
RESIDUAL VISION TRANSFORMER (RESViT) BASED SELF-SUPERVISED LEARNING MODEL FOR BRAIN TUMOR CLASSIFICATION

Meryem Altin Karagoz
Center for Diabetes Technology
University of Virginia
Charlottesville, VA 22904, USA,
ssy4uh@virginia.edu

O. Ufuk Nalbantoglu
Department of Computer Engineering
Erciyes University
Kayseri, Turkey
nalbantoglu@erciyes.edu.tr

Geoffrey C. Fox
Biocomplexity Institute and Initiative
University of Virginia
Charlottesville, VA 22904, USA,
vxj6mb@virginia.edu

November 21, 2024

ABSTRACT

Deep learning has proven very promising for interpreting MRI in brain tumor diagnosis. However, deep learning models suffer from a scarcity of brain MRI datasets for effective training. Self-supervised learning (SSL) models provide data-efficient and remarkable solutions to limited dataset problems. Therefore, this paper introduces a generative SSL model for brain tumor classification in two stages. The first stage is designed to pre-train a Residual Vision Transformer (ResViT) model for MRI synthesis as a pretext task. The second stage includes fine-tuning a ResViT-based classifier model as a downstream task. Accordingly, we aim to leverage local features via CNN and global features via ViT, employing a hybrid CNN-transformer architecture for ResViT in pretext and downstream tasks. Moreover, synthetic MRI images are utilized to balance the training set. The proposed model performs on public BraTs 2023, Figshare, and Kaggle datasets. Furthermore, we compare the proposed model with various deep learning models, including A-UNet, ResNet-9, pix2pix, pGAN for MRI synthesis, and ConvNeXtTiny, ResNet101, DenseNet12, Residual CNN, ViT for classification. According to the results, the proposed model pretraining on the MRI dataset is superior compared to the pretraining on the ImageNet dataset. Overall, the proposed model attains the highest accuracy, achieving 90.56% on the BraTs dataset with T1 sequence, 98.53% on the Figshare, and 98.47% on the Kaggle brain tumor datasets. As a result, the proposed model demonstrates a robust, effective, and successful approach to handling insufficient dataset challenges in MRI analysis by incorporating SSL, fine-tuning, data augmentation, and combining CNN and ViT.

Keywords Self-supervised learning · transformer · convolutional neural network · deep learning · brain tumor classification

1 Introduction

Brain tumors occur due to abnormal and uncontrolled growth of cells in the brain. Brain tumors increase mortality risk and negatively affect life quality due to several medical conditions, such as hearing, vision, and sense disorders. The most prevalent types of brain tumors can be categorized into meningioma, glioma, and pituitary, with incident rates of 15%, 45%, and 15%, respectively, among all brain tumors [1, 2]. The National Brain Tumor Society reports that 94,390 Americans were diagnosed with tumors in 2023, with a survival rate of 35.7% (Brain Tumor Facts). Therefore, early diagnosis is crucial in potentially reducing mortality rate, improving treatment, implementing interventions, and improving the quality of life. For this purpose, various imaging technologies are applied to assist experts in the diagnosis of brain tumors, such as PET (positron emission tomography), MRI (magnetic resonance imaging),

ultrasound screening, X-ray screening, and CT (computed tomography) [3]. MRI is a noninvasive, widely used, and effective imaging technique for early brain tumor diagnosis, enabling pain-free, high-quality 2D and 3D imaging [1]. Furthermore, MRI provides various sequences to capture different aspects of the tissues. The most commonly used sequences in brain MRI are T1-weighted, T1-weighted contrast enhancement, T2-weighted and fluid-attenuated inversion recovery (FLAIR) [4]. MRI assumes a pivotal role in the early and precise diagnosis of brain tumors owing to precise imaging, high soft tissue contrast, and comprehensive information from several sequences [5].

On the other hand, the examination of MRI images for brain tumor diagnosis requires an expert radiologist. The interpretation of MRI images varies subjectively between radiologists according to their experience and education. Furthermore, MRI annotation is a challenging, extensive labor, error-prone, and time-consuming process. Additionally, identifying some brain tumor cases can be challenging because of their location, characteristics, size, and visibility. Integrating artificial intelligence (AI) into computer-assisted diagnosis (CAD) systems in medical imaging has shown remarkable results in addressing these challenges. Consequently, AI-based CAD systems enhance accuracy, speed, efficiency, and consistency of diagnosis and treatment by assisting expert radiologists [1, 5, 6].

Deep learning-based CAD models have emerged as a robust and successful tool in the automated classification of brain tumors. We present the related studies of deep learning-based models for brain tumor diagnosis via CNNs and hybrid models in the Background section. Furthermore, Table 1 presents a summary of previous studies. According to these studies, the integration of deep learning-based CAD models has significantly enhanced the efficiency and accuracy of brain tumor diagnosis, providing automated and valuable support to clinicians across various domains. However, brain tumor diagnosis with deep learning faces a shortage of annotated datasets. Publicly releasing a substantial dataset is challenging due to the need for expert annotations, privacy concerns, extensive labor, time-cost requirements, and accurate labeling [7]. The large amount of data is substantial for Deep learning models to learn a substantial number of parameters effectively and to capture diverse features. Otherwise, an insufficient dataset can lead to an over-fitting problem of deep models, thereby reducing their generalizability. Several strategies have been operated to mitigate overfitting problems such as data augmentation, synthetic data generation, multi-task learning, and transfer learning models. Developing new deep-learning strategies is still an essential subject to eliminate dataset problems, especially for medical imaging tasks.

Self-supervised learning (SSL) models have emerged as a powerful approach to addressing limited labeled data problems with prior studies explained in Section 2. Thus, self-supervised learning models provide a data-efficient approach without relying on explicit labels or large datasets by leveraging unlabeled datasets during pretext tasks [8]. From this standpoint, this study proposes a new generative self-supervised learning model for brain tumor classification on small brain MRI datasets in two stages. The proposed generative SSL model includes the pretraining stage via a Residual Vision Transformer (ResViT) for MRI synthesis as a pretext task, followed by the fine-tuning stage with a ResViT-based classifier model as a downstream task. While the pretext task network of the proposed model enables the extraction of distinct features from the self-distribution of the MRI dataset, the learned features during the pretext task are applied to the downstream classification task through fine-tuning. The proposed SSL model is constructed by combining Residual CNN and transformer blocks to leverage local and global features simultaneously for MRI image synthesis and brain tumor classification. Additionally, synthesized MRI images by ResViT are utilized as data augmentation to balance train datasets in classification. Consequently, the proposed model enables a robust, accurate, and data-efficient approach by combining various strategies to deal with overfitting problems on small brain MRI datasets: self-supervised learning, fine-tuning, data augmentation, and hybrid architecture with CNN and ViT. The proposed model has been evaluated on Kaggle, Figshare brain tumor, and BraTs dataset for each T1, T2, and Flair sequences. This study demonstrates the comparative results of the proposed model against various deep learning models for pretext and classification tasks. Moreover, the pre-trained proposed model on BraTs has been compared with several pre-trained models on the ImageNet dataset. The Residual Vision Transformer (ResViT)-based generative self-supervised learning model provides the following contributions:

- The proposed SSL model enables the extraction of local features via Residual CNN modules and global contextual features via ViT to enhance classification performances. The proposed model exhibits superior performance compared to using Residual CNN and ViT separately. Furthermore, the proposed model surpasses previous models (in Section 3.1.) for each dataset and sequence.
- The proposed SSL model facilitates the learning of data distribution of MRI images in an unsupervised manner (regardless of tumor type) thanks to the pretext task strategy. Implementation of transfer learning from the pre-trained network on the MRI dataset in the pretext task to the ResViT-based classifier model yields superior performance compared to state-of-the-art pre-trained models on ImageNet.
- This study assesses the transferability of the proposed SSL model between different MRI datasets, such as from pre-trained models on BraTs to Kaggle and Figshare Brain MRI datasets. Thus, the proposed model

presents a powerful solution for processing MRI data due to its flexibility, robustness, adaptability to diverse datasets, and strong generalization ability.

- Utilizing both a self-learning strategy and synthesized MRI images enhances tumor classification results.

1.1 Background

1.1.1 CNN-based Deep Learning Models for Brain Tumor Diagnosis

Deep learning-based CAD systems have demonstrated promising results for brain tumor diagnosis, particularly convolutional neural networks (CNN) [9, 10, 11]. Convolutional neural networks [12] are able to capture robust, discriminative, and local features from raw datasets owing to the convolutional mechanism [13]. Therefore, Swati *et al.* [2] propose a block-wise VGG19-based brain tumor classification model via feature extraction framework and transfer learning mechanisms. Deepak *et al.* [14] present a transfer learning-based brain tumor classification model by pre-trained GoogLeNet for feature extraction of brain MRI images, then fed into a classifier model. Ghassemi *et al.* [15] introduce DCGAN (deep convolutional generative adversarial network) to extract robust features of MRI during the pretraining stage. The main idea of this study [15] is to leverage hierarchical representations of the brain MRI images via DCGAN. The last layer of the DCGAN replaces the fully connected layer in a classification task, followed by transfer learning and fine-tuning. Badža *et al.* [16] utilize a pre-trained CNN model and data augmentation to enhance brain tumor classification performance. Ayadi *et al.* [7] suggest a new CNN-based classification model for brain tumor identification on small MRI datasets. Alshayegi *et al.* [17] implement optimization for hyperparameters tuning of two-path CNN. Kakarla *et al.* [18] present average-pooling CNN as a lightweight model to address computational cost. Kumar *et al.* [19] suggest a ResNet-50-based brain tumor classification model and utilize a global average pooling layer to eliminate overfitting problems on small datasets. Abd *et al.* [20] propose BTC-fCNN for fast and efficient identification with a lightweight classification network of brain tumors. They [20] also utilize transfer learning and fine-tuning to improve classification performance. According to these studies, CNN-based brain tumor classification models are commonly utilized with transfer learning, fine-tuning, lightweight network, and average pooling strategies to avoid overfitting on a small dataset.

1.1.2 Hybrid Deep Learning Models via CNNs and Transformer Models for Brain Tumor Diagnosis

Although CNNs are able to capture robust discriminative local features, CNNs have difficulties modeling long-range dependencies for medical image classification tasks [21]. The meaning of long-range in a transformer is to capture dependencies between distant tokens through self-attention mechanisms. On the other hand, CNNs are more limited in this regard, because they focus on the local area of the input operated by the filter size. Although the usage of larger filter sizes in CNNs can allow the extraction of wider features than smaller filter sizes, they still have intrinsic limitations in capturing long-range features across transformers. Medical images contain contextual relationships across both healthy and pathological tissues. Therefore, extracting robust long-range pixel features is essential for identifying medical images. For this reason, transformer models have presented an attractive and effective solution by capturing long-range feature representations owing to the self-attention mechanism. However, transformer-based models require large datasets, time, and resources to perform well [21, 22, 23, 24]. Therefore, recent studies focus on developing a hybrid model by combining CNN and transformer networks to capture both local and global features of MRI images simultaneously and reduce overfitting on small MRI datasets by enhancing the performance of brain tumor diagnosis. From this perspective, Dai *et al.* [25] propose TransMed as a novel multi-modal medical image classification approach. TransMed leverages both CNN and transformer advantages to extract low-level image features efficiently and establish long-range dependencies between modalities. TransMed indicates remarkable improvements by surpassing other state-of-the-art CNN models. Aloraini *et al.* [26] introduce TECNN by combining a transformer and CNN for brain MRI classification on BraTS 2018 and Figshare public dataset. TECNN obtains higher performance than both CNN-based models and ViT models. Tabatabaei *et al.* [27] introduce a cross-fusion model with a parallel two-branch network that has been constructed by a transformer with a self-attention unit and lightweight CNNs for brain tumor classification. The cross-fusion model exhibits a high accuracy of 99.30% due to combining lightweight CNNs and transformers. Ferdous *et al.* [28] propose LCDEiT by utilizing a teacher-student strategy. While they use CNN as the teacher model to extract local features by reducing computational costs and dependencies of large datasets, the student model is designed by a transformer with multi-head attention mechanisms. All noted studies and their results are summarized in Table 1. All these studies indicate that utilizing CNN and transformer together has made significant progress and promising results in coping with data scarcity problems in brain tumor diagnosis. Although these methods yield highly accurate results, improving deep models is still required to deal with the lack of medical dataset problems.

Table 1: The summary of previous studies regarding their publication dates, dataset utilization, model types, partitioning configurations, and accuracy results.

References	Year	Dataset	Model	train:val:test split	Accuracy
Swati <i>et al.</i> [2]	2019	Figshare [50]	Block-Wise VGG19	5-fold cross-val.	94.82
Deepak <i>et al.</i> [14]	2019	Figshare [50]	pre-trained GoogLeNet	5-fold cross-val.	97.1
Ghassemi <i>et al.</i> [15]	2020	Figshare [50]	DCGAN+ConvNet	5-fold cross-val.	95.6
Badža <i>et al.</i> [16]	2020	Figshare [50]	CNN	10-fold cross-val.	96.56
Ayadi <i>et al.</i> [7]	2021	Figshare [50] Radiopaedia REMBRANDT [51]	CNN	5-fold cross-val.	Fig.: 94.74 Rad.: 93.71 REM.: 95.72
Alshayji <i>et al.</i> [17]	2021	Figshare [50]	Aggregation of two paths CNN	70:0:30	97.37
Kakarla <i>et al.</i> [18]	2021	Figshare [50]	average pooling+CNN	5-fold cross-val.	97.42
Kumar <i>et al.</i> [19]	2021	Figshare [50]	ResNet-50+global average pooling	5-fold cross-val.	97.48
Dai <i>et al.</i> [25]	2021	Parotid Gland Tumors	TransMed: hybrid model via CNN and transformer	80:20	88.9
Abd <i>et al.</i> [20]	2023	Figshare [50]	BTC-fCNN	5-fold cross-val.	98.86
Aloraini <i>et al.</i> [26]	2023	BraTS 2018 Figshare [50]	TECNN; hybrid model with transformer-enhanced CNN	70:20:10	Fig.:99.10 BraTS :96.75
Tabatabaei <i>et al.</i> [27]	2023	Kaggle [49] and Figshare [50]	cross-fusion model combining CNNs and transformers	60:20:20	99.06
Ferdous <i>et al.</i> [28]	2023	Figshare[50] and BraTS-21	LCDEIT	10-fold cross-val.	Fig.:98.11 Brats:93.69

2 SSL Models in Medical Image Analysis

Haghighi *et al.*[29] introduce DiRA as a self-supervised learning model on unlabeled medical image datasets. DiRA consists of a discrimination component for acquiring advanced discriminative representations, a restoration component for preserving detailed information, and an adversary component for enhancing feature learning of the restoration. DiRA is evaluated on various 2D and 3D medical image datasets for classification and segmentation downstream tasks. DiRA demonstrates improvement in both classification and segmentation performance against supervised models. Furthermore, DiRA provides an efficient solution and robust model for the limited annotated medical image dataset due to collaborative self-supervised learning.

Taleb *et al.* [30] propose self-supervised models through five techniques for 3D medical image segmentation and detection of downstream tasks on different medical image datasets. The pretext tasks are generated by jigsaw puzzles, contrastive predictive coding, relative patch location, rotation prediction, and exemplar networks. All of the self-supervised techniques help to learn representations of the unlabeled datasets in the first stage. Subsequently, pre-trained models in the pretext task stage are utilized for transfer learning and fine-tuning in the downstream tasks. The self-supervised strategies outperform state-of-the-art models by providing data and cost-efficient models for various medical imaging downstream tasks.

Zhou *et al.*[31] introduce a self-supervised learning model, called Model Genesis, for addressing limited 3D medical image datasets. They apply four efficient image transformation techniques to learn semantic representations of the 3D medical image dataset during the image restoration task. They present a comparative study of Model Genesis against previous self-supervised and supervised models on natural image datasets for classification and segmentation. Model Genesis surpasses supervised transfer learning models on ImageNet by reducing annotation requirements of medical imaging tasks.

Srinidhi *et al.* [32] offer self-supervised and semi-supervised learning models to deal with the lack of annotated dataset problems in histopathology. A self-supervised learning model is designed to capture contextual information in an unsupervised manner as a pretext task. Then, a teacher-student model is generated for consistency training via fine-tuning in classification and regression downstream tasks. Thus, their proposed model exhibits accurate, robust, and data-efficient models on limited histopathology datasets owing to the leveraging of both self and semi-supervised learning techniques.

Wang *et al.* [33] present a novel self-supervised learning model by combining CNN and Swin transformer networks, called the SRCL-CTransPath, to overcome insufficient data problems in histopathological images. Initially, CTransPath is trained on large unlabeled histopathological images to extract local and global features. Then, pre-trained CTransPath performs various downstream tasks such as classification, detection, segmentation, and patch retrieval for nine public histopathology datasets. SRCL-pre-trained CTransPath model surpasses the performance of prior self-supervised models and ImageNet-pre-trained models.

Yan *et al.* [34] suggest a Self-supervised Anatomical embedding model (SAM) for 2D (Xray)and 3D (CT) medical image analysis. SAM is based on a pixel-level contrastive learning and coarse-to-fine strategy to embed global and local anatomical features from unlabeled medical images. SAM outperforms well-known registration algorithms by providing

fast registration. As a result, SAM exhibits a robust model with high generalization ability to apply various medical imaging modalities and medical imaging tasks such as registration, detection, and classification on small datasets.

Kapse *et al.* [35] introduce a Diversity-inducing Representation Learning (DiRL) for slide-level and patch-level histopathology classification. Digital pathology images have complex and intermixed biological components, unlike natural images. Therefore, they design pretext task for cell segmentation through a vision transformer, followed by prior and disentangle blocks, to learn context-rich representations between histopathology views. Moreover, they demonstrate the effect of attention scarification in digital pathology classification tasks. According to qualitative results, DiRL enables high achievement by capturing comprehensive and context-rich representation thanks to attention de-sparsification mechanisms.

Chen *et al.* [36] present UNI as a self-supervised learning model for pathology image analysis downstream tasks. UNI utilizes DinoV2 model in the pretraining stage to extract features from the unlabeled pathology dataset. Before the downstream task, UNI is trained on a huge and diverse pathology dataset containing over 100 million tiles from various major tissue types. UNI has great potential to be used as a foundational model in anatomic pathology due to its successful outcomes and capacity to generalize and transfer across multiple tasks.

Vorontsov *et al.* [37] propose Virchow as the largest model with 632 million parameters and trained 1.5 million WSIs for pathology image analysis tasks. Virchow utilizes vision transformer (ViT)-Huge and DinoV2 models in the pretraining stage on a large pathology dataset. Thus, they leverage the advantages of self-supervised learning and a teacher-student model of semi-supervised learning, as in the study of Chen *et al.* [36]. Virchow is evaluated on various downstream tasks, such as prediction and classification, on different pathology datasets. Virchow gains superior and robust performance against state-of-the-art models for each dataset and downstream task owing to pretraining on large pathology image datasets.

In summary, Self-supervised learning exhibits immense potential to overcome the lack of dataset problems in various medical imaging analyses. While the SSL model emerged as a notable advancement in many medical imaging analysis tasks, there is a requirement for further progress in developing SSL models for brain tumor classification. Consequently, this study focuses on acquiring an SSL model to bridge the existing gaps and achieve more robust and effective solutions for limited dataset problems in brain tumor classification.

3 Details of Proposed Approach

This paper proposes a generative self-supervised learning model (SSL) for brain tumor classification in two stages. The pretext task stage includes the pretraining model before the downstream classification task for brain MRI image synthesis to learn the distribution of MRI data during synthesizing. The second stage is classification through fine-tuning from the pre-trained generative SSL model to the ResViT-based classifier model. Furthermore, synthesized MRI data are utilized as a data augmentation in the second stage to enhance classification performance. The proposed SSL model is given in Figure 1. Finally, the proposed SSL model has been compared with several state-of-the-art image synthesis and classification methods listed in Section 3.3.

3.1 Pre-training Stage: ResViT for Brain MRI Synthesis

The first stage of the proposed model is designed for brain MRI image synthesis as a pretraining stage before the downstream task. Accordingly, the proposed model aims to capture the distribution of MRI images during the synthesized image to enhance the performance of brain tumor classification. Moreover, generated synthetic MRI data is operated as a data augmentation in the classification process. For this purpose, ResViT is utilized for brain MRI image synthesis from one sequence to the other sequence, such as from T1 to T2, T2 to T1, Flair to T1, and T1 to Flair. Residual Vision Transformers (ResViT) have been proposed by Dalmaz *et al.* [38] as a generative adversarial model with hybrid architecture for medical image synthesis. ResViT combines deep residual convolutional networks and transformer blocks to capture local and global contextual features simultaneously. ResViT consists of four main subnetworks, which are the encoder, information bottleneck, decoder blocks, and discriminator. Encoder and decoder blocks are built with 2D residual CNN layers to learn the structural local features of source MRI images. On the other hand, the information bottleneck is located in the center of the network and is generated with Aggregated Residual Transformer (ART) blocks. Thus, ART blocks enable gathering information from residual CNN layers to capture structural local features and transformer branches to capture global contextual features. Finally, the discriminator is constructed by the PatchGAN [39] to ensure the generation of a realistic image by evaluating the realism of local patches in the generated (target) image compared to the corresponding patches in the real (target) image. The architecture of ResViT and ART blocks is given in Figure 1.

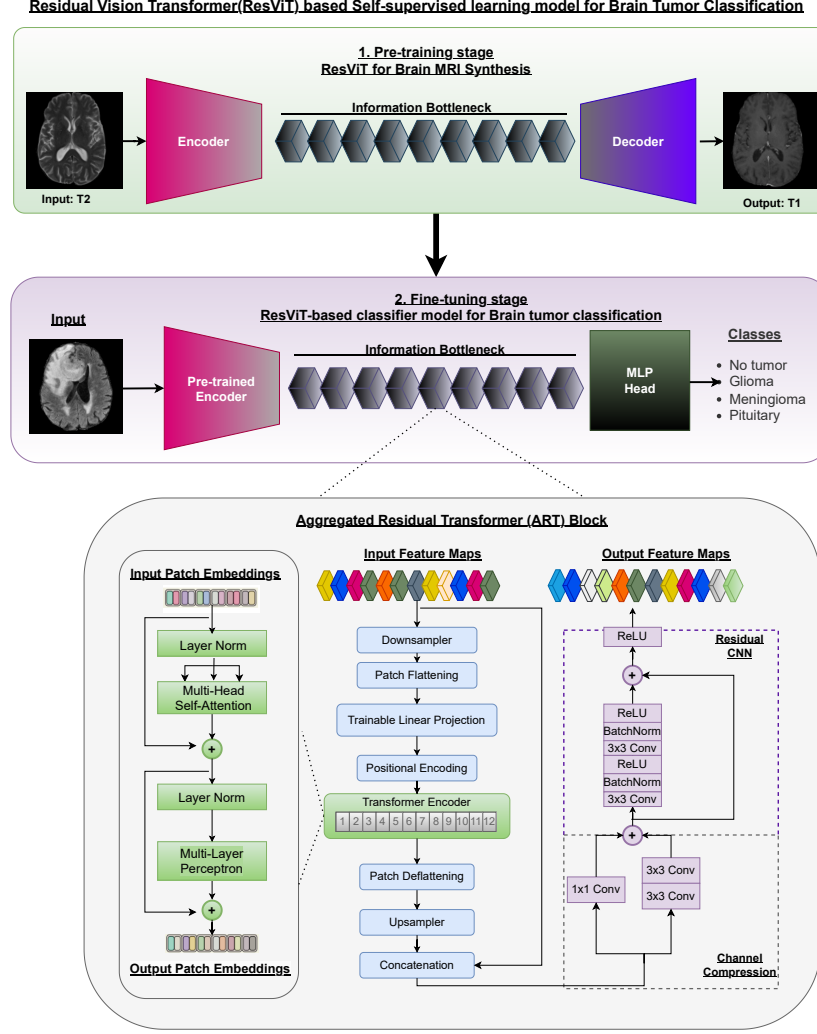


Figure 1: The proposed Residual Vision Transformer (ResViT) based Self-supervised learning model.

The encoder maps multi-channel input X ($256 \times 256 \times 3$) onto the lower-dimensional embedded latent feature map ($f \in \mathbb{R}^{N_c, H, W}$), where f , N_c , H , W are the feature map, number of channels, height, and width of the feature map, respectively. The encoded latent features are directly fed into the ART blocks via a vision transformer, the first layer of the information bottleneck. ART block consists of downsampler, patch flattening, trainable linear projection, positional encoding, the transformer encoder for patch embeddings with multi-head self-attention (MSA) [41], and multi-layer perceptron's (MLP)[42], patch deflattening, upsampler, concatenation layers, channel compression module and a residual CNN (ResCNN), respectively. Convolutional layers are utilized as a downsampler from ($f \in \mathbb{R}^{N_c, H, W}$) to ($f' \in \mathbb{R}^{(nc, h, w)}$), where $nc = N_c/M$, $h = H/M$, $w = W/M$, M is downsampling factor. Then, the downsampled feature map (f') is divided into non-overlapping patches with a patch size of (P, P) and is flattened. Trainable linear projection and positional encoding are applied to the patch embeddings onto an ND-dimensional space as follows:

$$z_0 = [f^1 P_E; f^2 P_E; f^3 P_E; \dots; f^{NP} P_E] + P_E^{pos} \in \mathbb{R}^{NP, ND}. \quad (1)$$

where z_0 , f^p , P_E , and P_E^{pos} are input patch embedding, pth patch of the downsampled feature map, embedding projection, and learnable positional encoding, respectively. Following that, the transformer encoder handles patch embeddings with MSA and MLP. The output of l^{th} layer belonging to the transformer encoder is given as follows, where LN is layer normalization:

$$z'_l = MSA(LN(z_{l-1})) + z_{l-1}. \quad (2)$$

$$z_l = MLP(LN(z'_l)) + z'_l. \quad (3)$$

The deflating layer is applied for processing the upsampling layer with transposed 2D convolutions. Then, channel-wise concatenation combines the global contextual features captured by the transformer (upsampled feature map is g) with the local features captured by convolutional (input feature map is f). Channel compression is utilized for compressed channels of the concatenate feature maps ($concat(g, f)$). The final feature maps are fed into Residual CNN to obtain the output feature map of the ART block. Finally, the decoder synthesizes high-resolution MRI images from low-dimensional feature maps of the ART block via transposed convolutional layers. ResViT utilizes a linear combination of pixel-wise loss (5), reconstruction loss (6), and adversarial loss (7), to calculate the loss function (8), by the following equations.

$$X^G = a_i \cdot m_i. \quad (4)$$

$$a_i = \begin{cases} 1, & \text{if } m_i \in \text{source sequence} \\ 0, & \text{if } m_i \in \text{target sequence} \end{cases}. \quad (5)$$

$$L_{pix} = \sum_{i=1}^I (1 - a_i) E [\|G(X^G)_i - m_i\|_1]. \quad (6)$$

$$L_{rec} = \sum_{i=1}^I (a_i) E [\|G(X^G)_i - m_i\|_1]. \quad (7)$$

$$L_{adv} = E [D(X^D(\text{acquired}))^2] - E [D(X^D(\text{synthetic}) - 1)^2]. \quad (8)$$

Where E is expectation, G is the generator network of ResViT, D is discriminator, $m_i (i \in \{1, 2, \dots, I\})$ is the i th image, X^D (*synthetic*) is the concatenation of source and synthetic images, X^D (*acquired*) is the concatenation of the source and acquired images. Finally, the loss function of ResViT is calculated by linearly combining loss functions and their weights (λ) in (9).

$$L_{ResViT} = \lambda_{pix} L_{pix} + \lambda_{rec} L_{rec} + \lambda_{adv} L_{adv}. \quad (9)$$

The encoder blocks of ResViT are constructed by stacking three convolutional layers in which the kernel size is 7,3,3 and output feature map size is 256, 64, and 64, respectively. The decoder of ResViT is the inverse of the encoder, which has three 2D transposed layers with kernel sizes 3,3,7 respectively. The information bottleneck consists of nine ART blocks (A_1, A_2, \dots, A_9) where A_1 and A_6 blocks have transformer modules. While the downsampler of transformer ART blocks has two convolutional layers and is split into patches with a patch size of (16,16), the up sampler contains two 2D transposed layers for inverse processing of downsampler layers. The down sampler and upsampler are set with kernel size 3, stride 2, and downsampling factor $M=4$. Next, a channel compression module is applied to reduce the number of channels from 512 to 256. Furthermore, the pre-trained base (R50+ViT-B_16, ViT-B_16) and large (ViT-L_16) transformer models on ImageNet (github.com/google-research) are used in transformer encoder blocks. While the base model has 12 layers with 12 attention heads, dimension of latent feature (ND)=768, and 3073 hidden units, the large model has 24 layers with 16 attention heads, dimension of latent feature (ND)=1024, 4096 hidden units for each MLP layer.

3.2 Fine-tuning stage: ResViT-based Classifier Model for Brain Tumor Classification

The second stage of the proposed model is designed for brain tumor classification as a downstream task. The proposed ResViT-based classifier model enables end-to-end fine-tuning by transferring weight from the pre-trained ResViT to the ResViT-based classifier model and leveraging synthesized MRI images in the training phase of the classifier model. Thus, the fine-tuning stage aims to improve classification performance through self-learning. The proposed ResViT-based classifier model has three main components: the pre-trained encoder, ART blocks, and the MLP head. The proposed classifier model aims to learn local structural features via Residual CNNs and global contextual features via ViTs for brain tumor classification as in the first stage. The encoder and ART blocks of ResViT are constructed with identical setups and architecture as in the first stage, explained in the last paragraph of Section 3.1, to enable end-to-end fine-tuning. Furthermore, the decoder blocks of ResViT in the first stage are replaced with a multi-layer perceptron (MLP) for classification. The MLP head consists of a dense layer, a normalization layer, a dropout layer to reduce overfitting with a dropout rate of 0.5, and the fully connected layer with SoftMax for classifying brain tumors into "no tumor," "glioma," "meningioma," and "pituitary."

3.3 Alternative Deep Learning Models for Comparative Study

This study includes several state-of-the-art models for image synthesis and classification to compare with the proposed model. Attention U-Net (A-UNet) [43], ResNet-9 [44], pix2pix [39], and pGAN [40] models are applied for MRI image synthesis and comparison with ResViT in the pretraining stage. The image synthesis methods are the type of conditional GAN for paired image-to-image translation. The conditional GAN models primarily consist of a generator subnetwork and a discriminator subnetwork. While A-UNet and pix2pix are constructed by Unet-based generator interconnected skip connections, ResNet-9-based conditional GAN and pGAN are generated by ResNet blocks. Furthermore, the discriminator subnetworks of the generative models have been set by the PatchGAN discriminator to assess the realism of synthesized MRI images identical to ResViT for fair comparison. Additionally, the pGAN uses pre-trained VGG16 to extract feature maps and calculate perceptual loss, unlike other models. Consequently, this study assesses the ResViT model against various state-of-the-art image-to-image translation methods.

On the other hand, the proposed ResViT-based classifier model has been compared with ConvNeXtTiny [45], Resnet101 [46], and Densenet121 [47]. ConvNeXt has been introduced as a modernized version of ConvNet by adapting standard ResNet architectures into a hierarchical vision transformer. While ResNet-101 consists of residual convolutional blocks with 101 layers, DenseNet-121 is constructed as a series of dense blocks with 121 layers. The methods are pre-trained on the ImageNet dataset [48]. Thus, this study observes the effects of the proposed pre-trained and fine-tuning strategy compared to other pre-trained models on ImageNet. The proposed model combines the Residual CNN and ViT models to capture both local and global features of brain tumors. Residual CNN and ViT are individually employed for brain tumor classification to demonstrate the effectiveness of the proposed ResViT-based classifier model. Residual CNN and ViT models are constructed identically to the encoder block of ResViT and the first ART block of ResViT, respectively, for fair comparison. Furthermore, the pre-trained transformer models (R50+ViT-B_16, ViT-B_16, ViT-L_16) are implemented to assess the transferability and effectiveness of leveraging pre-trained on both the natural ImageNet dataset and a dataset specific to brain MRI. In summary, we demonstrate a comprehensive comparison to observe the effectiveness of the proposed ResViT-based generative SSL model and its contributions.

4 Experimental Study and Results

4.1 Dataset

The proposed and alternative deep learning methods have been trained and evaluated on BraTS 2023 Glioma and Meningioma challenges, Kaggle brain tumor MRI dataset [49], and Figshare brain tumor datasets [50]. BraTS is a well-known benchmark dataset that has mainly used brain tumor segmentation for 12 years. BraTS 2023 Glioma [52, 53, 54, 55, 56] and BraTS 2023 Meningioma [57] datasets are utilized for synthesizing MRI and brain tumor classification tasks. BraTS 2023 Glioma and Meningioma challenges provide pre-gadolinium T1-weighted (T1), post-gadolinium T1-weighted (T1CE), T2-weighted (T2), T2-weighted fluid-attenuated inversion recovery (FLAIR) and segmented masks. The segmented tumors are labeled enhancing tumor, tumor core, and whole tumor. Axial samples of MRI slices including tumor and non-tumor regions are given in Figure 2. While the first MRI slices contain all tumor labels with maximum coverage, the slices in the 2nd rows do not completely present whole tumor labels because of smaller coverage. Therefore, we select slices of MRI from each glioma and meningioma case with maximum tumor region coverage to learn the most relevant information for the pretraining and classification stages. On the other hand, 3rd and the last row of Figure 2 show the healthy samples of MRI slices. While the presented samples of each slice do not contain any tumor region, the slices in the 3rd row provide more contextual and relevant information about brain tissue. Therefore, we selected a healthy slice closer to the center slice of the brain MRI to capture more relevant information. Consequently, we arranged two sub-datasets for the pretraining stage and classification step. The top five slices with maximum tumor coverage have been selected from 1251 glioma cases and 1000 meningioma cases, along with the top five healthy samples based on their proximity to the center slice during the pretraining stage from a total of 2251 glioma and meningioma cases for the pretraining stage, called a BRaTS dataset (5x). In the classification stage, a singular MRI slice per case has been selected based on its depiction of the most relevant information for each tumor, and the dataset is called a basic BRaTS dataset. Additionally, to ensure a balanced training dataset during the classification stage, data augmentation is employed through MRI image synthesis using ResViT. This dataset is referred to as the augmented BRaTS dataset. Thus, the number of training datasets for glioma and meningioma classes is effectively doubled through the synthesis of images in the classification stage. This augmentation is accomplished by utilizing ResViT, pre-trained on the BRaTS dataset(5x) in the pretext task stage. On the other hand, Kaggle and Figshare brain tumor MRI datasets have also been used to demonstrate the learning process and effectiveness of the proposed model on different brain MRI datasets. The Figshare dataset contains 3064 T1-weighted images for glioma, meningioma, and pituitary cases, and the Kaggle dataset includes 7023 MRI images for no tumor, glioma, meningioma, and pituitary

cases. The datasets have been split into a train set and a test set at a ratio of 80:20. The number of images for each dataset is presented in Table 2.

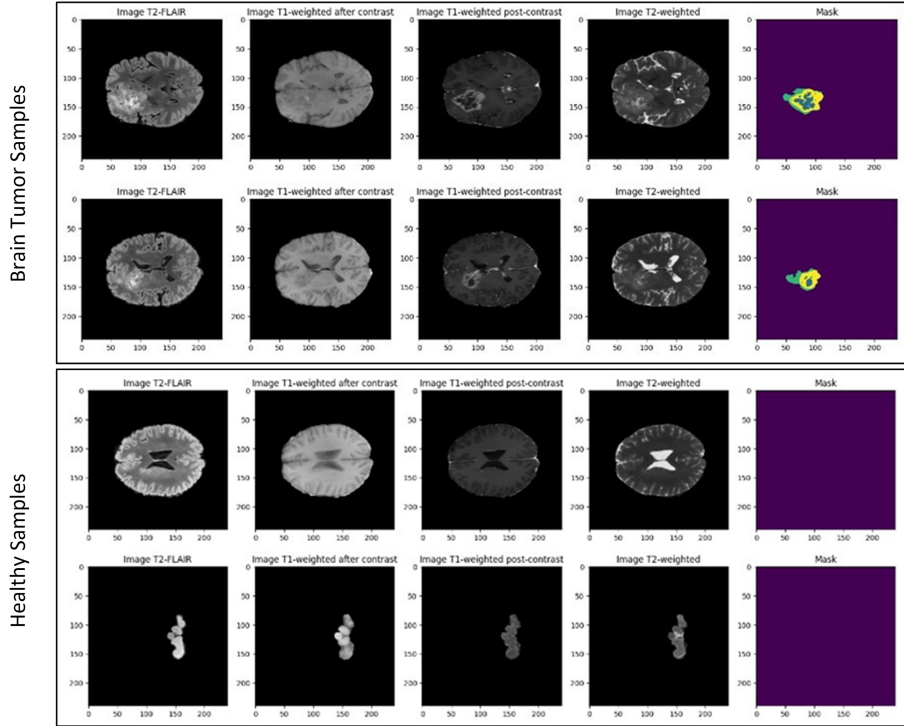


Figure 2: The several axial slices of MRI include labels and information of tumor region with coverage that all belong to the same case. The enhancing tumor (ET), the tumor core (TC), and the whole tumor (WT) are shown yellow, blue, and green, respectively.

Table 2: The number of MRI images for each dataset, built with T1, T2, or Flair, used in the pretraining and classification stages.

classes	subset	Pretraining stage		Classification Stage		
		BRaTS dataset(5x)	Basic BRaTS	Augmented BRaTS	Kaggle	Figshare
No tumor	Train	9005	1801	1801	1595	-
	Test	450	450	450	405	-
Glioma	Train	5005	1001	2002	1321	1,141
	Test	250	250	200	300	285
Meningioma	Train	4000	800	1600	1339	566
	Test	200	200	200	306	142
Pituitary	Train	-	-	-	1457	744
	Test	-	-	-	300	186
Total	Train	18,010	3602	5403	5712	2451
	Test	900	900	900	1311	613

4.2 Experimental Setup

The proposed deep learning model has been built on PyTorch Library. The deep models have been performed on A100-40GB GPU. While The BraTS dataset typically has a volume dimension of 240×240×155 voxels in NiftI format, the Figshare dataset provides images with a resolution of 512×512 pixels in MATLAB (.mat) format. However, the Kaggle dataset provides 2D MRI images in various sizes. To standardize the resolution across datasets, the MRI image resolution is fixed at 256×256 pixels. The proposed and selected alternative deep learning models for MRI image synthesis have been implemented by the following settings: 256×256 image size, 1e-4 learning rate, Adam optimizer, and 100 epochs. The proposed and transfer learning models (mentioned in section 3. C) for classification have been set up with 256×256 image size, 16 batch sizes, Adam optimizer, 2e-5 learning rate, and 100 epochs. The pre-trained ResViT model has been used for end-to-end fine-tuning and data augmentation in the classification stage. The proposed classifier model has been trained and tested separately on each sequence (T1, T2, and Flair). Therefore, pre-trained

ResViT from T1 to T2, T2 to T1, and Flair to T1 models are transferred into the classifier model on T1, T2, and Flair, respectively. In addition, pre-trained ResViT from T1 to T2 model have been utilized for transfer learning of the proposed classifier model on Figshare and Kaggle Brain Tumor Datasets. The synthesis quality between source and synthetic MRI was calculated using peak signal-to-noise ratio (PSNR), structural similarity index [58], and mean square error (MSE), of which the mean and standard deviations (std) were reported. The classification performance models were evaluated by accuracy, precision, recall, and F1 metrics, which were reported with weighted averages.

4.3 Brain MRI Image Synthesis Results

The MRI synthesis results are reported in Table 3. ResViT achieves PSNR of 25.391, SSIM of 0.878, MSE of 0.004 for T2 to T1 synthesis, PSNR of 25.663, SSIM of 0.884, MSE of 0.003 for T1 to T2 synthesis, and PSNR of 25.617, SSIM of 0.873, MSE of 0.004 for Flair to T1, and PSNR of 22.063, SSIM of 0.839, MSE of 0.008 for T1 to Flair. The ResViT demonstrates higher quality for synthesizing MRI images than other generative models (explained in Alternative Deep Learning Models for Comparative Study Section), especially the generator network constructed by Unet (A-Unet and pix2pix). The samples for MRI synthesis are given in Figure 3. The image-to-image translation models constructed by the ResNet, particularly ResViT, effectively captured and represented the tumor region with lower artifact and sharper tissue depiction than Unet-based models.

On the other hand, ResViT and alternative image-to-image translation deep learning models are evaluated on various MRI sequences (T1, T2, and Flair). T1, T2, and Flair provide distinct information about brain tissue characteristics. While T1 highlights anatomical details such as white matter, gray matter, and cerebrospinal fluid, T2 provides differences in water content. Flair is a type of T2 sequence used to improve the visualization of lesions by suppressing cerebrospinal fluid (CSF) signals. The image-to-image translation models perform well across different T1 and T2 synthesis tasks. The quality of synthesis T1 and T2 images is slightly higher by approximately 3dB for PSNR than Flair synthesis.

In summary, this pretraining stage is designed to enhance classification performance through fine-tuning and data augmentation with synthesized MRI images. For this purpose, ResViT is selected to adapt to the classification stage because ResViT outperforms and has a unique fusion of Residual CNN and transformer blocks to capture local and global features simultaneously. Moreover, ResViT synthesizes MRI images with lower artifacts, higher quality, and sharper brain tissue depiction, further supporting the brain classification task.

Table 3: The results of generating synthetic images with generative deep models are evaluated by the mean and standard deviation (std) of signal-to-noise ratio (psnr), structural similarity index, and mean square error (mse) across the test images.

model	T2 ->T1			T1 ->T2			Flair->T1			T1 ->Flair		
	PSNR std	SSIM std	MSE std	PSNR std	SSIM std	MSE std	PSNR std	SSIM std	MSE std	PSNR std	SSIM std	MSE std
A-Unet	24.383	0.856	0.005	23.968	0.852	0.005	23.574	0.832	0.006	20.830	0.803	0.010
	±3.149	±0.042	±0.005	±2.046	±0.057	±0.003	±3.204	±0.052	±0.006	±2.812	±0.062	±0.007
Resnet-9	25.089	0.870	0.004	25.105	0.875	0.004	24.961	0.867	0.004	22.145	0.838	0.008
	±3.194	±0.039	±0.005	±2.150	±0.052	±0.002	±3.158	±0.039	±0.005	±2.848	±0.051	±0.006
pix2pix	24.023	0.821	0.005	24.619	0.860	0.004	24.945	0.854	0.005	21.469	0.831	0.010
	±3.087	±0.045	±0.006	±2.113	±0.052	±0.003	±3.269	±0.041	±0.006	±3.088	±0.071	±0.010
pGAN	24.943	0.870	0.004	24.208	0.868	0.004	24.327	0.863	0.005	20.355	0.835	0.009
	±3.212	±0.039	±0.005	±2.193	±0.053	±0.003	±3.256	±0.041	±0.005	±3.210	±0.053	±0.009
ResViT	25.391	0.878	0.004	25.663	0.884	0.003	25.617	0.873	0.004	22.063	0.839	0.008
	±3.289	±0.039	±0.005	±2.237	±0.052	±0.002	±3.268	±0.040	±0.005	±2.945	±0.051	±0.008

4.4 Classification Results

The proposed ResViT-based classifier and transfer learning models have been performed and evaluated on the basic BraTS, augmented BraTS, Figshare, and Kaggle datasets. Furthermore, the classification models train and test separately for each MRI sequence (T1, T2, and Flair). The results for each MRI sequence on the basic BraTS dataset have been reported in Table 4. ResViT model demonstrates superior performance compared to ConvNeXtTiny, Resnet101, Densenet121, Residual CNN, and ViT models for each Brain MRI sequence. The proposed ResViT-based classifier model achieves an accuracy of 88.89% for T1, 86.22% for Flair, and 83.44% for T2 on the basic Brats dataset. The results clearly indicate that the generative self-supervised learning strategy significantly improved classification performance. In terms of fine-tuning models, transfer learning of the pre-trained ResViT model on the BraTS MRI dataset into the proposed classifier model provides higher performance for each dataset and sequence than other pre-trained models on ImageNet. Although Densenet121 obtains the best scores among the state-of-the-art transfer learning models pre-trained on ImageNet (ConvNeXtTiny, Resnet101, Densenet121), the proposed model has slightly higher accuracy by approximately 9 points for T1, 7 points for Flair, 8 points for T2. Furthermore, the proposed

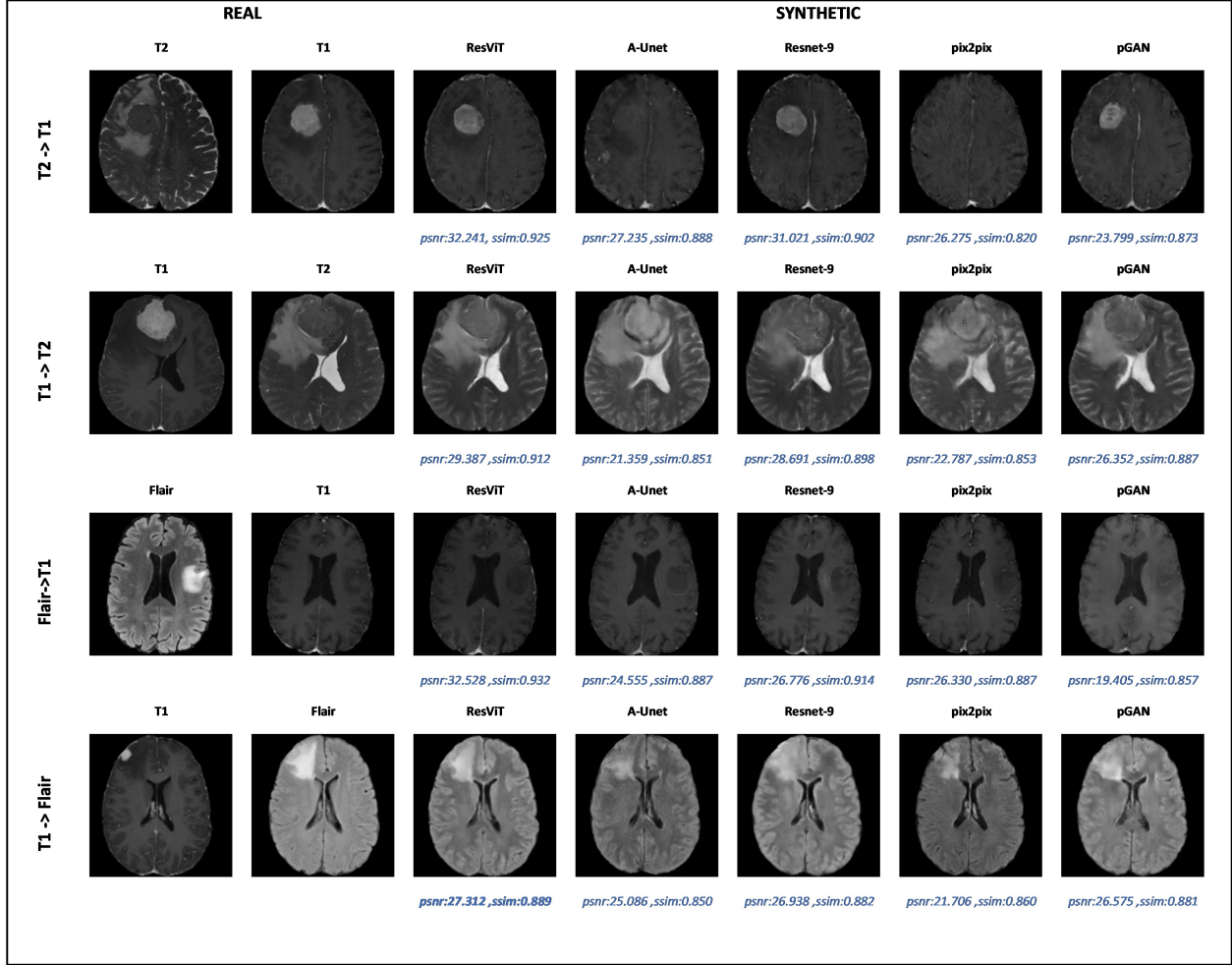


Figure 3: The samples of synthesis MRI from T2 to T1, T2 to T1, Flair to T1, and T1 to Flair.

ResViT-based generative self-supervised learning achieves better success than the pre-trained base (R50+ViT-B_16, ViT-B_16) and large (ViT-L_16) transformer models on ImageNet.

Table 4: The classification results of deep learning models for T1, T2, and Flair sequence on the basic Brats dataset.

Model	pretraining	T1				FLAIR				T2			
		acc	precision	recall	F1	acc	precision	recall	F1	acc	precision	recall	F1
ConvNeXtTiny	ImageNet	59.67	61.28	59.56	54.83	60.67	59.22	60.22	59.22	55.56	57.83	55.56	56.22
Resnet101	ImageNet	64.44	67.83	64.61	63.33	71.67	71.83	71.67	70.39	62.78	62.39	63.11	61.83
Densenet121	ImageNet	79.78	80.50	79.89	79.33	79.00	79.78	78.94	79.33	75.33	75.33	75.33	75.17
Residual CNN	w/o fine-tuning	80.56	80.33	80.89	80.11	77.11	76.56	77.00	75.83	69.22	68.17	69.00	68.22
ViT	w/o finetuning	79.44	79.17	79.50	79.33	79.00	78.33	79.17	78.28	71.67	70.50	71.67	70.83
	R50+ViT-B_16	83.22	83.78	83.06	83.06	78.33	77.61	78.33	77.61	74.44	75.17	74.61	74.39
	ViT-B_16	79.56	79.78	79.33	79.44	80.78	80.28	80.83	80.11	69.56	69.00	69.61	69.17
	ViT-L_16	81.89	82.00	82.17	81.94	79.44	78.78	79.39	78.17	76.56	75.33	76.94	75.56
The proposed ResViT-based classifier	w/o fine-tuning	81.44	80.89	81.28	80.83	77.56	76.83	77.44	76.67	75.67	74.67	75.56	74.39
	R50+ViT-B_16	83.56	83.67	83.83	83.39	78.78	78.50	78.94	78.61	75.67	75.94	77.11	76.17
	ViT-B_16	84.11	84.50	84.00	83.72	78.56	78.83	78.61	78.22	76.67	75.72	76.67	76.33
	ViT-L_16	85.56	85.06	85.89	85.33	78.33	77.33	78.72	77.44	76.78	76.50	76.94	76.22
	ResViT-Proposed	88.89	89.22	88.67	88.83	86.22	85.78	86.22	86.17	83.44	82.89	83.33	83.00

On the other hand, the synthetic MRI images by ResViT have been utilized as data augmentation in the classification stage. Therefore, the augmented BraTS dataset includes synthesized MRI images for glioma and meningioma cases. Table 5 presents the results of classification models on the augmented Brats dataset. The results demonstrate that using synthetic images by ResViT enhances the classification performance of the proposed model. The accuracy of the proposed model increased from 88.89% to 90.56% for T1, from 86.22% to 88.44% for Flair, and from 83.44% to

Table 5: The classification results of deep learning models for T1, T2, and Flair on the augmented Brats dataset.

Model	pretraining	T1				FLAIR				T2			
		acc	precision	recall	F1	acc	precision	recall	F1	acc	precision	recall	F1
ConvNeXtTiny	ImageNet	36.56	53.39	36.72	28.89	44.11	54.44	43.94	42.11	55.56	57.83	55.56	56.22
Resnet101	ImageNet	60.67	66.94	60.56	61.11	71.78	73.72	71.67	72.22	62.89	63.44	63.06	62.33
Densenet121	ImageNet	78.78	78.89	78.94	79.06	74.44	75.83	74.44	74.89	72.67	74.22	72.50	72.94
Residual CNN	w/o fine-tuning	80.22	80.17	80.33	80.39	76.11	74.94	75.83	75.06	76.89	76.89	76.56	76.72
ViT	w/o finetuning	76.56	77.50	76.72	76.61	77.44	77.39	77.50	77.44	77.44	76.89	77.72	77.06
	R50+ViT-B_16	80.00	79.28	80.00	79.50	80.22	79.61	80.33	79.50	78.11	77.78	78.39	77.83
	ViT-B_16	80.11	79.11	80.11	79.39	79.67	79.00	79.44	79.28	77.78	79.72	80.17	80.44
	ViT-L_16	85.56	80.17	80.61	80.00	79.56	79.06	79.44	79.11	79.11	78.94	78.94	78.83
The proposed ResViT-based classifier	w/o fine-tuning	81.67	82.83	81.56	81.83	80.00	79.28	79.83	79.44	80.89	81.44	80.78	81.22
	R50+ViT-B_16	83.11	82.83	83.22	82.67	78.44	77.78	78.39	77.83	79.22	79.28	79.39	78.83
	ViT-B_16	86.44	86.61	86.67	86.39	80.78	80.72	80.83	80.28	79.44	79.39	79.61	79.22
	ViT-L_16	83.44	83.39	83.39	83.28	80.33	80.44	79.67	79.56	79.67	79.17	79.56	78.89
	ResViT-Proposed	90.56	91.06	90.72	90.78	88.44	88.61	88.28	88.33	88.89	89.17	88.94	89.06

88.89% for T2. Furthermore, combining Residual CNN and ViT obtains more effective performance than their separate usage (Residual CNN, ViT). Moreover, the proposed model exhibits the highest achievement on the T1 sequence among various sequence types (T1, Flair, T2) with an accuracy of 88.89% for the basic Brats dataset and 90.56% for the augmented Brats dataset. Consequently, utilizing a self-learning strategy and synthesized MRI together images distinctly enhances the classification performance of the proposed ResViT-based classifier model.

The classification models have been performed on Figshare and Kaggle Brain Tumor Dataset to assess the effectiveness of the proposed model on different brain MRI datasets. Table 6 shows the results of classification models on Figshare and Kaggle Brain Tumor Dataset. The pre-trained ResViT from T1 to T2 model has been selected owing to the highest performance on T1 for BraTs dataset, then transferred into the classifier model on Figshare and Kaggle datasets. The ResViT-based classifier model exhibits remarkable performance, surpassing other pre-trained models on natural datasets, with an accuracy of 98.53% on the Figshare and 98.47% on the Kaggle brain tumor datasets. These results demonstrate that leveraging a pre-trained model on the BraTs dataset rather than the natural dataset is an effective approach for transfer learning on various MRI datasets, indicating the versatility and generalizability of the ResViT-based classifier. The brain tumor classification studies commonly evaluate on Figshare dataset. Table 7 compares the proposed model with previous studies for Figshare dataset. The proposed model achieves a higher accuracy score of 98.53 on Figshare dataset against previous studies that are Swati *et al.* [2], Deepak *et al.*[14], Alshayegi *et al.* [17], Kakarla *et al.* [18], Kumar *et al.*[19] and Ferdous *et al.* [28].

Table 6: The classification results of deep learning models on Figshare and Kaggle Brain Tumor Dataset.

Model	pretraining	Figshare dataset				Kaggle dataset			
		acc	precision	recall	F1	acc	precision	recall	F1
ConvNeXtTiny	ImageNet	81.89	81.52	82.08	81.19	89.93	89.83	89.95	89.93
Resnet101	ImageNet	84.67	85.89	84.46	84.86	87.87	89.63	87.84	87.31
Densenet121	ImageNet	90.54	90.61	90.65	90.13	96.72	96.98	96.78	96.77
Residual CNN	w/o fine-tuning	95.43	95.70	95.52	95.38	94.66	94.53	94.45	94.61
ViT	w/o finetuning	93.31	93.08	93.27	93.29	94.81	94.84	94.92	94.84
	R50+ViT-B_16	95.27	95.47	94.83	95.38	96.80	96.91	96.54	96.99
	ViT-B_16	91.84	92.76	91.80	91.97	94.89	95.07	94.92	94.84
	ViT-L_16	93.47	93.31	93.43	93.22	93.59	93.67	93.55	93.53
The proposed ResViT-based classifier	w/o fine-tuning	94.94	95.16	95.13	95.15	95.12	95.14	95.38	94.99
	R50+ViT-B_16	95.11	94.93	95.13	95.15	94.89	94.84	94.90	94.98
	ViT-B_16	92.99	92.70	92.97	92.99	96.11	96.15	96.30	95.99
	ViT-L_16	94.94	94.63	95.13	95.15	94.20	94.59	93.93	94.15
	ResViT-Proposed	98.53	98.54	98.54	98.54	98.47	98.45	98.61	98.53

Table 7: Comparative study in terms of the accuracy with the previous studies on Figshare dataset.

references	year	method	accuracy
Swati <i>et al.</i> [2]	2019	pre-trained VGG19	94.82
Deepak <i>et al.</i> [14]	2019	transfer learning with GoogLeNet	97.10
Alshayegi <i>et al.</i> [17]	2021	CNN by using optimization for hyperparameters	97.37
Kakarla <i>et al.</i> [18]	2021	CNN and average pooling	97.42
Kumar <i>et al.</i> [19]	2021	ResNet-50 and average pooling	97.48
Ferdous <i>et al.</i> [28]	2023	LCDEiT by combining transformer and CNN	98.11
The proposed model		a Residual Vision Transformer-based generative SSL model by combining transformer and ResNet	98.53

5 Conclusion

This study introduced a self-supervised learning model consisting of a pre-trained ResViT model for MRI synthesis and fine-tuning a ResViT-based classifier model for brain tumor identification. In addition, the synthesized MRI images by ResViT have been included to enhance classification performance. The proposed model has various important components. Firstly, the self-supervised learning strategy allows for the learning distribution of MRI datasets during MRI synthesis in an unsupervised manner without tumor-type labels. Secondly, The ResViT-based model enables the utilization of Residual CNN and ViT together to gather local and global features from MRI datasets. Moreover, fine-tuning and data augmentation via synthetic MRI strategies enable a more effective and data-efficient approach to overcome overfitting on a small number of brain tumor datasets. In summary, the proposed model combines various strategies, self-supervised learning, hybrid architecture with CNN and ViT, fine-tuning, and data augmentation for a more effective and data-efficient approach.

The proposed model compares state-of-the-art models for MRI synthesis and classification, fine-tuning models, sequence types, and datasets. The proposed model performs evidently better than other state-of-the-art models for each dataset and each sequence. Implementing fine-tuning via a pre-trained model on an MRI dataset in the pretext step and synthetic MRI image has increased the accuracy of the proposed model rather than pre-trained models on ImageNet. The results show that T1 is a more convenient sequence to identify tumor types. Furthermore, the pretrained proposed model can easily transferred to other MRI datasets for fine-tuning. As a result, the proposed model emerges as a powerful approach to diagnosing brain tumors by providing high performance, flexibility, robustness, adaptability to diverse datasets, and strong generalization ability.

Acknowledgment

This work was supported by The Scientific and Technological Research Council of Turkey (TUBITAK-BIDEB 2214/A) under project number 1059B142201736. The first author would like to thank The Scientific and Technological Research Council of Turkey (TUBITAK) and Biocomplexity Institute and Initiative, University of Virginia.

References

- [1] S. A. Abdelaziz Ismael, A. Mohammed, and H. Hefny, "An enhanced deep learning approach for brain cancer MRI images classification using residual networks," *Artif Intell Med*, vol. 102, Jan. 2020, doi: 10.1016/j.artmed.2019.101779.
- [2] Z. N. K. Swati *et al.*, "Content-based brain tumor retrieval for MR images using transfer learning," *IEEE Access*, vol. 7, pp. 17809–17822, 2019.
- [3] R. Microwave *et al.*, "A Lightweight Deep Learning Based Microwave Brain Image Network Model for Brain Tumor Classification Using," 2023.
- [4] G. S. Tandel, A. Tiwari, O. G. Kakde, N. Gupta, L. Saba, and J. S. Suri, "Role of Ensemble Deep Learning for Brain Tumor Classification in Multiple Magnetic Resonance Imaging Sequence Data," *Diagnostics*, 2023.
- [5] S. Solanki and U. P. Singh, "Brain Tumor Detection and Classification Using Intelligence Techniques: An Overview," vol. 11, no. January, 2023.
- [6] P. Rani, V. Ashish, and K. Bhandari, "Role of Deep Learning in Classification of Brain MRI Images for Prediction of Disorders: A Survey of Emerging Trends," *Archives of Computational Methods in Engineering*, no. 0123456789, 2023, doi: 10.1007/s11831-023-09967-0.
- [7] W. Ayadi, W. Elhamzi, I. Charfi, and M. Atri, "Deep CNN for Brain Tumor Classification," *Neural Process Lett*, vol. 53, no. 1, pp. 671–700, 2021, doi: 10.1007/s11063-020-10398-2.
- [8] S. C. Huang, A. Pareek, M. Jensen, M. P. Lungren, S. Yeung, and A. S. Chaudhari, "Self-supervised learning for medical image classification: a systematic review and implementation guidelines," *npj Digital Medicine*, vol. 6, no. 1. Nature Research, Dec. 01, 2023. doi: 10.1038/s41746-023-00811-0.
- [9] E. U. Haq, H. Jianjun, K. Li, H. U. Haq, and T. Zhang, "An MRI-based deep learning approach for efficient classification of brain tumors," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–22, 2021.
- [10] M. F. Alanazi *et al.*, "Brain Tumor/Mass Classification Framework Using Magnetic-Resonance-Imaging-Based Isolated and Developed Transfer Deep-Learning Model," *Sensors*, vol. 22, no. 1, 2022, doi: 10.3390/s22010372.

- [11] A. Sekhar, S. Biswas, R. Hazra, A. K. Sunaniya, A. Mukherjee, and L. Yang, "Brain Tumor Classification Using Fine-Tuned GoogLeNet Features and Machine Learning Algorithms: IoMT Enabled CAD System," *IEEE journal of biomedical and health informatics*, vol. 26, no. 3, pp. 983–991, 2022, doi: 10.1109/JBHI.2021.3100758.
- [12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [13] D. R. Sarvamangala and R. V Kulkarni, "Convolutional neural networks in medical image understanding: a survey.," *Evolutionary intelligence*, vol. 15, no. 1, pp. 1–22, 2022, doi: 10.1007/s12065-020-00540-3.
- [14] S. Deepak and P. M. Ameer, "Brain tumor classification using deep CNN features via transfer learning," *Computers in Biology and Medicine*, vol. 111, Aug. 2019, doi: 10.1016/j.compbimed.2019.103345.
- [15] N. Ghassemi, A. Shoeibi, and M. Rouhani, "Deep neural network with generative adversarial networks pre-training for brain tumor classification based on MR images," *Biomedical Signal Processing Control*, vol. 57, p. 101678, 2020, doi: 10.1016/j.bspc.2019.101678.
- [16] M. M. Badža and M. C. Barjaktarović, "Classification of brain tumors from mri images using a convolutional neural network," *Applied Sciences (Switzerland)*, vol. 10, no. 6, Mar. 2020, doi: 10.3390/app10061999.
- [17] M. Alshayegi, J. Al-Buloushi, A. Ashkanani, and S. Abed, "Enhanced brain tumor classification using an optimized multi-layered convolutional neural network architecture," *Multimedia Tools and Applications*, vol. 80, no. 19, pp. 28897–28917, Aug. 2021, doi: 10.1007/s11042-021-10927-8.
- [18] J. Kakarla, B. V. Isunuri, K. S. Doppalapudi, and K. S. R. Bylapudi, "Three-class classification of brain magnetic resonance images using average-pooling convolutional neural network," *International Journal of Imaging Systems and Technology*, vol. 31, no. 3, pp. 1731–1740, Sep. 2021, doi: 10.1002/ima.22554.
- [19] R. L. Kumar, J. Kakarla, B. V. Isunuri, and M. Singh, "Multi-class brain tumor classification using residual network and global average pooling," *Multimedia Tools and Applications*, vol. 80, no. 9, pp. 13429–13438, Apr. 2021, doi: 10.1007/s11042-020-10335-4.
- [20] B.S. Abd El-Wahab, M. E. Nasr, S. Khamis, and A. S. Ashour, "BTC-fCNN: Fast Convolution Neural Network for Multi-class Brain Tumor Classification," *Health Information Science and Systems*, 2023, doi: 10.1007/s13755-022-00203-w.
- [21] O. N. Manzari, H. Ahmadabadi, H. Kashiani, S. B. Shokouhi, and A. Ayatollahi, "MedViT: a robust vision transformer for generalized medical image classification," *Computers in Biology and Medicine*, vol. 157, p. 106791, 2023.
- [22] K. Han *et al.*, "A survey on vision transformer," *IEEE Trans Pattern Anal Mach Intell*, vol. 45, no. 1, pp. 87–110, 2022.
- [23] W. Wang *et al.*, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 568–578.
- [24] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [25] Y. Dai, Y. Gao, and F. Liu, "Transmed: Transformers advance multi-modal medical image classification," *Diagnostics*, vol. 11, no. 8, Aug. 2021, doi: 10.3390/diagnostics11081384.
- [26] M. Aloraini, A. Khan, S. Aladhadh, S. Habib, M. F. Alsharekh, and M. Islam, "Combining the Transformer and Convolution for Effective Brain Tumor Classification Using MRI Images," *Applied Sciences*, 2023.
- [27] S. Tabatabaei, K. Rezaee, and M. Zhu, "Attention transformer mechanism and fusion-based deep learning architecture for MRI brain tumor classification system," *Biomedical Signal Processing Control*, vol. 86, Sep. 2023, doi: 10.1016/j.bspc.2023.105119.
- [28] G. J. Ferdous, K. A. Sathi, M. A. Hossain, M. M. Hoque, and M. Ali Akber Dewan, "LCDEiT: A Linear Complexity Data-Efficient Image Transformer for MRI Brain Tumor Classification," *IEEE Access*, vol. 11, pp. 20337–20350, 2023, doi: 10.1109/ACCESS.2023.3244228.
- [29] F. Haghghi, M. Reza, H. Taher, M. B. Gotway, and J. Liang, "DiRA: Discriminative, Restorative, and Adversarial Learning for Self-supervised Medical Image Analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20824–20834.
- [30] A. Taleb *et al.*, "3D Self-Supervised Methods for Medical Imaging," *Advances in neural information processing systems*, vol. 33, pp. 18158–18172, 2020.
- [31] Z. Zhou, V. Sodha, J. Pang, M. B. Gotway, and J. Liang, "Models Genesis," *Medical Image Analysis*, vol. 67, Jan. 2021, doi: 10.1016/j.media.2020.101840.

- [32] C. L. Srinidhi, S. W. Kim, F. Der Chen, and A. L. Martel, “Self-supervised driven consistency training for annotation efficient histopathology image analysis,” *Medical Image Analysis*, vol. 75, Jan. 2022, doi: 10.1016/j.media.2021.102256.
- [33] X. Wang *et al.*, “Transformer-based unsupervised contrastive learning for histopathological image classification,” *Medical Image Analysis*, vol. 81, Oct. 2022, doi: 10.1016/j.media.2022.102559.
- [34] K. Yan *et al.*, “SAM: Self-Supervised Learning of Pixel-Wise Anatomical Embeddings in Radiological Images,” *IEEE Transactions on Medical Imaging*, vol. 41, no. 10, pp. 2658–2669, Oct. 2022., doi: 10.1109/TMI.2022.3169003.
- [35] S. Kapse *et al.*, “Attention De-sparsification Matters: Inducing diversity in digital pathology representation learning,” *Medical Image Analysis*, vol. 93, pp. 103070, 2024.
- [36] R. J. Chen *et al.*, “A General-Purpose Self-Supervised Model for Computational Pathology,” Aug. 2023, [Online]. Available: <http://arxiv.org/abs/2308.15474>
- [37] E. Vorontsov *et al.*, “Virchow: A Million-Slide Digital Pathology Foundation Model,” Sep. 2023, [Online]. Available: <http://arxiv.org/abs/2309.07778>
- [38] O. Dalmaz, M. Yurt, and T. Cukur, “ResViT: Residual Vision Transformers for Multimodal Medical Image Synthesis,” *IEEE Transactions on Medical Imaging*, vol. 41, no. 10, pp. 2598–2614, 2022, doi: 10.1109/TMI.2022.3167808.
- [39] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.
- [40] S. U. H. Dar, M. Yurt, L. Karacan, A. Erdem, E. Erdem, and T. Çukur, “Image synthesis in multi-contrast MRI with conditional generative adversarial networks,” *IEEE Transactions on Medical Imaging*, vol. 38, no. 10, pp. 2375–2388, Oct. 2019
- [41] A. Vaswani *et al.*, “Attention is all you need,” *Adv Neural Inf Process Syst*, vol. 30, 2017.
- [42] S. Haykin, *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1998.
- [43] O. Oktay *et al.*, “Attention u-net: Learning where to look for the pancreas,” *arXiv preprint arXiv:1804.03999*, 2018.
- [44] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [45] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A ConvNet for the 2020s,” *arXiv preprint arXiv:2201.03545*, 2022.
- [46] K. He, X. Zhang, S. Ren, and J. Sun, “Identity mappings in deep residual networks,” in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, Springer, 2016, pp. 630–645.
- [47] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.
- [48] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*, Jun. 2009.
- [49] M. Nickparvar, “Brain Tumor MRI Dataset,” Kaggle, 2021. [Online]. Available: <https://doi.org/10.34740/KAGGLE/DSV/2645886>.
- [50] J. Cheng, “brain tumor dataset”. figshare, 2017, [Online]. Available: <https://doi.org/10.6084/m9.figshare.1512427.v5> Accessed on: Apr. 2, 2017.
- [51] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, and L. Tarbox, “The cancer imaging archive (TCIA): maintaining and operating a public information repository,” *Journal of digital imaging*, vol. 26, no. 6, pp. 1045–1057, Jul. 2013.
- [52] U. Baid *et al.*, “The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification,” *arXiv preprint arXiv:2107.02314*, 2021.
- [53] B. H. Menze *et al.*, “The multimodal brain tumor image segmentation benchmark (BRATS),” *IEEE Transactions on Medical Imaging* 2014, vol. 34, no. 10, pp. 1993–2024,.
- [54] S. Bakas *et al.*, “Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features,” *Sci Data*, vol. 4, no. 1, pp. 1–13, 2017.

- [55] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, *et al.*, “Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-GBM collection”, *The Cancer Imaging Archive*, 2017. DOI: 10.7937/K9/TCIA.2017.KLXWJJ1Q
- [56] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, *et al.*, “Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-LGG collection”, *The Cancer Imaging Archive*, 2017. DOI: 10.7937/K9/TCIA.2017.GJQ7R0EF
- [57] D. LaBella *et al.*, “The ASNR-MICCAI Brain Tumor Segmentation (BraTS) Challenge 2023: Intracranial Meningioma,” *arXiv preprint arXiv:2305.07642*, 2023.
- [58] Z. Wang, A. C. Bovik, H.R. Sheikh, E. P. Simoncelli EP. “Image quality assessment: from error visibility to structural similarity,” *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004.