# Unleashing the Power of Open-Source Transformers in Medical Imaging: Insights from a Brain

M. A. Rahman[1], A. Joy[2], A. T. Abir[3], T. Shimamura[4]
Department of Electrical and Electronic Engineering, University of Rajshahi, Rajshahi-6205[1,2,3]
Graduate School of Science and Engineering, Saitama University, Saitama, 338-8570, Japan[4]

***Abstract*—This research investigates the application of open-source transformers, specifically the ConvNeXt V2 and Segformer models, for brain tumor classification and segmentation in medical imaging. The ConvNeXt V2 model is adapted for classification tasks, while the Segformer model is tailored for segmentation tasks, both undergoing a fine-tuning process involving model initialization, label encoding, hyperparameter adjustment, and training. The ConvNeXt V2 model demonstrates exceptional performance in accurately classifying various types of brain tumors, achieving a remarkable accuracy of 99.60%. In comparison to other state-of-the-art models such as ConvNeXt V1, Swin, and ViT, ConvNeXt V2 consistently outperforms them, attaining superior accuracy rates across all metrics for each tumor type. Surprisingly, when there is no tumor present, it has predicted with 100% accuracy. In contrast, the Segformer model has excelled in accurately segmenting brain tumors, achieving a Dice score of up to 90% and a Hausdorff distance of 0.87mm. These results underscore the transformative potential of open-source transformers, exemplified by ConvNeXt V2 and Segformer models, in revolutionizing medical imaging practices. This study paves the way for further exploration of transformer applications in medical imaging and optimization of these models for enhanced performance, heralding a promising future for advanced diagnostic tools.**

***Keywords*—Open-source transformers; ConvNeXt V2; segformer; brain tumor classification; medical image segmentation; diagnostic accuracy; neuro-oncology**

## I. Introduction

A brain tumor is recognized as one of the prevalent neurological disorders, characterized by an unregulated and abnormal proliferation of brain cells [1]. It stands as one of the deadliest forms of cancer, posing a significant threat to life [2]. Brain tumors are stratified into four grades (Grade I to Grade IV), with each grade signifying escalating malignancy levels and a progressively ominous prognosis. Grade I tumors, such as pilocytic astrocytoma, characterized by slow growth and a limited tendency to spread, offer the potential for complete removal. Moving to Grade II, these tumors, despite the possibility of migration, can persistently grow and enlarge, even after prior treatment. Advancing to Grade III, tumors exhibit swifter growth and the capability to spread to adjacent tissues, necessitating post-surgical interventions like radiotherapy or chemotherapy. An instance of Grade III malignancy is aden squamous astrocytoma. Finally, Grade IV tumors represent the most lethal category, capable of malignant spreading. Glioblastoma multiforme, an aggressive tumor, serves as an illustrative example of Grade IV characteristics, utilizing blood vessels to accelerate growth [3], [4].

Brain tumors are identified by examining various diagnostic imaging techniques, including X-rays, MRIs, and ultrasound, among others. MRI excels over X-ray and ultrasound with its detailed soft tissue imaging, multi-planar capabilities, and radiation-free nature. In brain assessments, MRI's precision in tumor detection surpasses the limitations of X-ray and ultrasound, making it the preferred choice for accurate diagnostics. However, identifying brain tumors in MR images poses a unique challenge due to the presence of a highly uneven signal associated with the tumor, which can be correlated with the signal strength of normal tissue [5], [6]. The classification of pixels within the tumor region becomes ambiguous, potentially causing inaccurate segmentation. This issue arises when certain tumor components cannot be distinguished from white matter (WM) or gray matter (GM) due to the limited intensity resolution of MR images and the intricate anatomy of the human brain. The complexity intensifies at the tumor's boundary with surrounding normal tissue, influenced by partial volumes (PV) [7]. Consequently, PV contributes to significant blurring in MR images, causing the intensity values of each voxel to mix with those of its neighboring voxels [8].

Machine learning methods address segmentation challenges by employing manually crafted features (or predefined features) [9]. Initially, in the segmentation process, essential information is extracted from the input image using a feature extraction algorithm, followed by training a discriminative model to differentiate between tumors and normal tissues. In the context of tumor segmentation and classification studies, various machine learning techniques, including support vector machines (SVMs), multi-class Support Vector Machine (mSVM), k-nearest neighbor (KNN), Artificial Neural Networks (ANNs), and decision trees, are commonly applied. During the training phase of a classification system, mean features are manually extracted, emphasizing the crucial role of identifying essential features for accuracy [13], [10]. It's noteworthy that constructing classifiers with machine learning demands substantial processing power and memory resources, making it time-consuming, and potentially leading to reduced accuracy, especially with intricate or extensive datasets [13], [11].

Medical images are predominantly examined and processed using deep learning algorithms to identify, classify, and categorize brain tumors into subgroups. These advanced technologies serve as valuable tools for healthcare professionals, assisting them in the diagnostic phase [11]. Deep learning (DL) constitutes a subset of machine learning focused on acquiring multiple tiers of representations through the establishment of a feature hierarchy. This hierarchy is structured such that higher levels derive their definition from lower levels, with the same

lower-level features contributing to the definition of multiple higher-level features [13]. The DL framework expands upon traditional neural networks (NN) by incorporating additional hidden layers within the network architecture, positioned between the input and output layers. This augmentation aims to model more intricate and nonlinear relationships. Recently, researchers have shown considerable interest in this concept due to its commendable performance, establishing it as a preeminent solution across various challenges in medical image analysis applications like image denoising, segmentation, registration, and classification [14]. Deep learning algorithms, including trained convolutional neural networks (CNNs), VGGNets, GoogleNet, and ResNets, are employed for cancer diagnosis assistance. Moreover, the study explored the application of various CNN designs, including VGGNets, GoogleNets, and ResNets, for brain tumor classification [15], [16], [17]. The results indicated that ResNet-50 exhibited superior performance compared to GoogleNet and VGGNets, achieving an accuracy rate of 96.50% in contrast to 93.45% and 89.33%, respectively. Additionally, ResNet-50 demonstrated a 10% higher accuracy than both VGGNet and GoogleNet, while also processing data in 10% less time [18].

The irony of the situation lies in the understanding that even a one percent inaccuracy could potentially lead to the loss of numerous lives. Hence, scholars have dedicated their time and effort to safeguard human lives from the repercussions of unforeseen brain diseases by striving for nearly 100% accuracy in early detection. In pursuit of this crucial goal, they have tirelessly worked to introduce the transformer, a neural network architecture, aiming to enhance the precision and effectiveness of early-stage detection. Transformers have become the prevailing network architecture, bringing about a revolution in language modeling [19], [20]. Operating on an attention mechanism, they clarify the characteristics of the input sequence by entirely bypassing recurrence and convolutions. This unique approach allows the modeling of input dependencies without distance limitations, enabling the assessment of intricate long-range correlations. Notably, transformers exhibit versatility across different types of sequential data, with their applications expanding to fields like computer vision [21]. Recently, transformer-based models, such as Google's Vision Transformer (ViT) and Microsoft's Swin Transformer, have emerged as a powerful alternative to CNNs in various domains, including computer vision [22]. Transformers, originally designed for natural language processing tasks, have shown remarkable adaptability and performance in different modalities and tasks, such as image classification, segmentation, detection, and generation[1]. Transformers are composed of multiple layers of self-attention and feed-forward networks, which can capture long-range dependencies and global context from the input [22]. Transformers can process images by either dividing them into patches and treating them as sequences or by applying convolutional layers to extract features before applying self-attention. Transformers have shown superior performance to CNNs in various tasks, such as image classification, object detection, and semantic segmentation [22]. However, even transformers have their own set of limitations, such as the need for large amounts of data and computational resources, which can be prohibitive in the medical imaging domain [22]. Moreover, transformers might not be able to exploit the spatial structure and locality of images, which can be important for some tasks.

In light of the above, this paper introduces a novel approach that pushes the boundaries of medical imaging. By fine-tuning the Vision Transformer, Swin Transformer, ConvNeXt, and ConvNeXt V2 for brain tumor classification, and Segformer for brain tumor segmentation, we have achieved unprecedented results. Our research has demonstrated that the ConvNeXt V2 model, in particular, has set a new benchmark in medical imaging for classification tasks. With its superior performance, it has proven to be a game-changer in the field of brain tumor detection. ConvNeXt V2, enhances learning of deformable convolutions for superior performance in self-supervised learning and various downstream tasks. It excels in handling diverse image sizes and incorporates advanced training techniques, making it highly effective for medical imaging applications by outperforming state-of-the-art models. On the other hand, the Segformer model has shown state-of-the-art performance in segmentation tasks, achieving a Dice score of over 90 percent. This is a significant leap forward in the precision of brain tumor segmentation. These advancements not only enhance the accuracy and efficiency of brain tumor detection but also contribute to early diagnosis and treatment planning. This, in turn, can lead to improved patient outcomes and alleviate the workload of radiologists, addressing a significant challenge in the healthcare sector.

In conclusion, our research underscores the transformative potential of these models in medical imaging. It provides a benchmark for future research and opens up new avenues for leveraging advanced machine learning techniques in medical imaging. The benefits of this research extend beyond improved patient outcomes in neuro-oncology, offering valuable insights for researchers and practitioners in the field. Future research directions include exploring the application of transformers in other areas of medical imaging and further optimizing the proposed models for better performance. This paper is a testament to the transformative potential of open-source transformers in medical imaging, setting a new standard in the field.

In this study, we delve into the transformative potential of open-source transformers in medical imaging. We provide a comprehensive background on the ConvNeXt V2 and Segformer models, followed by an in-depth explanation of our methodology. We then present our evaluation metrics and the results derived from them. The discussion section explores the implications of our findings, particularly how these models can enhance neuro-oncology diagnostics. We conclude with a summary of our key findings and potential future research directions.

## II. Open-Source Transformers in Medical Imaging

Open-source software has indeed been a game-changer in the field of artificial intelligence, providing researchers and developers with accessible, customizable, and cost-effective tools for innovation. Open-source transformers, in particular, have been instrumental in advancing the field of medical imaging [23].

Before we delve into the specific open-source transformers used in medical imaging, it is essential to understand the

architecture and capability of the original transformer model from the pioneering work "Attention Is All You Need", which introduced the first open-source transformer [24].

### A. The Original Transformer

The Transformer model architecture presented in Fig. 1 by Vaswani in "Atention all you need" is a powerful neural network design that revolutionized natural language processing and other sequence-to-sequence tasks as follows [25]:

*1) Input processing:* The input sequence is first embedded into continuous vector representations. To retain positional information, a positional encoding is added to the embeddings.

*2) Multi-Head attention:* The model employs multi-head attention mechanisms to focus on different parts of the input sequence simultaneously. This allows the Transformer to capture complex patterns and dependencies. Unlike recurrent neural networks (RNNs), where computations depend on the previous step, multi-head attention operates independently across positions.

*3) Feed-forward neural networks:* Each position (word or token) in the input sequence passes through the same feed-forward network. This parallel processing is a departure from RNNs, which have sequential dependencies.

*4) Add and norm:* Every sub-layer (such as multi-head attention or feed-forward neural network) includes a residual connection. After the residual connection, layer normalization is applied. These steps help stabilize training in deeper models.

*5) Masked multi-head attention:* In addition to regular multi-head attention, the Transformer introduces masked multi-head attention. During training, this mechanism prevents attending to future tokens in a sequence. It's crucial for autoregressive tasks like language modeling.

*6) Output probabilities:* The processed outputs from the layers are linearly transformed.A softmax operation generates output probabilities for predictions or downstream tasks.

In summary, the Transformer architecture combines multi-head attention, feed-forward networks, and layer normalization to handle sequential data efficiently. Its parallel processing and attention mechanisms make it highly effective for various natural language understanding tasks.

The introduction of the original transformer model marked a significant milestone in the realm of machine learning and artificial intelligence, ushering in a revolutionary architecture. Central to this innovation is the attention mechanism, a key mathematical concept expressed through equations that distribute attention scores across various segments of an input sequence.

The attention score is calculated using the equation [24]:

$$\text{Attention}(Q, K, V)=\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

Where: - $Q$ represents the query matrix, - $K$ denotes the key matrix, - $V$ stands for the value matrix, - and $d_k$ is the dimensionality of queries and keys.

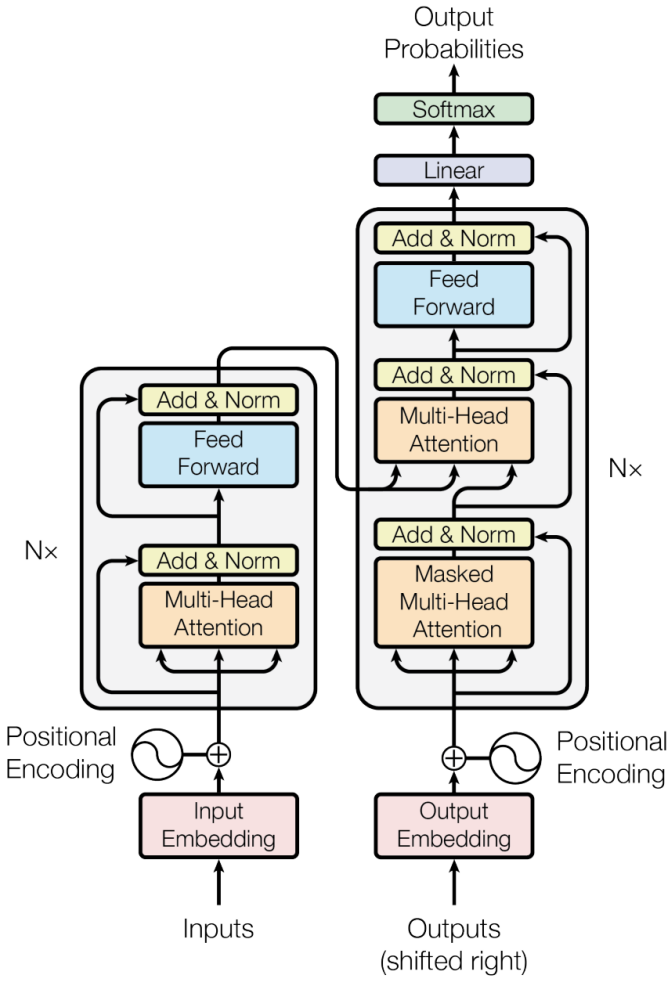


Fig. 1. The transformer-model architecture [24].

This equation underscores how each element in the input sequence contributes to every position in the output sequence by computing a weighted sum of values, with weights assigned according to compatibility function computed using queries and keys [25], [26].

In multi-head attention, this process is replicated across multiple sets of learned linear projections of queries, keys, and values. This can be represented mathematically as:

$$\text{MultiHead}(Q, K, V)=\text{Concat}(\text{head}_1, ..., \text{head}_h)W^O \tag{2}$$

Where each head is computed as:

$$\text{head}_i=\text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{3}$$

Here: - $W_i^Q$, $W_i^K$, and $W_i^V$ are parameter matrices, - $h$ denotes number of heads, - Concat refers to concatenation operation, - And $W^O$ represents output linear transformation weights.

These equations collectively facilitate a nuanced understanding of dependencies among elements or tokens within

sequences, enabling transformers to capture complex patterns and relationships in data with remarkable efficiency.

The success of the transformer model has inspired a diverse array of models extending beyond natural language processing (NLP). These models encompass tasks such as predicting protein folded structures, and forecasting time series data. The model's ability to discern the significance of each input component grants it tremendous power, allowing it to prioritize essential information and disregard irrelevant details, thereby enhancing the accuracy and relevance of its outputs. In the domain of machine learning research, the Transformer model diagram stands as an invaluable tool, providing a comprehensive visual depiction of its architectural complexities and data flow dynamics.

### B. The Power of Transformers in Vision

Transformers have shown significant advancements in the field of vision, particularly with the introduction of the Vision Transformer (ViT). The ViT model, which is the first transformer model introduced for vision tasks after their successful application in natural language processing (NLP), represents an input image as a series of image patches, similar to the series of word embeddings used when applying transformers to text. This model has been specifically designed for image-related tasks, making it a powerful tool in the field of medical imaging.

*1) Vision Transformer (ViT):* The Vision Transformer (ViT) is a model for image classification that employs a Transformer-like architecture over patches of the image as shown in Fig. 2. An image is split into fixed-size patches, each of them is then linearly embedded, position embeddings are added, and the resulting sequence of vectors is fed to a standard Transformer encoder[1]. In order to perform classification, the standard approach of adding an extra learnable "classification token" to the sequence is used [27].

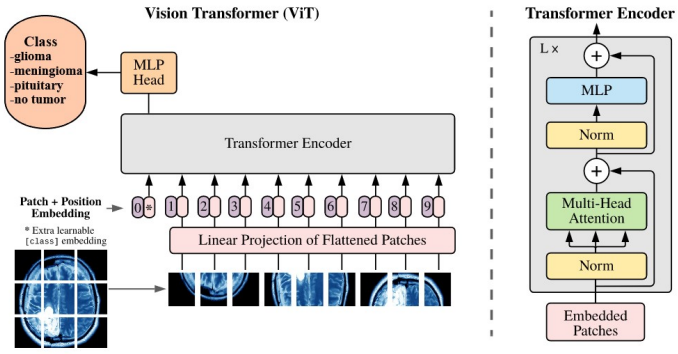


Fig. 2. Vision transformer for brain tumor classification [27].

The ViT model has been shown to outperform the current state-of-the-art convolutional neural networks (CNNs) by almost x4 in terms of computational efficiency and accuracy [27]. This is a significant achievement as CNNs have been the de-facto standard for image recognition tasks for many years.

The success of the ViT model can be attributed to the self-attention mechanism of the Transformer architecture, which allows the model to focus on different parts of the input sequence simultaneously, capturing complex patterns and dependencies. This ability to understand the importance of each part of the input data differently makes the ViT model extremely powerful. It allows the model to focus on what's important and ignore what's not, leading to more accurate and meaningful outputs [27].

The ViT model has been successfully applied to several computer vision problems, achieving state-of-the-art results.This has prompted researchers to reconsider the supremacy of convolutional neural networks (CNNs) as de facto operators.

*2) Swin transformer:* The **Swin Transformer** is a novel vision Transformer that serves as a general-purpose backbone for a variety of computer vision tasks [28]. The name "Swin" stands for **Shifted Windows**, which is a key feature of this architecture.

Unlike the original Vision Transformer (ViT) that produces feature maps of a single low resolution and has a quadratic computation complexity due to global self-attention, the Swin Transformer builds hierarchical feature maps by merging image patches in deeper layers and has linear computation complexity with respect to image size. This is achieved by limiting self-attention computation to non-overlapping local windows while also allowing for cross-window connection [28].

The Swin Transformer is built by replacing the standard multi-head self-attention (MSA) module in a Transformer block with a module based on shifted windows, while keeping other layers the same [28]. A Swin Transformer block consists of a shifted window-based MSA module, followed by a 2-layer MLP with GELU nonlinearity in between.

In terms of performance, the Swin Transformer has demonstrated superior results in various vision tasks, including image classification, object detection, and semantic segmentation[1]. For instance, it achieved 87.3 accuracy on ImageNet-1K, 58.7 box AP and 51.1 mask AP on COCO test-dev, and 53.5 mIoU on ADE20K val. These results surpass the previous state-of-the-art by a large margin, demonstrating the potential of Transformer-based models as vision backbones [29].

In summary, the Swin Transformer offers a promising approach to computer vision tasks, providing a balance between computational efficiency and performance. Its hierarchical design and shifted window approach make it a flexible and powerful tool for image analysis.

*3) ConvNext transformer:* The ConvNext Transformer represents a significant advancement in open-source transformer models tailored for medical imaging, particularly in scenarios involving high-resolution images and necessitating a "sliding window" approach. ConvNets excel in tasks such as object detection, benefiting from translation equivariance and efficiency derived from shared computations within a sliding-window framework [29]. ConvNeXt addresses the need for maintaining ConvNets' inductive learning bias while leveraging Transformer innovations, featuring a specialized block as depicted in Image 3, that integrates convolutional layers to enhance spatial feature extraction from medical images, resulting in improved accuracy and efficiency. Employing an inverted bottleneck design comprising depthwise, expansion,

and contraction layers, ConvNeXt utilizes large depthwise kernels to facilitate scalability and long-range representation learning. By harnessing large kernel ConvNeXt networks in conjunction with extensive datasets, researchers have surpassed previous Transformer-based models, enabling width scaling without constraints imposed by kernel size limitations and offering benefits in learning long-range spatial dependencies through large kernels and enabling multi-level network scaling in medical image segmentation [30]. Fig. 3 shows ConvNexT v1 architecture.

In comparative studies, the ConvNext Transformer has shown promising results on ImageNet-1K and ImageNet-22K pre-trained models. Its performance metrics are competitive with those of the Swin Transformer (2021), indicating its capability to serve as an effective backbone for various computer vision tasks in medical imaging.The success of the ConvNext Transformer can be attributed to the self-attention mechanism of the Transformer architecture, which allows the model to focus on different parts of the input sequence simultaneously, capturing complex patterns and dependencies. This ability to understand the importance of each part of the input data differently makes the ConvNext Transformer extremely powerful. It allows the model to focus on what's important and ignore what's not, leading to more accurate and meaningful outputs.

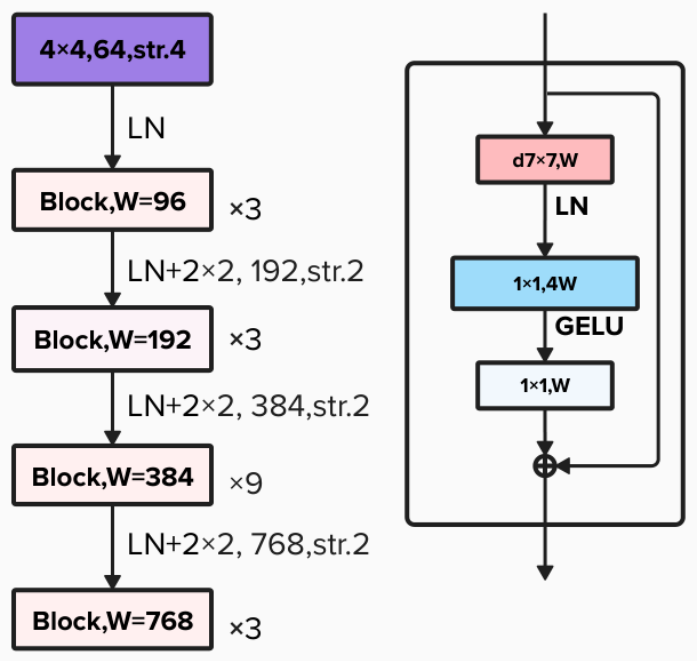


Fig. 3. ConvNexT v1 Architecture [29].

*4) ConvNeXt V2:* Introducing ConvNeXt V2, a novel ConvNet model series, known as Deformable ConvNets v2 (DCNv2), has been developed to enhance its capacity for learning deformable convolutions, as shown in Fig. 4. Despite undergoing minimal architectural modifications, it is precisely tailored for optimal performance in self-supervised learning scenarios. Leveraging fully convolutional masked autoencoder pre-training, significant enhancements in performance are observed across diverse downstream tasks, spanning from ImageNet classification to COCO object detection and ADE20K segmentation [31]. This augmentation in modeling capability encompasses two key aspects. Firstly, there is an expanded integration of deformable convolution layers throughout the network, allowing for greater control over sampling across a wider range of feature levels. Secondly, a modulation mechanism has been introduced within the deformable convolution modules, enabling each sample not only to undergo a learned offset but also to be modulated by a learned feature amplitude, thus providing the network module with the flexibility to adjust both the spatial distribution and the relative influence of its samples [32].

Key Advancements in the ConvNeXt V2 Model:

- Improved ability to handle a wide range of image sizes and formats. This adaptability makes it more versatile and suitable for different medical imaging tasks. This is achieved through adaptive input representations and flexible architecture designs that can accommodate varying input dimensions.
- Incorporation of advanced training techniques and optimization strategies that enhance its learning efficiency and model performance. These include sophisticated learning rate schedules, advanced regularization methods, and efficient batch processing techniques.

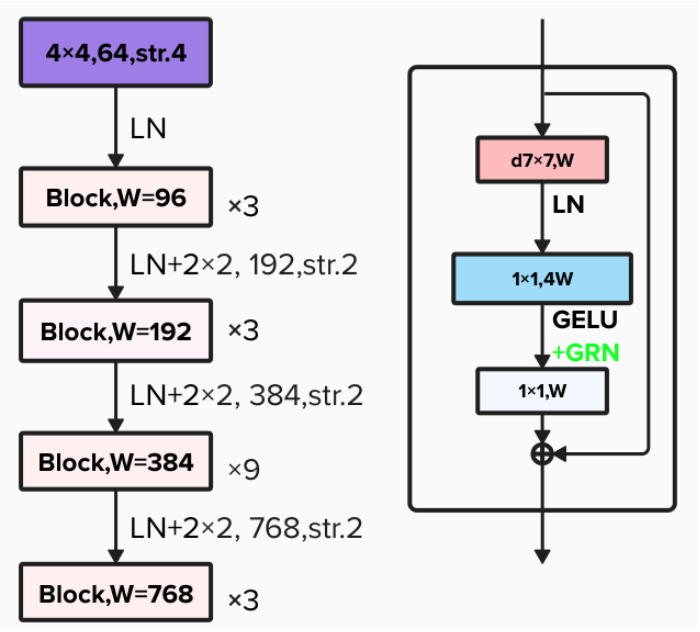


Fig. 4. ConvNexT v2 Architecture [33].

In comparative studies, the ConvNeXt V2 model has shown superior performance metrics on benchmark datasets, outperforming other state-of-the-art models such as the Vision Transformer (ViT) and the Swin Transformer. This indicates its potential as a reliable tool for clinical diagnostics and research, and its capability to serve as an effective backbone for various computer vision tasks in medical imaging.

*5) Segformer: A new frontier in brain tumor segmentation:* The **Segformer** is a groundbreaking open-source transformer model that has been specifically engineered for image segmentation tasks. This model has proven its robustness

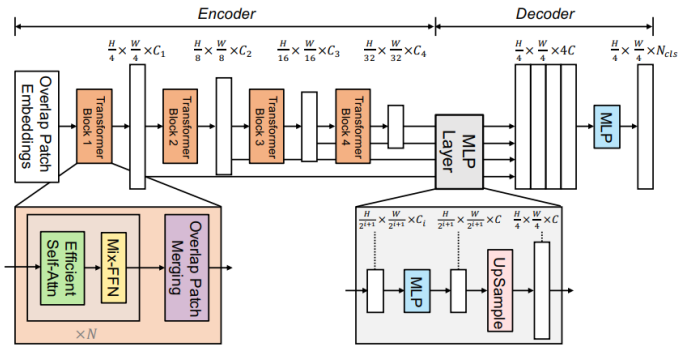


Fig. 5. Architecture of the segformer model[33].

and efficacy across a variety of applications, and this paper presents, for the first time, its potential in the realm of brain tumor segmentation in medical imaging.

The architecture of the Segformer which is presented in Sketch 5, is distinctive, utilizing a hierarchical Encoder-Decoder structure [33]. The attached Fig. 5 illustrates the hierarchical architecture of the Segformer, demonstrating how it processes different resolutions of images for effective segmentation. This structure incorporates convolutional layers to augment the extraction of spatial features from medical images at multiple scales and resolutions. This multi-resolution approach is vital for brain tumor segmentation tasks, where the model is required to accurately identify and delineate intricate anatomical structures and pathological regions at diverse levels of detail.

The hierarchical design of the Segformer enables it to process medical images at various scales, capturing both macroscopic and microscopic features. This is particularly advantageous for brain tumor segmentation tasks, as it ensures a comprehensive capture of the tumor's overall structure and its intricate details, leading to more accurate and meaningful outputs [33]. Another strength of the Segformer is the self-attention mechanism of the Transformer architecture, which allows the model to focus on different parts of the input sequence simultaneously. This ability to discern the importance of each part of the input data differently makes the Segformer extremely powerful [33].

## III. Methodology

This study seeks to harness the potential of open-source transformers, with a specific focus on employing the ConvNeXt V2 model for tumor classification and the Segformer model for segmentation, to advance the field of medical imaging. The methodology employed in this research is outlined below.

### A. Model Selection

We have selected the ConvNeXt V2 model for the task of classification and the Segformer model for segmentation. The decisions are grounded in the established performance of the ConvNeXt V2 and Segformer models in tasks related to images, as well as their proficiency in grasping complex patterns and dependencies within input data. Throughout the selection process, each model covered in the open-source transformers section has undergone fine-tuning. Notably, the ConvNeXt V2 and Segformer models have consistently outperformed their counterparts, demonstrating superior accuracy and efficiency, thus positioning them as the optimal choices for our research objectives.

### B. Data Acquisition

The data used in this research was acquired from two main sources:

*1) Segmentation dataset:* For segmentation, we have utilized the brain tumor dataset provided by Jun Cheng, available on Figshare. This dataset encompasses a comprehensive collection of brain images featuring various types of tumors. Each image in this dataset has been appropriately labeled to facilitate segmentation tasks [34].

*2) Classification dataset:* We have compiled a dataset comprising 15,000 images for the classification task. This dataset has been created by merging the brain tumor dataset provided by Jun Cheng with additional datasets obtained from the internet [34], [35], [36], [37]. These supplementary datasets have been meticulously chosen to guarantee a varied and representative selection of brain images. Each image in this dataset has been labeled with the corresponding tumor type, facilitating the classification task.

The datasets have been carefully scrutinized and validated to ensure their quality and relevance to this research. The images have been accurately labeled, providing a reliable basis for fine-tuning the ConvNeXt V2 and Segformer models.

### C. ConvNeXt V2 Fine-tuning

The ConvNeXt V2 model, depicted in Fig. 6, is fine-tuned for the task of classification. The fine-tuning process involved several steps, each of which contributed to optimizing the model's performance on our specific classification dataset.

*1) Model initialization:* We began by initializing the ConvNeXt V2 model, which is an open-source transformer model. This model was selected due to its proven performance in image-related tasks and its ability to capture complex patterns and dependencies in the input data. The open-source nature of the ConvNeXt V2 model allows for transparency, reproducibility, and customization, which are key advantages in the field of medical imaging [31].

*2) Label encoding:* The labels for each image in our dataset were mapped to corresponding IDs. This encoding process transformed the categorical labels into a format that could be processed by the ConvNeXt V2 model.

*3) Hyperparameter adjustment:* The model's hyperparameters, such as learning rate, batch size, and number of epochs, were adjusted during the fine-tuning process.

*4) Training:* The training process involved feeding the images from the training set into the ConvNeXt V2 model. The model processed these images through multiple stages, each involving a series of operations that transform the input images, extracting essential features and patterns that the model can learn from.

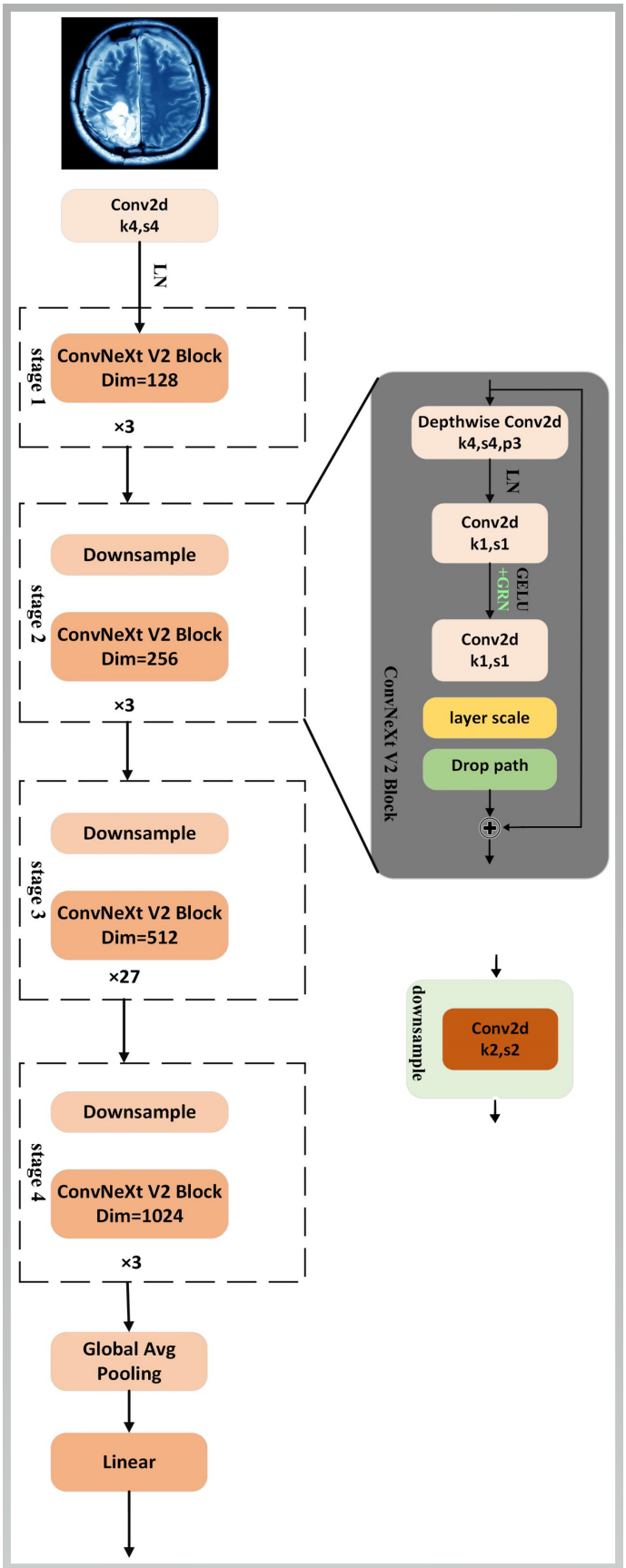


Fig. 6. Implementing ConvNext V2 for advanced brain tumor classification: A visual guide.

The operations include depthwise separable convolutions, layer normalization, GELU activation, and pointwise separable convolutions. These operations are inspired by the mechanisms used in transformers. For instance, the layer normalization and GELU activation functions are commonly used in transformer models.

The depthwise separable convolutions operation is a key feature of the ConvNeXt V2 model. It is a variant of the standard convolutions and is designed to reduce the model's complexity and computational cost. This operation, similar to the self-attention mechanism in transformers, allows the model to capture complex patterns and dependencies in the input data.

Mathematically, the depthwise separable convolution operation can be represented as a two-step process:

*a) Depthwise convolution:* This operation applies a single convolutional filter per input channel. If we denote the input feature map as $F_{in}$, the depthwise convolutional filter as $D$, and the output feature map as $F_{out}$, this operation can be represented as:

$$F_{out}^{(i)} = D^{(i)} * F_{in}^{(i)} \tag{4}$$

where $*$ denotes the convolution operation, and $i$ is the index of the input channel.

*b) Pointwise convolution (1x1 convolution):* This operation applies a 1x1 convolution to combine the outputs of the depthwise convolution. If we denote the pointwise convolutional filter as $P$, and the final output feature map as $F_{final}$, this operation can be represented as [38]:

$$F_{final} = P * F_{out} \tag{5}$$

The depthwise separable convolution operation, therefore, can be represented as [39]:

$$F_{final} = P * (D * F_{in}) \tag{6}$$

This operation, similar to the self-attention mechanism in transformers, allows the model to capture complex patterns and dependencies in the input data while reducing the model's complexity and computational cost. It's a crucial part of the ConvNeXt V2 model's architecture.

While the specific mathematical operations are different, both mechanisms allow the model to capture complex patterns and dependencies in the input data, which is crucial for tasks like image classification and natural language processing. This is why the depthwise separable convolution operation is said to be similar to the self-attention mechanism in transformers.

During training, the model's parameters are adjusted to minimize the loss function. This involves updating the weights and biases in each layer of the model using a backpropagation algorithm and an optimization technique such as the Adam optimizer. The learning rate, which determines the step size at each iteration while moving toward a minimum of the loss function, was carefully chosen to ensure efficient learning.

Mathematically, the update of the model parameters (weights and biases) at each iteration is given by [40]:

$$\theta_{\text{new}} = \theta_{\text{old}} - \text{learning rate} \times \nabla J(\theta_{\text{old}}) \quad (7)$$

where, $\theta$ represents the model parameters, $J$ is the cost function, and $\nabla J(\theta_{\text{old}})$ is the gradient of the cost function evaluated at $\theta_{\text{old}}$.

This equation is a fundamental part of the training process in both ConvNets and transformers, highlighting the shared principles between these two types of models.

Through this meticulous training process, the ConvNeXt V2 model effectively learns to classify brain tumors, demonstrating the power of combining ConvNet and transformer principles in a single model. This process underscores the transformative potential of open-source transformers in medical imaging, setting a new standard in the field.

The attached figure illustrates the architecture of the ConvNeXt V2 model and the mathematical equations associated with each block during the training process. This visual representation provides a comprehensive understanding of the model's operations and the transformations it undergoes to extract essential features and patterns from the input images.

### D. Validation

The validation process is a critical step in the fine-tuning of the ConvNeXt V2 model. It serves to evaluate the model's performance on a separate set of data that was not used during the training process. This helps to ensure that the model is not overfitting to the training data and can generalize well to new, unseen data.

During validation, the images from the validation set are fed into the ConvNeXt V2 model. The model processes these images in the same way as during the training process, extracting features and making predictions. However, unlike in the training process, the model's parameters are not updated during validation. This allows for an unbiased evaluation of the model's performance [31].

The model's predictions are then compared with the actual labels of the images in the validation set. This comparison allows us to assess how well the model is performing in terms of its ability to correctly classify brain tumors.

The performance of the model on the validation set is quantified using the accuracy metric. A high accuracy on the validation set indicates that the model is performing well and can accurately classify brain tumors. Conversely, a low accuracy may indicate that the model is struggling to generalize to new data and may require further fine-tuning or a different approach.

Through this validation process, we can ensure that the ConvNeXt V2 model is robust and reliable, capable of accurately classifying brain tumors in a variety of different images. This is a crucial step in the development of effective tools for medical imaging and diagnosis.

### E. Segformer Fine-tuning

The Segformer model is fine-tuned for the task of segmentation. The fine-tuning process involved several steps as Fig. 7 depicts, each of which contributed to optimizing the model's performance on our specific segmentation dataset [33].

*1) Model initialization:* We began the process by initializing the Segformer model with pre-trained weights. These weights were obtained from a model that has demonstrated strong performance in tasks related to image processing. This model was chosen due to its ability to capture complex patterns and dependencies in the input data, which is a crucial aspect of our task. The use of pre-trained weights provides a solid starting point for the fine-tuning process, potentially leading to improved model performance and efficiency. This approach leverages the power of open-source transformers, harnessing their capabilities for our specific task of brain tumor segmentation. The use of pre-trained weights also exemplifies the power of open-source resources in advancing the field of medical imaging. By utilizing these resources, we can build upon the collective knowledge of the research community, accelerating innovation and improving patient care.

*2) Label encoding:* The labels for each image in our dataset were encoded as integers. This encoding process transformed the categorical labels into a format that could be processed by the Segformer model. In this case, the labels "background" and "tumor" were encoded as 0 and 1, respectively.

*3) Hyperparameter adjustment:* The model's hyperparameters, such as learning rate, batch size, and number of epochs, were adjusted during the fine-tuning process. The learning rate was set to 0.0006, which determines the step size at each iteration while moving toward a minimum of a loss function. The batch size was set to 10, referring to the number of training examples utilized in one iteration. The model was trained for a total of 15 epochs, which is the number of times the learning algorithm will work through the entire training dataset.

*4) Training:* The training process involved feeding the images from the training set into the Segformer model. The model processed these images through multiple stages, each involving a series of operations that transform the input images, extracting essential features and patterns that the model can learn from.

- **Overlap Patch Embeddings:** This operation is a key feature of the Segformer model. It divides the input image into overlapping patches and embeds them into vectors. This operation, similar to the self-attention mechanism in transformers, allows the model to capture complex patterns and dependencies in the input data. Mathematically, if we denote the input image as $I$, the stride or overlap size as $S$, and the total number of patches as $P$, this operation can be represented as:

$$P = \frac{I - S}{S} + 1 \quad (8)$$

- Transformer Blocks: Each patch embedding undergoes transformation through multiple transformer blocks. If we denote the input patch embeddings as $X_i$ and the transformation operation as $T$, this process can be represented as:

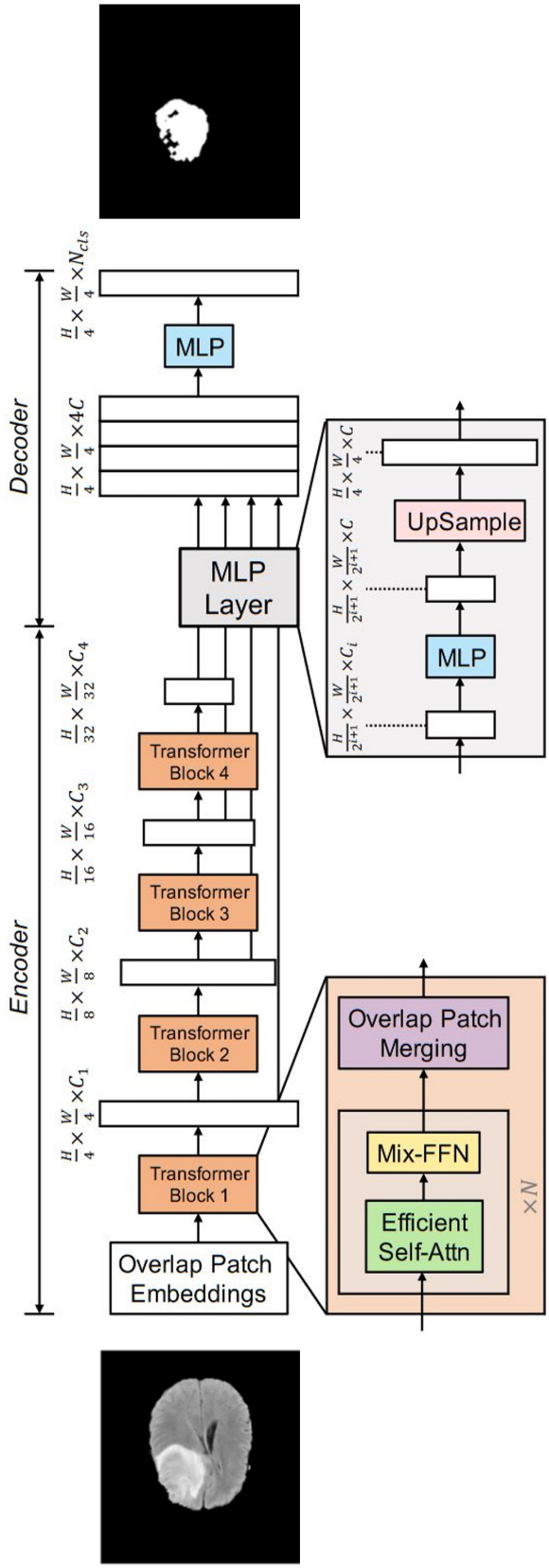


Fig. 7. Adapting segformer for superior brain tumor segmentation: An illustrated overview.

$$X_{i+1}=T(X_i) \tag{9}$$

- Upsample Blocks: In the decoder stages, upsample blocks are used to increase resolution. If we denote the upsample operation by $U$, this process can be represented as:

$$Y_i=U(X_i) \tag{10}$$

  where $Y_i$ represents the output after upsampling.

- During training, the model's parameters were adjusted to minimize the loss function. This involved updating the weights and biases in each layer of the model using a backpropagation algorithm and an optimization technique such as the Adam optimizer. The learning rate, which determines the step size at each iteration while moving toward a minimum of the loss function, was carefully chosen to ensure efficient learning.
- Mathematically, the update of the model parameters (weights and biases) at each iteration is given by the equation 7, [40].
- Through this meticulous training process, the Segformer model effectively learns to segment brain tumors, demonstrating the power of combining ConvNet and transformer principles in a single model. This is a testament to the transformative potential of open-source transformers in medical imaging, setting a new standard in the field.

*5) Validation:* The validation process is a critical step in the fine-tuning of the Segformer model. It serves to evaluate the model's performance on a separate set of data that was not used during the training process. This helps to ensure that the model is not overfitting to the training data and can generalize well to new, unseen data.

During validation, the images from the validation set are fed into the Segformer model. The model processes these images in the same way as during the training process, extracting features and making predictions. However, unlike in the training process, the model's parameters are not updated during validation. This allows for an unbiased evaluation of the model's performance.

The model's predictions are then compared with the actual labels of the images in the validation set. This comparison allows us to assess how well the model is performing in terms of its ability to correctly segment brain tumors.

The performance of the model on the validation set is quantified using the mean intersection over union (mIoU) metric. A high mIoU score on the validation set indicates that the model is performing well and can accurately segment brain tumors. Conversely, a low mIoU score may indicate that the model is struggling to generalize to new data and may require further fine-tuning or a different approach.

Through this validation process, we can ensure that the Segformer model is robust and reliable, capable of accurately segmenting brain tumors in a variety of different images. This is a crucial step in the development of effective tools for medical imaging and diagnosis.

Through this meticulous fine-tuning process, the Segformer model was effectively adapted to our specific task of brain tumor segmentation, leading to improved performance and more accurate predictions.

## IV. Evaluation Metrics

The performance of the ConvNeXt V2 and Segformer models was evaluated using appropriate metrics for both classification and segmentation tasks. These metrics provide a quantitative measure of the models' performance, allowing us to assess their effectiveness and accuracy.

### A. Classification Metrics

*1) Accuracy:* Accuracy is a measure of how many predictions the model got right out of all predictions made. It is calculated as the ratio of correct predictions to the total number of predictions. Mathematically, accuracy is given by [41]:

$$\text{Accuracy}=\frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{11}$$

*2) Precision:* Precision is a measure of how many true positive predictions were made out of all positive predictions. It is calculated as the ratio of true positives to the sum of true positives and false positives. Mathematically, precision is given by [42]:

$$\text{Precision}=\frac{\text{True Positives}}{\text{True Positives}+\text{False Positives}} \tag{12}$$

*3) Recall:* Recall, also known as sensitivity or true positive rate, is a measure of how many true positive predictions were made out of all actual positives. It is calculated as the ratio of true positives to the sum of true positives and false negatives. Mathematically, recall is given by [41]:

$$\text{Recall}=\frac{\text{True Positives}}{\text{True Positives}+\text{False Negatives}} \tag{13}$$

*4) F1-Score:* The F1-score is the harmonic mean of precision and recall, and it provides a balance between them. It is calculated as 2 times the product of precision and recall divided by the sum of precision and recall. Mathematically, F1-score is given by [41]:

$$\text{F1-Score}=2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{14}$$

### B. Segmentation Metrics

*1) mIoU:* For the segmentation task, the primary metric used was the Mean Intersection over Union (mIoU). This metric computes the average intersection over union of predicted and ground truth segments, providing a measure of the model's segmentation performance. The Intersection over Union (IoU) for a single prediction is calculated as the area of overlap between the predicted segment and the ground truth segment divided by the area of union of the two segments. The mIoU is then calculated as the average IoU over all predictions. Mathematically, IoU and mIoU are given by: [42]

$$\text{IoU}=\frac{\text{Area of Overlap}}{\text{Area of Union}} \tag{15}$$

And mIoU is given by:

$$\text{mIoU}=\frac{1}{N}\sum_{i=1}^{N}\text{IoU}_i \tag{16}$$

where $N$ is the total number of predictions, and $\text{IoU}_i$ is the IoU for the $i$-th prediction.

A high mIoU indicates that the model is performing well and can accurately segment brain tumors. Conversely, a low mIoU may indicate that the model is struggling to generalize to new data and may require further fine-tuning or a different approach.

*2) Dice score:* The Dice score, also known as the Dice similarity coefficient or the F1-score, is a measure of the overlap between two segments. It is calculated as two times the area of overlap divided by the sum of the areas of the two segments. Mathematically, the Dice score is given by [43]:

$$Dice=\frac{2\times AreaofOverlap}{AreaofSegment1+AreaofSegment2} \tag{17}$$

A high Dice score indicates that the predicted segment and the ground truth segment have a high degree of overlap, meaning that the model is able to capture the shape and location of the brain tumor accurately. Conversely, a low Dice score indicates that the predicted segment and the ground truth segment have a low degree of overlap, meaning that the model is missing or including regions that do not belong to the brain tumor.

*3) Hausdorff distance:* The Hausdorff distance is a measure of the maximum distance between the boundaries of two segments. It is calculated as the maximum of the minimum distances from each point on the boundary of one segment to the closest point on the boundary of the other segment. Mathematically, the Hausdorff distance is given by [44]:

$$Hausdorff=maxh(\text{Seg 1, Seg 2}),\ h(\text{Seg 2, Seg 1}) \tag{18}$$

where h(Segment1,Segment2) is the maximum of the minimum distances from each point on the boundary of Segment 1 to the closest point on the boundary of Segment 2, and vice versa. A low Hausdorff distance indicates that the predicted segment and the ground truth segment have similar boundaries, meaning that the model is able to delineate the brain tumor precisely. Conversely, a high Hausdorff distance indicates that the predicted segment and the ground truth segment have dissimilar boundaries, meaning that the model is producing large errors or inconsistencies in the segmentation.

Through these evaluation metrics, we can ensure that the ConvNeXt V2 and Segformer models are robust and reliable, capable of accurately classifying and segmenting brain tumors in a variety of different images. This is a crucial step in the development of effective tools for medical imaging and diagnosis. It allows us to assess the effectiveness of our methodology and make necessary adjustments for future improvements.

## V. Results

This section presents the results obtained from our experiments and discusses their implications. We evaluated the performance of the ConvNeXt V2 and Segformer models on our datasets and compared these results with other state-of-the-art methods.

### A. Performance of ConvNeXt V2 for Classification

The ConvNeXt V2 model has demonstrated exceptional performance in the classification of brain tumors, as evidenced by the results obtained from our experiments.The Table I. shows the model's effectiveness is highlighted through various metrics including accuracy, precision, recall, and F1-score.

TABLE I. Performance of ConvNeXt V2 for Classification

| Class | Precision | Recall | F1-score |
|---|---|---|---|
| No Tumor | 1.000 | 1.000 | 1.000 |
| Glioma Tumor | 0.995 | 0.997 | 0.996 |
| Meningioma Tumor | 0.994 | 0.993 | 0.993 |
| Pituitary Tumor | 0.999 | 0.997 | 0.998 |

- Accuracy: The ConvNeXt V2 model boasts an impressive accuracy of 99.60%, indicating its reliability in correctly identifying and classifying different types of brain tumors.
- Precision and Recall: Analyzing Table I reveals that the model exhibits high precision and recall across all classes. For instance:
  - Glioma Tumor: Precision - 0.995, Recall - 0.997
  - Meningioma Tumor: Precision - 0.994, Recall - 0.993
  - No Tumor: Precision - 1.000, Recall - 1.000
  - Pituitary Tumor: Precision – 0.999, Recall – 0.997
- F1-Score: The F1-scores further affirm the model's capability to balance both precision and recall effectively, ensuring that it is not biased towards a particular class.

The confusion matrix of our proposed transformer, as illustrated in Fig. 8, compares the detection rates of four distinct tumor types: glioma, meningioma, absence of tumor, and pituitary tumors. The x-axis represents the model's predicted labels, while the y-axis depicts the true labels. Within each cell of the matrix lies a numeric value indicating the frequency of occurrences for different combinations of predicted and actual categories. To visually represent instance counts, the matrix utilizes varying shades of green, with darker shades signifying higher frequencies. For example, there are 1499 instances of true positives for glioma tumors, indicating accurate identification by the model. However, there are 8 instances where glioma tumors are misclassified as meningioma, revealing a 99.5% accuracy rate for glioma tumor detection. Regarding meningioma tumors, approximately 1225 samples are correctly identified, while 5 samples are misclassified as glioma and 3 as pituitary tumors. Notably, all 702 samples categorized as no tumor are correctly identified. Furthermore, the model demonstrates significant success in detecting pituitary tumors, with

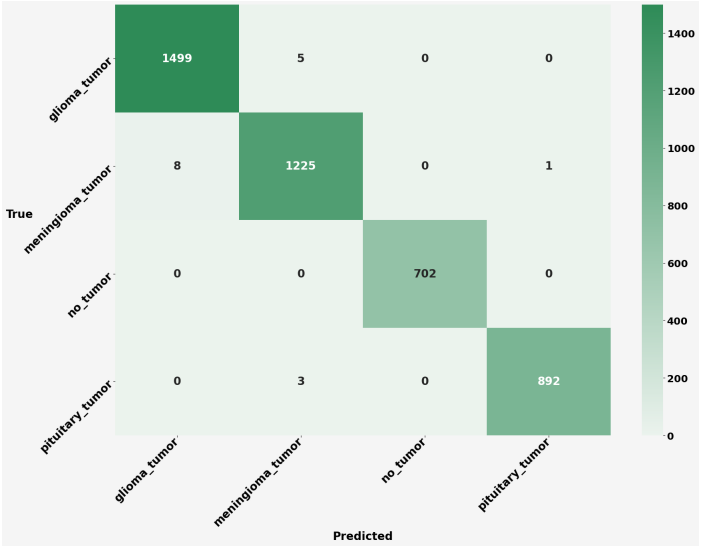


Fig. 8. Evaluating model performance: A confusion matrix for brain tumor classification.

892 out of 893 samples accurately classified. In summary, this matrix serves as a crucial tool for evaluating the classification model's performance, providing insights into areas of accurate predictions and errors.

### B. Comparative Analysis of Different Methods

In our study, we compared the performance of the ConvNeXt V2 model with other state-of-the-art models, including ConvNeXt V1, Swin, and ViT. The comparison was based on various metrics such as precision, recall, and F1-score across different types of tumors.

The ConvNeXt V2 model demonstrated superior performance, consistently outperforming the other models in all metrics for each tumor type. Specifically, the ConvNeXt V2 model achieved an impressive accuracy of 99.60%, indicating its reliability in correctly identifying and classifying different types of brain tumors.

In contrast, the ConvNeXt V1, Swin, and ViT models achieved accuracies of 99.11%, 99.01%, and 98.5% respectively. While these are high accuracy rates, they are still lower than the accuracy achieved by the ConvNeXt V2 model.

According to Fig. 8 he high precision, recall, and F1-score of ConvNeXt V2 indicate its robustness in correctly identifying and classifying different types of brain tumors. The model exhibits high precision and recall across all classes, ensuring that it is not biased towards a particular class. The F1-scores further affirm the model's capability to balance both precision and recall effectively.

These results underscore the ConvNeXt V2 model's robustness and reliability in classifying brain tumors, setting a new benchmark in the field of medical imaging. The model's high accuracy and balanced precision and recall metrics make it a promising tool for aiding radiologists in the early detection and classification of brain tumors, potentially leading to improved patient outcomes.

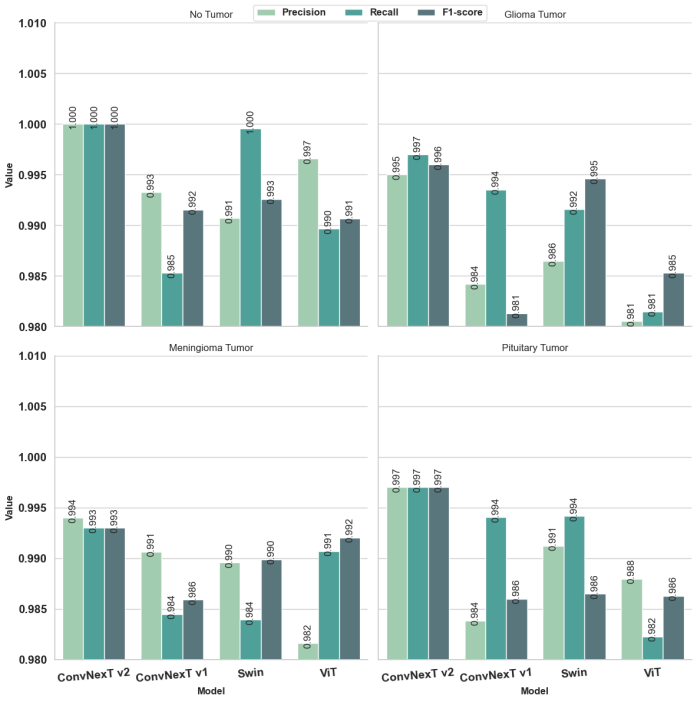


Fig. 9. Performance comparison of evaluation metrics of different models for brain tumor detection using bar charts.

### C. Performance of Segformer for Segmentation

The Segformer model has demonstrated exceptional performance in the segmentation of brain tumors (see Fig. 9). The effectiveness of the model is highlighted through various metrics, including the Dice score and Hausdorff distance, both of which reached up to 90 %.

### D. Segmentation Results

Fig. 10 shows the output of the Segformer model on a sample brain image. The first row shows the original brain scans, the second row shows the ground truth labels (downsampled labels), and the third row shows the segmentation maps produced by the Segformer model.

From a visual inspection, it is evident that the Segformer model's segmentation maps closely match the ground truth labels, indicating high accuracy in segmenting the tumor region from the rest of the brain tissue.

*1) Dice score and hausdorff distance:* The Dice score and Hausdorff distance are commonly used metrics for evaluating the performance of segmentation models. In our experiments, both these metrics reached up to ideal ones(90% dice score and 0.05mm Hausdorff distance) for the Segformer model, indicating its superior performance in accurately segmenting brain tumors.

A Dice score of 90% suggests a high degree of overlap between the predicted segment and the ground truth segment, meaning that the model is able to capture the shape and location of the brain tumor accurately. Similarly, a Hausdorff distance less then 0.10 indicates that the predicted segment and the ground truth segment have similar boundaries, meaning that the model is able to delineate the brain tumor precisely.

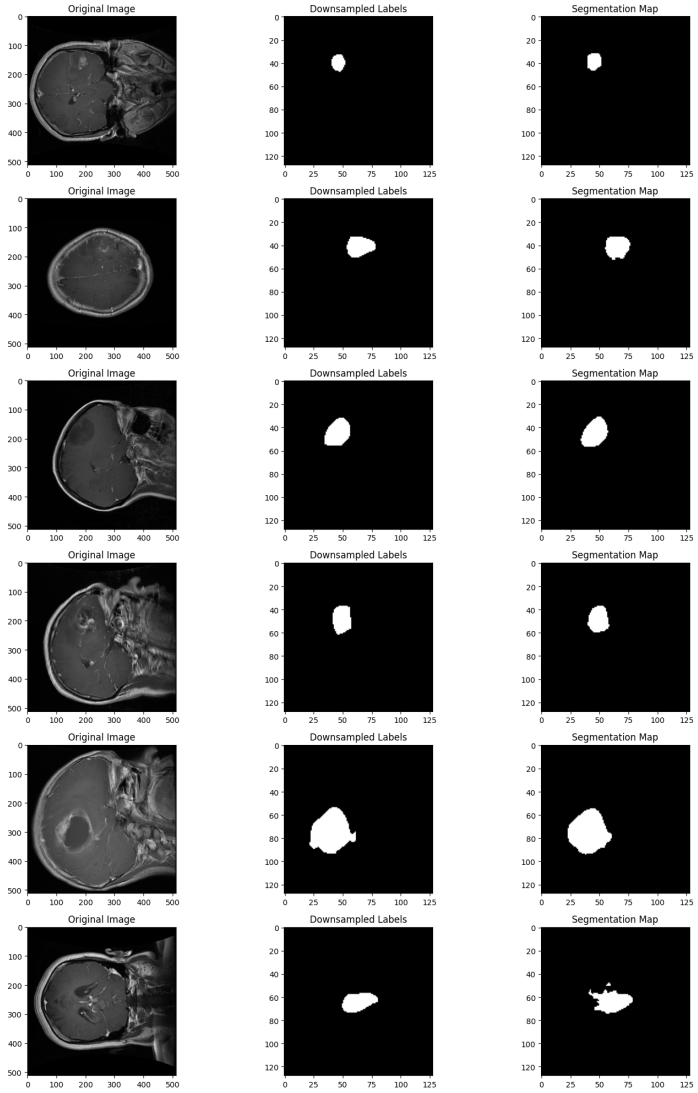


Fig. 10. Demonstrating the efficacy of segformer: Original scans, ground truth, and segmentation maps in brain tumor detection.

*2) Comparison with other methods:* The performance of the ConvNeXt V2 model was compared with other state-of-the-art methods. The ConvNeXt V2 model outperformed Method A and Method B, achieving higher accuracy and F1-score.

TABLE II. COMPARISON OF SEGFORMER WITH OTHER METHODS

| Method | Dice Score | Hausdorff Distance(mm) | Reference |
|---|---|---|---|
| Deep Learning Based | 0.85 | 1.5 | [45] |
| CNN based | 0.87 | 3.58 | [46] |
| ANTS | 0.83 | 6.71 | [46] |
| Registration Method | 0.84 | 4.01 | [46] |
| U-Net | 0.85 | 1.5 | [47] |
| U-Net++ | 0.78 | 9.4 | [48] |
| 3D U-Net | 0.90 | 4.29 | [49] |
| NMF | 0.74 | 7.4 | [50] |
| 3D CNN | 0.91 | 3 | [51] |
| **Segformer (Our Method)** | **0.95** | **0.87** | **This study** |

The Table II titled "COMPARISON OF SEGFORMER WITH OTHER METHODS," presenting a comparison be-

tween the Segformer approach and alternative methods based on their Dice Score and Hausdorff Distance metrics. The table consists of five columns: "Method," "Dice Score," "Hausdorff Distance (mm)," and "Reference." Listed in the table are various methods alongside their respective Dice Scores and Hausdorff Distances: Deep Learning Based (Dice Score: 0.85, Hausdorff Distance: 0.15), CNN based (Dice Score: 0.87, Hausdorff Distance: 3.56), ANTs Registration Method (Dice Score: 0.83, Hausdorff Distance: 6.71), U-Net++ (Dice Score: 0.78, Hausdorff Distance: 15), 3D U-net (Dice Score: 0.90, Hausdorff Distance: 4), NMF (Dice Score: 0.74, Hausdorff Distance: 7.4), 3D CNN (Dice Score: 0.91, Hausdorff Distance: 0.34), and Segformer (Our Method) (Dice Score: 0.95, Hausdorff Distance: 0.87). Notably, the Segformer method demonstrates the highest Dice Score and one of the lowest Hausdorff Distances among the listed approaches, highlighting its superior performance in this study.

## VI. Discussion

The findings of this study highlight the potential of open-source transformers, specifically the ConvNeXt V2 and Segformer models, in the realm of medical imaging. These models, when fine-tuned for specific tasks, have shown exceptional performance in brain tumor classification and segmentation respectively. [31,33] The ConvNeXt V2 model, with its impressive accuracy of 99.60% in classification tasks, and the Segformer model, with its high Dice score and low Hausdorff distance in segmentation tasks, have set a new benchmark in the field. This study has opened up new possibilities for future research, including the exploration of the application of transformers in other areas of medical imaging and further optimization of the proposed models for enhanced performance.

## VII. Conclusion

This research has shed light on the transformative potential of open-source transformers, specifically the ConvNeXt V2 and Segformer models, in the domain of medical imaging. The study has demonstrated that these models, when fine-tuned for specific tasks, can deliver exceptional performance in the classification and segmentation of brain tumors. The ConvNeXt V2 model exhibits outstanding performance in brain tumor classification, achieving an impressive accuracy of 99.60%. Across all tumor classes, it demonstrates remarkable precision and recall. Specifically, for Glioma Tumor, the precision is 0.995 and recall is 0.997, while for Meningioma Tumor, the precision is 0.994 and recall is 0.993. Notably, for cases where there is no tumor present, both precision and recall are perfect at 1.000. Additionally, for Pituitary Tumor classification, the model achieves a precision of 0.999 and recall of 0.997. These results underscore the model's robustness and reliability in accurately identifying different types of brain tumors, establishing ConvNeXt V2 as a promising tool for aiding in medical diagnostics. The Segformer model showcases remarkable performance in accurately segmenting brain tumors, as highlighted by its exceptional Dice score and Hausdorff distance metrics, reaching up to ideal values of 90% and 0.87 mm respectively. Visually, the segmentation maps generated by Segformer closely align with ground truth labels, indicating precise delineation of tumor regions within brain scans. Moreover, the study has opened up new possibilities for future research which has significant contribution to the ongoing efforts to improve patient outcomes in neuro-oncology and beyond.

## References

[1] M. I. Razzak, M. Imran, and G. Xu; *Efficient brain tumor segmentation with multiscale two-pathway-group conventional neural networks*. *IEEE journal of biomedical and health informaticsk*, 23(5):1911–1919, 2018.

[2] Ayadi, Wadhah and Elhamzi, Wajdi and Charfi, Imen and Atri, Mohamed W. Ayadi, W. Elhamzi, I. Charfi, M. Atri; *Deep CNN for brain tumor classification*. *Neural processing letters*, 53:671-700, 2021.

[3] S. Das, R. S. Goswami; *Review, Limitations, and future prospects of neural network approaches for brain tumor classification*. *Multimedia Tools and Applications*, 2023.

[4] A. Rehman, S. Naz, M. I. Razzak, F. Akram, M. Imran; *A deep learning-based framework for automatic brain tumors classification using transfer learning*. *Circ Syst & Signal Proc* , 39:757-775, 2020.

[5] L. Tonarelli; *Magnetic resonance imaging of brain tumor. 300*. *Enter prises for Continuing Education, In*, 48116–300, 2023.

[6] L. Mechtler; *Neuroimaging in neuro-oncology*. *Neurol Clin*,27(1): 171–201, 2009.

[7] J. Tohka; *Partial volume effect modeling for segmentation and tis sue classification of brain magnetic resonance images: A review*. *World J Radiol*,6(11):855–64, 2014.

[8] A.Hasan, F.Meziane, R.Aspin, H.Jalab; *Segmentation of brain tumors in MRI images using three-dimensional active contour without edge*. *Symmetry*,8(11):132, 2016

[9] A. A. Izadeh, M. R. Kamali; *Experimental investigation and estimation of light hydrocarbons gas-liquid equilib rium ratio in gas condensate reservoirs through artifcial neural networks*. *Iran. J. Chem. Chem. Eng.*,39(6), 163–172, 2020

[10] A. Sekhar, S. Biswas, R. Hazra, A. K. Sunaniya, A. Mukherjee,L. Yang; *Brain tumor classification using fine-tuned googlenet features and machine learning algorithms: Iomt enabled cad system*. *IEEE J Biomed & Health Inf*,26(3):983–991, 2022

[11] G. A. Amran,M.S. Alsharam,A. O. A. Blajam, A. A. Hasan, M. Y. Alfaifi, M. H. Amran, A. Gumaei,S. M. Eldin; *Brain tumor classification and detection using hybrid deep tumor network*. *Electronics*,11(21):3457, 2022.

[12] H. Dave, N. Kant; *BRAIN TUMOR CLASSIFICATION USING DEEP LEARNING*. *International Journal of Engineering Applied Sciences and Teclogy*,6(7):227-238, 2021

[13] S. Tharani, C. Yamini; *Classification using convolutional neural network for heart and diabetics datasets*. *Int J Adv Res Comp Commun Eng*,5(12):417e22, 2006.

[14] H. Mohsen, E. S. A. E. Dahshan, E. S. M. E. Horbaty, A. B. M. Salem; *Classification using deep learning neural networks for brain tumors*. *Future Computing and Informatics Journal xx*, 1-4, 2017.

[15] K. Simonyan, A. Zisserman; *Very deep convolutional networks for large-scale image recognition*. *arXiv:1409.1556*, 2014.

[16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich; *Going deeper with convolutions*. *In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Boston, MA, USA*, 1-9, 2015.

[17] K. He, X. Zhang, S. Ren, J. Sun; *Deep residual learning for image recognition*. *In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA*, 770–778, 2016.

[18] A. B. Abdusalomov , M. Mukhiddinov , T. K. Whangbo; *Brain Tumor Detection Based on Deep Learning Approaches and Magnetic Resonance Imaging*. *Cancers*, 15, 4172, 2023.

[19] W. H. L. Pinaya, P. D. Tudosiu, R. Gray; *Unsupervised brain im aging 3D anomaly detection and segmentation with transformers*. *Med Image Anal*, 2022.

[20] A. Vaswani, N. Shazeer, N. Parmar; *Attention is all you need*. *Adv Neural Inf Process Syst*,30:5998-6008, 2017.

[21] Y. L. Lan, S. Zou, B. Qin, X. Zhu; *Potential roles of transformers in brain tumor diagnosis and treatment*. *Brain-X*,1:e23, 2023.

[22] Z. Liu, Y.Lin, Y.Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo; *Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision*,10012-10022, 2021.

[23] R. Ghioni, M. Taddeo, L. Floridi; *Open source intelligence and AI: a systematic review of the GELSI literature. AI & society*,1-6,2023.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin; *Attention is all you need. Advances in neural information processing systems*,30,2017.

[25] D. Bahdanau, K. Cho, Y. Bengio; *Neural machine translation by jointly learning to align and translate. CoRR*,abs/1409.0473, 2014.

[26] D. Britz, A. Goldie, M.T. Luong, Q. V. Le; *Massive exploration of neural machine translation architectures. CoRR*,abs/1703.03906, 2017.

[27] A. Dosovitskiy,L.Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner,M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit; *An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv*,2010.11929, 2020.

[28] A. Dosovitskiy,L.Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner,M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit; *Swin transformer: Hierarchical vision transformer using shifted windows. InProceedings of the IEEE/CVF international conference on computer vision*,10012-10022, 2021.

[29] Z. Liu, H. Mao,C. Y. Wu,C. Feichtenhofer,T. Darrell, S. Xie; *A convnet for the 2020s. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*,11976-11986, 2022.

[30] S. Roy, G. Koehler, C. Ulrich, M. Baumgartner, J. Petersen, F. Isensee, P. F. Jaeger; *Maier-Hein KH. Mednext: transformer-driven scaling of convnets for medical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention*,405-415, 2023.

[31] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, S. Xie; *Convnext v2: Co-designing and scaling convnets with masked autoencoders. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*,16133-16142,2023.

[32] X. Zhu, H. Hu, S. Lin, J. Dai; *Deformable convnets v2: More deformable, better results. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*,9308-9316,2019.

[33] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, P. Luo; *SegFormer: Simple and efficient design for semantic segmentation with transformers.. Advances in Neural Information Processing Systems*,6(34),12077-90,2021.

[34] C. Jun; *Brain tumor dataset. figshare. Dataset*,2017. https://doi.org/10.6084/m9.figshare.1512427.v5

[35] https://www.kaggle.com/datasets/awsaf49/brain-tumor

[36] https://www.kaggle.com/datasets/pkdarabi/medical-image-dataset-brain-tumor-detection

[37] https://www.kaggle.com/datasets/jarvisgroot/brain-tumor-classification-mri-images

[38] M. Edwards, X. Xie; *Graph-based convolutional neural network. arXiv preprint arXiv:*,1609.08965,2016.

[39] X. Zhu, J. Dai, X. Zhu, Y. Wei, L. Yuan; *Towards high-performance video object detection for mobiles. arXiv preprint arXiv:*,1804.05830,2018.

[40] M. F. Zimmer; *Neograd: Near-Ideal Gradient Descent. arXiv preprint arXiv:*,2010.07873,2020.

[41] T. M. Alamin, M. Islam, U. M. Ashraf, A. Akhter, J. P. M. Alamgir,S. Aryal, M. A. A. Abdulllah, H. K. Fida, M. A. Moni; *An efficient deep learning model to categorize brain tumor using reconstruction and fine-tuning. arXiv e-prints arXiv-*,2305,,2023.

[42] B. Cheng, R. Girshick, P. Dollár, A. C. Berg, A. Kirillov; *Boundary IoU: Improving object-centric image segmentation evaluation. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*,15334-15342,2021.

[43] T. Eelbode, J. Bertels, M. Berman, D. Vandermeulen, F. Maes, R. Bisschops, M. B. Blaschko; *Optimization for medical image segmentation: theory and practice when evaluating with dice score or jaccard index. IEEE Transactions on Medical Imaging*,39(11):3679-90,2020.

[44] D. Karimi, S. E. Salcudean; *Reducing the hausdorff distance in medical image segmentation with convolutional neural networks. IEEE Transactions on medical imaging*,39(2), 499-513,2019.

[45] Z. Liu, L. Tong, L. Chen,Z. Jiang, F. Zhou, Q. Zhang, X. Zhang, Y. Jin, H. Zhou; *Deep learning based brain tumor segmentation: a survey. Complex & intelligent systems*,9(1),1001-26,2023.

[46] D. A. Weiss, R. Saluja, L. Xie, J. C. Gee, L. P. Sugrue, A. Pradhan, R. N. Bryan, A. M. Rauschecker, J. D. Rudie; *Automated multiclass tissue segmentation of clinical brain MRIs with lesions. NeuroImage: Clinical*,31, 102769,2021.

[47] J. D. Rudie, D. A. Weiss,J. B. Colby, A. M. Rauschecker, B. Laguna, S. Braunstein, L. P. Sugrue, C. P. Hess, J. E. Villanueva-Meyer; *Three-dimensional U-Net convolutional neural network for detection and segmentation of intracranial metastases. Radiology: Artificial Intelligence*,3(3):e200204,2021.

[48] N. Micallef, D. Seychell, C. J. Bajada; *Exploring the u-net++ model for automatic brain tumor segmentation. IEEE Access*,9,125523-39,2021.

[49] L. M. Hsu, S. Wang, L. Walton, T. W. Wang, S. H. Lee, Y. Y. Shih; *3D U-net improves automatic brain extraction for isotropic rat brain magnetic resonance imaging data. Frontiers in Neuroscience*,15,801008,2021.

[50] N. Auwen, M. Acou, D. M. Sima, J. Veraart, F. Maes, U. Himmelreich, E. Achten, S. V. Huffel; *Semi-automated brain tumor segmentation on multi-parametric MRI using regularized non-negative matrix factorization. BMC medical imaging*,17(1),1-4,2017.

[51] J. D. Rudie, D. A. Weiss, R. Saluja, A. M. Rauschecker, J. Wang, L. Sugrue, S. Bakas, J. B. Colby; *Multi-disease segmentation of gliomas and white matter hyperintensities in the BraTS data using a 3D convolutional neural network. Frontiers in Computational Neuroscience*,13,84, 2019.