BrainScan AI

(Brain Tumor Segmentation)

Done by:

Ahmad Ghassan AbuDyak – 0207744

Abdallah Nader Abdelhamid – 0206921

Moayad Bilal Rabaa – 0203353

Mohammad Jehad Ali – 0207684

Supervisors:

Dr. Tamam Al-Sarhan – Dr. Ali Al-Rodan

Table of Contents

Table of Figures

1. Introduction

A brain tumor is an abnormal and uncontrolled growth of cells in the brain that can be either benign (non-cancerous) or malignant (cancerous), potentially causing various neurological symptoms and health issues. Brain tumors are categorized into primary and secondary types. Primary brain tumors originate from brain cells, while secondary tumors metastasize to the brain from other organs. The most common primary brain tumors are gliomas, which arise from brain glial cells and are classified into low-grade (LGG) and high-grade (HGG) subtypes. High-grade gliomas are an aggressive type of malignant brain tumor that grows rapidly, typically requiring surgery and radiotherapy, and are associated with a poor survival prognosis.

Early detection of brain tumors is needed to get proper and accurate treatment. Since there are many types of brain tumors that affect the human brain and as we enter the era of Artificial Intelligence (AI) for healthcare, AI-based intervention for diagnosis and surgical pre-assessment of tumors became a necessity rather than a luxury. As a reliable diagnostic tool, Magnetic Resonance Imaging (MRI) is essential for the monitoring and surgical planning of brain tumors. A variety of complementary 3D MRI techniques are used to highlight different aspects of the brain tissue and the extent of the tumor. These include T1, T1 with contrast (T1c), T2, and Fluid Attenuated Inversion Recovery (FLAIR) scans. In particular, the use of a contrast agent like gadolinium in the T1c MRI helps in identifying highly active areas of the tumor, as can be seen on Figure 1.

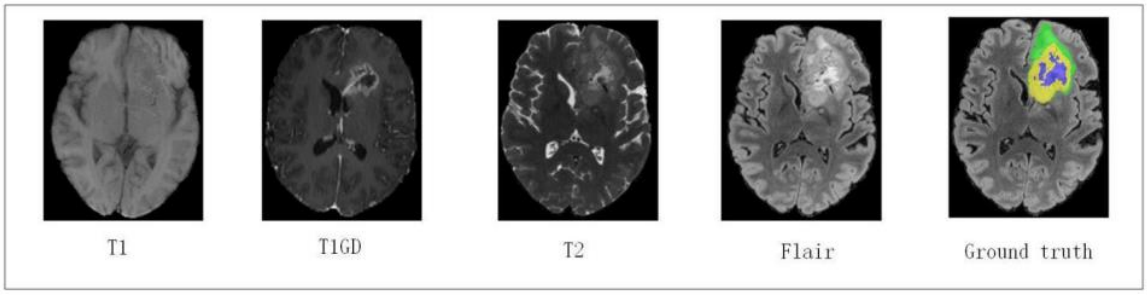

***Figure 1**: An example of multimodal MRI volumes for brain tumor segmentation.*

Benefiting from the development of deep learning, computer vision technology has been widely used in medical image analysis. Image segmentation is an important part of medical image analysis. In particular, accurate and reliable medical image segmentation

can play a cornerstone role in computer-aided diagnosis and image-guided clinical surgery.

In recent years, deep learning methods have shown promising results in medical image segmentation, exceeding traditional approaches based on manually developed features and machine learning algorithms. Among these deep learning models, U-Net [1] has emerged as one of the most popular and efficient architectures for biomedical image segmentation. In addition, the original goal of U-Net is to solve the problem of medical image segmentation. This network mainly rely on fully convolutional neural network (FCNN) with U-shaped structure which consists of a symmetric Encoder-Decoder with skip connections. With this design, U-Net has a strong ability of learning discriminating features in medical images.

Despite its success, U-Net has some limitations in handling large-scale dependencies and fine details of medical images. Image segmentation is still a challenge task in medical image analysis. Since convolution operations fundamentally focus on local information, CNN-based methods often struggle to effectively learn and represent global and long-range semantic information. Recently, inspired by Transformer's great success in the nature language processing (NLP) domain [2], researchers have tried to bring Transformer into the vision domain. As part of these efforts, a hierarchical Swin Transformer has been developed.

Inspired by the success of the Swin Transformer [3], we propose to combine U-Net with Swin Transformer a recently proposed vision transformer architecture that utilizes a shifted windows mechanism to model local and global interactions in visual data. In this work, we attempt to use Swin Transformer block as basic unit to build an Encoder which extracts features to be then fed to a CNN-decoder with skip connections for the brain tumor segmentation task.

We will employ the BraTS 2023 dataset to train and validate our model for the task of 3D brain tumor segmentation. This dataset comprises 1251 case each with four distinct 3D MRI modalities: native (T1), T1-weighted with gadolinium contrast enhancement (T1-Gd or T1c), T2-weighted (T2), and T2 Fluid-Attenuated Inversion Recovery (T2-FLAIR).

2. Related Work

In recent years, convolutional neural networks (CNNs) have become the most popular method in image classification and are widely used in medical image analysis. Myronenko et al. [4] proposed a method for multi-modal semantic segmentation specifically designed for 3D brain tumors, based on the U-Net encoder-decoder architecture with integrated variable autoencoder (VAE) branches. This inclusion aimed at concurrently reconstructing input images and performing segmentation, thereby regularizing the shared encoder. Jiang et al. [5] designed a novel two-level cascade U-Net for brain tumor segmentation, which trains the substructures of brain tumors in an end-to-end manner, progressing from a coarse to a fine level of detail. In the second stage, both the initial coarse segmentation map and the original image are input into another U-Net. This second U-Net, with more network parameters, generates a more precise and detailed segmentation map. Recently, Isensee et al. [6] utilized the nnU-Net network for brain tumor segmentation and implemented specific modifications. These modifications included the integration of brain tumor segmentation techniques such as post-processing and region-based training which effectively improved the accuracy of brain tumor segmentation. Luu et al. [7] enhanced brain tumor segmentation accuracy by applying an advanced nnU-Net network, replacing batch normalization with group normalization, and integrating axial attention into the decoder, which further improved the accuracy of brain tumor segmentation. Moreover, Xie et al. [8] introduced a framework that employs a backbone CNN for extracting features, a transformer for processing these encoded representations, and a CNN decoder for predicting segmentation outcomes. In a related approach, Wang et al. [9] proposed the use of a transformer within the bottleneck of a 3D encoder-decoder CNN specifically for semantic brain tumor segmentation. Chen et al. [10] developed a methodology for multi-organ segmentation that incorporates a transformer as an additional layer within the bottleneck of a U-Net architecture. Hatamizadeh et al. [11] introduced the UNETR architecture. This design features an encoder based on the Vision Transformer (ViT) that directly processes 3D input patches, coupled with a CNN-based decoder. The UNETR architecture has demonstrated encouraging outcomes in the segmentation of brain tumors, particularly when applied to the MSD dataset.

Since the introduction of the U-Net [1], CNN-based networks have achieved state-of-the-art results on various 2D and 3D medical image segmentation tasks. Subsequently, Zhou et al. [12] proposed the U-Net++ network which enhances the standard U-Net by adding a dense jump connection layer and incorporating a series of built-in U-Nets at various depths. This innovation results in improved segmentation of objects of varying sizes by effectively combining multiple U-Net networks for simultaneous image segmentation training. Additionally, they refined the U-Net++ model through deep supervision-driven pruning, resulting in substantial speed gains with a minimal decrease in performance. Additionally, Milletari et al. [13] introduced the V-Net, a variation of the U-Net network, which integrates residual connections in both the contraction and expansion layers. This approach enhances the convolutional neural network through these residual connections. V-Net employs convolution layers instead of pooling layers for down-sampling, optimizing weights more effectively, and introduces a dynamically adjustable dice loss function, particularly useful when dealing with unbalanced sample categories. This function dynamically adjusts weights, eliminating the need for sample reweighting during training, thereby addressing sample imbalance issues. Huang et al. [14] posited that the skip connection layer in U-Net++ causes a loss of features by merging the encoder's low-level features with the decoder's high-level ones. To address this, they proposed the UNet3+ network, employing full-scale feature fusion by combining decoder features with adjacent, lower, and higher features of the encoder and decoder, respectively. Qin et al. [15] developed an enhanced version of the UNet3+ network. They replaced the conventional convolution layer with a phase residual network and the batch normalization layer with a Filter Response Normalization (FRN) layer, improving feature extraction and reducing the impact of batch size on performance. Furthermore, Zhou et al. [16] proposed a multi-modality segmentation network guided by a novel tri-attention fusion. Dual attention was first used to reweight the features among the modal and spatial paths. Then a correlation attention module is introduced, using correlation description blocks to learn inter-modal correlations. This module employs correlation-based constraints to guide the network in learning features that are more relevant to segmentation. Jia et al. [17] proposed a BiTr-Unet model for brain tumor segmentation in multimodal MRI scans, using TransBTS as the core component. They

incorporated a 3D Convolutional Block Attention Module (CBAM) into the encoder and modified the skip connection between the fourth and fifth levels. This modification involved replacing the traditional skip connection with linear projection, a vision transformer block, and feature mapping replacement.

Transformer-based models have recently gained significant attention in computer vision and medical image analysis. Subsequently, Cao et al. [18] introduced a purely Swin Transformer-based U-Net model integrating elements from U-Net, including an encoder, decoder, skip connections, and a connection layer. The Swin Transformer in the encoder extracts multi-scale features, with a patch expanding module in the decoder to increase image resolution and reduce feature channels, thereby improving image segmentation accuracy. Wu et al. [19] proposed a transformer model built on the 3DUNet architecture. This model introduces a new local attention mechanism (LSM) and a global attention mechanism (GSM). The global attention mechanism, GS-MSA, mimics the dilated convolution model by selecting patches at fixed intervals to form global attention units and using the remaining patches to extract global feature information. Moreover, Hatamizadeh et al. [20] integrated Swin Transformers into the U-Net network. They used Swin Transformers as the encoder for feature extraction, then input these features into a convolutional network for image restoration via up-sampling. During this process, encoder features are passed to the decoder through skip connections, and residual blocks are employed at each level of the decoder. Li et al. [21] embedded a Transformer into U-Net++ and used a dense feature extractor to model global feature information and remote dependencies. Leveraging the symmetrical build of the U-shape, this setup enables the feature maps to establish a one-to-one matching relationship from shallow to deep layers making semantic features denser and resulting in higher resolution outputs. Vatanpour et al. [22] addressed the limitations of Fully Convolutional Neural Networks (FCNNs) in segmenting tumors of varying sizes by developing the TransDoubleU-Net, which merges U-Net's efficiency with the global information processing abilities of Transformers. This model features a dual-scale Swin Transformer for encoding and a dual-level decoder combining CNNs and Transformers, emphasizing improved feature localization and segmentation accuracy. Moreover, Cai et al. [23] introduces a model that combines CNNs and Vision Transformers for 3D medical image segmentation. Swin Unet3D effectively

captures both local and long-distance dependencies, achieving a balance between model complexity and segmentation accuracy. This innovative approach leverages the strengths of both convolutional and transformer architectures for improved brain tumor segmentation.

2.1 Problems of existing systems and our solution

Existing systems for brain tumor segmentation, predominantly based on convolutional neural networks (CNNs), face several challenges. CNN-based methods such as U-Net and its variants (e.g., U-Net++, V-Net, and UNet3+) have demonstrated state-of-the-art performance in various medical image segmentation tasks. However, these approaches struggle with capturing long-range dependencies and global context due to their intrinsic locality in convolution operations. This limitation is particularly problematic for accurately segmenting tumors of varying sizes and shapes within 3D brain MRI images. Additionally, enhancements like dense skip connections and residual connections, while improving segmentation accuracy, often lead to increased model complexity and computational demands.

In recent years, Transformer-based models have shown promise in overcoming these limitations by leveraging self-attention mechanisms to capture global and long-range dependencies. However, pure Transformer models or simple combinations with CNNs have yet to fully realize their potential in medical image segmentation due to issues such as computational inefficiency and difficulty in preserving fine details.

Our proposed solution, Swin-Unet, addresses these challenges by integrating the strengths of both CNNs and Transformers. Swin-Unet employs a U-shaped architecture with a Swin Transformer-based encoder and a CNN-based decoder connected via skip connections. This hybrid approach allows the model to effectively capture both local and global features, leading to more precise and effective segmentation. By using the Swin Transformer in the encoder, we efficiently model long-range dependencies and fine details, while the CNN decoder ensures computational efficiency and effective reconstruction of high-resolution segmentation maps. This design not only improves segmentation accuracy but also

mitigates the issues of model complexity and resource constraints faced by existing systems.

3. Methodology

3.1 Architecture Overview

Figure 2 illustrates the overall architecture of our proposed Swin-Unet model. Swin-Unet consists of encoder, bottleneck, decoder and skip connections. The input is a multimodal MRI medical image $X \in \mathbb{R}^{H \times W \times D \times C}$, where $H \times W \times D$ is the image size and C the number of channels. It is divided into non-overlapping patches, which are fed to transformer-encoder. Then the transformer-encoder uses those patches to encode the feature representations via the Swin transformer blocks. The feature representations are then fed to the CNN-decoder via skip connections at multiple resolutions to obtain the final output segmentation. In the following, we describe the individual components of the proposed architecture in details.

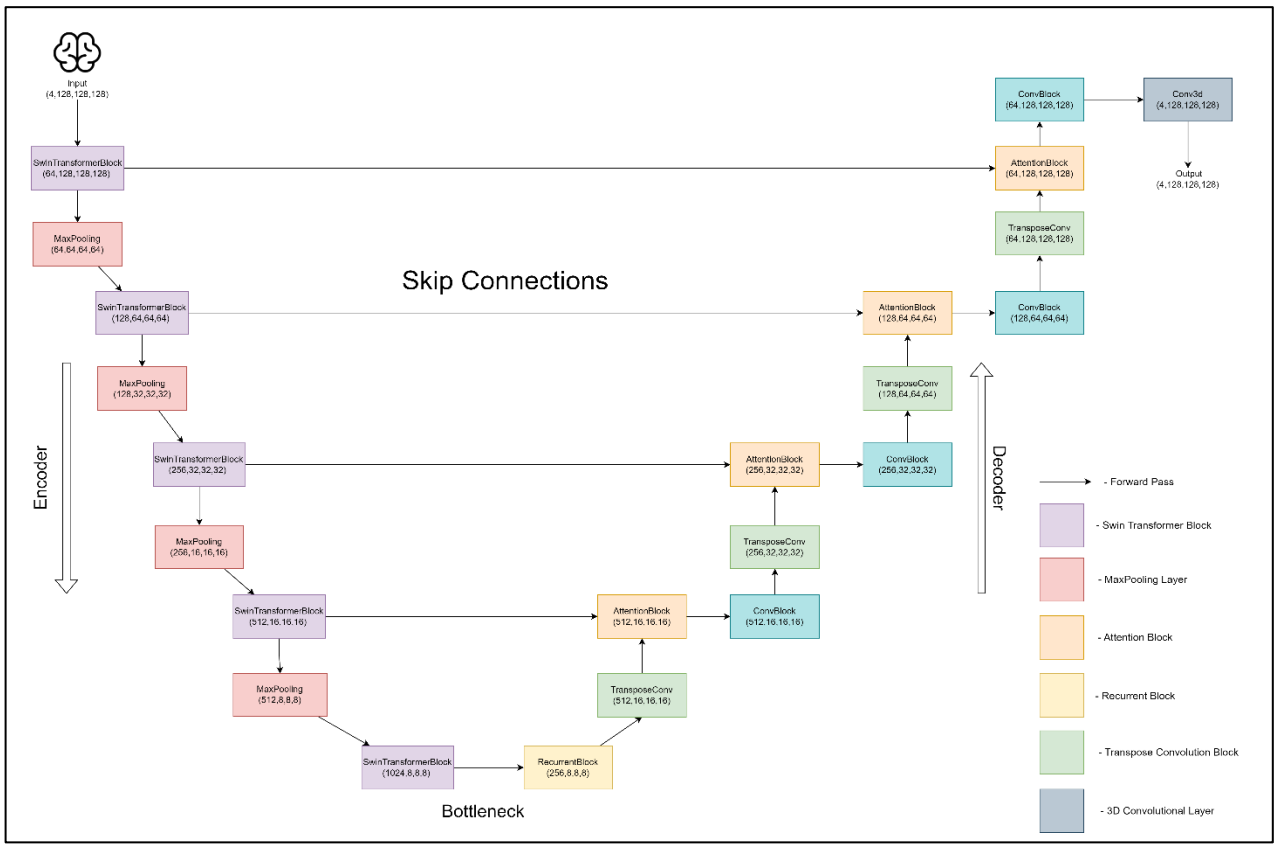

Figure 2: *Our Swin-Unet Model Architecture*

3.2 Swin Transformer block

Different from the conventional multi-head self attention (MSA) module, Swin transformer [3] is a hierarchical ViT that computes self-attention in an efficient shifted window partitioning scheme. In Figure 3, two consecutive Swin transformer blocks are presented. Each Swin transformer block is composed of LayerNorm (LN) layer, multi-head self attention module, residual connection and 2-layer MLP with GELU non-linearity. The window-based multi-head self attention (W-MSA) module and the shifted window-based multi-head self attention (SW-MSA) module are applied in the two successive transformer blocks, respectively. Based on such window partitioning mechanism, continuous Swin transformer blocks can be formulated as:

$$\hat{z}^{l} = W\text{-}MSA\left(LN(z^{l-1})\right) + z^{l-1}, \tag{1}$$

$$z^{l} = MLP\left(LN(\hat{z}^{l})\right) + \hat{z}^{l}, \tag{2}$$

$$\hat{z}^{l+1} = SW\text{-}MSA\left(LN(z^{l})\right) + z^{l}, \tag{3}$$

$$z^{l+1} = MLP\left(LN(\hat{z}^{l+1})\right) + \hat{z}^{l+1}, \tag{4}$$

where $\hat{z}^{l}$ and z^{l} represent outputs of the (S)WMSA and MLP modules for the l^{th} block, respectively. Consistent with preceding works, self-attention is computed as follows:

$$\text{Attention}(\mathrm{Q_i}, \mathrm{K_i}, \mathrm{V_i}) = \text{softmax}\left(\frac{\mathrm{Q_i K_i^T}}{\sqrt{\mathrm{dk}}}\right)\mathrm{V}_i, \tag{5}$$

In which $\mathrm{Q}, \mathrm{K}, \mathrm{V}$ denote queries, keys, and values respectively; d represents the size of the query and key. This hierarchical approach allows for efficient computation and improved modeling of local and global dependencies. The integration of Swin Transformer blocks significantly enhances the performance of brain tumor segmentation models, leading to more accurate and reliable results.

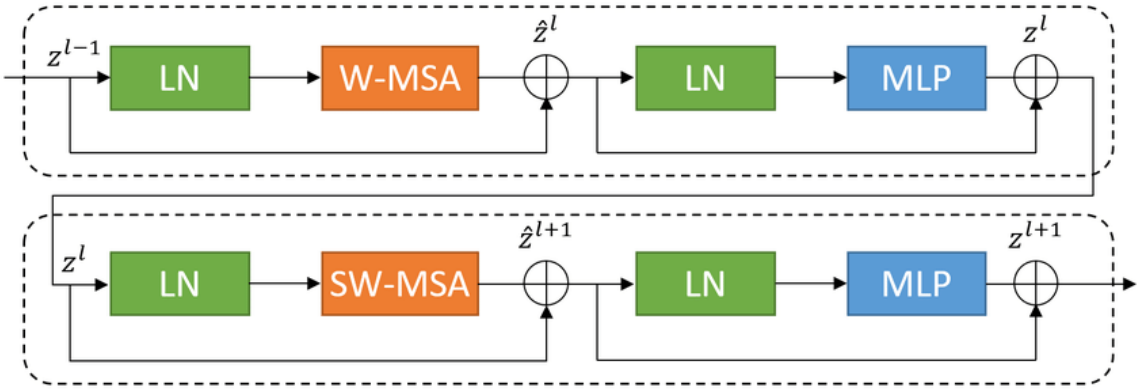

Figure 3: *Swin Transformer block*

3.3 Encoder

In the Swin-Unet architecture, the fundamental unit is the Swin Transformer block [3]. To facilitate the transformation of inputs into sequence embeddings in the encoder, images are divided into non-overlapping patches with a size of 2 x 2 x 2. Consequently, each patch's feature dimension becomes 2 x 2 x 2 x 4 = 32, considering the multi-modal MRI images with 4 channels. In our encoder, the embedding space size, denoted as C, is set to 64. These modified patch tokens then undergo processing via a sequence of Swin Transformer blocks and Max Pooling layers, crucial for constructing hierarchical feature representations. Each Max Pooling layer performs a pivotal role in down-sampling the feature maps, reducing the dimensions to $\frac{H}{2} x \frac{W}{2} x \frac{D}{2}$. Simultaneously, the Max Pooling layer reduces the number of tokens (2× down-sampling). This iterative process occurs five times within the encoder.

3.4 Recurrent block

Following the encoder, the output passes through a Recurrent Block designed to capture temporal dependencies effectively. This block contains two 3x3 convolutional layers that process the features sequentially, enhancing the model's ability to handle dynamic changes in the input data across frames or sequences.

3.5 Decoder

The Decoder architecture initiates by upscaling the output from the Recurrent Block (2× up-sampling) through transposed convolutions. The upscaled features are subsequently concatenated with corresponding features from the encoder, which were retained via skip connections. This pivotal step reinstates finer details into the processed data. After concatenation, an Attention Block is employed to the merged features, enhancing relevant encoded features selectively, thereby ensuring the model concentrates on the most crucial aspects of the image. The attended features undergo further refinement through Convolutional Blocks (ConvBlocks) in preparation for the final output segmentation.

3.6 Skip connections

Skip connections play a crucial role in our Swin-Unet architecture , establishing direct links between encoder outputs and corresponding decoder layers. These connections help preserve detailed spatial information that might otherwise be lost during deep processing, enabling more precise segmentation. Moreover, also they stabilize training by alleviating the gradient vanishing problem, improving model performance.

3.7 Loss Function

The utilized loss function, known as the soft Dice loss, plays a crucial role in our segmentation task. It evaluates dissimilarities between predicted (G) and ground truth (Y) segmentations on a voxel-wise basis. It is computed as:

$$L(G,Y) = 1 - \frac{2}{J}\sum_{j=1}^{J}\frac{\sum_{i=1}^{I} G_{i,j}Y_{i,j}}{\sum_{i=1}^{I} G_{i,j}^2 + \sum_{i=1}^{I} Y_{i,j}^2}. \quad (6)$$

where I denotes voxels numbers; J is classes number; $G_{i,j}$ and $Y_{i,j}$ denote the probability of output and one-hot encoded ground truth for class j at voxel i, respectively. This formulation accurately quantifies segmentation discrepancies while normalizing the loss across classes and voxels, thus ensuring effective and precise segmentation results in medical imaging applications.

4. Experiments

4.1 Dataset

Our model was trained and validated on BraTS2023 dataset, a public dataset provided by the BraTS challenge, which includes 1251 cases of brain images for training, each with four 3D MRI modalities T1-weighted (T1), T1-enhanced contrast (T1c), T2-weighted (T2), and T2 fluid-attenuated inversion recovery (FLAIR) and its correspondent segmentation label. The input image size of each mode is 240 × 240 × 155, which has been aligned and resampled to a 1 × 1 × 1 mm isotropic resolution and skull-stripped. The labels include a background (Label 0) and three tumor categories, namely necrotic and non-enhancing tumors (Label 1), peritumoral edema (Label 2), and enhancing tumors (Label 4). The three categories were combined into three nested subregions: whole tumor (WT, Labels 1, 2, 4), tumor core (TC, Labels 1, 4), and enhancing tumor (ET, Label 4), as shown in Figure 4.

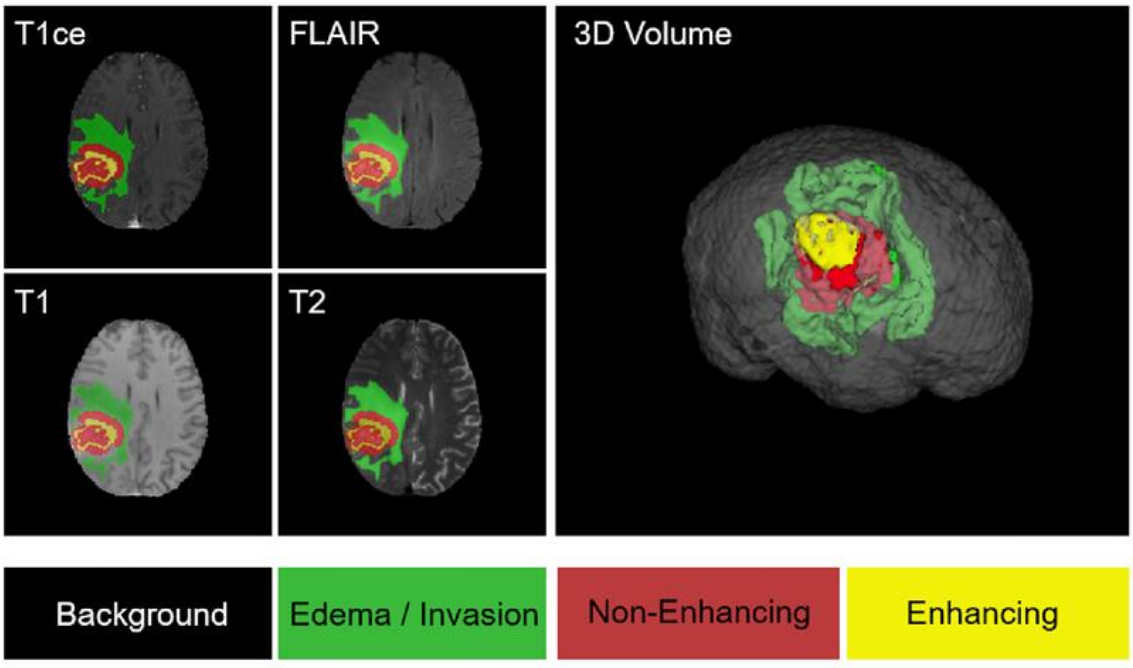

Figure 4:*Dataset different labels and modalities*

4.2 Preprocessing and Implementation details

Swin-Unet, implemented using PyTorch and trained on an NVIDIA Tesla T4 GPU, employs the Adam optimizer with a learning rate of 0.00005 and a batch size of 1. To enhance model performance, patches are first scaled to a normalized range using MinMaxScaler, ensuring consistency in input data scale. Prior to data conversion for compatibility with PyTorch, preprocessed image data is converted to tensors and transferred to the GPU for efficient computation. The processed images and corresponding labels are then saved in '.npy' format for subsequent training and testing phases. This streamlined preprocessing procedure optimally formats input data for deep learning, facilitating accurate and efficient image segmentation. The input image is cropped from size (240,240,155) to (128,128,128), reducing image volume and expediting training without significant performance compromise due to empty space margins in original images. The model is trained for a total of 150 epochs, with the dataset partitioned into an 80% training set, 10% validation set, and 10% test set.

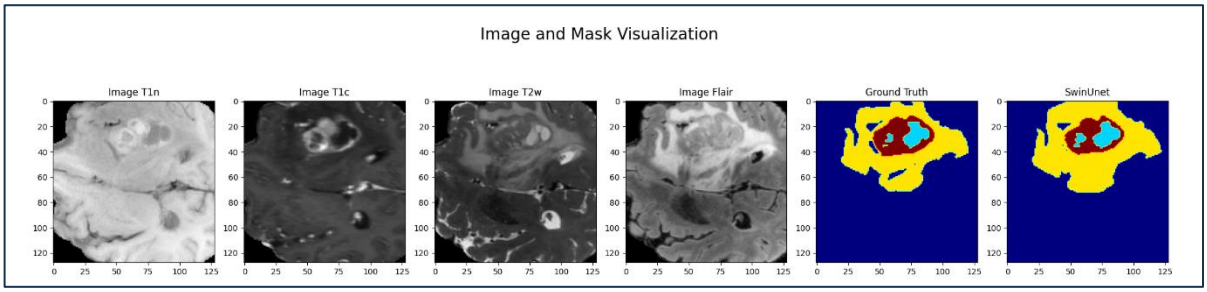

Figure 5: *Sample output from our Swin-Unet model*

5. Results and Discussion

5.1 Results

In Table 1, we compared the performance of our model, Swin-Unet, against traditional CNN methodologies for brain tumor segmentation tasks on the BraTS 2023 dataset. Since there is no enough methods to compare with because of the novelty of the dataset, one of the methods we compared with is our U-net model, which has been specifically optimized for high-contrast tumor regions. While U-net has historically shown high performance in segmenting medical images, our Swin-Unet introduces an improved architecture that integrates hierarchical transformers, allowing for the capture of more nuanced features.

Our architecture yields outstanding results with Dice scores of 0.807, 0.839, and 0.864 for the Enhancing Tumor (ET), Tumor Core (TC), and Whole Tumor (WT) classes respectively, on the validation dataset. This comparison aims to highlight the advancements Swin-Unet offers in terms of accuracy, providing insights into its potential benefits over conventional U-net in handling complex brain tumor segmentation tasks. Figure 5 shows a sample of the segmentation output produced by our Swin-Unet model on the BraTS 2023 dataset. Additionally, Figure 6 illustrates another output visualized using ImFusion's GUI [26].

	Dice			
Models	ET	TC	WT	Avg.
Optimized U-Net [24]	0.752	0.774	0.825	0.783
ReFuSeg [25]	0.786	0.832	**0.910**	0.842
Our U-net	0.718	0.743	0.775	0.745
Our Swin-Unet	**0.807**	**0.839**	0.864	0.836

Table 1: Comparison of our model Swin-Unet on BraTS 2023 validation dataset in terms of the Dice Loss values

5.2 Limitations

Limited access to GPUs and financial constraints posed significant challenges, impeding our capacity to effectively train and optimize the intricate Swin-Unet model. This scarcity of computational resources restricted extensive experimentation and slowed the iterative refinement process of the model. Moreover, the inherent complexity of Swin-Unet demanded meticulous parameter tuning, a task rendered challenging under such resource limitations. These challenges underscore the critical need for enhanced support and resources for student-led research endeavors in advanced machine learning applications, particularly in domains like medical image analysis.

5.3 Future work

To enhance the model's performance and applicability, we can implement several strategies. Firstly, expanding the dataset and employing data augmentation techniques can generate anatomically accurate training samples, aiding the model in generalizing across

diverse medical imaging scenarios. Furthermore, reducing the model's complexity can improve efficiency and deployment ease. Additionally, developing a custom graphical user interface (GUI) can enhance user interaction and accessibility, making the model more practical for widespread application in clinical environments.

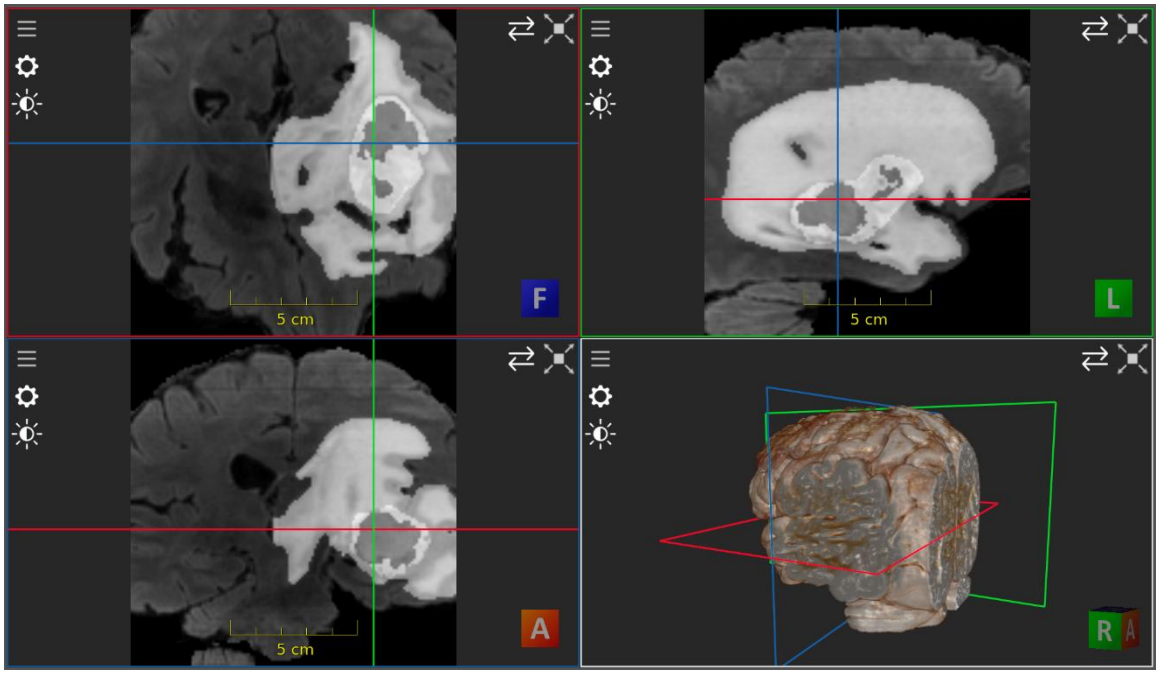

Figure 6: *Output displayed using ImFusion's GUI.*

6. Conclusion

In this paper, we proposed our Swin-Unet model which is a new architecture for semantic segmentation of brain tumors using multi-modal MRI images. Our proposed model has a U-shaped network design and uses a Swin transformer as the encoder and CNN-based decoder that is connected to the encoder via skip connections at different resolutions. We have validated the effectiveness of our approach by in the BraTS 2023 challenge dataset. Our architecture yields outstanding results with Dice scores of 0.807, 0.839, and 0.864 for the Enhancing Tumor (ET), Tumor Core (TC), and Whole Tumor (WT) classes respectively, on the validation dataset. In future, we plan to make strategies such as dataset expanding and augmentation, model complexity reduction, and GUI development to enhance the performance and usability of the model for widespread application in medical imaging.

7. References

[1] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox., "U-net: Convolutional networks for biomedical image segmentation," *Medical Image Computing and Computer-Assisted Intervention--MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18,* pp. 234--241, 2015.

[2] Vaswani Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin., "Attention is all you need.," Advances in neural information processing systems,, 2017.

[3] Liu Ze, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo., "Swin transformer: Hierarchical vision transformer using shifted windows," Proceedings of the IEEE/CVF international conference on computer vision, 2021.

[4] Myronenko, Andriy, "3D MRI brain tumor segmentation using autoencoder regularization," 2019.

[5] Jiang, Zeyu, Changxing Ding, Minfeng Liu, and Dacheng Tao., "Two-stage cascaded u-net: 1st place solution to brats challenge 2019 segmentation task," Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part I 5, 2020.

[6] Isensee Fabian, Paul F. Jäger, Peter M. Full, Philipp Vollmuth, and Klaus H. Maier-Hein., "nnU-Net for brain tumor segmentation," Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers, Part II 6, 2021.

[7] Luu, Huan Minh, and Sung-Hong Park., ""Extending nn-UNet for brain tumor segmentation."," In International MICCAI brainlesion workshop, pp. 173-186. Cham: Springer International Publishing,, 2021.

[8] Xie, Yutong, Jianpeng Zhang, Chunhua Shen, and Yong Xia, ""Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation."," In Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24, pp. 171-180. Springer International Publishing,, 2021.

[9] Wang, Wenxuan, Chen Chen, Meng Ding, Hong Yu, Sen Zha, and Jiangyun Li., "Transbts: Multimodal brain tumor segmentation using transformer," 2021, Medical Image Computing and Computer Assisted Intervention--MICCAI 2021: 24th International Conference, Strasbourg, France, September 27--October 1, 2021, Proceedings, Part I 24.

[10] Chen, Jieneng, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou., "Transunet: Transformers make strong encoders for medical image segmentation," arXiv preprint arXiv:2102.04306, 2021.

[11] Hatamizadeh, Ali, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R. Roth, and Daguang Xu., "Unetr: Transformers for 3d medical image segmentation," Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2022.

[12] Zhou, Zongwei, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang., "Unet++: A nested u-net architecture for medical image segmentation," Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, P, 2018.

[13] Milletari, Fausto, Nassir Navab, and Seyed-Ahmad Ahmadi., "V-net: Fully convolutional neural networks for volumetric medical image segmentation," 2016 fourth international conference on 3D vision (3DV), 2016.

[14] Huang, Huimin, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu., "Unet 3+: A full-scale connected unet for medical image segmentation," ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2020.

[15] Qin, Chuanbo, Yujie Wu, Wenbin Liao, Junying Zeng, Shufen Liang, and Xiaozhi Zhang, "Improved U-Net3+ with stage residual for brain tumor segmentation," BMC Medical Imaging, 2022.

[16] Zhou T., Ruan S., Vera P., & Canu S., ""A Tri-Attention fusion guided multi-modal segmentation network."," Pattern Recognition 124 (2022): 108417., 2021.

[17] Jia Q., & Shu H., ""BiTr-Unet: A CNN-Transformer Combined Network for MRI Brain Tumor Segmentation.," arXiv 2021." arXiv preprint arXiv:2109.12271., 2021.

[18] Cao, Hu, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," European conference on computer vision, 2022.

[19] Wu, Yixuan, Kuanlun Liao, Jintai Chen, Jinhong Wang, Danny Z. Chen, Honghao Gao, and Jian Wu, "D-former: A u-shaped dilated transformer for 3d medical image segmentation," Neural Computing and Applications, 2023.

[20] Hatamizadeh, Ali, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R. Roth, and Daguang Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," International MICCAI Brainlesion Workshop, 2021.

[21] ZongRen, Li, Wushouer Silamu, Wang Yuzhen, and Wei Zhe, "DenseTrans: Multimodal Brain Tumor Segmentation Using Swin Transformer," IEEE Access, 2023.

[22] Vatanpour, Marjan, and Javad Haddadnia., "TransDoubleU-Net: Dual Scale Swin Transformer With Dual Level Decoder for 3D Multimodal Brain Tumor Segmentation," IEEE Access, 2023.

[23] Cai, Yimin, Yuqing Long, Zhenggong Han, Mingkun Liu, Yuchen Zheng, Wei Yang, and Liming Chen, ""Swin Unet3D: a three-dimensional medical image segmentation network combining vision transformer and convolution."," BMC medical informatics and decision making 23, no. 1 (2023): 33., 2023.

[24] Ren T., Honey E., Rebala H., Sharma A., Chopra A., & Kurt M., "An Optimization Framework for Processing and Transfer Learning for the Brain Tumor Segmentation.," arXiv preprint arXiv:2402.07008., 2024.

[25] Kasliwal, Aditya, Sankarshanaa Sagaram, Laven Srivastava, Pratinav Seth, and Adil Khan., ""ReFuSeg: Regularized Multi-Modal Fusion for Precise Brain Tumour Segmentation."," arXiv preprint arXiv:2308.13883 (2023)., 2023.

[26] NVIDIA, "imfusion," NVIDIA, [Online]. Available: https://www.imfusion.com/products/imfusion-suite.