Array 25 (2025) 100376

Contents lists available at ScienceDirect

Array

journal homepage: www.sciencedirect.com/journal/array

An attention based residual U-Net with swin transformer for brain MRI segmentation

Tazkia Mim Angona, M. Rubaiyat Hossain Mondal *

Institute of Information and Communication Technology, Bangladesh University of Engineering and Technology (BUET), Dhaka, 1205, Bangladesh

ARTICLE INFO

Keywords:
Brain tumor segmentation
MRI
Deep learning
Convolutional neural networks
Attention
Swin transformer

ABSTRACT

Brain Tumors are a life-threatening cancer type. Due to the varied types and aggressive nature of these tumors, medical diagnostics faces significant challenges. Effective diagnosis and treatment planning depends on identifying the brain tumor areas from MRI images accurately. Traditional methods tend to use manual segmentation, which is costly, time consuming and prone to errors. Automated segmentation using deep learning approaches has shown potential in detecting tumor region. However, the complexity of the tumor areas which contain various shapes, sizes, fuzzy boundaries, makes this process difficult. Therefore, a robust automated segmentation method in brain tumor segmentation is required. In our paper, we present a hybrid model, 3-Dimension (3D) ResAttU-Net-Swin, which combines residual U-Net, attention mechanism and swin transformer. Residual blocks are introduced in the U-Net structure as encoder and decoder to avoid vanishing gradient problems and improve feature recovery. Attention-based skip connections are used to enhance the feature information transition between the encoder and decoder. The swin transformer obtains broad-scale features from the image data. The proposed hybrid model was evaluated on both the BraTS 2020 and BraTS 2019 datasets. It achieved an average Dice Similarity Coefficients (DSC) of 88.27 % and average Intersection over Union (IoU) of 79.93 % on BraTS 2020. On BraTS 2019, the model achieved an average DSC of 89.20 % and average IoU of 81.40 %. The model obtains higher DSC than the existing methods. The experiment result shows that the proposed methodology, 3D ResAttU-Net-Swin can be a potential for brain tumor segmentation in clinical settings.

1. Introduction

The human brain is a complex organ with around 100 billion neurons and non-neuronal glial cells. Glioblastoma (GBM) is a very invasive and fatal form of brain tumor that poses a significant health risk [1]. One of the worst types of cancer in the world is brain tumors [2]. Brain tumors arise in the central nervous system due to abnormal cell respiration and proliferation. If not accurately identified, brain tumors can be fatal. However, not all brain tumors are cancerous; only malignant ones are. Malignant tumors grow faster than non-cancerous tumors [3]. Gliomas are the most typical type of malignant brain tumor which originates from glial cells. There are two types of gliomas: high-grade (HGG) and low-grade (LGG); HGG is considered to be more aggressive and has a worse prognosis [4]. Precise segmentation of brain tumor is very crucial for diagnosis and treatment planning. The usual course of treatment for gliomas is a combination of radiation, chemotherapy, and surgery. However, radiologists have difficulty detecting gliomas because of their variable size and location. The location, type, and grade of the tumor all influence how well it responds to treatment; therefore, segmentation is crucial to determining the best course of action [5].

Medical image segmentation is essential in determining the tumorous regions from the normal tissues in brain image data. The typical method used in the diagnosis is magnetic resonance imaging (MRI). MRI has different modalities, most used ones in brain tumor diagnosis are T1-weighted, T2-weighted, fluid-attenuated inversion recovery (FLAIR), and contrast-enhanced T1-weighted sequences. These modalities allow for a more detailed analysis of the tumor by providing information about the tumor regions, healthy tissues and edema [6]. Radiologists manually segment the brain tumor traditionally, but this is a tedious and error-prone procedure. Automated segmentation algorithms can help radiologists to identify brain tumors from MRI images accurately and quickly. Deep learning (DL) methods have facilitated many new segmentation methods. This process has helped in the early identification of malignant tumors and preventive measures can be implemented to save lives [7]. Brain tumor segmentation is complex as the brain cancerous cells vary in size, shape, and appearance. Many

* Corresponding author.
E-mail address: rubaiyat97@iict.buet.ac.bd (M.R.H. Mondal).

https://doi.org/10.1016/j.array.2025.100376
Received 8 October 2024; Received in revised form 27 January 2025; Accepted 30 January 2025
Available online 30 January 2025

researchers have developed DL algorithms for this, but due to the variability and complexity of brain structures, identifying tumors from 3D MRI scans remains a challenging task. The existing models often struggle to capture both fine-grained and broader contextual information. Some of these existing models have a disadvantage in that they rely on 2D patches which may result in the loss of some 3D spatial information. Furthermore, these models are often found ineffective at adjusting to newer data and perform inconsistently across diverse datasets, which means their performance depends on the datasets considered. While transformers have shown promising results in computer vision tasks, their application in 3D medical imaging is still underexplored. A robust and reliable solution has yet to be developed for brain tumor segmentation. In this paper, we introduce a 3D Attention-based Residual U-Net with Swin Transformer (3D ResAttU-Net-Swin), which combines residual U-Net, attention-based skip links, and a swin transformer to improve brain tumor segmentation performance by addressing these challenges. The architecture effectively segments 3D brain MRI images. Residual U-Net uses residual blocks as encoder and decoder in the state-of-the-art method, U-Net. Attention-based skip connections and a swin transformer block in the bottleneck is introduced in the model. This hybrid structure helps the network to learn both global and local features and improves segmentation performance. The contributions of our work are stated below.

- *The proposed hybrid model, Res-AttU-Net-Swin is able to deal with 3D images directly and improves segmentation accuracy by extracting both low- and high-level features effectively from brain MRI images.*
- *4 level encoder and decoder blocks with residual connections enhance feature learning and address the vanishing gradient problem. Attention based skip connections encourage feature recovery while narrowing the semantic gap between the encoder and decoder. A 3D Swin Transformer is introduced in this model to capture long-range dependencies, which is necessary for precise segmentation of brain tumors of different sizes.*
- *The model's performance is evaluated on the publicly available BraTS-2020 and BraTS 2019 datasets where it has demonstrated competitive results against state-of-the-art. This achievement highlights the proposed model's potential for practical applications in real-world settings for accurate brain tumor diagnosis.*

The article's remaining sections are arranged as follows: The review of the most recent Brain Tumor Segmentation techniques is covered in Section 2. The details of the dataset and the proposed model are described in Section 3. The evaluation result and comparison with existing work are mentioned in Section 4. Finally, Section 5 summarizes the paper.

2. Literature review

Recent advancements in DL have shown positive outcomes in medical image segmentation tasks. Numerous scholars have explored various machine learning and DL techniques, particularly convolutional methodologies, to find the most effective solution. This section describes some of the existing work [8–18] of automated methods of brain tumor segmentation.

2.1. Convolutional neural network (CNN) based methods

Pereira et al. [8] introduced a convolutional neural network (CNN) architecture for brain tumor segmentation. The architecture used small kernels (3x3) to build a deeper CNN network to address overfitting. The design aimed to extract features on 2D patches of MRI images. While the small patches reduce computational load, it sacrifices the global context, such as tumor boundaries by focusing on smaller regions. The method also lacks 3D spatial information, which is crucial for correct tumor segmentation across multiple MRI slices. A multi-task CNN model was implemented by Zhou et al. [11] to address some of these issues, which analyzed all tumor regions concurrently in MRI scans. The shared encoder layers extract common features, while separate decoders focus on different tumor regions, therefore improving segmentation by learning shared and task specific features. Despite the improvements, the lack of 3D context is a major drawback as CNNs are generally not suited to capture complex spatial relationships across slices. These DL methods improved the segmentation performance over traditional ones, but their inability to learn 3D spatial features necessitates the development of more advanced architectures.

2.2. U-Net based methods

Many researchers subsequently applied CNN models for brain tumor segmentation. The most popular DL model in medical image segmentation, U-Net was introduced by Ronneberger et al. [9]. U-Net is a U-shaped neural network model with two main parts: the encoder and the decoder, connected by skip connections, highly efficient in learning spatial information. This structure showed promising results in medical image segmentation and was adapted as a basis for many research studies. Brain tumor segmentation method uses 2D or 3D convolutions. 2D convolutions do not fully leverage the spatial features of medical images while being computationally efficient. On the other hand, 3D convolutions provide richer spatial information but are computationally expensive and need a lot of memory. To address this trade-off, Chen et al. [10] introduced S3D-UNet model that used separable 3D convolutions. This method reduces the computational cost and memory requirement by using separable 3D convolutions. The model splits the 3D convolution into three parallel branches focusing on features across multiple perspectives simultaneously. Each 3D convolution is replaced with a one 2D and one 1D convolution layer to extract spatial and temporal features effectively. However, the simplification in convolutional layers limited the model's ability to learn complex features. Qamar et al. [16] proposed HI-Net, built upon the U-Net architecture. This model used hyperdense inception block as encoders, which has stacked factorized 3D convolutional layers with residual connections. The model utilizes hyperdense connections to extract contextual information. The dense connections and multiple branches increase the model complexity and computational load. Raza et al. [17] presented dResU-Net, a hybrid model of U-Net and residual network. This model deals with the vanishing gradient problem by using residual network as encoder and extracts contextual information from multimodal MRI. The framework captured low-level features and transmitted them to the decoder layers with skip connections. The model generalized well on an external dataset. Despite improved segmentation accuracy, the model's reliance on 3D convolution makes it less efficient at handling small and irregularly shaped tumor regions located near complex structures, because of limited spatial information.

2.3. Transformer based methods

CNNs struggle to capture long-range relationships from an image and gradients for low-level features become low in deeper layers. This affects the segmentation results. Vision Transformers (ViTs) address these issues by processing the entire image and calculates the relationship between pixels dynamically. This enables ViTs to capture both local and global dependencies. Thus, improving the segmentation accuracy. This progress has led to the creation of many transformer-based networks. A TransBTS architecture was developed by Wang et al. [12] which integrated Transformers into a 3D encoder-decoder architecture. Initially, the encoder used 3D CNNs to extract local features, which are then processed by the Transformer module to capture global information. This framework can process 3D MRI image slices at a time. The decoder subsequently used these enriched features to generate a detailed tumor segmentation map. The hybrid design improves detailed segmentation but is still limited by computational cost of transformers. Li et al. [13] introduced a TransU2 model which utilized transformers in the skip

connections of the U-Net structure. U2-Net allows capturing multi-scale contextual information while transformers process long-range dependencies that aids to better segmentation results. However, due to the limitations of the vision transformer, this architecture is unable to capture feature maps of the channel dimension. Moreover, the transformers with CNN models increase architecture complexity hindering ease of implementation. Inspired by the success of vision transformers and their variants, Hatamizadeh et al. [14] proposed a novel segmentation model called Swin UNETR. The encoder in this structure was a swin transformer which was less complex and more efficient than ViTs. Swin transformer encoder extracts feature from the image by computing self-attention using shifted windows. These features were then fed to an FCNN-based decoder. The model was able to extract long range dependencies from the images and demonstrated competitive performance on challenging brain tumor segmentation tasks. However, like other transformer-based models, it still faces information loss as it segments the image into non-overlapping patches. This limits the model's ability to capture fine-grained features.

2.4. Deep learning with attention based methods

Attention mechanisms were adapted in many models to improve the segmentation results by focusing on important parts. Zhang et al. [15] introduced a 2D attention-based residual U-Net (AResU-Net), which utilized the attention modules to reduce the semantic gap and recover features between the residual encoder and decoder. The inclusion of residual connections improved training stability and accuracy without introducing extra parameters. Furthermore, attention and squeeze excitation (ASE) block is added on the skip connections to enhance local responses from the encoder. These enriched feature maps were utilized by the decoder to enhance feature recovery. However, the model's DSC score was suboptimal for the Whole Tumor (WT) region. This was due to the use of 2D slices as input, which led to context information loss. Another study by Jawad MT et al. [18] overcame this challenge by using 3D images as input within an attention-based skip connection in the U-Net structure. This approach facilitated multi-class segmentation of gliomas, enabling the identification of tumor heterogeneity. The hybrid structure of convolutional layers and attention mechanism balances local and global feature extraction. The scores for Enhancing Tumor (ET) regions were lower compared to Whole Tumor (WT) and Tumor Core (TC) regions due to the class imbalance. Additionally, the computational resources required to train the model were significant, which might be challenging in real-world settings. However, the model demonstrated quick segmentation capabilities without extensive pre-processing. The performance results of these methods on brain MRI segmentation using benchmark datasets are shown in Table 1.

Medical images in real-world clinical settings often have lower resolution and fragmented contextual details. Despite significant advancements made by recent studies outlined above in brain tumor segmentation using DL, capturing the 3D spatial, detailed, and global context of complex MRI scans remains a challenging task. Traditional CNN face challenges in modeling complex 3D MRI scans with intricate details (local features) as well as larger-scale information (global features). U-Net based methods stated in Section 2.2 are robust but need to balance between computational efficiency and proper feature extraction. Moreover, deep networks may be affected by vanishing gradients which gives rise to problems in training lower layers and results in improper segmentation. The transformer-based approaches stated in Section 2.3 are promising but in many cases struggle to converge on small datasets, limiting their clinical application. Swin transformer-based models within U-Net have not fully been explored yet. There is scope for 3D models with less complexity that can handle data through adaptation and provide reliable segmentation. Our proposed model, 3D ResAttU-Net-Swin, addresses these challenges and improves segmentation by integrating the swin transformer with residual U-Net and attention mechanism. The challenges our study overcame are stated below.

Table 1
Comparison of various methods across different BraTS dataset.

Dataset	Year	DSC Scores (WT, TC, ET)	Method	Used in
BraTS	2013	0.88, 0.83, 0.77	CNN	Pereira et al. [8]
	2015	0.78, 0.65, 0.75	CNN	Pereira et al. [8]
		0.87, 0.75, 0.64	One-pass multi-task CNN	Zhou et al. [11]
	2017	0.9128, 0.8250, 0.8084	One-pass multi-task CNN	Zhou et al. [11]
		0.892, 0.853, 0.825	AResU-Net	Zhang et al. [15]
	2018	0.89353, 0.83093, 0.74932	S3D-UNet	Chen et al. [10]
		0.876, 0.810, 0.773	AResU-Net	Zhang et al. [15]
		0.9048, 0.8759, 0.7956	GSNet	Jawad MT et al. [18]
	2019	0.9000, 0.8194, 0.7893	TransBTS	Wang et al. [12]
		0.8977, 0.8698, 0.7907	GSNet	Jawad MT et al. [18]
	2020	0.900, 0.817, 0.787	TransBTS	Wang et al. [12]
		0.874, 0.837, 0.794	HI-Net	Qamar et al. [16]
		0.8660, 0.8357, 0.8004	Residual network with 3D U-Net	Raza et al. [17]
		0.9239, 0.9103, 0.8139	GSNet	Jawad MT et al. [18]
	2021	0.9230, 0.8632, 0.8588	TransU2	Li et al. et al. [13]
		0.926, 0.885, 0.858	Swin UNETR	Hatamizadeh et al. [14]

- *3D images processing: The hybrid model handles 3D images directly, overcoming the limitation of 2D approaches.*
- *Integration Residual blocks within U-Net: Residual connections instead of CNN layers are introduced here to alleviate vanishing gradients - by allowing the skip-connection layer to receive gradient during back-propagation as well - and make training easier through passing information about low-level features.*
- *Attention Mechanism: Attention mechanisms improve feature recovery and lessen the semantic gap between an encoder-to-decoder, thereby resulting in better segmentation accuracy.*
- *Incorporation of Swin Transformer: Unlike methods relying solely on CNNs, our model incorporates a 3D Swin Transformer block. This block excels at capturing long-range dependencies and global context in MRI scans, which is crucial for the accurate segmentation of tumors with variable sizes. This is also less complex than vision transformers.*

By combining these elements, the 3D ResAttU-Net-Swin model facilitated both local and global feature extraction, resulting in a more robust brain tumor segmentation and higher DSC compared to previous methods.

3. Research methodology

This section describes the datasets, the proposed model, the proposed model's architecture and details of the implementation.

3.1. Dataset

Publicly available challenge datasets, BraTS 2019 [19–21] and BraTS 2020 [19–21] are utilized to evaluate the proposed model. These are obtained from "*Medical Image Computing and Computer- Assisted Intervention (MICCAI) Multimodal Brain Tumor Segmentation Challenge (BraTS) 2020*" [19]. These datasets contain brain MRI scans affected by

HGG and LGG glioma as NIfTI files (. nii.gz) for segmentation which are obtained by following various clinical protocols. The distribution of the subjects in training datasets is given in Table 2. Each scan is 3D ($1 \times 1 \times 1$ mm^3) and volume shape of $240 \times 240 \times 155$. These scans describe four modalities: a) T1-weighted, b) T1-weighted contrast-enhanced, c) T2-weighted, and d) FLAIR images. These segmentations have been done manually by neurologists. Ground truth annotations use four label values (0, 1, 2, 4) for different regions. These tumor regions include.

- the tumor core, TC (labels 1, 4)
- the whole tumor, WT (labels 1–4)
- and the enhancing tumor core, ET (label 4)

The validation dataset contains 125 cases without publicly available ground truth. So, we have used the training set of these datasets. For BraTS 2020, 220 samples are selected from the total 369 training samples. For BraTS 2019, only the HGG samples are used, totaling 259 samples. We performed a 5-fold stratified k-fold method for the train-test split. Particularly, for BraTS 2020, 176 samples are allocated for training and 44 for testing. For BraTS 2019, 192 samples are allocated for training and 48 for testing.

3.2. Preprocessing

The following preprocessing steps were taken on the raw images of the dataset to optimize the model's performance.

Cropping: The MRI scans with the tumor and corresponding areas are in the center. Cropping approach can save computational time, cost and improves segmentation accuracy [18]. So, we used center cropping to every image in order to improve the network's learning capabilities and to save computational expenditure. This way the surroundings which are not important and can have a negative impact on the network's performance are ignored. We focus on the ROI. The original size of the images is 240×240 which are resized to 128×128. This helped the model to focus on the important details of the image and achieve better results.

Combining Modalities: The dataset contained four modalities, T1, T2, T1ce and FLAIR. These modalities were combined to form a multi-modal brain MRI image to input into the model easily. This step ensures that the data from all the modalities can be used by the model for better segmentation results.

Normalization: Furthermore, the pixel intensity values of the MRI images were normalized to range [0,1], ensuring consistency and better convergence during training. The min-max normalization technique was applied here using Eq. (1) [22].

$$X_n = (X - X_{min})/(X_{max} - X_{min}) \quad [1]$$

X_n is the normalized value, X denotes as original pixel intensity value, X_{min} is the minimum and X_{max} is the maximum pixel intensity value.

Mask Labeling: The segmentation masks contain four label values (0,1,2,4) corresponding to different tumor regions. These labels were preprocessed into three binary masks: Whole Tumor (WT), which combines labels 1,2,4; Tumor Core (TC), which combines labels 1 and 4; and Enhancing Tumor (ET), which uses label 4. These multi-channel masks helped the model to segment the different tumor regions accurately.

After these steps, the data was fed to the model for training. No data augmentation technique was utilized here. The brain images and masks had shapes of (4,128,128,128) and (3,128,128,128), respectively, where 4 denotes the combined four modalities and 3 corresponds to the three new labels.

Table 2
Distribution of subjects in training datasets.

Training Dataset	Total Samples	HGG	LGG
BraTS 2020	369	293	76
BraTS 2019	335	259	76

3.3. Network fundamentals

3.3.1. Residual U-Net

U-Net is a CNN based "U-shaped" architecture which is mainly used for segmentation tasks. The U-Net model consists of an encoder followed by a decoder. The encoder is responsible for extracting features from the high-resolution input image and the decoder is responsible for up sampling intermediate features and producing the final output. The encoder and decoder are symmetrical and connected by paths [9]. One of the major problems in training a neural network is vanishing gradient problem. Gradient refers to the loss function with respect to the weights and is calculated using backpropagation to update the weights. Gradient in earlier layers of the network becomes extremely small. When the gradient gets vanishingly small, the updated weights barely change from their initial value since the weights update proportionately to the gradient. As a result, the weights become stuck and never updates to its optimal value. Consequently, it impairs the ability of the network to learn. This issue can be addressed by residual U-Net, mainly by using residual blocks in the U-Net structure [23]. A residual block is a set of layers that have a shortcut connection which skips one or more layers. The output of the block is given by Eq. (2) [17].

$$y = F(x, \{W_i\}) + x \quad (2)$$

Here, $\boldsymbol{y}$ is the output and $\boldsymbol{x}$ is the input of the residual block. $\boldsymbol{F}(\boldsymbol{x}, \{\boldsymbol{W_i}\})$ is the function of the residual block, typically a series of convolutional layers, where $\boldsymbol{W_i}$ is the weights of the ith layer at the residual unit. This structure permits gradients to pass straight through the network's layers via shortcut connection. As a result, convergence is faster and consistent while training deeper networks.

3.3.2. Swin transformer

CNNs struggle to capture global information and relationships within an image. Vision transformers can help with this issue in this situation [24]. Vision transformers perform better than CNNs in visual contexts because they use self-attention techniques to extract long-range relationships from raw input. But vision transformer suffers from limitations when it comes to high resolution images. To solve this problem, swin transformer is built upon the success of vision transformer architecture. Swin transformer architecture was first introduced in Ref. [25] in 2021. Swin transformers outperform vision transformers due to their ability to handle large images with lower computational complexity.

The first step of the swin transformer architecture is pacification. The input image is initially divided into patches. Then linear embedding layers convert the image pixels into a numerical representation or a vector. These vectors then are fed to the transformer blocks. The structure of a swin transformer block is described in Fig. 1.

Swin transformer blocks consist of two subunits. Each subunit is composed of a normalization layer, attention layer, another normalization layer and a multi-layer perceptron layer. The first subunit contains a window multi-head self-attention (W-MSA) where attention is calculated within non-overlapping windows. This reduces the computational complexity by refining the self-attention calculation to local areas. In Eq. (3) [26] and (4) [26], computational complexity of the MSA and W-MSA is described. W-MSA is linear whereas the MSA is quadratic to patch number.

$$\Omega(MSA) = 4hwC^2 + 2(hw)^2C \quad (3)$$

$$\Omega(W-MSA) = 4hwC^2 + 2M^2hwC \quad (4)$$

As the number of patches $\boldsymbol{hw}$ rises, the original MSA's quadratic term, $(\boldsymbol{hw})^2\boldsymbol{C}$, becomes costly, particularly for high-resolution images.

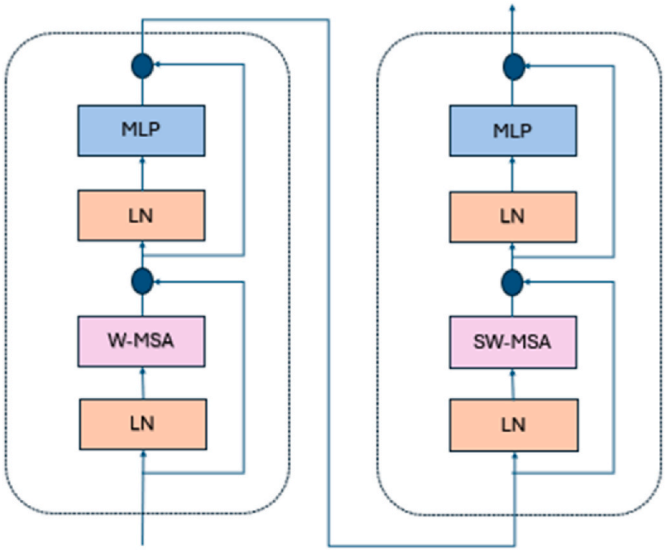

Fig. 1. Swin transformer block.

In comparison, $2M^2hwC$, the linear term in W-MSA, is far more scalable. Since the window size M is fixed, the term $2M$ remains constant and does not increase as the number of patches hw increases. The second subunit applies a shifted window multi-head self-attention (SW-MSA). Here, a cycle shift method is used to see the cross-window connections which helps the model to capture global context. After the cyclic shift, self-attention is applied within the shifted windows. The self-attention mechanism in Swin Transformer improves the capturing of positional relationships between patches by using a relative positional bias. The attention function is defined in Eq. (5) [26].

$$Attention\ (Q, K, V) = SoftMax\left(\frac{QK^T}{\sqrt{d}} + B\right)V \tag{5}$$

Here, Queries (Q), Keys (K), Values (V) are derived from input patches to compute attention weights, d is the dimension of the vectors, and B is relational positional bias matrix which accounts for the positional relationships between patches. The multi-layer perceptron layer is composed of two fully connected layers with non-linear activation. This captures non-linear relationships between features. To stabilize the training process, layer normalization is applied both before and after each MSA and MLP layer. By introducing a residual connection, the input patches are added to output without going through the full block. This prevents vanishing gradients and aids in information preservation. Lastly, swin transformer selectively merges the adjacent patches to capture global information effectively. Patch merging is performed hierarchically, concatenating M neighboring patches along the channel dimension, down sampling the image by a factor of N [25].

3.4. Proposed model

The proposed model is a hybrid model, ResAttU-Net-Swin, consisting of residual U-Net with an attention mechanism in skip connection and swin transformer in the bottleneck. The first element is a 3D 4 level residual U-Net where 4 residual blocks are used to build the encoder, and 4 residual blocks are used to build the decoder. Attention blocks are embedded in each skip connection between the encoder and decoder block. This helps the network to look at the important parts of the feature map. Moreover, a 3D swin transformer block is embedded in the bottleneck which helps the model to understand the low-level features.

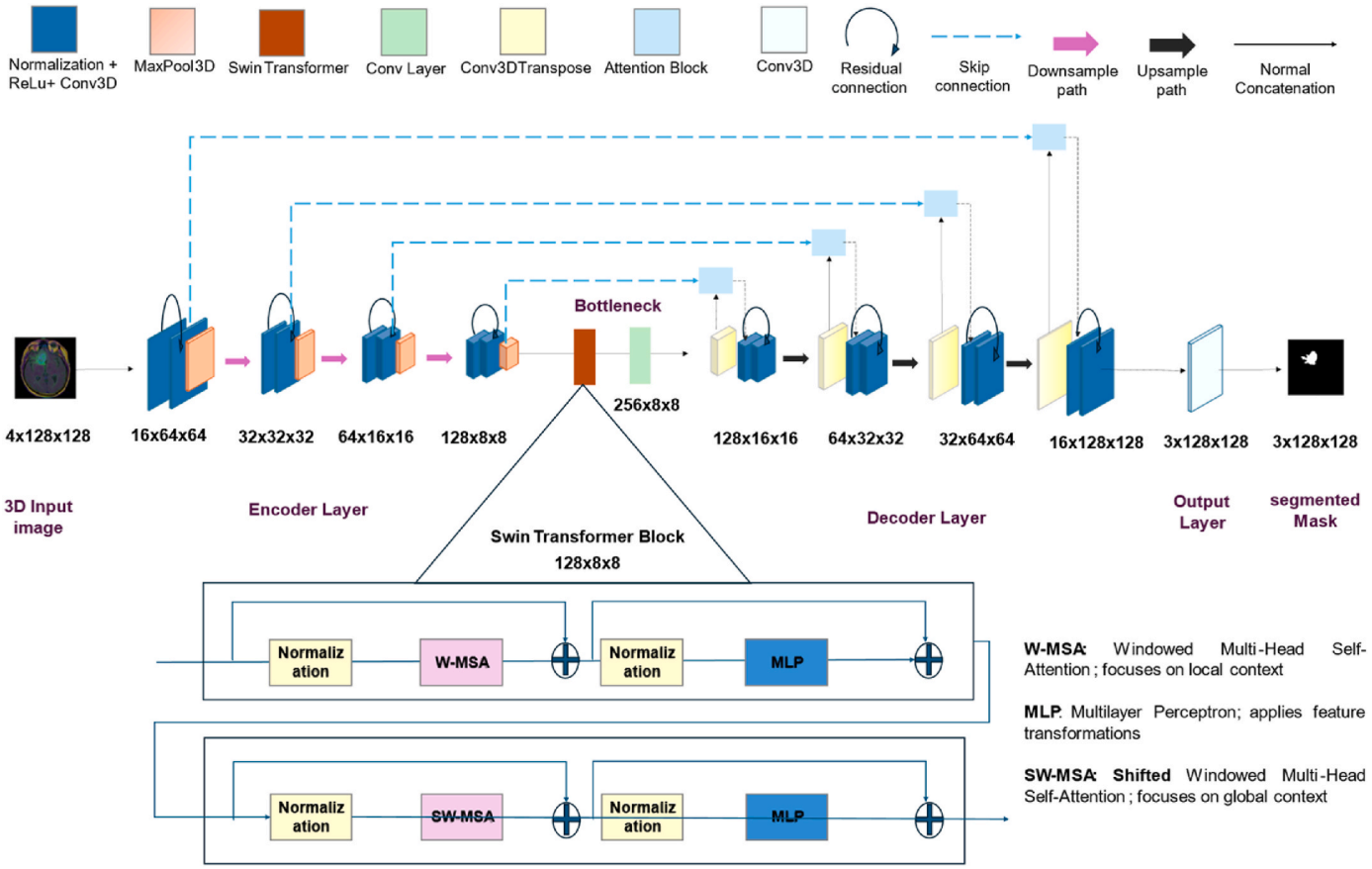

Fig. 2. Proposed ResAttU-Net-swin architecture.

The model architecture is shown in Fig. 2.

In each encoder block illustrated in Fig. 3 (Left), a 3D max pooling is first applied to reduce the input factor by 2 in each spatial dimension. This pooling operation is then followed by two sequences of group normalization, leaky ReLU activation, and a 3D convolutional layer. Moreover, there is a skip connection that normalizes and performs a separate convolution on the input before adding the output of the double convolution. This residual connection helps to mitigate the vanishing gradient problem. At every stage, the encoder blocks twice the number of channels: 16 to 32, 32 to 64, and 64 to 128 channels. The encoder blocks down sample the input and outputs abstract features in order for the network to learn complex features.

A 3D swin transformer block, located in the bottleneck of the architecture, is used to extract comprehensive information and wide-range dependencies. This block uses 128 channels to process the previous encoder's output. The block is composed of one transformer stage with 6 attention heads and uses a patch size of $2 \times 2 \times 2$ and embedding dimension of 96. The attention mechanism with shifted window method helps the network to understand the relationship between different parts of the image. The functionality of this block is mentioned in Section 3.3.2. The transformer's output is then fed through a 3D convolutional layer using a $3 \times 3 \times 3$ kernel, increasing the number of channels to 256 such that it satisfies the decoder's input requirements.

The decoder part takes the feature maps and refines them to output the final segmentation mask. A single decoder block is shown in Fig. 3 (Right). In the decoder part, a 3D up sampling layer (transposed convolution) is utilized to double the spatial dimension of the feature maps in each stage. Each up-sampling layer is followed by an attention block which enhances the features maps while emphasizing the important parts and suppressing the irrelevant ones. The attention block receives the up-sampled features and the corresponding encoder's output at the same resolution. The combined feature maps are subjected to a residual double convolution after concatenation. This process entails two sequences of group normalization, leaky ReLU activation, and convolution in addition to a skip connection. Finally, a $1 \times 1 \times 1$ convolution layer produces the final 3D segmentation mask.

The ResAttU-Net-Swin architecture contains 6,188,677 number of parameters. The model is a powerful framework for 3D medical image segmentation that combines the advantages of residual connections, attention mechanisms, and a 3D SwinTransformer block. The model learns low-level and high-level attributes by using such components.

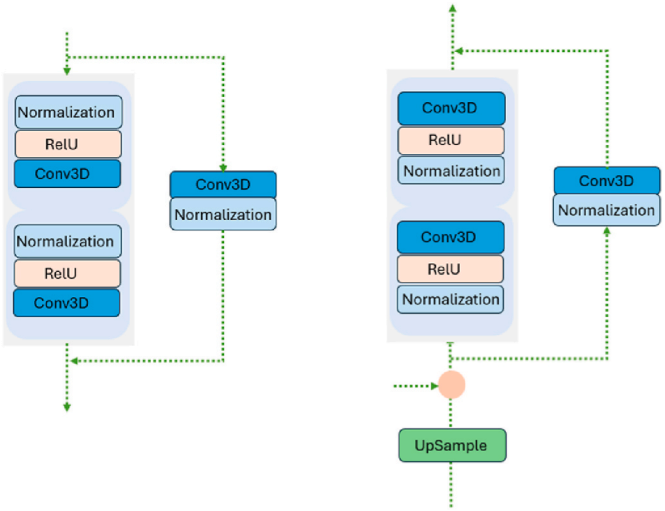

Fig. 3. (Left) encoder block (Right) decoder block.

4. Result analysis

This section describes the evaluation metrices, performance of the proposed model and its comparison with the existing methods.

4.1. Evaluation measures

The proposed model was assessed using Dice Coefficient Score (DSC), Intersection over Union (IoU), Recall and Precision metrics.

The Dice Similarity Coefficient Score (DSC) is a metric that penalizes false positives and false negatives based on how closely the anticipated segmentation mask and ground truth overlap. Better segmentation is indicated by higher values on the Dice score, which runs from 0 to 1. This is particularly used for segmentation tasks. It is calculated as given by Eq. (6) [27].

$$Dice\ Score(X, Y) = \frac{2\,|X \cap Y|}{|X| + |Y|} \tag{6}$$

In this case, the predicted segmentation masks and ground truth are denoted by X and Y, respectively.

Intersection over Union (IoU), also known as Jaccard's Index, is a critical metric for evaluating segmentation models since it measures the model's ability to differentiate between items in an image from their backgrounds. It determines how much two images—a predicted segmentation result and a ground truth—overlap. It is calculated by Eq. (7) [27].

$$Intersection\ over\ Union(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} \tag{7}$$

In the context of this research, recall measures how well the model captures all tumor regions, while precision measures how many of the predicted tumor pixels are correct. The recall and precision can be calculated by Eqs. (8) and (9), respectively [28].

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{8}$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{9}$$

4.2. Training

The proposed model was trained and evaluated using BraTS 2020 and BraTS 2019 dataset's training set. 5- fold stratified k-fold have been used to split the dataset into training and testing sets. In this case, 220 samples from 369 samples of BraTS 2020 and 240 samples (only HGG) of 335 samples of BraTS 2019 were selected. The training and testing split details are defined in Table 3.

Three distinct binary masks from the MRI segmentation masks were created. The Whole Tumor (WT) mask combines all tumor regions into one, the Tumor Core (TC) mask includes only the core tumor regions (excluding edema), and the Enhancing Tumor (ET) mask isolates just the enhancing tumor region. The model was trained using these three masks.

Pytorch (v1.12.0) library, Keras (v2.9.0) library, TensorFlow (v2.9.2) and Python (v3.7.12) Programming Language were used to implement the model. The model was trained and tested on Kaggle platform. The kernel used 4 CPU cores and NVIDIA Tesla P100-PCIE GPU with 16 GB memory. The system provided 33.66 GB RAM to

Table 3
Train test set distribution of the datasets.

Training Dataset	Total Samples	Selected	Train	Test
BraTS 2020	369	220	176	44
BraTS 2019	335	240	192	48

handle data during training. Adam optimizer with a learning rate of 0.0005 was used. Adam optimizer was selected after experiments, where Adam demonstrated superior performance. The low learning rate was determined considering the local minima problem [29]. The other parameters related to the optimizer are kept default values without tuning. The model was trained for 100 epochs on a batch size of 2. The batch size was set to 2 due to the memory constraints, with gradient accumulation over 4 steps to simulate a batch size of 8, to stabilize training. The model's performance metrics and validation loss were monitored across different ranges such as 50,70,100,150. This indicated minimal improvement beyond epoch 100. This configuration ensured optimization with less computational load.

The model consists of 4 encoders and 4 decoders, each applying a kernel size of $3 \times 3 \times 3$ with a padding of 1. The encoder blocks decrease the spatial dimension of the input, while the decoder blocks increase them by a factor of 2. The channels or number of feature maps are doubled by each encoder to capture the abstract features. The number of channels decreases by the decoders to reconstruct to the original segmentation map. The swin transformer was placed in the bottleneck with 6 attention heads, $7 \times 2 \times 2$ window size, $2 \times 2 \times 2$ patch size and embedding dimension of 96. A convolutional layer with a $3 \times 3 \times 3$ kernel size was applied after the swin transformer, increasing the number of channels to meet the decoders requirements. The model uses LeakyRelu activation function in encoder and decoder blocks for better gradient flow, while the sigmoid function is used at the output layer for better segmentation. The parameter configuration at each stage during training and evaluation are given in Table 4. To tune the hyperparameters, several experiments were performed on the model and the optimum hyperparameters are presented in Table 5.

4.3. Evaluation

The ResAttU-Net-Swin model was evaluated on the 44 samples of BraTS 2020 and 48 samples of BraTS 2019. The model segments three tumor regions: Enhancing Tumor (ET), Tumor Core (TC), and Whole Tumor (WT). The model's performance metrics on both the datasets are documented in Table 6. The DSC and IoU scores show strong segmentation across all tumor regions. The model can segment WT and TC region efficiently, achieving approximately 90 % DSC for both datasets. This highlights the model's capability to segment tumors with defined boundaries. The model showed slightly lower performance in segmenting ET region, 82.6 % and 84.0 % DSC for BraTS 2020 and 2019 datasets, respectively. This indicates the model struggles slightly to identify the enhancing tumor due to its complex nature. Despite this, the model still performs competitively for the ET region. The model achieves an average IoU of 79.93 % and 81.40 % for BraTS 2020 and BraTS 2019, respectively, demonstrating its reliability. Strong recall and precision scores reflect that the model effectively identifies relevant pixels with very low false positives, which is crucial for accurate diagnosis. Fig. 4 represents the DSC, IoU and Loss results obtained from training the proposed model on BraTS 2020 dataset for 100 epoch. The Dice and IoU plot show that both the metrices improved steadily over the training period. The training and validation scores reached high levels gradually, which suggests the model is learning and generalizing well to unseen data. The loss curve portrays a decreasing trend in both training and validation loss. The close alignment of validation loss and training loss indicates that the model is not overfitting. There is a noticeable spike in loss plot and drop in DSC, IoU plots around epoch 50, however, the model quickly recovers.

Table 4
Parameter configuration of each block of the proposed model.

Layer	Input Channel	Output Channel	Input Size	Output Size	Kernel Size
Encoder 1	4	16	128x128	64x64	3x3
Encoder 2	16	32	64x64	32x32	3x3
Encoder 3	32	64	32x32	16x16	3x3
Encoder 4	64	128	16x16	8x8	3x3
Swin Transformer	128	128	8x8	8x8	–
Convolution Layer	128	256	8x8	8x8	3x3
Decoder 1	256	128	8x8	16x16	3x3
Decoder 2	128	64	16x16	32x32	3x3
Decoder 3	64	32	32x32	64x64	3x3
Decoder 4	32	16	64x64	128x128	3x3
Output Layer	16	3	128x128	128x128	1x1

Table 5
Hyperparameters of the proposed model.

Hyperparameters	Value
Optimizer	Adam
Learning Rate	0.0005
Batch Size	2
Epochs	100
Loss Function	Binary cross entropy, Dice, Focal, Tversky Loss
Activation function in the internal convolutional layers	LeakyRelu
Activation function in the output layer	Sigmoid
Number of Attention heads in Swin Transformer	6
Embedding Dimension in Swin Transformer	96
Patch Size in Swin Transformer	2 × 2 x 2
Window Size in Swin Transformer	7 × 2 x 2

Table 6
Evaluation scores (%) on BraTS 2020 and BraTS 2019.

Dataset	Metrics	WT	TC	ET	Avg.
BraTS 2020	**DSC**	92.5	89.7	82.6	**88.27**
	IoU	86.3	82.2	71.3	**79.93**
	Recall	92.3	85.0	83.5	**86.93**
	Precision	93.0	96.0	84.3	**91.10**
BraTS 2019	**DSC**	92.4	91.2	84.0	**89.20**
	IoU	86.1	84.4	73.7	**81.40**
	Recall	90.9	89.8	85.4	**88.70**
	Precision	94.4	93.4	85.1	**90.97**

Examples of the segmented output by the proposed model and the ground truth are shown in Figs. 5 and 6. The images include multiple slices of a single MRI, with the corresponding ID mentioned below each comparison. The model demonstrates strong alignment in recognizing the boundaries of WT region with minimal mismatch. However, in some slices under segmentation is observed at the edges missing smaller tumor regions. The core regions in TC segmentation are effectively captured by the model due to the residual and attention block that enhanced the model's ability to focus on intricate patterns. However, there are marginal discrepancies in segmenting the complex regions. In case of the ET region, the model segments successfully in many slices but false negatives and occasional false positives can be observed.

4.3.1. Modality wise performance analysis

MRI gives better contrast compared to other medical imaging methods for the brain [30]. The proposed model was trained and tested using the four modalities (T1, T2, T1CE, Flair) both individually and with all the modalities combined using consistent ground truth across tests. Each modality emphasizes different tumor properties. For example, T1 and T1CE reveal Tumor Core (TC) information, while T2 and Flair focus on edema regions [31]. The performance of the model across modalities can be observed from Table 7. T1CE modality performed relatively well compared to individual modalities, particularly for TC and ET. However, the model showed poor performance in segmenting ET region, particularly with T2 and Flair modality, likely due to its limited ability to identify complex tumor regions. T2 exhibited the lowest performance in TC and ET region as it is primarily used to detect

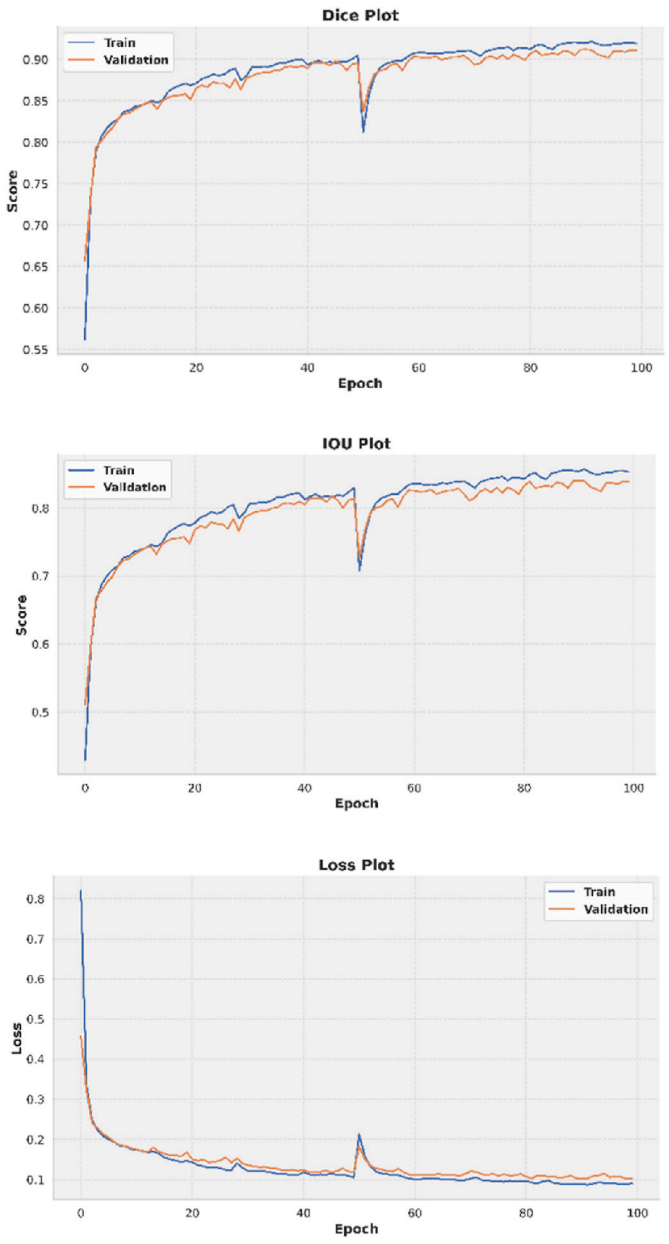

Fig. 4. DSC, IoU and Loss Curve obtained from training the proposed model on BraTS 2020 dataset.

abnormalities, but it struggles to differentiate tumors from normal tissue. T1 gave suboptimal results as it lacks contrast to tumor subregions. WT segmentation performed well across modalities, with Flair and T2 providing strong results. The irregular-shaped tumors are challenging to identify using a single modality as missing modalities mean missing information, leading to poor segmentation [32]. The model gave the best result when all the modalities are combined, extracting all the significant features, enabling the model to recognize complex tumors.

4.3.2. Ablation study

To assess the contribution of each component of the proposed model, we evaluated the segmentation performance by incrementally adding the components to the baseline U-Net model. This showed the influence of each modification in segmentation performance on the BraTS 2020 dataset. From Tables 8 and 9, the results for the DSC, IoU scores, and Wilcoxon signed-rank test trained over 50 epochs are visible. It can be seen that for the case of the U-Net model, an average DSC and IoU of 84.97 % and 75.23 % are obtained, respectively. The model performed well, but its relatively low performance, especially for the TC and ET regions, indicates its limited capacity to extract complex features. Incorporating residual blocks as encoder and decoder improved the DSC and IoU score for WT and ET region, suggesting the significance of using residual connections that helps in better information flow and prevent vanishing gradient problem. The addition of swin transformer resulted in performance improvement with an average DSC and IoU of 86.73 % and 77.70 %, respectively. This indicates that the global context capture by swin transformer improved the segmentation accuracy. The overall model demonstrated improvement having an average DSC of 87.23 % and an average IoU of 78.33 % across all regions. The combination of residual connection, swin transformer and attention method shows the progressive benefits of each component in the tumor segmentation performance. While the proposed model comes with higher complexity than baseline U-Net model, it significantly improves performance especially for challenging regions like TC and ET. The Wilcoxon signed-rank test was performed using DSC and IoU scores for the 44 test sample pairs to evaluate the statistical significance of the proposed architecture over the baseline models. The statistical analysis from Tables 8 and 9 highlights that the proposed model's results are statistically significant ($p < 0.05$) than the U-Net and Residual U-Net models. This emphasizes the importance of the modifications in improving segmentation accuracy. The p-values for the proposed model were not statistically significant compared to the Residual U-Net + Swin Transformer model, suggesting that the attention mechanism did not yield a substantial performance improvement. However, the small improvements observed, especially in the TC and ET regions, can substantially impact clinical decision-making.

To compare the efficiency of swin transformer with vision transformer, the model was initially trained with a vision transformer for 100 epoch and then evaluated. The results are mentioned in Tables 10 and 11. The quantitative results show that swin transformer improved the segmentation performance with around 64 % less parameters. The swin transformer improves performance by 0.05 %, 1.5 % in terms of average DSC and by 0.23 %, 1.7 % in terms of average IoU in BraTS 2020 and BraTS 2019 dataset, respectively. This indicates that the proposed model attains better segmentation accuracy while maintaining a less complex structure compared to the Vision Transformer-based model.

4.3.3. Comparison with existing methods

A comparison of the DSC scores and improvement % of the proposed ResAttU-Net-Swin model over existing architectures on BraTS 2020 and 2019 dataset are presented in Tables 12 and 13, respectively.

The proposed model, ResAttU-Net-Swin achieved DSC of 0.925 for Whole Tumor (WT), 0.897 for Tumor Core (TC) and 0.826 for Enhancing Tumor (ET) on BraTS 2020. The proposed model was compared with five methods for BraTS 2020: U-Net with residual convolutional encoder [17], U-Net with attention-based skip connection [18], U-Net with Transformers [33], CNN-Transformer based method [34] and a hybrid network of U-Net, V-Net and Transformers [35]. Our proposed model, ResAttU-Net-Swin achieved 4.87 % improvement in average DSC compared to Ref. [17], that incorporated residual blocks as encoder within the U-Net structure. In Ref. [18], the model implemented attention blocks in a 5-level encoder-decoder structure for better feature extraction. However, ResAttU-Net-Swin achieved 0.11 % and 1.21% higher DSC for WT and ET, respectively, and showed competitive performance for TC. The model in Ref. [33] featured a U-Net and transformer pipeline in parallel with transformers integrated in the last two skip connections, yet ResAttU-Net-Swin improved the average DSC by 4.3 %. Similarly, the CNN-Transformer method in Ref. [34] and Residual Inception based U-Net in Ref. [35] was outperformed by 1.46 % and 2.79 % respectively in average DSC. The ResAttU-Net-Swin model performed better in segmenting WT than other models. The disparities in DSC for TC between ResAttU-Net-Swin and [18], and for ET between ResAttU-Net-Swin and [34] are marginal.

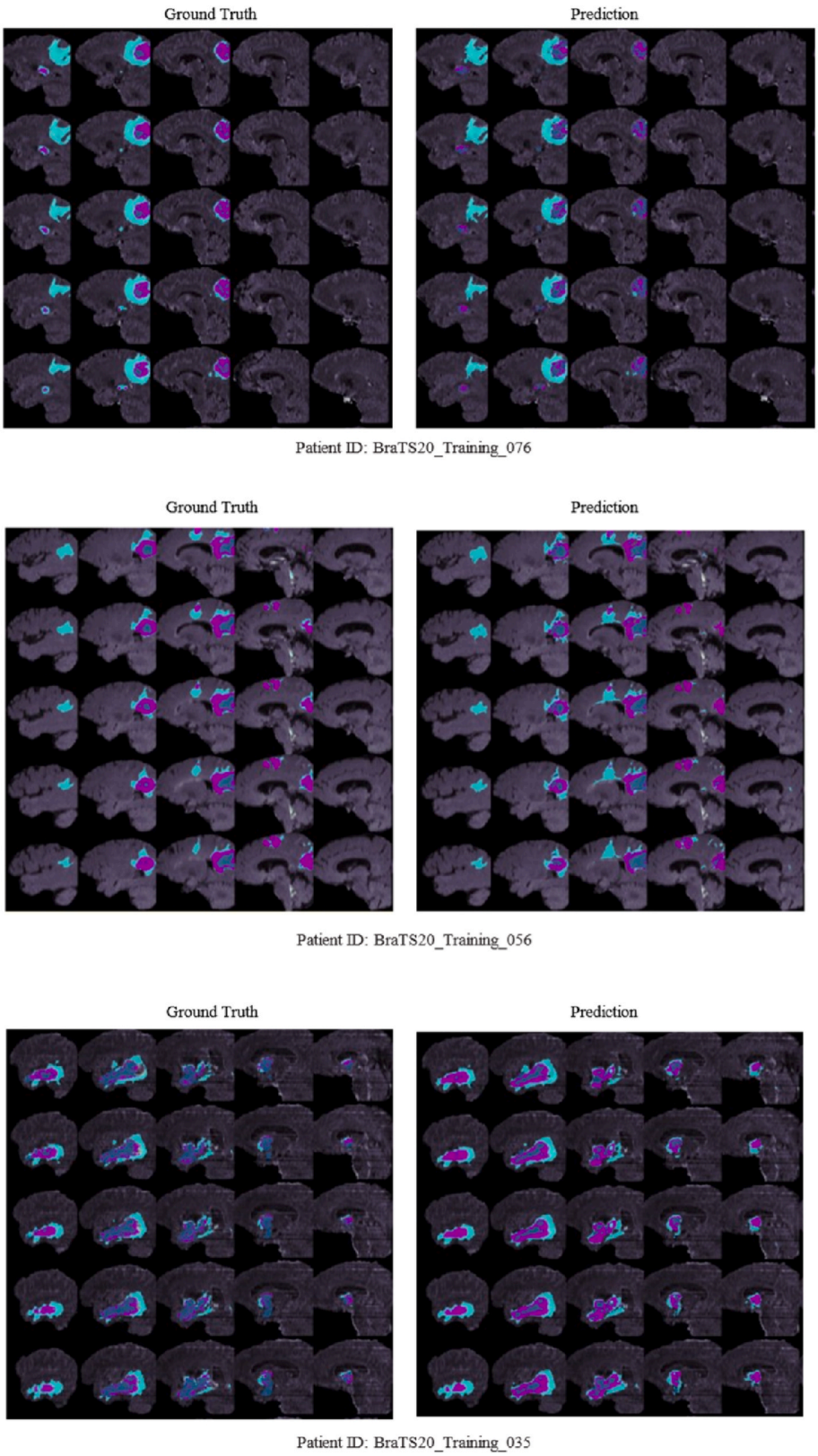

Fig. 5. Ground Truth and Predicted Mask obtained by the proposed model for ID: BraTS20 Training 076, 056, 035 (WT: cyan, TC: navy blue, ET: magenta).

For BraTS 2019, the model was compared with three methods, U-Net with attention-based skip connection [18], U-Net with residual decoder and attention gates [36] and U-Net with attention and transformer [37]. The ResAttU-Net-Swin achieved DSC of 0.924 for Whole Tumor (WT), 0.912 for Tumor Core (TC) and 0.840 for Enhancing Tumor (ET) on BraTS 2019. ResAttU-Net-Swin achieved 3.93 % higher DSC than [18]. The model in Ref. [36] featured a residual U-Net and attention blocks in the skip connection with guided loss function at decoder level for generating better feature maps. The ResAttU-Net-Swin gave superior performance than [36] across all regions due to the incorporation of

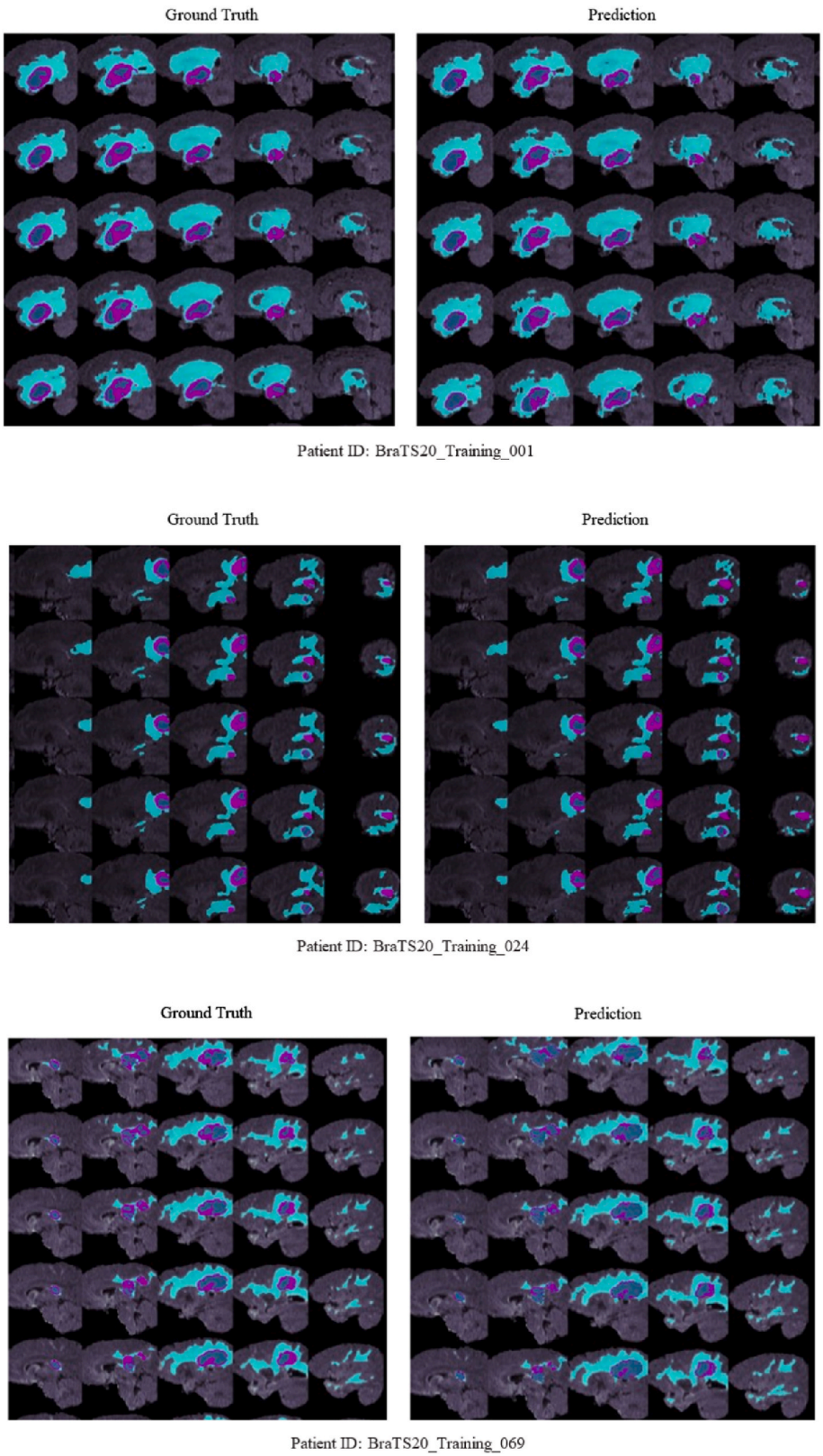

Fig. 6. Ground Truth and Predicted Mask obtained by the proposed model for ID: BraTS20 Training 001, 024, 069 (WT: cyan, TC: navy blue, ET: magenta).

swin transformer, which helped to obtain global dependencies. Against [37], which utilized attention method and transformer in the U-Net structure, the ResAttU-Net-Swin scored 3.67 % higher DSC scores. As shown in Table 13, the proposed framework outperforms the other three methods in terms of DSC for BraTS 2019 dataset. This shift from a traditional encoder-decoder approach to the Swin Transformer allowed the model to capture the complex spatial patterns in brain tumor segmentation tasks. This led the proposed model to outperform the other methods.

Compared to studies in Refs. [17,18], and [36] that use residual

Table 7
Evaluation Scores (%) across modalities after 50 epochs of training on BraTS 2020.

Modality	Metrices	WT	TC	ET	Avg.
T1CE	DSC	75.60	84.90	78.80	**79.77**
	IoU	63.30	76.30	67.40	**69.00**
FLAIR	DSC	87.70	63.30	43.50	**64.83**
	IoU	79.10	50.70	30.80	**53.53**
T1	DSC	78.10	65.50	49.30	**64.30**
	IoU	65.30	53.30	35.70	**51.43**
T2	DSC	83.40	46.90	25.80	**52.03**
	IoU	72.40	35.10	16.70	**41.40**
All Modality	DSC	91.00	89.70	81.00	**87.23**
	IoU	83.80	81.80	69.40	**78.33**

Table 8
Impact of Components (Residual Network, Swin Transformer, Attention Mechanism) on DSC (%) and Wilcoxon Signed Rank Test after 50 epochs of training on BraTS 2020.

Architecture	WT	TC	ET	Avg.	p-Value	z-Value
U-Net	89.70	86.70	78.50	**84.97**	0.0000042	−4.6447
Residual U-Net	90.10	85.20	79.60	**84.97**	0.0000025	−4.388
Residual U-Net + Swin Transformer	90.50	88.90	80.80	**86.73**	0.3693529	−0.910
Residual U-Net + Attention + Swin Transformer (Proposed)	91.00	89.70	81.00	**87.23**	-	-

Table 9
Impact of components (Residual Network, Swin Transformer, Attention Mechanism) on IoU (%) and Wilcoxon Signed Rank Test after 50 epochs of training on BraTS 2020.

Architecture	WT	TC	ET	Avg.	p-Value	z-Value
U-Net	81.70	77.90	66.10	**75.23**	0.0000002	−4.738
Residual U-Net	82.40	76.20	67.60	**75.40**	0.0000005	−4.609
Residual U-Net + Swin Transformer	83.20	80.90	69.00	**77.70**	0.3523800	−0.933
Residual U-Net + Attention + Swin Transformer (Proposed)	83.80	81.80	69.40	**78.33**	-	-

Table 10
Performance Comparison (%) of Swin Transformer and Vision Transformer after 100 epochs training on BraTS 2020.

Architecture	Parameters	Metric	WT	TC	ET	Avg.
ResAttU-Net with Vision Transformer	17.32M	DSC	92.5	89.4	82.7	**88.22**
		IoU	86.2	81.5	71.4	**79.70**
ResAttU-Net Swin Transformer	6.19M	DSC	92.5	89.7	82.6	**88.27**
		IoU	86.3	82.2	71.3	**79.93**

blocks or attention method in the U-Net, the proposed ResAttU-Net-Swin utilizes residual blocks as both encoder and decoder. This allowed the model to capture low-level features while avoiding vanishing gradient problems. Additionally, adding attention blocks helps the network focus on important regions while reducing the semantic gap between tumor regions and background, which enhances the segmentation performance. While transformers were used in Refs. [33–35,37], swin transformer outperforms the normal transformers in vision tasks and is less complex, as analyzed in Section 4.3.2, which helped our model to learn

Table 11
Performance Comparison (%) of Swin Transformer and Vision Transformer after 100 epoch training on BraTS 2019.

Architecture	Parameters	Metric	WT	TC	ET	Avg.
ResAttU-Net with Vision Transformer	17.32M	DSC	92.1	89.2	81.9	**87.70**
		IoU	85.7	82.3	71.3	**79.70**
ResAttU-Net with Swin Transformer	6.19M	DSC	92.4	91.2	84.0	**89.20**
		IoU	86.1	84.4	73.7	**81.40**

Table 12
DSC scores and improvement % of existing hybrid models with proposed model on BraTS 2020.

Model Architecture	WT	TC	ET	Avg.	Improvement %
4 level U-Net with **residual convolutional encoder** and convolutional decoder [17]	0.8660	0.8357	0.8004	0.8340	4.87 %
GSNET: 5 level convolutional encoder- decoder with **attention-based skip connections** [18]	0.9239	0.9103	0.8139	0.8827	WT: 0.11 % ET: 1.21%
UNet and **Transformer pipelines in parallel**, with Transformers in the last two skip connections [33]	0.9000	0.8360	0.7830	0.8397	4.30 %
CNN-Transformer based encoders and edge sharpening models in decoder with Sobel and Laplacian filters [34]	0.9084	0.8638	0.8322	0.8681	1.46 %
3DUV-NetR+: Hybrid architecture of **U-Net, V-Net and transformers** [35]	0.9195	0.8280	0.8170	0.8548	2.79 %
Proposed Model	**0.9250**	**0.8970**	**0.8260**	**0.8827**	-

global context more effectively. The combination of residual network, attention block and swin transformer helps the network to learn low level features, long range dependencies and deals with vanishing gradient problem. This leads to better segmentation performance than the other hybrid networks mentioned above.

The comparison clearly highlights that the ResAttU-Net-Swin model achieves exceptional performance in segmenting the WT and TC regions while maintaining a competitive DSC for the ET region. Due to the imbalanced dataset, the DSC for ET region is low compared to WT and TC. This performance underlines the model's robustness and efficacy, making it a reliable approach for brain tumor segmentation tasks.

The model achieved promising segmentation results overall. However, the lower DSC in the ET region suggests the model does not perform very well in segmenting more complex tumor regions. The model also struggles to detect the boundaries of the small tumor regions. The combination of residual U-Net, swin transformer and attention blocks in the proposed model may increase the model complexity compared to the stand-alone U-Net model.

Table 13
DSC scores and improvement % of existing hybrid models with proposed model on BraTS 2019.

Model Architecture	WT	TC	ET	Avg.	Improvement %
GSNET: 5 level convolutional encoder- decoder with **attention-based skip connections** [18]	0.8977	0.8698	0.7907	0.8527	3.93 %
Hybrid network with **Attention Gates and Res-UNet backbone**; decoder uses individual loss function at each layer [36]	0.9110	0.8760	0.8010	0.8627	2.93 %
Hybrid Architecture of **U-Net, Attention method and Transformer** [37]	0.8920	0.8450	0.8290	0.8553	3.67 %
Proposed Model	**0.9240**	**0.9120**	**0.8400**	**0.8920**	-

5. Conclusion and future directions

We introduced a hybrid brain segmentation model that is a combination of residual convolutional encoder-decoder structure with attention methods in skip connections and swin transformer in the bottleneck. This hybrid structure contributed significantly to improving segmentation by capturing local - global information and refining feature maps. The model was trained and tested on the benchmark BraTS 2020 and BraTS 2019 datasets. The model achieved high DSC for whole tumor (WT) and tumor core (TC) regions and a moderate DSC for the enhancing tumor (ET) region. It is demonstrated that the proposed model yields better results for several state-of-art methods, particularly in segmenting the WT and TC regions.

There are some limitations of the research that should be addressed in the future. The lower DSC score in the ET region suggests that there is room for improvement of the proposed model in segmenting complex tumor regions. The combination of residual U-Net, swin transformer and attention blocks increases the model complexity compared to standalone U-Net model. Data augmentation can be used to address dataset imbalances and improve accuracy for small tumor areas. Incorporating approaches like Conditional Random Fields (CRF) may also help with the detection of complicated tumor boundaries.

For future work, a large dataset with multiple modalities, brain structure segmentation and other organ segmentation can be added to the model in order to make it diverse. This will help to increase the generality of the model. Applying advanced data augmentation and post-processing techniques may further refine segmentation outcomes. Exploring 3D methods that balance computational cost with optimal contextual information is another promising direction. Furthermore, the model can be fine-tuned for the real-time segmentation for the clinical applications to provide fast and accurate segmentation for treatment planning.

CRediT authorship contribution statement

Tazkia Mim Angona: Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **M. Rubaiyat Hossain Mondal:** Writing – review & editing, Supervision, Project administration, Conceptualization.

Ethical statement

All authors have reported that they have no relationships relevant to the contents of this paper to disclose. All the ethical guidelines were followed during the research work.

Funding

The work received no funding.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research work is part of the MSc (ICT) thesis of the author Tazkia Mim Angona under the supervision of M. Rubaiyat Hossain Mondal at the Institute of Information and Communication Technology (IICT) of Bangladesh University of Engineering and Technology (BUET), Bangladesh. Hence, the authors would like to thank IICT, BUET for its support.

Data availability statement

The data used to support the findings of this study are available from the corresponding author upon request.

References

[1] Bleeker F, Molenaar R, Leenstra S. Recent advances in the molecular understanding of glioblastoma. Journal of neuro-oncology 2012;108:11–27.
[2] Bauer S, Wiest R, Nolte L, Reyes M. A survey of mri-based medical image analysis for brain tumor studies. Phys Med Biol 2013;58:R97–129.
[3] Abdalla HEM, Esmail MY. Brain tumor detection by using artificial neural network. In: Proceedings of the 2018 International conference on computer, control, electrical, and electronics engineering (ICCCEEE). IEEE; 2018. p. 1–6.
[4] Khan AR, Khan S, Harouni M, Abbasi R, Iqbal S, Mehmood Z. Brain tumor segmentation using k-means clustering and deep learning with synthetic data augmentation for classification. Microscopy Research and Technique 2021;84(7):1389–99.
[5] Kumar EK, Ajay A, Vardhini KH, Vemu R, Padmanab- ham AA. Residual edge attention in u-net for brain tumour segmentation. International Journal on Recent and Innovation Trends in Computing and Communication 2023;11(4):324–40.
[6] Wadhwa A, Bhardwaj A, Verma VS. A review on brain tumor segmentation of mri images. Magnetic resonance imaging 2019;61:247–59.
[7] Hinton G. Deep learning a technology with the potential to transform health care. Jama 2018;320(11):1101–2.
[8] Pereira S, Pinto A, Alves V, Silva CA. Brain tumor segmentation using convolutional neural networks in mri images. IEEE transactions on medical imaging 2016;35(5):1240–51.
[9] Ronneberger O, Fischer P, Brox T. U-net: convolutional net- works for biomedical image segmentation. In: Proceedings of the Medical image computing and computer-assisted intervention-MICCAI 2015: 18th international conference; 2015. Munich, Germany, 5-9 October.
[10] Chen W, Liu B, Peng S, Sun J, Qiao X. "S3d-unet: separable 3d u-net for brain tumor segmentation," brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries: 4th international work- shop, BrainLes 2018, held in conjunction with MICCAI 2018, granada, Spain, September 16, 2018. Revised selected papers, Part II 4. Springer; 2019. p. 358–68.
[11] Zhou C, Ding C, Lu Z, Wang X, Tao D. One-pass multi-task convolutional neural networks for efficient brain tumor segmentation. In: Medical image computing and computer assisted intervention-MICCAI 2018: 21st international conference, granada, Spain, 16-20 september, 2018, proceedings, Part III, vol. 11. Springer; 2018. p. 637–45.
[12] Wenxuan W, Chen C, Meng D, Hong Y, Sen Z, Jiangyun L. Transbts: multimodal brain tumor segmentation using transformer. In: Proceedings of the inter- national conference on medical image computing and computer-assisted intervention. Springer; 2021. p. 109–19.
[13] Li X, Fang X, Yang G, Su S, Zhu L, Yu Z. Transu^2-net: an effective medical image segmentation framework based on transformer and u^2- net. IEEE Journal of Translational Engineering in Health and Medicine 2023;11:441–50.
[14] Hatamizadeh A, Nath V, Tang Y, Yang D, Roth HR, Xu D. Swin unetr: swin transformers for semantic segmentation of brain tumors in mri images. In:

Proceedings of the International MICCAI brainlesion workshop. Springer; 2021. p. 272–84.
[15] Zhang J, Lv X, Zhang H, Liu B. Aresu-net: attention residual u-net for brain tumor segmentation. Symmetry 2020;12(5):721.
[16] Qamar S, Ahmad P, Shen L. Hi-net: hyperdense inception 3 d unet for brain tumor segmentation. In: Proceedings of the brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries: 6th international workshop, brain- les 2020, held in conjunction with MICCAI 2020, Lima, Peru, october 4, 2020, revised selected papers, Part II 6. Springer; 2021. p. 50–7.
[17] Raza R, Bajwa UI, Mehmood Y, Anwar MW, Ja- mal MH. dresu-net: 3d deep residual u-net based brain tumor segmentation from multimodal mri. Biomedical Signal Processing and Control 2023;79:103861.
[18] Jawad MT, Yeafi A, Halder KK. Gsnet: a multi-class 3d attention-based hybrid glioma segmentation network. Optics Express 2023;31(24):40881–906.
[19] Menze BH, et al. The multimodal brain tumor image segmentation benchmark (brats). IEEE Transactions on Medical Imaging 2015;34(10):1993–2024.
[20] Bakas S, Akbari H, Sotiras A, Bilello M, Rozycki M, Kirby J, Freymann J, Farahani K, Davatzikos C. Advancing the cancer genomeatlas glioma mri collections with expert segmentation labels and radiomic features. Scientific Data 2017;4. 09.
[21] Bakas S, Reyes M, Jakab A, Bauer S, Rempfler M, Crimi A, Shinohara R, Berger C, Ha S, Rozycki M, Prastawa M, Alberts E, Lipkova J, Freymann J, Kirby J, Bilello M, Fathallah-Shaykh H, Wiest R, Kirschke J, Chen Z. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge 2019;38.
[22] Kordon F, Lasowski R, Swartman B, Franke J, Fischer P, Kunze H. Improved x-ray bone segmentation by normalization and augmentation strategies. In: Bildverarbeitung für die Medizin 2019: Algorithmen–Systeme–Anwendungen. Proceedings des Workshops vom 17. bis 19. März 2019 in Lübeck. Springer; 2019. p. 104–9.
[23] Alwan AH, Ali SA, Hashim AT. Medical image segmentation using enhanced residual u-net architecture. Mathematical Modelling of Engineering Problems 2024;11(2).
[24] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR); 2021. CoRR 2020. abs/2010.11929.
[25] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision; 2021. p. 10012–22.
[26] Pacal I. A novel swin transformer approach utilizing residual multi-layer perceptron for diagnosing brain tumors in mri images. International Journal of Machine Learning and Cybernetics 2024:1–19.
[27] Ghosh S, Santosh K. Tumor segmentation in brain mri: U-nets versus feature pyramid network. In: Proceedings of the 2021 IEEE 34th international symposium on computer-based medical systems (CBMS). IEEE; 2021. p. 31–6.
[28] Rabby SF, Arafat MA, Hasan T. Bt-net: an end-to-end multi-task architecture for brain tumor classification, segmentation, and localization from mri images. Array 2024;22:100346.
[29] Gori M, Tesi A. On the problem of local minima in backpropagation. IEEE Transactions on Pattern Analysis and Machine Intelligence 1992;14(1):76–86.
[30] Yeasmin MN, Al Amin M, Joti TJ, Aung Z, Azim MA. "Advances of ai in image-based computer-aided diagnosis: a review,". Array 2024:100357.
[31] De Sutter S, Wuts J, Geens W, Vanbinst A-M, Duerinck J, Vandemeulebroucke J. Modality redundancy for mri-based glioblastoma segmentation. International journal of computer assisted radiology and surgery 2024;19(10):2101–9.
[32] Xin B, Hu Y, Zheng Y, Liao H. Multi-modality generative adversarial networks with tumor consistency loss for brain mr image synthesis. In: 2020 IEEE 17th international symposium on biomedical imaging (ISBI). IEEE; 2020. p. 1803–7.
[33] Soh WK, Yuen HY, Rajapakse JC. Hut: hybrid unet transformer for brain lesion and tumour segmentation. Heliyon 2023;9(12):e22412.
[34] She D, Zhang Y, Zhang Z, Li H, Yan Z, Sun X. Eoformer: edge- oriented transformer for brain tumor segmentation. In: Proceedings of the international conference on medical image computing and computer-assisted intervention. Springer; 2023. p. 333–43.
[35] Aboussaleh I, Riffi J, el Fazazy K, Mahraz AM, Tairi H. 3duvnetr+: a 3d hybrid semantic architecture using transformers for brain tumor segmentation with multimodal mr images. Results in Engineering 2024;21:101892.
[36] Maji D, Sigedar P, Singh M. Attention res-unet with guided decoder for semantic segmentation of brain tumors. Biomedical Signal Processing and Control 2022;71:103077.
[37] Nguyen-Tat TB, Nguyen T-QT, Nguyen H-N, Ngo VM. Enhancing brain tumor segmentation in mri images: a hybrid approach using unet, attention mechanisms, and transformers. Egyptian Informatics Journal 2024;27:100528.