Journal of Applied Research and Technology

www.jart.icat.unam.mx

Journal of Applied Research and Technology 23 (2025) 341-349

Original

BT-GANformer: A generative ensemble transformer mechanism for brain tumor segmentation and classification

P. Mishra[a] • U. Jain[a] • A. Dash[a] • A. Pandey[b]*

[a]*School of Computer Engineering, KIIT, Deemed to be University, Odisha, India*
[b]*Mechatronics Lab, School of Mechanical Engineering, Kalinga Institute of Industrial Technology (KIIT), Deemed to be University, Bhubaneswar-751024, Odisha, India*

Received 09 03 2024; accepted 11 22 2024
Available 08 31 2025

Abstract: The segmentation task for brain tumors from magnetic resonance imaging (MRI) has been both challenging and crucial for radiologists in their decision-making process. Recent developments in attention mechanisms for natural language processing tasks have gained wide popularity and have shown potential applications in computer vision and related problems. This article proposes a generative ensembled vision transformer that achieves state-of-the-art (SOTA) performance in segmenting brain tumors from multiple modalities of MRI scans. The proposed method includes an encoder and decoder block with CNN and transformer components, forming the generative architecture. The discriminator distinguishes the predictions of the generator from the ground truth and consists of convolutional layers along with a softmax for the classification tasks. The model was trained using the BraTS 2021 Task 1 dataset for the segmentation, and the Task 2 dataset was applied to evaluate the classification task. The proposed model scores a DICE average of 91% in tumor-core (TC), enhancing-tumor (ET), and whole-tumor (WT) categories. The model also achieves a 99% ROC AUC score in the methylguanine-methyltransferase (MGMT) classification task.

Keywords: Brain tumor segmentation, machine learning, generative AI, transformers, deep learning, convolutional neural network, ensemble models, computer vision.

*Corresponding author.
E-mail address: anish06353@gmail.com (A. Pandey).
Peer Review under the responsibility of Universidad Nacional Autónoma de México.

1. Introduction

The tumor inside the brain is considered one of the most disastrous cancers in the world. It is a severe threat to human lives, with its variants spanning almost 100 categories. The disease is caused by the unnatural growth of glial and neural cells inside the skull. Thus, early detection of brain tumors is essential for proper planning of treatment, surgery, and follow-up appointments. Radiologists frequently utilize magnetic resonance imaging (MRI), a standard non-invasive method (Bauer et al., 2013), to identify brain malignancies since it produces noticeably different forms of tissue contrast. However, it might be difficult and time-consuming to segregate brain tumors from MRI scans manually. Therefore, creating automatic and reliable brain tumor segmentation techniques is particularly desirable.

Recent advances in deep learning techniques, especially CNNs (Pereira et al., 2016; Ronneberger et al., 2015) for medical image segmentation and image generation (Chen et al., 2021), have found an essential application in diagnosing tumors and pre-assessing surgical interventions. MRI produces images with variant contrasts, well-known in medical terms as 'modalities.' Four imaging modalities make up an entire MRI scan, including T2-Flair, T1-weighted (T1), T1-ce (enhanced contrast), and T2-weighted (T2). The underlying anatomical information of the brain is captured in specific ways by each of the four modalities. A sample for each type is presented below in Figure 1 from the BraTS 2021 dataset (Baid et al., 2021).

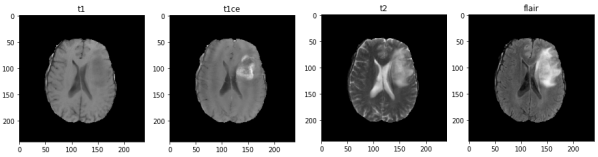

Figure 1. The four modalities of MRI scans - T1, TI-ce, T2, and T2-flair (left to right).

The method developed in this paper is an ensemble transformer-based generative encoder-decoder for image segmentation. The 3D convolutional neural networks (CNNs) and ensemble transformer blocks bridged between them help capture pixel-level information in the images. The attention mechanism (Vaswani et al., 2017) is the main reason behind the success of the transformer. The ensembling process is achieved by a novel bipolar attention mechanism proposed in this paper. Inception v3 (Szegedy et al., 2016) has been used as the backbone framework for the CNN. The decoder contains upsampling deconvolution layers. Skip connections have been added to facilitate the extraction of long- and short-range spatial features of the MRI.

The discriminative framework is inspired by SeGAN (Xue et al., 2018) and features an encoder module to extract features from the ground truth and the predictions from the Generator framework. The L1-Norm distance is calculated and used as one of the penalizing functions for the model. A sigmoid layer is added at the end for the tumor classification task. The flow diagram of the overall architecture of the method is shown below in Figure 2.

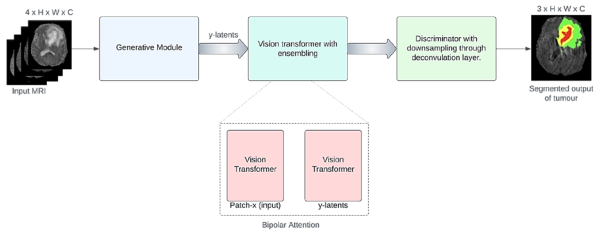

Figure 2. The flow diagram for the proposed method.

The rest of the work is well-presented and divided into the following sections:

(1) Section 2 explores the previous research conducted on the segmentation of tumors in the brain.
(2) Section 3 describes the proposed model in this paper in detail.
(3) Section 4 shows the training details, DICE and ROC AUC scores, and a comparison with previous works in the area.
(4) Section 5 elaborates further on the future work of the BT-transformer.

2. Related work

This section reviews previous research work related to the domain of our proposed architecture. These are mainly vision transformers, brain tumor segmentation, and GAN-related research work.

2.1. Brain tumor segmentation

In the past few years, deep learning has taken over the task of biomedical image segmentation in various modalities. CNNs have mainly dominated the domain of segmentation models. Researchers have also tried to introduce models to capture contextual information (Havaei et al., 2017) in images in 2D scenarios. To understand the 3D context, DeepMedic was introduced by Kamnitsas et al. (2017), which enabled the extraction of 3D patches but was slow during the inference process. This efficiency issue was overcome by fully connected CNN architectures such as U-Net (Weng & Zhu, 2021). However, these models were prone to class imbalance. To overcome the challenge, the MC strategy was adopted for cancer cell segmentation (Hu et al., 2017) and liver lesion segmentation tasks (Zhou et al., 2018). The strategy divides the segmentation tasks into multiple stages to enhance the tumor areas in the MRI. However, this strategy was more

complex and less convenient, ignoring the correlation between the stages. This issue was handled by Zhou et al. (2018) in their one-pass strategy, which takes only one-third of the total MC parameters.

2.2. Vision transformers

First introduced by Vaswani et al. (2017), they are used in NLP tasks to achieve SOTA results. In computer vision, the earliest research work included transformer-based CNN models, namely nnFormer (Zhou et al., 2021), which used transformers for encoding and decoding tasks, and CNN for up and downsampling of the images. Shortly after, the TransBTS (Wang et al., 2021) was introduced, which explores the 3D multi-modal aspect of brain tumor segmentation. It used a 3D CNN and transformers to extract local and global features. The BiTr-UNet (Jia & Shu, 2021) adds two vision transformers (VT) during the skip connections to ensure the modeling of the global features. It gave an excellent performance, but the extra layers of VTs resulted in increased parameters, making it sophisticated. To overcome this issue, a lightweight VT-Unet (Xie et al., 2021) was introduced, which uses just two self-attention layers during the encoding task, thus allowing hierarchical capturing of local and global information from the image.

2.3. Generative AI

The first generative model, GAN (Goodfellow et al., 2020), was introduced to generate images by training with a generator and discriminator. It is a type of machine learning model that has two neural networks – the generator and the discriminator. These networks compete with each other with the generator network producing synthetic data (images) while the former network evaluating it against real data, thus distinguishing between them. The continuous training of these networks results in a much-improved model capable of generating realistic images. The model has brought innovation in the area of training stability and diversified image generation. Afterward, GAN models were used in various applications (Mishra et al., 2022; Prasad et al., 2023) in the medical and robotics domains. However, in the problem of image segmentation, the original GAN is unable to balance the interaction between the generator and the discriminator. The SeGAN (Xue et al., 2018) model is based on a generative AI approach with multi-scale loss, which efficiently solves the issue by minimizing the distance between feature maps of masks and predictions in brain tumor segmentation tasks. A conditional GAN (CGAN) (Mirza & Osindero, 2014) based method was introduced by Ding et al., who implemented an encoder-decoder-based generator and CGAN discriminator for the brain tumor segmentation task. This method took an additional input, which was the image labels.

There are notable differences and improvements in the proposed model apart from the previous efforts:

1. The transformer bridge between the generator and discriminator is ensemble-based to fully utilize the power of vision transformers, which can work significantly on sparse datasets.
2. The generative layer enhances the performance by introducing randomness, which the transformers later train on the attention layer. It ensures that the model is free from overfitting issues.
3. The entire network is formed by a GAN and vision transformers; hence, it combines the advantages of both models and helps map the under-extracted features of the image.

3. Proposed model

The model architecture comprises an encoder-decoder-enabled generator network with an ensemble vision transformer bridge. The transformer follows a novel bipolar attention scheme and applies the attention mechanism to the input patches and the latent features from the generator encoder in a mutual manner. Skip connections have been added to capture the extended- and short-range spatial features of every encoder stage. These features flow through the decoder network after the transformer, which uses transposed convolutions and deconvolutions along with the concatenated features at the intermediate layers. A softmax classifier is added to the last layer to classify the tumor. Each network is explained in detail in the subsections below.

3.1. Generator network

The generator network comprises an encoder and a decoder. The encoder downsamples the images. After the downsampling, the extracted patches and latent features are fed to the ensemble vision transformer with the Inceptionv3 backbone, which implements the bipolar attention mechanism to extract global representations along with the context.

The output of the transformer layer is upsampled in the decoder block, and the softmax layer obtains the segmented image, which is fed to the discriminator network. The step-by-step process of the generator is described in the sections below, and the network diagram can be visualized in Figure 3.

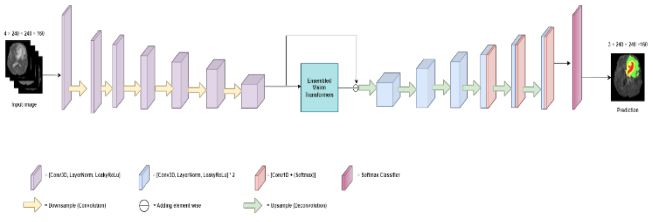

Figure 3. The generator network.

3.1.1. The encoder block

The encoder has seven convolutional layers that take an input of dimensions 240 × 240 × 160 input from the brain tumor dataset. The patches of input have four channels that represent each of the four modalities of the MRI scan. The next layers are downsampling layers, each consisting of Conv3D with dimensions 3 × 3 × 3 and stride 2. The layers are instantly normalized and activated by LeakyReLU (Maas et al., 2013) as it drops the zero-weighted features. Table 1 shows the details of each of the seven layers and the output size after passing through each layer of the downsampling stage in the encoder. The features, such as areas of the brain, edges of the image, and white areas, are extracted with the normalized convolution layers, which reduces the resolution due to the stride of 2.

Table 1. Encoder layer details with the output size at each layer.

Layer	Details of the Layer	Output size
1	[Conv3D, LayerNorm, LeakyReLu, Dropout] [Conv3D, LayerNorm, LeakyReLu, Dropout]	96 × 240 × 240 × 160
2	[Conv3D (*stride =2*), LayerNorm, LeakyReLu, Dropout] [Conv3D, LayerNorm, LeakyReLu, Dropout]	128 × 120 × 120 × 80
3	[Conv3D (*stride =2*), LayerNorm, LeakyReLu, Dropout] [Conv3D, LayerNorm, LeakyReLu, Dropout]	192 × 60 × 60 × 40
4	[Conv3D (*stride =2*), LayerNorm, LeakyReLu, Dropout] [Conv3D, LayerNorm, LeakyReLu, Dropout]	256 × 30 × 30 × 20
5	[Conv3D (*stride =2*), LayerNorm, LeakyReLu, Dropout] [Conv3D, LayerNorm, LeakyReLu, Dropout]	384 × 15 × 15 × 10
6	[Conv3D (*stride =2*), LayerNorm, LeakyReLu, Dropout] [Conv3D, LayerNorm, LeakyReLu, Dropout]	512 × 8 × 8 × 5
7	[Conv3D (*stride =2*), LayerNorm, LeakyReLu, Dropout] [Conv3D, LayerNorm, LeakyReLu, Dropout]	786 × 4 × 4 ×4

3.1.2. The ensembled vision transformer block

The downsampled image patches, as well as the latent output from the encoder, are fed into the transformer block. It works on the principle of the attention mechanism as described by Vaswani et al. (2017) but with mutual attention between the inputs and the latents. The input patches are fed into the attention layer as queries, keys, and values in three matrices. A single block of the transformer can be seen in Figure 4.

The first layer inside the transformer is Layer Normalization, followed by the attention block, which leverages the power of both CNN and second-order mappings. In the proposed model, the attention layer is modified from the original (Xue et al. 2018) research. Here, the queries, keys, and values, are passed through a series of convolutional layers, GeLU, and softmax operations, promoting a second-order mapping for the input patches. The modified attention layer is described in Figure 5.

As mentioned earlier, the Conv3D layer generates the required *Q, K*, and *V* matrices. The Q and K matrices are first linearly transformed and stacked with additional layers of Conv3D, GeLU, and Softmax. The *V* matrix is passed through the output of this operation through a skip connection. The element-wise product is calculated for *Q* and *K* matrices before sending the output to further layers. The overall operation can be summed up in the following equations:

$$y_i = Softmax(Conv(x_i))x_i + x_i \quad (1)$$

Here, x_i denotes the maps of features for the inputs and y_i for the output. The *Conv* here is a sequence of operations involving convolution that generates the matrices *Q, K,* and *V*.

$$(Q, K, V) \in y_i \quad (2)$$

After the creation of the matrices, the product is taken for *Q* and *K,* and the result is passed to further Conv3D, GeLU, and Softmax as represented below:

$$\hat{y_i} = Softmax(Conv(Q \star K))V + x_i \quad (3)$$

where $\star$ represents the elementwise product of the matrices.

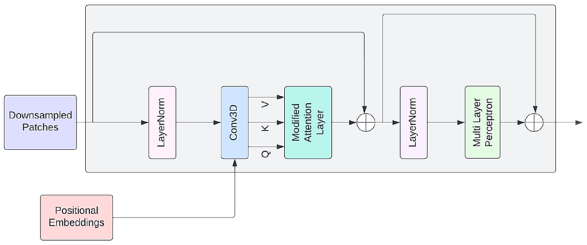

Figure 4. Overview of a single transformer layer.

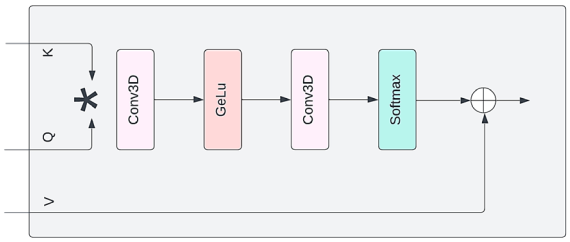

Figure 5. The modified attention block.

The overall output is again layer normalized and passed through a multilayer perceptron layer. Skip connections are added between the bottleneck formed by the previous two layers (LayerNorm and modified attention layer) and after the final layer of the transformer block.

The ensembling of the transformers is achieved by passing the input patches with the latents (positional embeddings). Here, the attention mechanism for a single pass differs from the one above. The attention is calculated as:

$$a = Attention\,(Q,K,V) = Softmax(\frac{QK^T}{\sqrt{d}})V \quad (4)$$

Here, *d* is the dimension of the matrices (which is the same for the three matrices).

The output *Y* of the bipolar attention is a key-value pair that is then calculated using the following update rule:

$$Y = a(Q,K,V) * \omega(X) + a(q,k,v) \quad (5)$$

where *q, k,* and *v* are the matrices obtained from the latents. ω is a normalization function, which can be calculated as

$$\omega = \frac{X - LayerNorm(X)}{LayerNorm(X)} \quad (6)$$

3.1.3. The decoder block

The decoder is made up of upsampling layers along with skip connections. These layers enhance the recovery of the semantic information and the resolution of the patches from the transformer output. The first layer is implemented using the interpolation technique, followed by progressive deconvolution layers. The deconvolution layers have the same stride (2) as the encoder block.

For better supervision of the contextual information, the outputs from the first three layers are processed with a 1 × 1 ×1 convolution and fused with the successive layers. These successive layers include the output of the first three layers, thereby preserving most of the information from the patches.

The last layer is the softmax classifier, which predicts the segmentation maps with varied resolutions. The whole encoder block ensures that the generated segmented images do not deviate significantly from the ground truth. Fusing the upsampled patches with the corresponding previous layers is useful in this step. The discriminator later ensures that the model does not overfit the task by using the loss to penalize the images segmented improperly. The discriminator is explained in detail in the next section. Table 2 contains information about the output sizes at each layer, including the outputs from the concatenated layers.

Table 2. The details of each layer in the decoder block, along with the output sizes of the image patches.

Layer	Details of the Layer	Output size
6	Upsample using Interpolation [Conv3D, LayerNorm, LeakyReLu, Dropout] x 2	512 × 8 × 8 × 5
5	DeConvolution Concatenation [Conv3D, LayerNorm, LeakyReLu, Dropout] [Conv3D, LayerNorm, LeakyReLu, Dropout]	384 × 15 × 15 × 10
4	DeConvolution Concatenation [Conv3D, LayerNorm, LeakyReLu, Dropout] [Conv3D, LayerNorm, LeakyReLu, Dropout]	256 × 30 × 30 × 20
3	DeConvolution [Conv3D, LayerNorm, LeakyReLu, Dropout] [Conv3D, LayerNorm, LeakyReLu, Dropout]	192 × 60 × 60 × 40
Output of Layer 3	[Conv1D + Softmax] DeConvolution Concatenation	4 × 60 × 60 × 40 128 × 120 × 120 × 80
2	[Conv3D, LayerNorm, LeakyReLu, Dropout] [Conv3D, LayerNorm, LeakyReLu, Dropout]	
Output of Layer 2	[Conv1D + Softmax] DeConvolution Concatenation	4 × 120 × 120 × 80 96 × 240 × 240 × 160
1	[Conv3D, LayerNorm, LeakyReLu, Dropout] [Conv3D, LayerNorm, LeakyReLu, Dropout]	
Output of Layer 1	[Conv1D + Softmax]	3 × 240 × 240 × 160

3.2. The discriminator network

The discriminator performs the task of differentiating the ground truth image from that generated by the generator block. It also implements the loss function, which penalizes the model to prevent premature convergence. The overall visualization of the discriminator is presented in Figure 6. The SeGAN (Xue et al. 2018) model largely inspired the loss function, which uses a multiscale L1 loss. The main difference between the latter and the proposed loss lies in using the squared difference between consecutive features. Hence, it is a multiscale L2 loss.

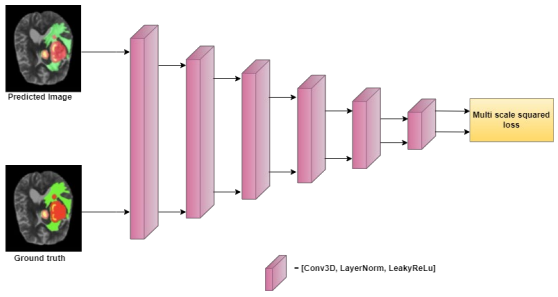

Figure 6. The discriminator network.

Each block of the discriminator contains 3D convolution layers with batch normalization and LeakyReLU activation functions. The stride for the convolution layers is 2. These blocks extract the features from the ground truth and the prediction and compute the squared norm distance between them. If one assumes the j-th feature extracted from the i-th layer from the feature space f, then for x' and x features, the calculation is as follows:

$$l_D(x, x') = \sum_{i=1}^{L} \sum_{j=1}^{M} (f_j^i(x) - f_j^i(x'))^2 \quad (7)$$

Here, L and M denote the number of layers and the number of features, respectively.

Hence, according to the original GAN (Goodfellow et al., 2020) and the above distance calculation, the overall loss can be calculated as shown in Equation (8) below:

$$min_{\theta_G} max_{\theta_D} \mathcal{L}(\theta_G, \theta_D) = l_D(G(x), y) + l_{Dice}(G(x), y) \quad (8)$$

where l_{Dice} denotes the Dice loss calculated on the segmentation maps of the generator *G*. Also, *x* and *y* represent the input and prediction, respectively. The dice loss is the weighted sum of the mask and prediction.

4. Experiments and evaluation

The developed model was trained on the benchmark BraTS dataset (Baid et al., 2021) for brain tumor classification tasks and evaluated using the Dice metric. The details about the implementation and evaluation are provided in the sub-sections below.

4.1. The BraTS 2021 dataset

This multi-modal dataset is provided for the prestigious BraTS challenge on brain tumor classification. The challenge aims to compare the best state-of-the-art (SOTA) models for 3D MRI images. It contains 1,251 patients' training images labeled by physicians and 219 patients' testing images, which are unlabeled. The model developed in this paper used the training dataset divided into 1,000 samples for actual training, 125 for validation, and the rest for testing purposes. The dataset has images for every patient for T1, T1ce, T2, and FLAIR modalities. The size of each image is 240 × 240 × 155. The label consists of 0 or 1, which represent scores in the O6-Methylguanine-DNA Methyltransferase (MGMT) classification. For segmentation, the labels include four categories: background portion (label 0), non-enhancing and necrotic tumors (label 1), peritumoral edema (label 2), and GB-enhanced tumors (label 4). The labels for segmenting the ET region (ET, label 4), the CT region (CT, labels 1 and 4), and the whole tumor region (WT, labels 1, 2, and 4) are used. Figure 7 below visualizes a few samples with MGMT values of 0 and 1.

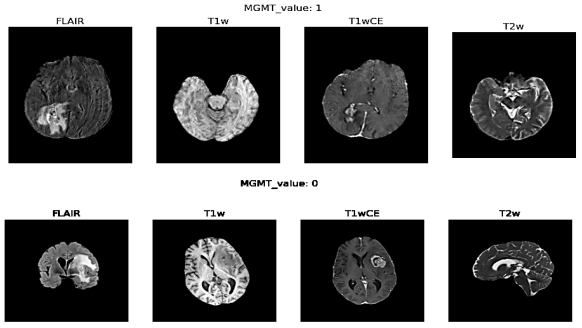

Figure 7. A few samples of both classes from the BraTS dataset (Baid et al. 2021).

4.2. Evaluation metrics

The most common metric used to evaluate brain tumor segmentation is the Dice score (Crum et al., 2006), along with the Positive Predictive Value (PPV) and Sensitivity. For the region predicted (P) and the corresponding ground truth (G), the Dice score is calculated as in Equation 9:

$$Dice(P, G) = \frac{1}{2} \times \frac{|P_t \cap G_t|}{|P_t| + |G_t|} \quad (9)$$

where the subscript *t* denotes the tumorous region in the segmented predictions and ground truths. The | | symbol denotes the voxel count inside the regions and ⋂ denotes the intersection between two regions.

The PPV, however, measures the intersected regions with respect to predictions only, as shown in Equation 10:

$$PPV(P,G) = \frac{|P_t \cap G_t|}{|P_t|} \quad (10)$$

The sensitivity, on the other hand, takes into account the non-tumorous regions inside the segmentation. It can be calculated as follows:

$$Sensitivity(P,G) = \frac{|P_0 \cap G_0|}{|P_0|} \quad (11)$$

where the subscript *0* represents other regions.

The above three metrics range from 0 to 1, where 1 means higher accuracy and better model performance.

For the classification task, the area under the receiver operating characteristic curve (ROC AUC Score) (Greiner et al., 2000) was calculated on the datasets. It plots the true positive rate (TPR) against the false positive rate (FPR) at various thresholds levels. This metric distinguishes the classes in the dataset across these thresholds. It is also measured from 0 to 1, where 1 means higher accuracy and better model performance. Equation 12 shows the calculation:

$$ROC\_AUC = \int_0^1 (TPR(FPR^{-1}(x))dx \quad (12)$$

4.3. Training details

The data structure was divided into training, validation, and test sets from the original labeled training set for BraTS 2021. The experiments were also conducted on the BraTS 2015 dataset. The experiments ran on a system configured with an NVIDIA 3080Ti, 32 GB RAM, and 1TB HDD. A cluster of seven systems was used for training. The images were resized to 240 × 240 × 160 from the original size of 240 × 240 × 155.

The learning rate was initially set at 0.001 with the AdamW optimizer (Loshchilov & Hutter, 2017), which is based on the Adam Optimizer (Kingma & Ba, 2014). The ReduceLROnPlateau (Chollet et al., 2015) Callback was used to reduce the learning rate gradually during the training procedure. The TensorFlow Keras Tuner was used to find the best model hyperparameters. The overall training was performed for 200 epochs.

4.4. Evaluation and scores

The model was trained intensively and tested on BraTS 2021 and 2015 datasets for the DICE score, PPV, and sensitivity values for the segmentation task and the ROC AUC score for the classification task. Tables 3 and 4 summarize the ET, Core Tumor, and WT results. The DICE score variation during these 200 epochs is plotted below in Figure 8.

The discriminator and generator loss values were also monitored using the Tensorboard callback. The losses for the generator and the discriminator showed decreasing and increasing patterns over the number of iterations, respectively, as shown in Figure 9 (a) and (b) below.

Table 3. The evaluation scores for the segmentation task on the 2021 and 2015 datasets.

Dataset	DICE			PPV			Sensitivity		
	ET	Core	WT	ET	Core	WT	ET	Core	WT
BraTS 2021	0.81	0.89	**0.91**	0.78	0.81	**0.89**	0.85	0.87	**0.93**
BraTS 2015	0.83	0.89	**0.93**	0.77	0.79	**0.87**	0.88	0.89	**0.95**

Table 4. The ROC AUC scores for the 2021 and 2015 datasets.

Dataset	ROC AUC Scores
BraTS 2021	97.35%
BraTS 2015	99.89%

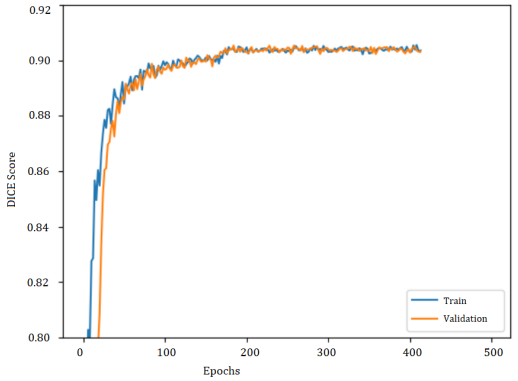

Figure 8. The DICE Score of train and validation set throughout the training process.

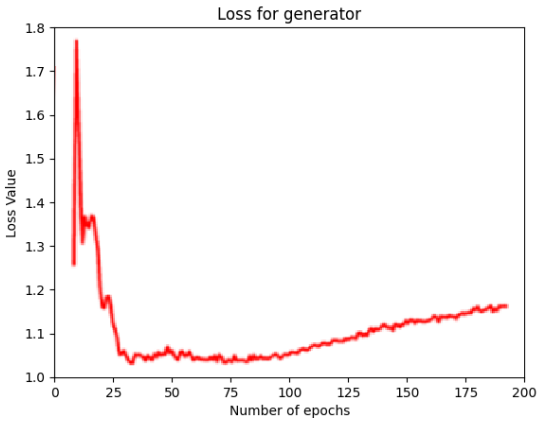

Figure 9 (a). The generator loss curve after training.

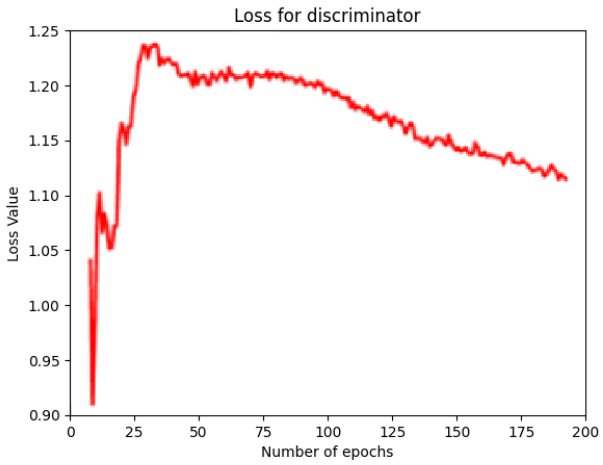

Figure 9 (b). The discriminator loss curve after training.

5. Conclusions

This paper presents and explores the application of generative AI and transformers for segmenting tumors inside the brain. The model uses a generator framework with ensemble vision transformers that act as an encoder to downsample and extract local and global spatial features. The transformer applies modified attention, and the decoder again upsamples the images to get segmented patches. The discriminator differentiates the predictions from the original image and applies the multiscale L2 loss. The model achieves SOTA performance in both segmentation and classification by scoring 91% and 97% in DICE and ROC AUC, respectively.

In the future, there is a scope to research the attention mechanism to improve overall performance. At the same time, the efficient combination of transformers and convolutional neural networks (CNNs) can be explored to reduce the model complexity. For instance, the downsampling of the ground truth and predictions can be done via a single encoder block that ignores positional embeddings when a ground truth image is passed, thus improving the stability of the model.

Conflict of interest

The authors have no conflict of interest to declare.

Funding

The authors received no specific funding for this work.

References

Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., ... & Bakas, S. (2021). The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*. https://doi.org/10.48550/arXiv.2107.0231

Bauer, S., Wiest, R., Nolte, L. P., & Reyes, M. (2013). A survey of MRI-based medical image analysis for brain tumor studies. *Physics in Medicine & Biology*, *58*(13), R97. https://doi.org/10.1088/0031-9155/58/13/R97

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., ... & Zhou, Y. (2021). Transunet: Transformers make strong encoders for medical image segmentation. https://doi.org/10.48550/arXiv.2102.04306

Chollet, F. (2015). keras, GitHub. https://github.com/fchollet/keras

Crum, W. R., Camara, O., & Hill, D. L., (2006). Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE transactions on medical imaging*, *25*(11), 1451–1461. https://doi.org/10.1109/TMI.2006.880587

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., & Bengio, Y., (2020). Generative adversarial networks. *Communications of the ACM*, 63(11), 139–144. https://doi.org/10.1145/3422622

Greiner, M., Pfeiffer, D., & Smith, R. D., (2000). Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Preventive veterinary medicine*, 45(1-2), 23–41. https://doi.org/10.1016/S0167-5877(00)00115-X

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., & Larochelle, H., (2017). Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35, 18–31. https://doi.org/10.1016/j.media.2016.05.004

Hu, H., Guan, Q., Chen, S., Ji, Z., & Lin, Y., (2017). Detection and recognition for life state of cell cancer using two-stage cascade CNNs. *IEEE/ACM transactions on computational biology and bioinformatics*, *17*(3), 887–898. https://doi.org/10.1109/TCBB.2017.2780842

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. https://doi.org/10.48550/arXiv.1711.05101

Mirza, M., & Osindero, S., (2014). Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784. https://doi.org/10.48550/arXiv.1411.1784

Mishra, P., Jain, U., Choudhury, S., Singh, S., Pandey, A., Sharma, A., & Gehlot, A., (2022). Footstep planning of humanoid robot in ROS environment using Generative Adversarial Networks (GANs) deep learning. *Robotics and Autonomous Systems*, 158, 104269. https://doi.org/10.1016/j.robot.2022.104269

Pereira, S., Pinto, A., Alves, V., & Silva, C. A., (2016). Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE transactions on medical imaging*, 35(5), 1240–1251. https://doi.org/10.1109/TMI.2016.2538465

Prasad, A. O., Mishra, P., Jain, U., Pandey, A., Sinha, A., Yadav, A. S., and Dixit, A. K., (2023). Design and development of software stack of an autonomous vehicle using robot operating system. *Robotics and Autonomous Systems*, 161, 104340. https://doi.org/10.1016/j.robot.2022.104340

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Cham: Springer international publishing. https://doi.org/10.1007/978-3-319-24574-4_2

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z., (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818–2826. https://doi.org/10.48550/arXiv.1512.00567

Jia, Q., & Shu, H. (2021). Bitr-unet: a cnn-transformer combined network for mri brain tumor segmentation. In International MICCAI Brainlesion Workshop, Cham: Springer International Publishing, pp. 3–14. https://doi.org/10.1007/978-3-031-09002-8_1

Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., & Glocker, B., (2017). Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical image analysis*, 36, 61–78. https://doi.org/10.1016/j.media.2016.10.004

Kingma, D. P., & Ba, J., (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. https://doi.org/10.48550/arXiv.1412.6980

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I., (2017). Attention is all you need. Advances in neural information processing systems, 30. https://doi.org/10.48550/arXiv.1706.03762

Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., & Li, J., (2021). TransBTS: Multimodal Brain Tumor Segmentation Using Transformer. In de Bruijne, M., et al. Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI Springer, Cham, pp 109–119. https://doi.org/10.1007/978-3-030-87193-2_11

Weng, W., & Zhu, X. (2021). INet: convolutional networks for biomedical image segmentation. *IEEE Access*, 9, 16591–16603. https://doi.org/10.1109/ACCESS.2021.3053408

Xie, Y., Zhang, J., Shen, C., & Xia, Y. (2021). 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, *Springer International Publishing*, pp. 171–180. https://doi.org/10.1007/978-3-030-87199-4_16

Xue, Y., Xu, T., Zhang, H., Long, L. R., & Huang, X. (2018). Segan: Adversarial network with multi-scale l 1 loss for medical image segmentation. *Neuroinformatics*, 16, 383–392. https://doi.org/10.1007/s12021-018-9377-x

Zhou, C., Ding, C., Lu, Z., Wang, X., & Tao, D., (2018). 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, September 16-20, 2018, Proceedings, Part III 11 Springer International Publishing, pp. 637-645. https://doi.org/10.1007/978-3-030-00931-1_73

Zhou, H. Y., Guo, J., Zhang, Y., Yu, L., Wang, L., & Yu, Y., (2021). nnformer: Interleaved transformer for volumetric segmentation. arXiv preprint arXiv:2109.03201. https://doi.org/10.48550/arXiv.2109.03201