BRAIN TUMOR SEGMENTATION BASED ON DEEP SUPERVISION AND CONTEXT FEATURE FUSION

Do Nhu Tai[1*], Vo Thanh Hoang Son[3*], Tran Minh Hai[3], Tran Nguyen Quynh Tram[3], Nguyen Thanh Huy[1], Nguyen Thi Ngoc Thanh[4], Nguyen Quoc Huy[5], Kim Soo Hyung[2 ¶]

[1]University of Economics Ho Chi Minh City — UEH, Vietnam; [2]Chonnam National University, South Korea; [3]Ho Chi Minh City University of Foreign Language - Information Technology, Vietnam; [4]Ho Chi Minh City Open University, Vietnam; [5]Sai Gon University, Vietnam

Email: taidn@ueh.edu.vn, hoangson.vothanh@gmail.com, tranminhhai1506@gmail.com, tramtnq@huflit.edu.vn, huynt@ueh.edu.vn, thanh.ntn@ou.edu.vn, nqhuy@sgu.edu.vn, shkim@jnu.ac.kr

[*]*Equal Distribution* ¶Corresponding: shkim@jnu.ac.kr

ABTRACT *— Glioma is a hazardous type of cancer; it grows very fast and is drug-resistant. The best treatment for this type of cancer is to remove the entire tumor. To do this, it requires expertise, time, and effort to analyze the tumor. Therefore, many deep-learning methods have been developed to assist doctors in brain tumor segmentation. However, this is still always a challenge for researchers because of the diversity and complexity of the tumor. In this paper, we propose the DeepDynUnet model. This model uses deep supervision combined with the Attention mechanism attached to the decoder process of the DynUnet model to exploit the standard features from each stage in the decoder stage of the model. In addition, this paper also proposes a Feature-Context Fusion method. This method will merge features from the tumor and features of the brain components. The goal is to exploit the features of the brain components to help the model identify the correct position and classify the components in the best way. Our experimental process has proven the performance of both proposed methods, where the feature-context fusion approach has achieved perfect results of 90.6% on the WT class, 84.3% on the TC class, and 79.2% on the ET class.*

Keywords— *Brain Tumor Segmentation, Fusion Method, Attention mechanism, Deep Supervison mechanism.*

I. INTRODUCTION

Glioma - one of the most common cancers in the brain, accounting for more than 50%. This type of tumor is formed from glial cells with altered properties [1]. The prognosis of this type of cancer is inferior, about 14-15 months after being diagnosed [2]. This type of cancer can be complicated to treat, as they are surrounded by many fine blood vessels called the blood-brain barriers, which block most radiation therapy [3]. Therefore, the most recommended treatment for this type of cancer is to remove as much of the cancerous area as possible [4]. To do that, the team of doctors has to do an in-depth analysis of the tumor and the strategy through the MRI image.

The analysis and diagnosis of gliomas through MRI images require high expertise, time, and effort. Many scientific studies have been conducted to address the problem of image-based glioma segmentation to assist physicians in treating gliomas. With the development of deep learning networks, the efficiency of cancer segmentation has achieved specific results such as [5],[6]. But, in these methods, the input feature processing of the model stacks the MRI image modes on top of each other, so it is impossible to take advantage of information from the multimodality.

Recently, researchers often used fusion methods to solve medical image-related problems. From the fusion method, the models have solved the problem of missing information from multimodal MRI through which there have been models that apply combination methods to achieve specific results such as [7] proposed a method to generate N paths corresponding to N MRIs methods at the encoder stage, then merge the features based on the dual attention method at the decoder stage. [8] proposed a model using two methods of incorporating at the pixel level to diversify input information and feature level to synthesize and select data. However, most of these models only essentially combine the tumor features. This paper proposes a DeepDynUnet model that combines the DynUnet model, deep supervision method, and attention method. This model aims to combine the tumor features and use the deep supervision mechanism to enhance their common characteristics. Additionally, we propose a method that combines the features of the tumor with the features of the brain components. Our approach involves training two models - the first on the BraTS 2018 [9] dataset to extract features from brain tumors and the second on the iSeg2019 [10] dataset to extract features from the brain context. The selected features will be merged through a combination of methods to achieve optimal results.

The paper has four sections: introduction, proposed method, experiment, and conclusion. The Introduction provides the context, motivation, and a paper summary. The proposed method section presents all the methods proposed in the paper. The experiment section describes the conducted experiments and their results. Finally, the conclusion summarizes the work and its achievements.

II. PROPOSED METHOD

Overview: The main goal of this work is to classify each voxel $v = (x, y, z)$ from input MRIs $I \in R^{w \times h \times d \times 4}$ to one in three parts of the tumor including whole tumor (WT), tumor core (TC), and enhancing tumor (ET). To solve this problem, we proposed two solutions. Once, we designed a new model called DeepDynUnet. We exploit the strength of the deep supervisor method and combine it with the attention mechanism. After that, we applied it to the

DynUnet model. Based on the proposed fusion method, the second solution is to take advantage of the context of brain composition features and fuse it with the brain tumor features.

A. DeepDynUnet – A new model used the deep supervisor method and attention mechanism.

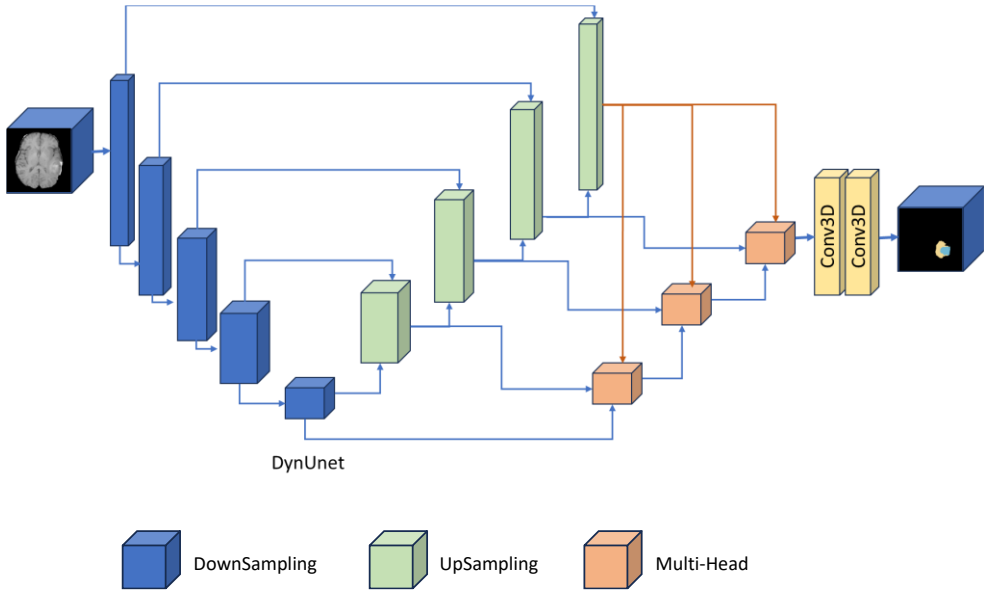

Figure 1. Overview of DeepDynUnet model, including DownSampling block, UpSampling block, and Multi-Head Attention Block.

Figure 1 shows the proposed model based on the DynUnet architecture [11] with three main components: (1) Encoder constructed primarily from DownSampling Blocks, (2) Decoder comprising UpSampling blocks, and (3) Deep Supervision where Multi-head Attention blocks are utilized. Figure 1 illustrates the overview of the DeepDynUnet model, which follows a U-shape net structure comprising DownSampling and UpSampling blocks. For each Down-Up Sampling pair, there is a corresponding skip connection. After each block in the Decoder stage, information is aggregated using the Multi-Head block and progressively synthesized in a cascaded manner. Finally, the Conv3D blocks are responsible for voxel-wise classification.

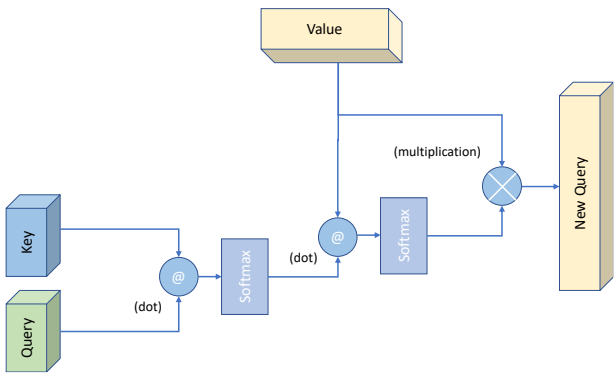

Figure 2. Describes the architecture of a Multi-head deep supervision block consisting of three inputs called Query, Key, and Value, respectively. The output of this block will act as the Query feature for the next step.

Multi-head deep supervision: The Multi-Head Attention Block shown in Figure 2 is inspired by the Scale-dot Attention Module . It takes three input feature maps: the feature extracted from the corresponding decoder block (Q), the feature aggregated from the previous stages (K), and the feature extracted from the output of the DynUnet model (V). The procedure is as follows: Firstly, the scalar product between Q and K is analyzed to generate a feature map containing prominent shared information from the two initial values. This is a common feature among decoder blocks. Subsequently, the same method is employed to extract feature maps between the newly derived map and V. The features between the decoder's stages will be enhanced through these two stages of scalar product analysis. It is important to note that after each scalar product analysis step, a Softmax layer is used to constrain the values to the range of 0 to 1 to prevent noise interference in this process. Finally, the feature maps found will be multiplied again with V on each element. At this stage, the feature maps play the role of a mask that helps to enhance and reinforce the information for V. The entire process is written in formula as below:

$$f(K,Q,V) = \sigma\left(\left(\frac{\sigma(Q.K^T)}{\sqrt{(d_k)}}\right).V\right) \otimes V \qquad (1)$$

The architecture of this model is presented in Figure 2. It consists of 3 inputs corresponding to 3 feature maps. The operator (@) represents the matrix multiplication between 2 feature maps, and the operator ($\otimes$) is the element-wise multiply between (V) and the newly standard feature map. This boosts the general efficiency of the decoder stages.

B. Context-Feature Fusion.

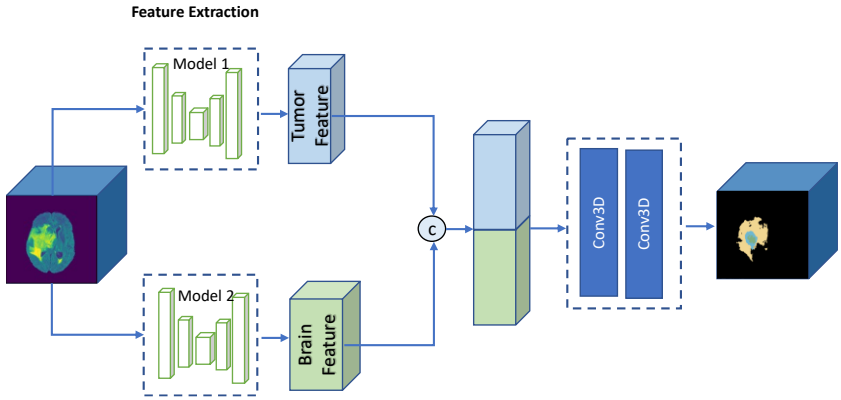

Figure 3. The Context-Feature Fusion method is based on feature maps extracted from 2 different models. Model 1 extracts features of brain tumors, while Model 2 extracts features from brain components.

The main problem to be solved in this paper is to classify each voxel from the input MRIs into one of four label types: whole tumor (WT), enhancing tumor (ET), tumor core (TC) and background (BG). The context-feature fusion module shown in Figure 3 will fuse the features of the tumor and the features of the brain components. The goal of this method is to take advantage of the information from the brain components to supplement the process of recognizing which voxel V belongs to which class in the tumor. We called the feature maps that extracted from brain tumor model is $F_1 \in R^{h\times w\times d\times f_1}$ and the feature maps extracted from brain segmentation model is $F_2 \in R^{h\times w\times d\times f_2}$. The process will be presented as below:

$$F(I) = \sigma\big(C(\mathrm{F}_1 \oplus \mathrm{F}_2)\big) \qquad (2)$$

where I is the input MRIs patch, F_1, F_2 are two models for feature extraction, C is the classify block, $\oplus$ represents for concatenate function and σ represents the sigmoid activation function.

III. EXPERIMENT

A. Dataset

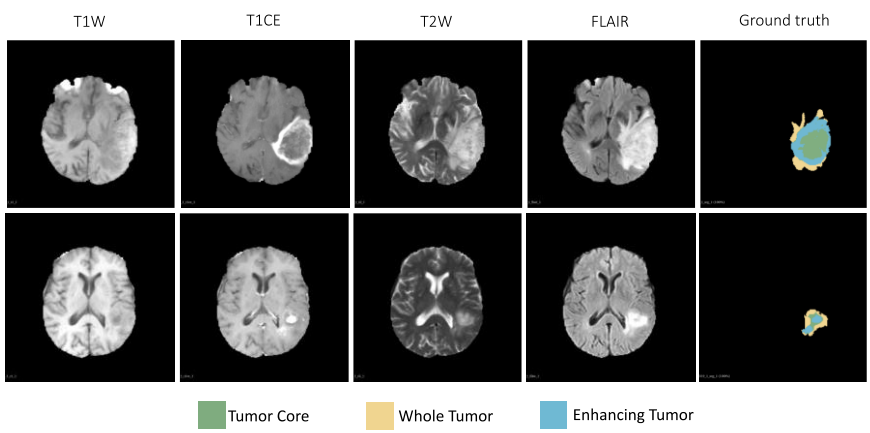

Figure 4. Image from BraTS 2018 included T1W, T1CE, T2W, FLAIR modality respectively. The ground-truth cconsists of 3 main ingredients are tumor core, whole tumor and enhancing tumor, which colors are shown in the note.

BraTS 2018: In this paper, we used the BraTS 2018 dataset shown in Figure 4, first introduced in the Brain Tumor Segmentation (BraTS) Challenge 2018, organized by MICCAI (Medical Image Computing and Computer Assisted Intervention Society). This annual international organization and conference focuses on the application of

artificial intelligence in medical imaging and healthcare. This ensures the credibility and popularity of this dataset. The BraTS 2018 dataset consists of a training set of 285 MRI scans, with 210 and 75 high-grade (HGG) and low-grade (LGG) brain tumor images, respectively. Each MRI volume includes four modalities: FLAIR, T1w, T2w, and T1ce, with a size of 240 ×240 ×155 for each imaging modality. The ground truth volume is a 3D image volume with 4-pixel values: whole tumor (WT), tumor core (TC), enhancing tumor (ET), and background (BG). The entire evaluation process will be carried out on the dataset's website[1].

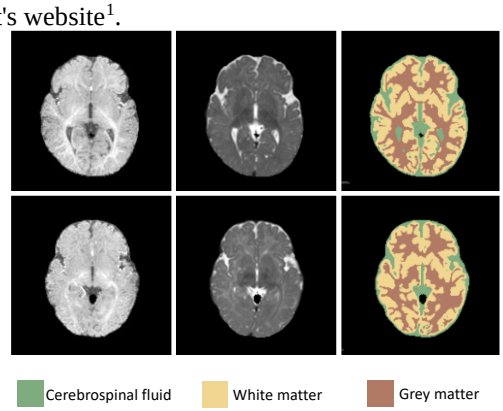

Figure 5. Examples of images from iSeg-2019 included T1w modality, T2w modality. and the ground truth. The ground truth include 2 type which shown follow the color in note.

iSeg 2019: In addition to the Context-Feature Fusion method, we used the 6-month-old brain MRI Segmentation dataset shown in Figure 5 to train a model for classifying the brain's components. This dataset consists of 10 training samples and 13 test samples. Each sample will include 2 MRI modalities, T1w and T2w, and with a label volume of $144 \times 192 \times 256$. Each pixel in the ground truth volume will belong to 1 of 4 values, including Cerebrospinal fluid (CSF), White matter (WM), and Background (everything outside the brain).

B. Experiment setup

Training Process. In both methods, we used a standard data preprocessing and splitting method. The training dataset was split into 80% for training and 20% for testing. The input data was cut into small 128 x 128 x 128 patches and then randomly flipped on the three planes (axial, sagittal, and coronal) adjusted randomly within 10% intensity and normalized on all channels. Similarly, the ground truth volume was also cut in proportion to the patch and converted to one-hot encoding.

The number of epochs was limited to 50 epochs. We trained the model using Adam optimization, with an initial learning rate 0.0001. The learning rate schedule was Co- sine Annealing, with a minimum learning rate of 0.00001. The experiments ran on the NVIDIA Tesla P100 16GB hardware and used Pytorch and Monai frameworks.

Loss and Evaluation Metric. We used the dice loss for training. Mean dice score [12] was used for the result evaluation representing as below:

$$DiceLoss = 1 - DiceScore$$
$$DiceScore = \frac{2\sum_{i=1}^{N} p_i g_i + \epsilon}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} g_i + \epsilon} \quad (3)$$

where p_i is the predicted i-th pixel, g_i is the grouth truth value at the i-th position too, ϵ will help the formula not to fall into undefined cases.

C. Result

First, to have a foundation to reference, we experimented with baseline modern deep learning models that have achieved specific results in the whole brain tumor segmentation task, including VNet, SegResNet, SwinUnetR, and DynUnet. The best performance on the Dice score was achieved by the DynUNet model shown in Table 1. Then, we applied the first solution presented to the DynUnet model called DeepDynUnet.

Table 1. The result of the baseline models and DeepDynUnet.

BraTS2018	**Dice Score ↑**		
Method	**WT**	**TC**	**ET**
Vnet [13]	0,801	0,701	0,684

[1] https://ipp.cbica.upenn.edu/

SwinUnetR [14]	0,8221	0,7482	0,7365
SegResNet [15]	0,8758	0,7847	0,7275
DynUnet [16]	0,8809	0,7857	0,7339
DeepDynUnet*	**0,8903**	**0,7914**	**0,7826**

Table 1 shows that the DeepDynUnet method has achieved very good performance, increasing the DiceScore by up to 5% on the ET class compared to the model that does not use the above method. In addition, the remaining two classes, WT, and TC, also increased by about 1% on each type.

Table 2. The result of the fusion method between DynUnet and DeepDynUnet models.

Fusion Method	**Dataset**		**Dice Score ↑**		
	BraTS2018	**Iseg2019**	**WT**	**TC**	**ET**
Early	DynUnet	DynUnet	0,9015	**0,8473**	0,7810
	DynUnet	DeepDynUnet	0,9021	0,8383	0,7579
	DeepDynUnet	DynUnet	0,8962	0,7779	0,7380
	DeepDynUnet	DeepDynUnet	0,8973	0,7781	0,7398
Joint	DynUnet	DynUnet	0,9005	0,8282	0,7821
	DynUnet	DeepDynUnet	**0,9057**	0,8426	**0,7923**
	DeepDynUnet	DynUnet	0,8963	0,7769	0,7412
	DeepDynUnet	DeepDynUnet	0,8964	0,7704	0,7351

After experimenting to prove the effectiveness of the DeepDynUnet model, we proceeded to test with the second solution, Context-Feature Fusion. Based on the results obtained in Table 1, we selected the two optimal models as the data extraction models for this method. Conduct an ablation study combining models and fusion methods. Here, we only use two fusion methods, Early Fusion and Joint Fusion, without using the Late Fusion method. Because the Late Fusion method only uses essential mathematical functions to calculate the results based on the classification value of the model's output and does not use the nature of the Feature maps. The results displayed in Table 2 show that The Joint Fusion method gives slightly better results than the Early Fusion method.

In particular, when the fusion uses the tumor feature extracted by the DynUnet model, it will give optimal results. It has given the best result, especially when combined with the features of the brain component extracted by the DeepDynUnet model using the Joint Fusion method. The results are 90.5% on WT, 84.2% on TC, and 79.2% on ET.

Table 3. The result of our method compairision to related works.

Method	**Year**	**Dice Score ↑**			
		WT	**TC**	**ET**	**AVG**
D. Zhang et al [5]	2020	0,866	0,769	0,744	0,793
Y. Zhang et al [6]	2022	0,896	0,857	0,776	0,843
T. Zhou et al [7]	2022	0,887	0,796	0,750	0,811
Ours		0,906	0,843	0,792	0,847

Moreover, our method has demonstrated its optimal performance on the WT and ET classes compared with previous related methods. The results are shown in Table 3, which shows that our approach increased by 1-4% in the WT class and about 2-5% in the ET class.

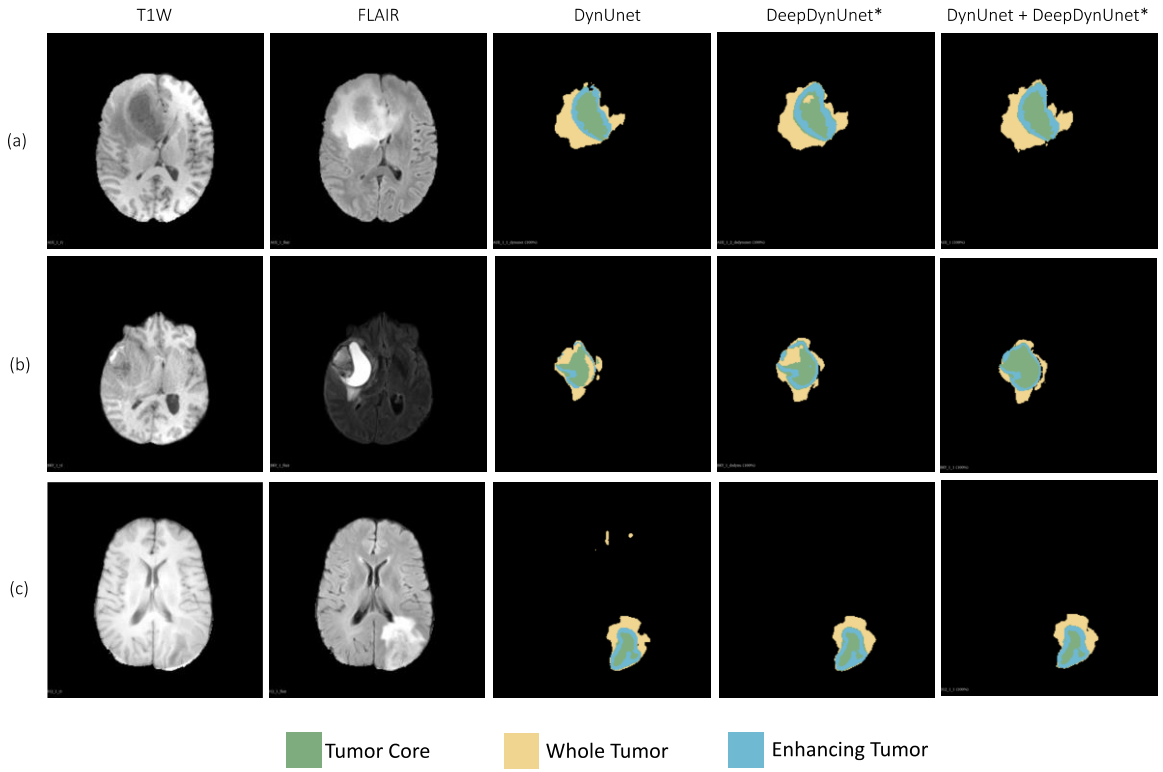

Figure 6. Brain tumor segmentation results of our proposed, mark * is ours proposed method.

Figure 6 shows a few sample predictions from the models in this paper, including DynUnet, DeepDynUnet, and the Feature-Context Fusion method. T1w and FLAIR are the two modalities, and the remaining three columns represent the results obtained from the methods, all of which are displayed as two-dimensional slices. The asterisk (*) marks our proposed method.

We can see that when using the DeepDynUnet model, the tumor boundaries are identified with higher coverage and accuracy than the conventional DynUnet method. Specifically, the size of the tumor identified by the DeepDynUnet model is more significant than that of the tumor identified by the DynUnet model. Additionally, in row (c), we can see that the model misidentified some areas of the image, which was fixed in the DeepDynUnet model. However, some limitations remain in distinguishing the areas of the tumor's components. For example, in rows (a) and (b), the model did not correctly classify the ET class. However, the results were significantly improved after using the Feature-Context Fusion method. The tumor coverage and accuracy were enhanced over the DynUnet model, and the gaps in the DeepDynUnet model were resolved.

IV. CONCLUSION

This paper presents two novel methods to solve the brain tumor segmentation problem to improve the performance of brain tumor segmentation. Our issue deals with uncertainties in morphology, tumor position, and unclear tumor border. The two methods are the DeepDynUnet model and the feature-context fusion module.

DeepDynUnet is a model that uses Attention and Deep supervision mechanisms to attach to the stages in the decoder stage of the DynUnet model. This helps to synthesize the standard features in each decoder stage, and thanks to the deep supervision mechanism, the features are enriched through each stage and are synthesized before going through the segmentation block. Experimental results have shown that this method has helped improve the model's performance, with the results of 89% on WT, 79% on TC, and 78% on WT.

In addition, the Feature-Context Fusion method is a method of combining features between the tumor segmentation model and the brain segmentation model. The features extracted by the brain segmentation model will help the model correctly position the tumor and simultaneously support the process of segmenting each tumor component. We have experimented with different fusion methods, and the best result is achieved when using the DynUnet model to extract tumor information with the DeepDynUnet model to extract brain segmentation information using the Joint Fusion method.

ACKNOWLEGEMENT

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2023-RS-2023-

00256629) grant funded by the Korea government(MSIT). This research was funded by University of Economics Ho Chi Minh City, Vietnam.

V. REFERENCES

[1] D. Hanahan and R. A. Weinberg, "Hallmarks of Cancer: The Next Generation," *Cell*, vol. 144, no. 5, pp. 646–674, Mar. 2011, doi: 10.1016/j.cell.2011.02.013.

[2] F. Hanif and K. Muzaffar, "Perveen kahkashan, Malhi S, Simjee S," *Glioblastoma Multiforme Rev. Its Epidemiol. Pathog. Clin. Present. Treat. APJCP*, vol. 18, no. 3, pp. 10–22034, 2017.

[3] M. A. Dymova, E. V. Kuligina, and V. A. Richter, "Molecular Mechanisms of Drug Resistance in Glioblastoma," *Int. J. Mol. Sci.*, vol. 22, no. 12, p. 6385, Jun. 2021, doi: 10.3390/ijms22126385.

[4] E. Belykh, K. V. Shaffer, C. Lin, V. A. Byvaltsev, M. C. Preul, and L. Chen, "Blood-Brain Barrier, Blood-Brain Tumor Barrier, and Fluorescence-Guided Neurosurgical Oncology: Delivering Optical Labels to Brain Tumors," *Front. Oncol.*, vol. 10, 2020, Accessed: May 13, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fonc.2020.00739

[5] Z. Zhou, Z. He, and Y. Jia, "AFPNet: A 3D fully convolutional neural network with atrous-convolution feature pyramid for brain tumor segmentation via MRI images," *Neurocomputing*, vol. 402, pp. 235–244, 2020, doi: https://doi.org/10.1016/j.neucom.2020.03.097.

[6] Y. Zhang *et al.*, "mmFormer: Multimodal Medical Transformer for Incomplete Multimodal Learning of Brain Tumor Segmentation." arXiv, Aug. 04, 2022. Accessed: Aug. 16, 2023. [Online]. Available: http://arxiv.org/abs/2206.02425

[7] T. Zhou, S. Ruan, P. Vera, and S. Canu, "A Tri-Attention fusion guided multi-modal segmentation network," *Pattern Recognit.*, vol. 124, p. 108417, Apr. 2022, doi: 10.1016/j.patcog.2021.108417.

[8] Y. Liu, F. Mu, Y. Shi, J. Cheng, C. Li, and X. Chen, "Brain tumor segmentation in multimodal MRI via pixel-level and feature-level image fusion," *Front. Neurosci.*, vol. 16, 2022, Accessed: Mar. 07, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2022.1000587

[9] S. Bakas *et al.*, "Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection," *Cancer Imaging Arch.*, vol. 286, 2017.

[10] Y. Sun *et al.*, "Multi-Site Infant Brain Segmentation Algorithms: The iSeg-2019 Challenge," *IEEE Trans. Med. Imaging*, vol. 40, no. 5, pp. 1363–1376, May 2021, doi: 10.1109/TMI.2021.3055428.

[11] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nat. Methods*, vol. 18, no. 2, Art. no. 2, Feb. 2021, doi: 10.1038/s41592-020-01008-z.

[12] A. Vaswani *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[13] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*, Ieee, 2016, pp. 565–571.

[14] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part I*, Springer, 2022, pp. 272–284.

[15] A. Myronenko, "3D MRI Brain Tumor Segmentation Using Autoencoder Regularization," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, and T. van Walsum, Eds., Cham: Springer International Publishing, 2019, pp. 311–320.

[16] M. B. M. Ranzini, L. Fidon, S. Ourselin, M. Modat, and T. Vercauteren, "MONAIfbs: MONAI-based fetal brain MRI deep learning segmentation." arXiv, Mar. 21, 2021. doi: 10.48550/arXiv.2103.13314.