A Comparative Analysis Between CNNs and ViTs for MRI-based Brain Tumor Classification

Yuteng Zhang *

School of Data Science, The Chinese University of Hong Kong, Shenzhen, Shenzhen, 518172, China

* Corresponding Author Email: yutengzhang@link.cuhk.edu.cn

Abstract. Brain tumor classification from Magnetic Resonance Imaging (MRI) images is an important task in medical imaging for the determination of appropriate treatment strategies and improvement in patient outcomes. Brain tumors, including gliomas, meningiomas, and glioblastomas, are of the most lethal forms of cancer. This research explored the potential of replacing convolutional neural networks (CNNs) by Vision Transformers (ViTs) on classifying brain tumors by MRI images. The paper focused on pretrained model Vit-B16, comparing it with traditional CNNs, including VGG16, ResNet-50, and EfficientNet-B0. ViT-B16, pretrained on ImageNet-21k, achieves improved accuracy, precision, recall, and F1-Score after being fine-tuned on the brain tumor dataset when applying data augmentation. The self-attention mechanism helps ViTs capture long-range dependencies and global context from the images, significantly improving the performance. As shown in the results, ViTs can efficiently handle complex dataset and become a useful tool for the area of medical imaging classification. This paper emphasizes the potential of Vision transformers in improving classification accuracy in the diagnosis of brain tumors. Future work can be focused on exploring better ViT architectures and data augmentation techniques.

Keywords: Brain tumor classification; vision transformer; convolutional neural network; magnetic resonance imaging.

1. Introduction

Brain tumors are of the most challenging and deadly forms of cancer. Brain tumors cause a remarkable percentage of cancer deaths across the world. These tumors can be classified into various types, which vary greatly in terms of malignancy and prognosis. The most common types of brain tumors include glioblastoma, meningioma, and gliomas [1]. For the purpose of choosing the best course of therapy and, consequently, improving patient outcomes, a brain tumor must be properly classified. Correct diagnosis is important to the treatment plan, which includes surgery, radiation, and chemotherapy.

A popular non-invasive imaging technique that produces helpful images of the architecture and disease of the brain is Magnetic Resonance Imaging (MRI) [2]. A crucial method for identifying and categorizing brain cancers is MRI. However, error rates in diagnostic by imaging is relatively high [3]. Radiological analysis of MRI images is a highly time-consuming and subjective work. As a result, it often leads to an erroneous diagnosis with attendant possibility of treatment delay. Therefore, automated and reliable methods for brain tumor classification are helpful and necessary.

The development of machine learning methods, particularly deep learning, has altered medical image analysis considerably in recent years [4]. Machine learning presents highly promising results with automation in the classification of brain tumors for timely and precise diagnosis [5,6]. Among these techniques, convolutional neural networks are in the leading position, holding their capabilities to capture hierarchical characteristics from the images efficiently [7].

Many image classification tasks are successfully completed by Convolutional Neural Networks (CNNs), and some of these tasks are related to medical imaging. In particular, these models can work best in tasks with image data since the models pick up vital features effectively through their convolutional layers. CNNs have been widely applied to different applications in medical imaging for tumor detection, segmentation, and classification [8]. For instance, CNN-based approaches give

great precision in using MRI images to detect brain cancers and typically outperform the classical machine learning approach.

Though having widespread use and success, CNNs suffer from certain limitations. The major drawback here is that a CNN tends to introduce bias into local patterns, arising directly from the nature of convolutional operations. At a basic level, convolutions are local operations that look just into small regions of the input image one at a time. While this allows CNNs to effectively capture detailed and localized features, it also limits their ability to capture long-range dependencies and global context within the data [9]. Moreover, CNNs are seen as a black box because the processes of making decisions are not transparent. This could be a very big problem in medical domains since an explanation for the models is strongly needed.

In order to solve the problems, Vaswani et al. proposed attention mechanism, encouraging the model to focus on important information of inputs [10]. Later, as the success of transformers in the area of Natural Language Processing, transformers were introduced to computer vision. Vision transformers (ViTs) are considered to be a promising alternative of CNNs [11, 12]. ViTs can capture global context and long-range dependencies more effectively than CNNs. By addressing some of the basic shortcomings of CNN-based models, this novel approach has the promise to greatly improve the performance of automated medical image analysis, including the categorization of brain cancers.

The basic goal of this research is to enhance the accuracy in the classification of brain tumors by using Vision Transformers for further refinement of treatment strategies and patient outcomes. By addressing the limitations of CNNs, ViTs offer a robust method for analyzing MRI scans to differentiate between gliomas, meningiomas, and glioblastomas. The methodology includes the collection and preprocessing of annotated MRI datasets; developing and fine-tuning of a ViT model, and comparison of its performance with a traditional approach based on CNNs. The goal is to evaluate ViTs' effectiveness in classifying brain tumors, exploring the feasibility of replacing CNNs with ViTs.

2. Methodology

2.1. Dataset

The dataset contains three kinds of brain tumors images: meningioma, glioma, and pituitary tumor [13]. The images are T1-weighted contrast-enhanced brain MRI images (3064 slices in total). The numbers of slices are 708, 1426, and 930 respectively. The images were collected from 233 patients.

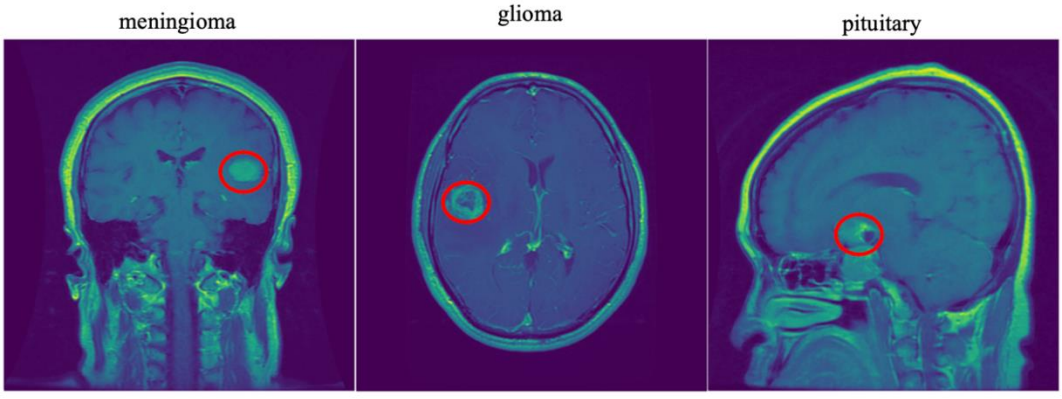


Fig 1. Representative examples of brain MRI images [13].

Fig. 1 shows one MRI images from each type of brain tumors. The red circle marked the tumors' locations. The original size of these images is 512 × 512 pixels.

2.2. Data Processing

First, in order to satisfy the input size of pretrained vision transformer model, the MRI images were resized from 512 × 512 pixels to 224 × 224 pixels. The pixel values were then normalized to the range [0, 1] to standardize the input data and facilitate the training process.

A few data augmentation methods were applied to increase the robustness of the model:

Flipping: Each image has a chance (25%) of being flipped horizontally. This mirrors the image along the vertical axis. It is the same for being flipped vertically, with a probability of 25%. This mirrors the image along the horizontal axis.

Rotation: images are randomly rotated at random angles in range from -30 to 30 degrees.

Noise: Gaussian noise is added to each image. This involves adding small, random variations to the pixel values.

2.3. Model Structure

The model used in this paper is ViT, which was first proposed by Dosovitskiy et al. in 2020. Vision Transformer divides images into patches of the same size which is fixed. Then Vit embeds these patches and applies a standard transformer architecture, which models the relationships between patches. The workflow helps Vit to process images. Vision Transformer architecture contains several main components:

Embedding Layer: The input image is divided into different patches that is non-overlapping. The size of each patch is 16 × 16 pixels. 196 pitches are got from the 224 × 224 pixels inputs. After that, a learnable linear projection is used to flatten and linearly embed these patches into a high-dimensional space. Then 1D sequences of embedding are gotten from the 2D images patches. The spatial information of the acquired patches is preserved by appending the learnable position embeddings to the patch embeddings. To help the model fully understand the spatial information, the relationship between pitches, the position embeddings encode the position of each patch of the image.

Transformer Encoder and Self-Attention Mechanism: The transformer encoder passes the embedded patches through a number of layers. The layers mainly comprised by multi-head self-attention and feed-forward neural networks.

Multi-Head Self-Attention: Each patch's attention score is calculated by the self-attention mechanism in relation to every other patch in the image. Attention is distributed across multiple representation subspaces possibly entangled by letting multiple heads look at different parts of the same image. With the help of this multi-head technique, the model may focus on several points in the input sequence at once, which helps it comprehend the image more fully.

Unlike convolutions in CNNs that focus on local features, self-attention captures global relationships within the entire image. This allows the model to understand the broader context and interactions between distant regions of the image. The attention mechanism dynamically calculates weights for each patch, enabling the model to prioritize more relevant parts of the image for the current task. This adaptiveness makes the architecture more capable of learning the important features. Self-attention is one of the main features of this architecture, allowing the model to learn the long-range dependencies and global context within the image.

Classification Head: To calculate the probability of class labels, the Transformer encoder's output is fed to a classification head, which is followed by a fully connected layer and a softmax layer.

2.4. Transfer Learning

In this work, pre-trained Vision Transformer models was utilized in transfer learning. Transfer learning fine-tunes a model that was pre-trained on data from a different tack on another dataset and adopted for training in another dataset but maintains the features learned from the initial training process.

Pre-trained Model: In the first stage, the Vision Transformer model is pretrained on the ImageNet dataset. This massive dataset helps the model understand various features and patterns across images, which may later be transferred to the task of brain tumor classification.

Fine-tuning: Fine-tune involves using a lower learning rate on a pre-trained ViT model on the brain MRI dataset. To be specific, fine-tuning will adjust the weights of the pre-trained model in fitting better with a new dataset. The general traits that the model learned from the original dataset

can therefore still be retained by the model, but the model is adjusted to the specific features of brain MRI images.

2.5. Evaluation Metrics

To comprehensively evaluate the performance of the Vision Transformer model, a variety of evaluation metrics are used. TP: true positive. TN: true negative. FP: false positive. FN: false negative.

Accuracy: The proportion of accurately anticipated cases to all instances. It provides an overall measure of model performance.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \tag{1}$$

Precision: The proportion of correctly predicted positives to all expected positives. It indicates the capacity of the model to correctly identify positive cases.

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

Recall: The proportion of all actual positives to original positive predictions. It assesses how well the model can recognize each positive case.

$$Recall = \frac{TP}{FN+TP} \tag{3}$$

F1 Score: The harmonic mean of precision and recall. It provides a combination of precision and recall.

$$F1\ Score = \frac{2\times Precision\times Recall}{Precision+Recall} \tag{4}$$

3. Experiments and Results

3.1. Experimental Settings

The experiments were conducted on an Nvidia RTX 4060 Graphics Processing Unit (GPU). The deep learning frameworks TensorFlow 2.10.0 was used, with NVIDIA CUDA v11.8. The version of Python is 3.8

ViT-B16 was used as pretrained model in the experiments. VGG-16, ResNet-50, EfficientNet-B0 were used as baseline models for comparison.

ViT-B16: ImageNet-21k served as pre-trained dataset and ImageNet 2012 served as fine-tuned dataset.

VGG-16: A well-known deep CNN model with 16 layers. Its main attribute is its architectural simplicity and regularity. Though simple in structure, VGG-16 was found highly effective for tasks in image classification and serves as a good baseline because it has established performance.

ResNet-50: A deep CNN model containing 50 layers that makes use of residual connections to make training deeper networks easier. The introduction of the residual blocks is useful, in that it helps to mitigate the vanishing gradient problem and let the network learn more complex features without degradation in performance. Its ability to train deeper networks while maintaining accuracy makes it a powerful model for comparison.

EfficientNet-B0: A model that balances network depth, width, and resolution using a compound scaling method, designed for efficient computation. The EfficientNet-B0 model shows high performance with considerably fewer parameters and very low computational cost. The architecture is optimized in a manner that brings strong performance but uses few resources.

The batch size was 12 for both the training and the validation. A fine-tuning learning rate of 0.0001 was used for all models. Default setting Adam optimizer. The models were trained for 15 epochs, with the patience set to 4 for early stopping.

3.2. Results

3.2.1. Model comparison

Table 1 shows the total parameters in millions (M) trained, memory used, and average training time per epoch needed by different models. The Vision Transformer (ViT-B16) has the largest number of parameters. Therefore, ViT-B16 also has the highest memory requirement. The number of parameters of Vit-B16 is greatly larger than VGG-16, ResNet-50, and EfficientNet-B0. Also, the memory requirement of ViT-B16 is more than twice as much as ResNet-50, which is higher than VGG-16 and EfficientNet-50. The mean time per epoch is comparable to that of VGG-16 and EfficientNet-B0 but slightly longer with the ResNet-50 model since the number of parameters in ViT-B16 is significantly bigger.

The difference in parameter counts as well as memory usage indicates the intricacy and depth of model ViT-B16. However, larger parameter counts typically enable learning more complex patterns and features within data sets but at the same time require more computational resources and memory thereby limiting their practical utility in environments with limited resources. On the other hand, EfficientNet-B0 stands out as a highly efficient model due to its much lower parameter count together with its memory usage.

Table 1. Model comparison.

Model	Parameters(M)	Memory(MB)	Avg Time per Epoch(s)
VGG-16	18.0	206	172
ResNet-50	36.8	421	136
EfficientNet-B0	4.2	48	163
ViT-B16	85.9	983	165

3.2.2. Result comparison without data augmentation

The performance of the models without data augmentation is shown in Table 2. It's clear that ResNet-50 yields the best performance with accuracy, precision, and recall, and F1-Score. This is mainly because its deep architecture and the effectiveness of the residual connections for the capturing and learning of the features embedded in the MRI images. Vit-B16 is the second-best model after ResNet-50, with an accuracy of 89.22%. These results also underline that Vision Transformers can process and learn very effectively through the image data, even without augmentation. These results also highlight that Vision Transformers can process and learn very effectively through the image data, even without augmentation.

EfficientNet-B0 and VGG-16 give low performance metrics compared with ResNet-50 and ViT-B16. For comparison, EfficientNet-B0 obtains an accuracy of 87.91%. It is seen that models like EfficientNet-B0, despite having much lower complexity and less parameter count, do manage to deliver quite well due to their highly optimized balance in network depth. VGG-16 reports the lowest among all performance. Its architecture simplicity may not eventually prove to have the robustness to match the performance exhibited by these advanced models such as ResNet-50 and ViT-B16.

Table 2. Result comparison without data augmentation.

Model	Accuracy	Precision	Recall	F1-Score
VGG-16	86.60%	88.63%	86.60%	86.80%
ResNet-50	**90.03%**	**90.40%**	**90.03%**	**90.14%**
EfficientNet-B0	87.91%	87.75%	87.91%	87.76%
ViT-B16	89.22%	89.32%	89.22%	89.26%

3.2.3 Result comparison with data augmentation

Using data augmentation in all models actually helps in improving performance, as shown in Table 3. Data augmentation includes flipping, rotation, and adding Gaussian noise to the images in the training dataset. As a result, the model can more effectively generalize to new, untested data. ViT-B16 performs best under data augmentation. It achieves the highest accuracy, precision, and recall,

and F1-Score. The self-attention mechanism of ViT-B16 benefits greatly from the increased data variability, allowing it to learn more robust and generalized features from the MRI scans. EfficientNet-B0 also improved significantly with data augmentation, surpassing ResNet-50 in performance metrics. It achieves an accuracy of 92.81%. Data augmentation with the efficient structure of EfficientNet-B0 can capture varied patterns within the data and is proving to achieve huge performance gains. However, data augmentation was still not enough for the underperforming ResNet-50 model compared to ViT-B16 and EfficientNet-B0. It achieves an accuracy of 91.01%. This indicates a diminishing return on performance improvements with data augmentation compared to the other models. VGG-16, although showing the least improvement, still benefits from data augmentation with increased metrics across the board. It records an accuracy of 87.25%. The relatively simpler architecture of VGG-16 limits its potential performance gains from data augmentation, but it still demonstrates notable improvements compared to its performance without augmentation.

Table 3. Result comparison without data augmentation.

Model	Accuracy	Precision	Recall	F1-Score
VGG-16	87.25%	88.70%	87.25%	86.95%
ResNet-50	91.01%	91.32%	91.01%	90.90%
EfficientNet-B0	92.81%	92.94%	92.81%	92.85%
ViT-B16	**94.28%**	**94.30%**	**94.28%**	**94.21%**

4. Discussion

The results of this work stress the great potential of ViTs in brain tumor classification from MRI images compared to conventional CNNs. Additional performance metrics shown by the novel ViT-B16 model underline its advantages in medical image analysis.

Among the tested models, ViT-B16 demonstrates higher classification accuracy, precision, recall, and F1-Score—especially with data augmentation. This is largely in part due to one of the inherent features of Vision Transformer uniqueness: the self-attention mechanism. In contrast to CNNs, which mainly focus on local patterns through convolutional operations, ViTs can capture long-range dependencies and global context within an image. Their ability to consider the whole image contextually allows the ViTs to develop much more comprehensive and nuanced representations of data, which is important for accurate tumor diagnosis.

Large increases in all performance metrics upon data augmentation further call ViTs robust. Data augmentation brings variation in the training data, providing the model with more generalization ability, hence reducing overfitting. In fact, the very fact that the ViT-B16 was able to make good use of the increased data suggests that it can adapt even to diverse and complicated patterns from MRI scans very well and is thus a very useful tool in the task of brain tumor classification.

Traditional CNNs, including VGG-16 and ResNet-50, over the last few years have been extensively applied to various image classification tasks as they can extract hierarchical features using convolutional layers. However, they are limited by their local receptive fields, which can overlook important global patterns within the data. The limitation is especially noticeable in medical imaging, since the relationships between distant regions of an image can be critical for accurate diagnosis.

The superior performance of ViTs in this study has some important implications for the field of medical imaging. Accurate and reliable brain tumor classification is essential for determining appropriate treatment strategies and improving patient outcomes. The ability of ViTs to provide higher accuracy and robustness in classification tasks means they could potentially reduce diagnostic errors and enhance the efficiency of medical professionals.

Furthermore, the success of ViTs in exploiting data augmentation techniques means that they can be effectively trained even on quite small and diverse medical datasets, a task that proves in particular extremely challenging in medical imaging. Large labeled datasets are difficult to get for medical imaging.

There is a future need to work on architectures of Vision Transformer models catering to the same tasks in medical imaging. It would be necessary to investigate various configurations of self-attention and experiment with different data augmentation techniques for optimal performance. Examining the interpretability of ViTs in medical imaging is another promising direction. As interpretability is a very key facet for application in medicine, where people have to understand how models come to their decisions, or even accurate ones, probably developing methods to visualize and come to an understanding of the decisions made by ViTs will take an important step in achieving trust from medical practitioners. This can likely reduce the interpretational barriers and make ViTs become acceptable in more medical settings.

In conclusion, the results from this experiment highlight that Vision Transformers might be the state-of-the-art alternative to traditional CNNs for the problem of brain tumor classification. Their ability to capture global context, coupled with their robustness to data augmentation, makes them highly effective for medical image analysis. Continuous research and development in this area are going to further contribute significantly toward automated medical diagnostics, enhancing the results and care of patients.

5. Conclusion

The high potential of Vision Transformers, and specifically the ViT-B16 model, for the categorization of brain cancers from MRI images was shown in this study. The results showed that ViTs have higher classification performance compared to conventional Convolutional Neural Networks like VGG-16 and ResNet-50 especially when data augmentation techniques are applied. The self-attention mechanism inherent in ViTs enables them to capture long-range dependencies and global context within images, which is essential for correctly classifying tumors. The superior performance of ViT-B16 highlights the model's robustness and generalization ability from complex, diverse datasets. This capability of efficiently handling augmented data suggests that ViTs can be trained efficiently even with limited and varied medical datasets, which presents a regular difficulty in medical imaging. The findings of this research suggest that Vision Transformers could enhance the accuracy and dependability of brain tumor classification, potentially reducing medical errors. However, further research on ViT architectures and data augmentation techniques is necessary to optimize their performance for medical applications. Future research should focus on optimizing ViT architectures for medical imaging and enhancing their interpretability to gain the trust of medical professionals. Developing methods to visualize and interpret the decisions made by ViTs will facilitate their integration into medical diagnostics.

References

[1] McFaline-Figueroa, J. Ricardo, Eudocia Q. Lee. Brain tumors. The American journal of medicine, 2018, 131(8): 874-882.

[2] Villanueva-Meyer, Javier E., Marc C. Mabray, Soonmee Cha. Current clinical brain tumor imaging. Neurosurgery, 2017, 81(3): 397-415.

[3] Itri Jason N, Tappouni Rafel R, McEachern Rachel, et al. Fundamentals of diagnostic error in imaging. Radiographics, 2018, 38(6): 1845-1865.

[4] Currie Geoff, Hawk K Elizabeth, Rohren Eric, et al. Machine learning and deep learning in medical imaging: intelligent imaging. Journal of medical imaging and radiation sciences, 2019, 50(4): 477-487.

[5] Aamir Muhammad, Rahman Ziaur, Dayo Zaheer Ahmed, et al. A deep learning approach for brain tumor classification using MRI images. Computers and Electrical Engineering, 2022, 101: 108105.

[6] Nazir, Maria, Sadia Shakil, Khurram Khurshid. Role of deep learning in brain tumor detection and classification (2015 to 2020): A review. Computerized medical imaging and graphics, 2021, 91: 101940.

[7] Sarvamangala, D. R., and Raghavendra V. Kulkarni. Convolutional neural networks in medical image understanding: a survey. Evolutionary intelligence, 2022, 15(1): 1-22.

[8] Salehi Ahmad Waleed, Khan Shakir, Gupta Gaurav, et al. A study of CNN and transfer learning in medical imaging: Advantages, challenges, future scope. Sustainability, 2023, 15(7): 5930.

[9] Azad Reza, Kazerouni Amirhossein, Heidari Moein, et al. Advances in medical image analysis with vision transformers: a comprehensive review. Medical Image Analysis, 2023: 103000.

[10] Vaswani Ashish, Shazeer Noam, Parmar Niki, et al. Attention is all you need. Advances in neural information processing systems, 2017, 30: 1-11.

[11] Dosovitskiy Alexey, Beyer Lucas, Kolesnikov Alexander, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv, 2020:2010.11929.

[12] Shamshad Fahad, Khan Salman, Zamir Syed Waqas, et al. Transformers in medical imaging: A survey. Medical Image Analysis, 2023, 88: 102802.

[13] Cheng Jun. Brain tumor dataset. Figshare. Dataset, 2017, URL: https://doi.org/10.6084/m9.figshare.1512427.v5. Last Accessed 2024/07/22