-Net (RAG^4-Unet) for the segmentation of tumors from brain MRI scans. RAG^4-Unet consists of four residual encoders, one bridge, and four residual decoders. All the residual encoders are connected to the attention gate to generate the attention maps, and the attention gates are concatenated with the decoders. The RAG^4-Unet accepts the input of size $256\times256\times3$.

a) Encoder Phase:

The first encoder consists of one residual block, max-pooling with stride 2, and one dropout layer. The dropout factor is 0.1. The residual block contains two convolutional layers with a 3×3 filter size, 64 filters, and a stride of 1. The residual encoder is desribed as follows:

$$\partial_1 = \emptyset_1(I) \tag{8}$$

$$\partial_1^{\psi} = \psi_1(\partial_1) \tag{9}$$

$$\partial_2 = \emptyset_2(\partial_1^{\psi}) \tag{10}$$

$$\partial_2^{\psi} = \psi_2(\partial_2) \tag{11}$$

$$\partial_3 = \emptyset_3(\partial_2^{\psi}) \tag{12}$$

$$\partial_3^{\psi} = \psi_3(\partial_3) \tag{13}$$

$$\partial_{\text{skip}} = \partial_3^{\psi} + \partial_2^{\psi} \tag{14}$$

$$\partial_{\text{ReLU}} = \lambda(\partial_{\text{skip}}) \tag{15}$$

$$\partial_{\boxplus} = \boxplus_{\text{Mpool}}(\partial_{\text{ReLU}}, s = 2) \tag{16}$$

$$\partial_{\boxminus} = \boxminus_{\text{drop}}(\partial_{\boxplus}, f = 0.1) \tag{17}$$

Where the $\emptyset_c$ represents the convolutional operation, ψ is the batch normalization, λ represents the ReLU activation, $\boxplus_{\text{Mpool}}$ is max pooling, and $\partial_{\boxminus}$ represents the dropout layer. The second and third encoders also consist of one residual block, max-pooling with stride 2, and one dropout layer with a 0.1 dropout factor. In the second residual block, the convolutional layer is configured with a 3×3 filter size, 128 filters, and a stride 1. In the third residual block, the convolutional operation is performed by employing a 3×3 kernel size, 256 filters, and a stride 1. In the last encoder, dropout factor is 0.2, and convolutional inside the fourth residual block is configured with a 1×1 kernel size, 512 filters, and a stride 1.

b) Bridge Phase:

The bridge between the encoder and decoder is the deepest point in the network. The bridge is configured by employing a residual block with 1024 depth and one dropout layer with a 0.3 drop factor.

c) Decoder Phase:

After the Bridge, the first decoder consists of one transpose convolutional layer configured with a 2×2 filter size, 512 depth, and 2×2 stride, one attention gate that is applied on the fourth encoder and transpose layer. The resultant feature map of attention-gated and transpose layers is further combined using the concatenation layer. After that, one residual block and dropout layer with a 0.2 drop factor is employed. The mathematical representation is:

$$\partial_{\text{Tconv}} = \emptyset^T(\beta, k = 2, s = 2, \text{ch} = 512) \tag{18}$$

$$\partial_{\text{AG}} = \text{AttGate}(\partial_{\text{end4}}, \partial_{\text{Tconv}}) \tag{19}$$

$$\partial_{\text{Con}} = \biguplus(\partial_{\text{AG}}, \partial_{\text{Tconv}}) \tag{20}$$

$$\partial_{d1} = \emptyset(\partial_{\text{Con}}) \tag{21}$$

$$\partial_{d1}^{\psi} = \psi_{d1}(\partial_{d1}) \tag{22}$$

$$\partial_{d2} = \emptyset_{d2}(\partial_{d1}^{\psi}) \tag{23}$$

$$\partial_{d2}^{\psi} = \psi_{d2}(\partial_{d2}) \tag{24}$$

$$\partial_{d3} = \emptyset_{d3}(\partial_{d2}^{\psi}) \tag{25}$$

$$\partial_{d3}^{\psi} = \psi_{d3}(\partial_{d3}) \tag{26}$$

$$\partial_{\text{skip}}^{d} = \partial_{d3}^{\psi} + \partial_{d2}^{\psi} \tag{27}$$

$$\partial_{\lambda}^{d} = \lambda(\partial_{\text{skip}}^{d}) \tag{28}$$

$$\partial_{\boxminus}^{d} = \boxminus_{drop}(\partial_{\lambda}^{d}, f = 2) \tag{29}$$

Where $\biguplus$ is the concatenation layer, AttGate is the attention mechanism, and $\emptyset^T$ represents the transpose convolutional operation. In the second decoder, the transpose convolutional has a 2×2 filter size, 256 depth, and 2×2 stride, and the remaining mechanisms are the same. The implementation phenomena of the third and fourth decoders are the same. However, the configurations of transpose convolutional and

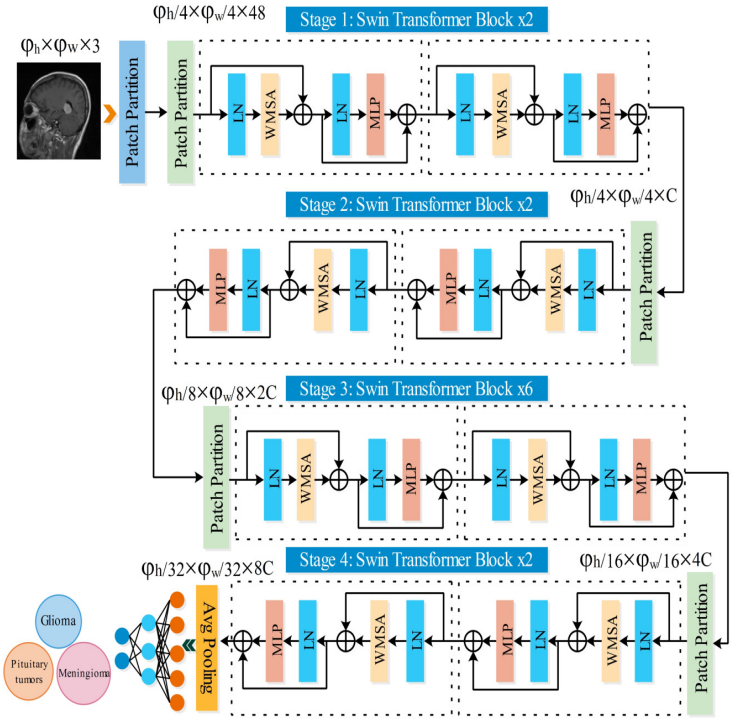

Fig. 2. Architecture of Swin Transformer for the classification of brain tumor

dropout layers are updated. The updated configurations are a 2×2 filter size, 128, 64 depth, and 2×2 stride, and the dropout factor is 0.1, respectively. The architecture of the proposed RAG^4-Unet is presented in Fig. 3.

The proposed model is developed using the TensorFlow framework and The proposed model has 99.45M parameters bringing the model size to around 379.37 MB of memory. 33.14M are trainable and 17.66K non-trainable parameters kept by the optimizer in memory 252.87 MB while training. The model inference complexity is evaluated with GFLOPS. The overall computation cost of a single forward pass is approximately 106.25 GFLOPs.

D. Novelty: Proposed RAG Module

In this work, We designed a novel hybrid feature enhancement module based on Residual and Attention gated mechanism. This module synergistically combines the residual learning to stabilizes the gradient flow, with attention gating, which focused on salient regions of the interest within the brain MRI image.The tumor regions often confused with healthy tissues. the RAG module addresses this problem by filtering irrelevant and low importance features while enhancing the high relevance activations related to tumor boundaries and cores. It enhances the boundary detail and localization of tumor objects, because the attention gated mechanism reduces irrelevant activations that strengthen the task of boundary detail and out-of-distribution activations that strengthens spatial detail of the tumor region when propagating features and helping to ensure gradient stability.

The sequence of RAG module begins with a series of convolutions to extract features from the input tensor, with the output then entering more convolutions and subsequently Attention Gated module, when performing spatial attention analysis, attention maps are created using extracted features and a gating signal is produced from a feature map. The attention maps are resampled, and modify the original feature map, it allows the network to learn where to increase and where to decrease specific spatial regions and a residual connection allows the network to skip non-linearity,if needed, thus minimizing the possibility of vanishing gradients strengthening the source of information and allowing for richer contextual experience for the network over numerous forward passes. The output of this module contains local enriched features and global semantic guiding features useful for precise identification of the tumor edges. the proposed RAG module is presented in Figure 4

III. Results

A. Experimental Setup

The dataset is divided into training, testing, and validation. 70% data is employed for training, 10% data is employed for the validation during the training process, and the 20% data is utilized for the testing process. The hyperparameters selected for Swin Transformer are batch size, number of workers, selected optimizer ADAM, learning rate, momentum, and epochs having values are 8, 4, 0.0004, 0.9, and 250. For RAG^4-Unet the utilized hyperparameters are learning rate is 0.0001, epochs is 100, optimizer is ADAM, batch size is 8,

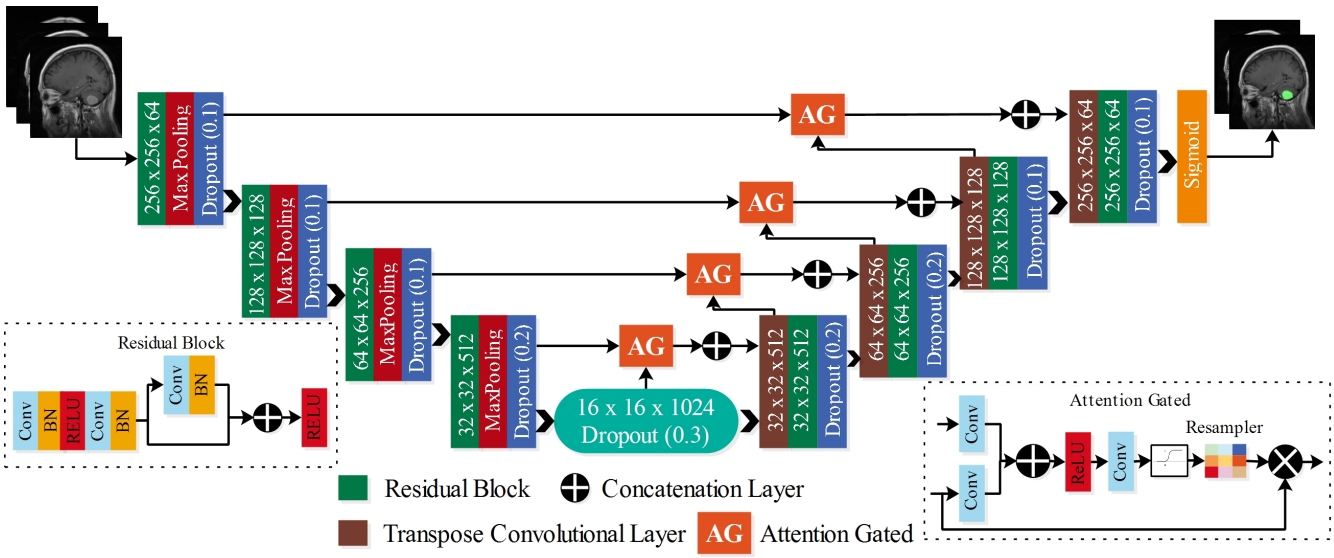

Fig. 3. Architecture of proposed RAG^4-Unet for brain tumor segmentation

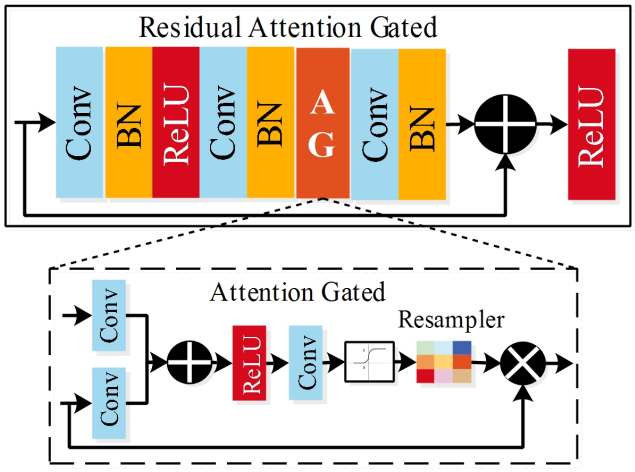

Fig. 4. Architecture of Residual Attention Gated mechanism (RAG) module

and early stopping is employed with learning decay is 0.2, patience is 5, min_lr is 0.00001. The evaluation metrics are accuracy, recall, precision, f1-score, Dice, Jaccard loss, and IoU for the segmentation and classification.

The experiments are conducted on MSI GL75 Leopard model configured with Core–i7 10 generation 2.59GHz processor, 16 GB of RAM, 512GB of SSD storage, and GTX GeForce 1660ti 6GB graphics card.

B. Results of Swin Transformer

The classification results of Swin Transformer on Figshare dataset has been presented in Table I. The model achieved 91.74% accuracy, 91.64% precision, 91.73% recall, 91.52% f1-score, 98.33% AUC, and 86.91% kappa index. The performance across the individual classes such as pituitary tumor gained the highest accuracy of 97.85%, precision of 95.13%, recall of 97.85%, and f1-score of 96.48% with 2.87 (sec) inference time. The confusion matrix gives more comprehensive details about the class's performance, as shown in Fig. 5. Glioma and pituitary tumor have the highest accuracy of 95.79%, and 97.85% respectively, because 205 samples of glioma and 137 samples of pituitary tumor class are correctly classified and 9 samples from the glioma and only 3 samples from the pituitary tumor are misclassified. The meningioma class has 75.47% of accuracy, 88.88% precision, 75.47% recall, and 81.63% f1-score. Meningioma tumor suffers from the considerable misclassification, the 5 samples are incorrect classified as pituitary tumor and 21 samples are misclassified as glioma. The overall misclassification rate of the meningioma class is higher than the other two classes. The overall confidence index of model is quite better which is 97.22%.

TABLE I
RESULTS OF SWIN TRANSFORMER ON FIGSHARE DATASET

Class-wise	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**
Glioma	95.794	90.707	95.794	93.181
Meningioma	75.471	88.888	75.471	81.632
Pituitary Tumor	97.857	95.138	97.857	96.478
Overall Performance				
Accuracy (%)	**Precision (%)**	**Recall (%)**	**F1-score (%)**	**AUC (%)**
91.74	91.64	91.73	91.52	98.33
Kappa (%)	**CI**	**Inference Time (sec)**		
86.91	97.22	2.306		

C. Results of proposed RAG^4-Unet

The segmentation is implemented using the proposed RAG^4-Unet model. The images and their masks are provided as input to the proposed model. After training the RAG^4-Unet model, the model is evaluated on the test data. The

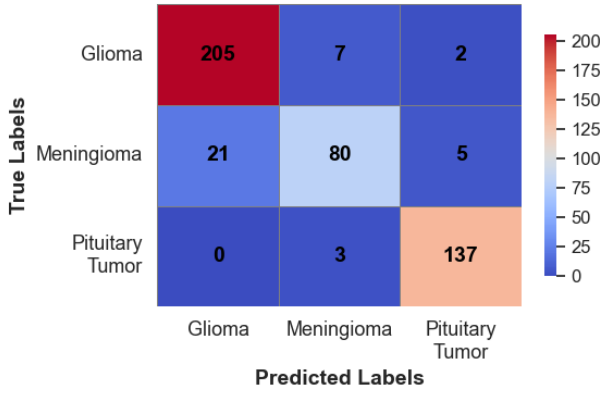

Fig. 5. Confusion matrix of Swin Transformer on Figshare dataset

overall performance and sample-wise results of the proposed RAG^4-Unet are presented in Table II. The proposed model achieved 91.37% Dice, 94.74% precision, 96.23% sensitivity, and 98.46% specificity. Some of the testing sample results are presented in Table II. Most test samples have a high Dice of 0.90, with consistent IoU and low Jaccard loss. The model could segment the tumor regions accurately and clearly distinguish the tumor portion from the surrounding information, such as samples 1, 2, 3, 4, 7, 8, 9, and 12 have more than 90% Dice score, 87-94% IoU, due to the clear and uniform morphology of the tumor region and a few samples, like 10, 11, 13, and 14, have quite better Dice scores and IoU with the small size of the tumor. However, the model is struggling with samples that do not clear the tumor boundary because the results are leading to under or over-segmentation, like in samples 5 and 6.

TABLE II
SEGMENTATION RESULTS OF PROPOSED RAG^4-UNET BASED ON FIGSHARE DATASET

Sr.	**Dice**	**Jaccard Loss**	**IoU**	**Sr.**	**Dice**	**Jaccard Loss**	**IoU**
1	0.906	0.171	0.828	2	0.931	0.127	0.872
3	0.943	0.106	0.893	4	0.948	0.097	0.902
5	0.649	0.519	0.481	6	0.782	0.357	0.642
7	0.971	0.055	0.944	8	0.945	0.10	0.896
9	0.950	0.094	0.905	10	0.957	0.081	0.918
11	0.960	0.076	0.923	12	0.971	0.054	0.945
13	0.929	0.131	0.868	14	0.970	0.056	0.943
Overall Performance							
Dice	**Dice Loss**	**Precision**	**Sensitivity**	**Specificity**			
0.9137	0.0863	0.9474	0.9623	0.9846			

Fig. 6 presents a visual comparison of the original ground truth and predicted ground truth with the overlap maps for further investigation of the above samples. In overlapping maps, the green region indicates the original mask, the red region demonstrates the predicted mask, and the yellow region indicates the perfect match between the predicted and original masks. In addition, in the last column of Fig. 6, the attention maps are generated by the proposed model to further evaluate the transparency. The generated attention maps highlighted the focus of the RAG^4-Unet during the segmentation process. These maps show the areas of the segmentation process where the model concentrates. The tumor locations are prominently highlighted in the attention maps, signifying that the model effectively suppresses background noise and prioritizes relevant areas. For samples 5 and 6, as shown in Fig. 6, the model generated a weaker or scattered focus, which indicated low performance. For the overall performance, generating attention maps are a suitable instrument for interpreting the decision-making process of the model.

D. Results of Yolo11 model

In this section, the Yolo11 model is implemented for the detection of tumor region from the brain MRI and the metrics are presented in Table III. The Yolo11 model achieved 89.6% boundary box precision, demonstrates that the model has high rate of correct detections with the less false positive. While, the recall box is 87.4% indicates that the model has quite number of missed detections and the mAP50 and mAP50-95 are 86.4% and 81.76% respectively. The inference time is also measure for the Yolo11 which is 1.3 (sec), reveals that the model is fast and responsive. The fitness score which is 0.7443 exposes the balance among the accuracy and computational cost. The overall pre and post processing of Yolo11 is 9.5 and 18.6265 (sec) respectively, reflecting the computation strength to arranged the brain MRI for detection. Table III also presents the speed, preprocessing, inference time, and confidence of the few individual cases. The each individual case the preprocessing time is lies between the 7.8 to 11.3 (sec) and the 1.3 (sec) is need for all the most cases. Few of samples such as 1,4,5, and 14 have high confidence scores which is 0.88, 0.96, 0.91, and 0.90, respectively, highlighting the effective predictions with tumor localization while the sample 3 achieved the confidence score of 0.00 which indicates the complete failure of detection of tumor region. Similarly, the samples that have overlapping visual features tend to results in low confidence score.

Fig. 6 shows visual comparison results between the Yolo11 detection and proposed RAG^4-Unet model. In this figure, Yolo11 model fails to align with the tumor boundaries, evidently, clearly visual in samples such as 5,6,and 9 and in some samples the boundary boxes has missed of the tumor and include non-tumor regions, indicating that the model faces the challenges when tumor has complex and irregular shape. In contrast, the proposed RAG^4-Unet segmentation maps indicating the higher boundary alignment. The segmented region by the RAG^4-Unet model are more closely to the original ground truth. In addition, the proposed model provides a detailed representation of tumor boundaries that Yolo11 boundary boxes cannot match and the Yolo11 is unable in detecting and localizing the tumor regions with the high precision such as in sample 3 and 10, the Yolo11 model missed and incorrectly detect the tumor region.

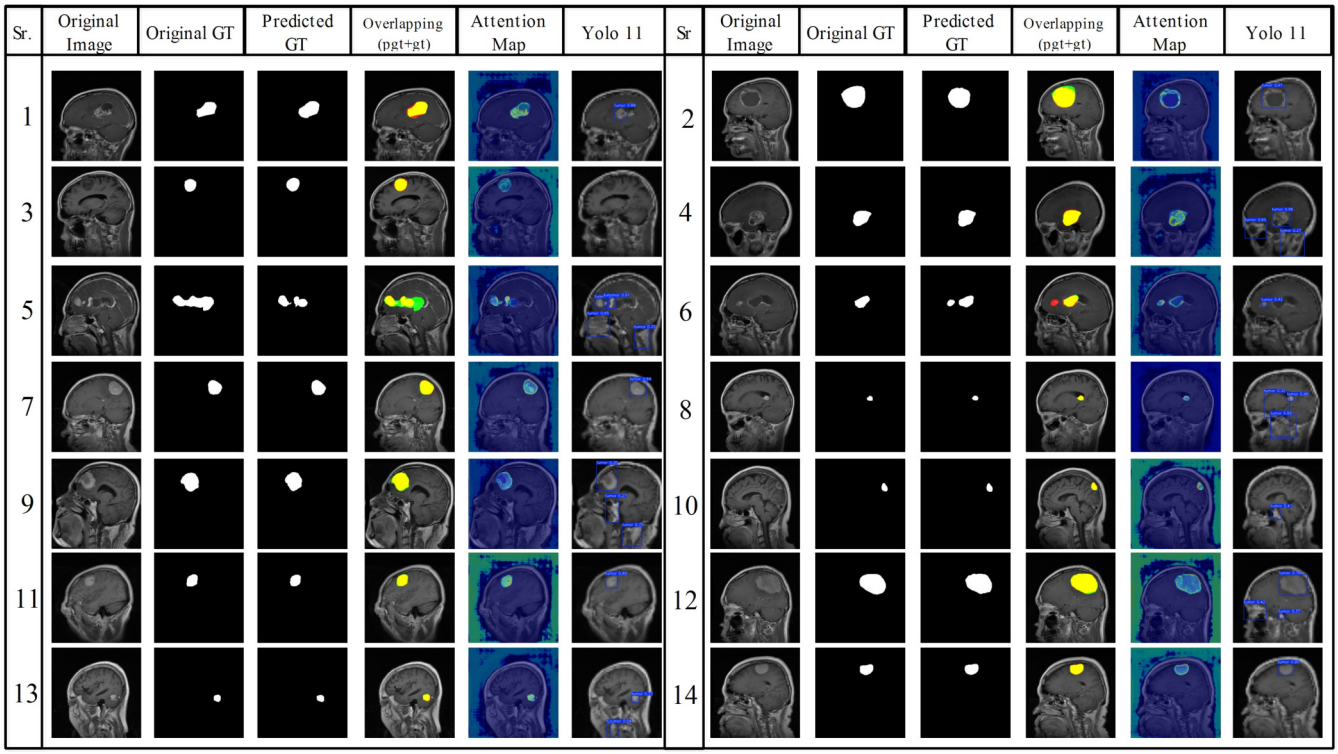

Fig. 6. Segmentation visualizations of predicted mask, overlapping maps, attention maps, and yolo 11 detection for analyzing the RAG^4-Unet

TABLE III
DETECTION RESULTS OF YOLO11 MODEL ON FIGSHARE DATASET

Sr.	Speed (ms)	Preprocess (ms)	Inference (ms)	Confidence (%)
1	2.4	7.8	1.2	0.88
2	2.5	8.7	1.3	0.47
3	2.7	10.4	0.6	0.00
4	2.5	9.8	1.5	0.96
5	2.9	10.6	1.3	0.91
6	2.6	8.9	1.4	0.42
7	2.6	9.8	1.3	0.64
8	2.7	8.2	1.2	0.60
9	2.7	9.5	1.3	0.30
10	2.6	10.0	1.4	0.41
11	2.7	10.7	1.3	0.80
12	2.7	9.5	1.3	0.78
13	2.6	11.3	1.3	0.77
14	2.5	9.3	1.4	0.90
Overall Performance				
Precision(B)	**Recall(B)**	**mAP50(B)**	**mAP50-95(B)**	
0.896	0.876	0.864	0.8176	
Preprocessing	**Inference**	**Fitness**	**Post process**	
0.494	4.260	0.7443	18.6265	

E. Comparison with SOTA

The comprehensive comparison has been conducted between the proposed and state-of-the-art methods, as shown in Table IV. Authors in [11] employed U-net architecture for the segmentation and conducted experiments on Figshare dataset. They achieved 88.1% of accuracy. In [14], the authors proposed customized CNN for the classification of tumor types using the Figshare dataset and they achieved 88% accuracy. Authors in [15] implemented ResNet50 model using deep transfer learning method on private dataset and they gained 90% of accuracy. In [16], the authors employed semi deep learning framework based on customized Unet and histogram features. The performed experiments on BITE dataset and they achieved 91%. However, our proposed methods achieved the highest accuracy of 91.74% using the swin transformer and 91.37% Dice score using proposed RAG^4-Unet in segmentation task.

TABLE IV
COMPREHENSIVE COMPARISON BETWEEN THE PROPOSED FRAMEWORK AND STATE-OF-THE-ART METHOD

Ref	Year	Dataset	Methodology	Accuracy
Ahsan et al. [11]	2024	Figshare	Unet architecture	88.1%
Asiri et al. [14]	2024	Figshare	Customized CNN	88%
Rajput et al. [15]	2024	Private	ResNet50	90%
Shiny et al. [16]	2024	BITE	Semi Deep learning	91%
Proposed Work		**Figshare**	**Swin Transformer**	**91.74%**
		Figshare	**RAG^4-Unet**	**91.37%**

IV. STATISTICAL ASSESSMENT

In order to fully assess the consistency and reliability of the proposed RAG^4-Unet model, we utilized Z-score method of the Dice similarity produced from the 14 test samples. the z-score for the each dice score is calculated using the equation 38.

$$\partial_z = \frac{d_i - \mu}{\sigma} \tag{30}$$

where d_i represents the Dice score of each sample, μ denotes the mean across all samples, and σ is the standard deviation. The value of the standard deviation is 0.0853 and the mean is 0.9216.

All samples (12 out of 14) showed Z-scores that fell within -1.0 and +1.0 which means that the most of the Dice values are close to the mean and demonstrate consistent segmentation performance across the samples in the test data. Sample 5 with a Dice score of 0.649 produced a Z-score of -3.20 indicating it was a significant outlier case, as shown in Figure 7. This Z-score indicated a material drop in performance relatively speaking for that one case, which could have been due to noise, complicated tumor morphology. Sample 6 had a moderately low Z-score of -1.635 which indicates that it did perform below the mean relative to the remaining samples. Sample 7, Sample 12 and Sample 14 had Z-scores that were above +0.5, indicating those samples performed above and beyond the average segmentation performance.

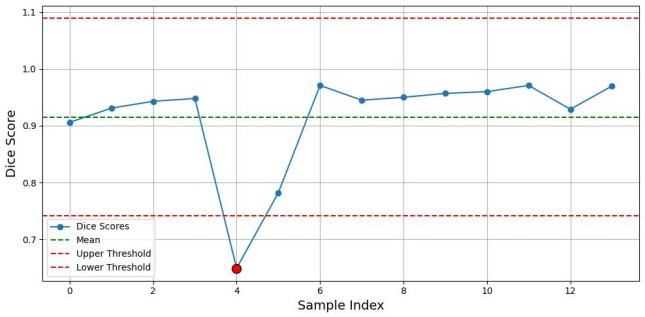

Fig. 7. Z-score Analysis of Dice Coefficients

V. Conclusion

In this work, we proposed a novel RAG^4-Unet architecture for the segmentation task integrated with swin transformer and Yolo11 for the classification and detection task. The proposed RAG^4-Unet architecture addresses the challenges of irregular shapes of boundaries and intersecting visual features of tumors by employing the residual attention gated mechanism. The proposed model achieved 91.73% of Dice coefficient, 94.74% of precision, 96.23% of sensitivity, and 98.46% specificity and swin transformer achieves 91.74% of accuracy, 91.64% of precision, 91.73% of recall, 91.52% of f1-score, 98.33 AUC, 86.91 kappa index, and 97.22% of confidence index with 2.306 (sec) inference time. The Yolo11 model achieves a boundary precision o 86.6%. The limitation of the proposed work is the proposed model goes under segmentation and low Dice when tumor size are small and Yolo11 lead to inaccurate boundary boxes when the tumor are complex.

In future work, we will focus on addressing these limitations using more diverse datasets and we will further explore and refine the attention mechanism to improve the tumor boundary delineation.

Acknowledgements

We acknowledge the support from COST Action "A Comprehensive Network Against Brain Cancer" (Net4Brain - CA22103).

References

[1] T. Rahman, M. S. Islam, and J. Uddin, "Mri-based brain tumor classification using a dilated parallel deep convolutional neural network," *Digital*, vol. 4, no. 3, pp. 529–554, 2024.

[2] H. A. Munira and M. S. Islam, "Hybrid deep learning models for multi-classification of tumour from brain mri," *J Inf Syst Eng Bus Intell*, vol. 8, pp. 162–74, 2022.

[3] N. Elazab, W. A. Gab-Allah, and M. Elmogy, "A multi-class brain tumor grading system based on histopathological images using a hybrid yolo and resnet networks," *Scientific Reports*, vol. 14, no. 1, p. 4584, 2024.

[4] P. Kuppler, P. Strenge, B. Lange, S. Spahr-Hess, W. Draxinger, C. Hagel, D. Theisen-Kunde, R. Brinkmann, R. Huber, V. Tronnier, *et al.*, "Microscope-integrated optical coherence tomography for in vivo human brain tumor detection with artificial intelligence," *Journal of Neurosurgery*, vol. 1, no. aop, pp. 1–9, 2024.

[5] Z. Rasheed, Y.-K. Ma, I. Ullah, M. Al-Khasawneh, S. S. Almutairi, and M. Abohashrh, "Integrating convolutional neural networks with attention mechanisms for magnetic resonance imaging-based classification of brain tumors," *Bioengineering*, vol. 11, no. 7, p. 701, 2024.

[6] S. Saket, Y. Nilipour, R. Taherian, and N. F. Marnaanni, "Evaluation of radiographic, neuropathological, and demographic findings in children aged 1 to 18 years with brain tumor," *Novelty in Biomedicine*, vol. 12, no. 2, pp. 55–59, 2024.

[7] M. S. I. Khan, A. Rahman, T. Debnath, M. R. Karim, M. K. Nasir, S. S. Band, A. Mosavi, and I. Dehzangi, "Accurate brain tumor detection using deep convolutional neural network," *Computational and Structural Biotechnology Journal*, vol. 20, pp. 4733–4745, 2022.

[8] D. Reyes and J. Sánchez, "Performance of convolutional neural networks for the classification of brain tumors using magnetic resonance imaging," *Heliyon*, vol. 10, no. 3, 2024.

[9] K. Singh, A. Kaur, and P. Kaur, "Computer aided detection of brain tumors using convolutional neural network based analysis of mri data," 2023.

[10] Y. Zhang, H. C. Ngo, Y. Zhang, N. F. A. Yusof, and X. Wang, "Imaging segmentation of brain tumors based on the modified u-net method," *Information Technology and Control*, vol. 53, no. 4, p. 1074 – 1087, 2024.

[11] R. Ahsan, I. Shahzadi, F. Najeeb, and H. Omer, "Brain tumor detection and segmentation using deep learning," *Magnetic Resonance Materials in Physics, Biology and Medicine*, pp. 1–10, 2024.

[12] T. Arumaiththurai and B. Mayurathan, "The effect of deep learning and machine learning approaches for brain tumor recognition," in *2021 10th International Conference on Information and Automation for Sustainability (ICIAfS)*, pp. 185–190, IEEE, 2021.

[13] J. Alyami, A. Rehman, F. Almutairi, A. M. Fayyaz, S. Roy, T. Saba, and A. Alkhurim, "Tumor localization and classification from mri of brain using deep convolution neural network and salp swarm algorithm," *Cognitive Computation*, vol. 16, no. 4, pp. 2036–2046, 2024.

[14] A. A. Asiri, A. Shaf, T. Ali, M. Aamir, M. Irfan, and S. Alqahtani, "Enhancing brain tumor diagnosis: an optimized cnn hyperparameter model for improved accuracy and reliability," *PeerJ Computer Science*, vol. 10, p. e1878, 2024.

[15] I. S. Rajput, A. Gupta, V. Jain, and S. Tyagi, "A transfer learning-based brain tumor classification using magnetic resonance images," *Multimedia Tools and Applications*, vol. 83, no. 7, pp. 20487–20506, 2024.

[16] K. Shiny, "Brain tumor segmentation and classification using optimized u-net," *The Imaging Science Journal*, vol. 72, no. 2, pp. 204–219, 2024.