ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Vision transformers for glioma classification using T1 magnetic resonance imaging

W. M. S. P. B. Wickramasinghe, and Maheshi B. Dissanayake*

Department of Electrical and Electronics Engineering, Faculty of Engineering, University of Peradeniya, Peradeniya, Sri Lanka

(This article belongs to the *Special Issue: Artificial intelligence for diagnosing brain diseases*)

Abstract

Automated image analysis and classification have increasingly advanced in recent decades owing to machine learning and computer vision. In particular, deep learning (DL) architectures have become popular in resource-limited and labor-restricted environments such as the health-care sector. Transformer architecture, a DL method with self-attention mechanism, excels in natural language processing; however, its application in image-based diagnosis in health-care sector remains limited. Herein, the feasibility, bottlenecks, and performance of transformers in magnetic resonance imaging (MRI)-based brain tumor classification were investigated. To this end, a vision transformer (ViT) model was trained and tested using the popular Brain Tumor Segmentation (BraTS) 2015 dataset for glioma classification. Owing to limited data availability, domain adaptation techniques were used to pretrain the ViT model and the BraTS 2015 dataset was used for its fine-tuning. With the model only trained for 100 epochs, the confusion matrix for the two-class problem of tumor and nontumor classification showed an overall classification accuracy of 81.8%. In conclusion, although convolutional neural networks are traditionally used for DL-based medical image classification owing to their attention mechanism and long-range dependency-capturing capability, ViTs can outperform them in MRI-based brain tumor classification.

Keywords: Vision transformers; Medical image analysis; Deep neural networks; Magnetic resonance imaging; Convolutional neural network; Glioma detection

***Corresponding author:**
Maheshi B. Dissanayake
(maheshid@eng.pdn.ac.lk)

Citation: Wickramasinghe, WMSPB and Dissanayake MB. Vision transformers for glioma classification using T1 magnetic resonance imaging. *Artif Intell Health*. 2025;2(1):68-80. doi: 10.36922/aih.4155

Received: July 5, 2024

1st revised: August 28, 2024

2nd revised: September 9, 2024

Accepted: September 19, 2024

Published Online: November 6, 2024

Copyright: © 2024 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Medical imaging is crucial in the health-care sector for noninvasive diagnostic procedures because it can provide functional and visual representations of internal organs. X-ray imaging, nuclear imaging, magnetic resonance imaging (MRI), mammography, computed tomography (CT), and ultrasound imaging are some popular imaging techniques.[1] The four primary phases of medical image analysis include image formation, reconstruction, processing, and analysis. These phases help to create two-dimensional and three-dimensional (3D) images and enhance them; quantitative data are used for segmentation, classification, and object identification.[2,3] Modern advancements in artificial intelligence (AI), computer vision (CV), machine learning, and deep learning (DL) techniques can qualitatively and quantitatively improve medical image analysis.

MRI is a noninvasive medical imaging technique used for diagnosing brain tumors, injuries, and bleedings[4]. Over 120 types of brain tumors have been identified using MRI, which begin with primary tumors and are followed by secondary tumors. Glioma is a type of brain tumor that originates from glial cells and can be visualized through MRI or CT. Hence, automated image analysis techniques can positively contribute to the diagnosis of glioma.

Automated image analysis has rapidly evolved in the past few years owing to the introduction of AI and CV into traditional image processing techniques.[2] DL has been used in medical imaging to recognize cells of various sizes and shapes, locate organs and body parts, and automatically identify local anatomical features.[3-5] Owing to the intrinsic locality of convolution operations, popular architectures such as convolutional neural network (CNNs) have shown limitations in modeling straightforward long-range relations. Therefore, CNNs with attention mechanisms that assist AI models to focus on specific pixels, regions, or features have gained research attention in image analysis.

Transformer architecture, a DL method with self-attention mechanism, has become vital in natural language processing (NLP) tasks.[6] Recently, it has considerably impacted text classification, machine translation, and query responding. However, its application in CV problems requires further research. In CV, attention can either work in tandem with CNNs or replace some of its components while maintaining the overall network structure. Thus, this architecture largely has the potential to provide promising results in object detection, video classification, image classification, and image generation.

1.1. Contribution

This study mainly focused on evaluating the transformer architecture for MRI image classification when applied directly to the sequences of image patches. It concentrates on the classification of MRI images based on the presence and absence of glioma while overcoming the persistent class imbalance within a dataset to obtain feasible and resource-optimized solutions. We focus on glioma as it is a malignant (cancerous) brain tumor, which is treatable with high prognosis if detected early. The classification of brain tumors before segmentation is beneficial for saving time and resources, improving accuracy, and providing valuable information for treatment planning. Moreover, organized data aid in analysis and model training. However, medical data are mostly biased toward the absence of disease (negative outcome) and require careful implementation of algorithms to avoid model overfitting. This study aims to present a comprehensive ground-up mechanism for glioma classification using vision transformers (ViTs).

The primary contributions of this study are as follows:

(a) A ViT was proposed as an alternative to CNN for glioma classification using MRI.
(b) A pretraining method was proposed for ViTs when working with small datasets.
(c) An effective approach of intensity uniformization for MRI images as a preprocessing technique was introduced.
(d) The performance of CNN models and ViTs was compared for two grades of glioma classification as well as tumorous and nontumorous MRI classification.
(e) The effects of class imbalance in the medical dataset were discussed.

2. Background literature review

With the advancement in AI technologies, computer-aided diagnoses have been extensively studied in medical sciences for different disease diagnoses. In particular, noninvasive image-based diagnosis has garnered the attention of researchers and medical practitioners owing to its high accuracy, high precision, and auxiliary capabilities in applications such as brain tumor classification and segmentation using AI and DL models. Moreover, this field has gained popularity among medical image analysis researchers owing to well-established open challenges such as the BraTS challenge and publicly available large MRI datasets.[7-9] For instance, in their brain tumor classification and segmentation study, Kaldera *et al.*[5] proposed a simple CNN-based classifier for classifying glioma, meningioma, and the absence of a tumor using MRI. One of the main bottlenecks faced when using DL architectures for medical domain are data scarcity. This bottleneck was addressed using general data augmentation techniques such as flipping, rotation, and translation. Alsaif *et al.*[8] presented an improved ResNet50 architecture, which incorporated data augmentation techniques for effective brain tumor classification.

Because of the intrinsic locality of convolution operations, CNN-based approaches are generally inadequate for modeling straightforward long-range relations. Therefore, CNN-based architectures exhibit weak performances, particularly for target structures with varying textures, shapes, and sizes across patients. In previous studies, self-attention mechanisms with CNN features were used to overcome these limitations.[9]

Transformers, intended for sequence-to-sequence prediction, have emerged as ideal candidates to replace CNNs. These were first proposed for machine translation by Vaswani *et al.*[6] It was then established as the state-of-the-art method for many NLP tasks. It has the capacity to substitute attention mechanisms in place of convolution.[9-13]

Transformers also exhibit superior transferability for downstream tasks through extensive pretraining and superior performance in modeling global contexts. In many applications of machine translation and NLP, long short-term memory and artificial neural network have been successfully substituted by transformers.[10]

The results of transformers have matched or surpassed those of the state-of-the-art methods in various image recognition tasks.[12-16] The original design of transformers presented by Dosovitskiy *et al.*[11] has undergone several changes for suitability with CV tasks. For instance, Parmar *et al.*[12] modified transformers that used the self-attention mechanism only in local neighborhood of each query pixel. A novel transformer model, known as sparse transformers, was proposed by Child *et al.*[14] which attained global self-attention using scalable approximations. Wu *et al.*[15] introduced convolutions into ViTs to achieve best results on both convolutions and Transformers.

In general, large amounts of data and powerful computers are required for training ViTs, limiting their application in medical imaging diagnostics.[17-20] Hence, the research presented exploits the possibility of utilizing transformer-based attention features along with DL for the classification of brain tumors with a relatively small clinical dataset. We proposed mechanisms to tackle data scarcity and high processing power requirements while achieving sufficient model performance.

After image classification, MRI images with tumor underwent segmentation. Although segmentation generally provides detailed information about the spatial extent of tumors, classification offers insights into their nature. Therefore, segmentation was not researched and only image classification was focused on herein. However, as segmentation and classification work in tandem to provide a comprehensive understanding of disease diagnosis, existing studies can be referred to for more information on medical image segmentation.[21-25]

2.1. ViT model

ViTs, as presented by Dosovitskiy *et al.*,[11] mimic the original transformer model developed for NLP tasks using image patches as words for the input. ViTs can be used for image classifications primarily because they reduce architectural complexity and have enhanced exploring scalability and training efficiency. Recent studies have shown that the direct application of transformers with global self-attention to input images provided excellent results on ImageNet classification.[15] Moreover, ViTs can achieve high training accuracy with less computational time.[16] The success of transformers in medical image segmentation and classification was proven in the diagnosis of breast cancer using biopsy images and an end-to-end holistic attention network.[17] ViT-based medical image classification and segmentation continues to be a popular topic among researchers.

ViT contains stacks of encoder and decoder layers in its core, which will be hereinafter referred to as an encoder and a decoder, respectively. The encoder comprises two sublayers, namely multihead attention and feed-forward layers. The decoder comprises three sublayers, where the masked multihead attention layer is followed by the multihead attention layer and feed-forward layer. The encoder maps an input sequence $x = (x_1, x_2,\ldots, x_n)$ to a sequence $z = (z_1, z_2,\ldots, z_n)$. Based on z, the decoder generates an output sequence $y = (y_1, y_2,\ldots, y_n)$, with one element at a time. The model is auto-regressive at every step and uses already generated data as the additional input to create a new data instance. For more detail on the implementation of the ViT architecture, please refer to "An image is worth 16×16 words" by Dosovitskiy *et al.*[11]

There are two approaches to ViTs: hybrid and transformer-only architectures.[20] Hybrid architectures use a CNN to produce an embedding for an image or subregion of an image (patch). Encoding is used as the input for a subsequent transformer. In hybrid method, a CNN was used to process lower-level features in the input. In transformer-only architectures, a trainable part of the architecture projects patches to an embedding space and a hand-coded or convolutional architecture is not used. The transformer architecture learns only lower- and higher-level features.[16] Herein, transformer-only architecture is focused on and the model developed by Dosovitskiy *et al.*[11] was used for image classification.

Transformers have been used for tumor analysis in several studies. For instance, Asiri *et al.* used fine-tuned ViT model with the CE-MRI dataset containing only 5712 images for brain tumor classification.[25] The lack of diversity and limited number of images in the dataset affected the generalizability of ViT to real-world scenarios, suggesting further research to improve its accuracy and reliability, particularly for complex cases. Overall, the current ViT model used for brain tumor classification might not be fully optimized, and further research is required to enhance its diversity, reliability, and accuracy. This study focused on addressing this research gap in brain tumor classification using diverse BraTS datasets that primarily contain glioma MRIs. This dataset offered a benchmarked set of ground truth labels for glioma classification, addressing the limitations of existing studies. Moreover, potential model optimization techniques and MRI preprocessing techniques were discussed for their use in improving the model results.

3. Methodology

This section presents in detail, dataset preparation, including data preprocessing, ViT architecture, and model training with special attention to pretraining and fine-tuning approaches.

3.1. Dataset preparation and preprocessing

The BraTS 2015 dataset[7] containing 220 MRI images of high-grade gliomas (HGGs) and 54 images of low-grade gliomas (LGGs) was used for model training, validation, and testing. The dataset also contained MRI images of a patient in four modalities: T1 (spin-lattice relaxation), T1Gd (postcontrast T1-weighted), T2 (spin–spin relaxation), and T2-Flair (fluid attenuation inversion recovery). The analysis was restricted to the axial plane images of T1-MRIs, and the file format of the dataset was ".mha," which primarily is associated with the insight segmentation and registration toolkit. The DL architecture used the ".png" as the input image format. Hence, the T1-MRIs of a patient were converted to ".png" using "mha2png." Each patient's record resulted in 154 independent ".png" files, corresponding to brain slices in the coronal plane. Therefore, this resulted in a ".png" image dataset containing 42,196 images. Using the tumor mask of the BraTS 2015 dataset, each slice was first labeled based on the presence or absence of brain tumors. Then, slices with tumors were categorized into HGG or LGG tumors using the auxiliary data available in the BraTS 2015 dataset.

Intensity uniformization is another essential step in the preprocessing of MRI images. The pixel intensity of MRI images in BraTS ranges from −1000 to +1000, with more than 2000 levels. To aid image handling in limited resource environment, this pixel intensity range was decreased and scaled to match the intensity levels of 0–255, i.e., 8 bits/pixel grayscale. During preprocessing, the values above the upper gray level (G_u) and below the lower gray level (G_d) were assigned white and black, respectively. The center, also known as the window level (WL) and window width (WW), was changed based on the upper and lower gray levels. The upper gray level (G_u) was calculated as $G_u = WL + \left(\frac{WW}{2}\right)$, and the lower gray level (G_d) was calculated as $G_d = WL - \left(\frac{WW}{2}\right)$. Table 1 summarizes the effect of different values for WL, WW, and Range (G_u, G_d) on the preprocessed images. For instance, input images preprocessed considering WL = 0, WW = 400, and Range (−200 – 200) failed to show fine details of brain MRI images. After few trial and error iterations, range (−200 – 100), WW = 1200, and WL = 400 were chosen as the best parameters for 8 bit/pixel grayscale conversion.

Moreover, as the data on one patient record holds 154 (or 155) images, each image was considered a single input in the analysis and classified into one of the three classes: HGG, LGG, and nontumorous. The dataset was first developed using 120 patient records comprising 18,480 images, which were subgrouped into 3 subsets with 40 patients each. The dataset was further separated into two subsets, namely training and testing; approximately 70% of data were used for training and 30% for testing.

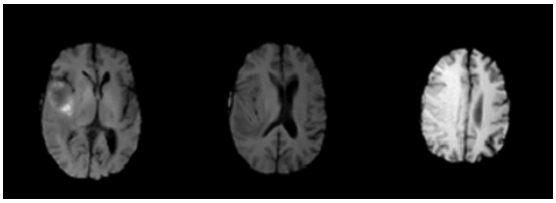
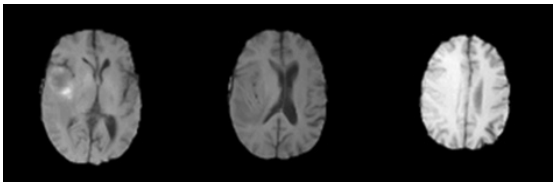
3.2. ViT architecture

ViTs are a group of neural network architectures that convert one sequence of images into another sequence. During preprocessing in ViTs, the input image is split into fixed-size patches and an input sequence is generated by linearly embedding each image into a sequence vector by adding position embedding information (Figure 1). The encoder transforms the input sequence into an embedding space, which is a vector representation of the image. Subsequently, the decoder receives the data in the embedding space and converts this into an output vector. An embedding layer generally proceeds each encoder or decoder to process their respective input, and an output layer is used toward the end of the architecture to generate the final output. ViTs perform classification using an extra learnable layer, i.e., classifier.[20] Figure 1 summarizes the process of image classification using the ViT for image recognition. Herein, a modified version of the model[26] was used for the classification of MRI images from the BraTS 2015 dataset. The classification operation flow of ViT is shown in Figure 2, and the performance of the proposed system was analyzed using accuracy, training, and validation loss and confusion matrix.

3.3. Model pretraining and fine-tuning

ViT is a DL model that requires considerably large dataset for model training. As BraTS is a relatively small dataset to train the ViT effectively, pretraining was performed to generate initial weights. CIFAR-10, a simple dataset, can serve as a foundation for pretraining models for medical image analysis.[27] The ViT was pretrained using the grayscale images of CIFAR-10, comprising 60,000 32 × 32 images belonging to 10 classes. All classes in CIFAR-10 are mutually exclusive, without any overlap between each class, which are well defined and bounded. For pretraining, the dataset was split into five training batches and one test batch, with each batch comprising 10,000 images. The test batch of CIFAR-10 was created using exactly 10,000 randomly selected images, and the training batches contained the remaining 50,000 images. Some training batches contained more images from one class than the other because the remaining images were added to the training batches in a random order.

Table 1. Comparison of preprocessed magnetic resonance imaging images with different intensity ranges, WLs, and WWs

Range	WW	WL	Image
from−1000 to−200	800	−600	
from−200 to 200	400	0	
from 200 to 1000	800	600	
from−200 to 1000	1200	400	

Abbreviation: WL: Window level; WW: Window width.

After the ViT model was successfully pretrained using CIFAR-10, transfer learning was used to initiate starting weights for the brain tumor classification task. The BraTS dataset with 15,000 images generated was split into training and testing datasets with a 70:30 ratio. Using the pretrained initial weights obtained using CIFAR-10, the model was warm started and its weights were fine tuned for brain tumor classification using BraTS dataset.

3.4. Statistical analysis

The analysis performed herein was simulated using Google Colab Jupyter notebook and Python 3.6 programming language. To evaluate the performance of the proposed ViT architecture, its training and validation accuracies and loss curves were analyzed. Thereafter, the model's performance was compared against a simple CNN network. Also, performance of the model was tested further using the accuracy, precision, and recall metrics. These metrics were calculated from the confusion matrix.[28]

4. Results

The performance of the ViT model in classifying glioma from MRI images was evaluated herein and compared with that of the conventional CNN. Its performance was evaluated for the task of handling two- and three-class problems under the class imbalance problem.

4.1. Training the ViT model

4.1.1. Pretraining the ViT model

In medical image analysis, collecting a considerably large dataset is a practically infeasible task. However, to achieve desirable performance with the ViT model, the DL architecture must be trained using a large dataset. To address this shortcoming, the customized ViT model was pretrained using a large general dataset, specifically CIFAR-10, and later fine-tuned with BraTS. Figure 3 shows the performance of the ViT model during pretraining using CIFAR-10, indicating that the model stabilized over time under 100 epochs.

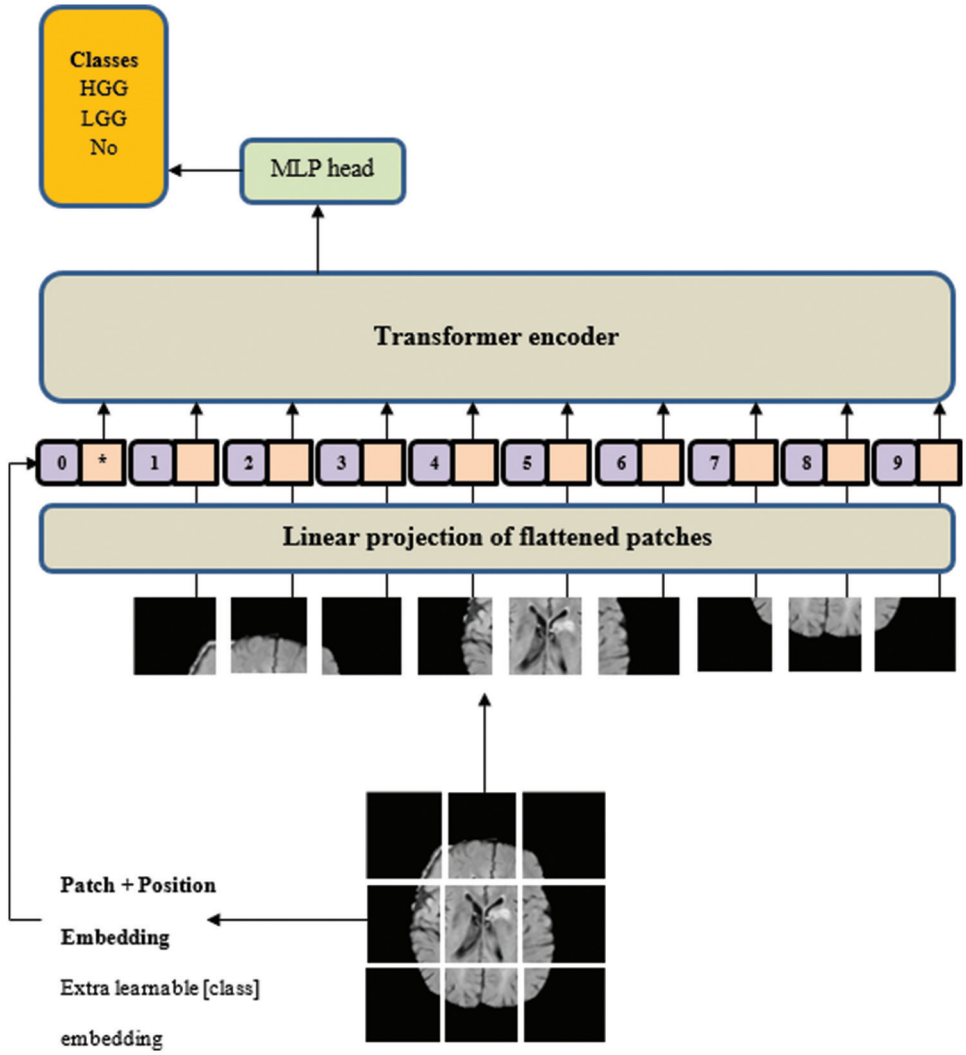


Figure 1. Basic block diagram of the vision transformer model

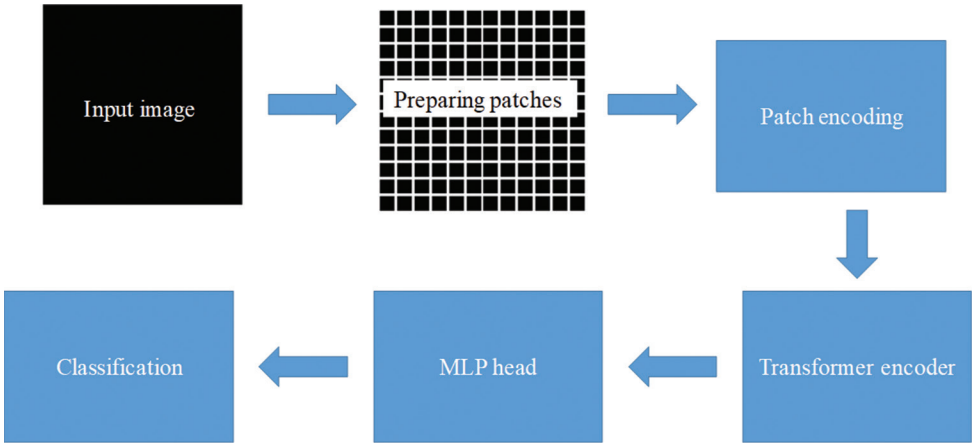


Figure 2. Operation flow of vision transformer-based classification

4.1.2. Fine-tuning the ViT model under different patch sizes

One of the distinct novelties associated with ViT model is the patch architecture. The pertained ViT model was fine tuned under different patch resolutions using the BraTs 2015 dataset. The objective of this approach was to find the most suitable patch size for a given application. The performance of each patch size was analyzed using model accuracy, loss performance, and confusion matrix. Figures 4-6 demonstrate the performance variation of ViT with patch sizes of 6 × 16, 8 × 8, and 4 × 4, respectively. In these figures, subplot (a) presents the training and validation loss, subplot (b) presents the training and validation accuracy while subplot (c) presents the confusion matrix, for the respective patch size. Table 2 summarizes the performance of ViT model under each patch size. As shown in Figures 4-6 and Table 2, the 4 × 4 patch resolution shows acceptable performance with 62.56% accuracy and lower level of fluctuation in the validation curves. The model could accurately detect the nontumorous MRI images, as shown in Figure 6C. However, the 4 × 4 patch resolution drastically increased the model tuning time.

4.2. Comparison of ViT model performance against CNN architecture

The traditional CNN architecture was used as the reference model for performance comparison of the ViT model

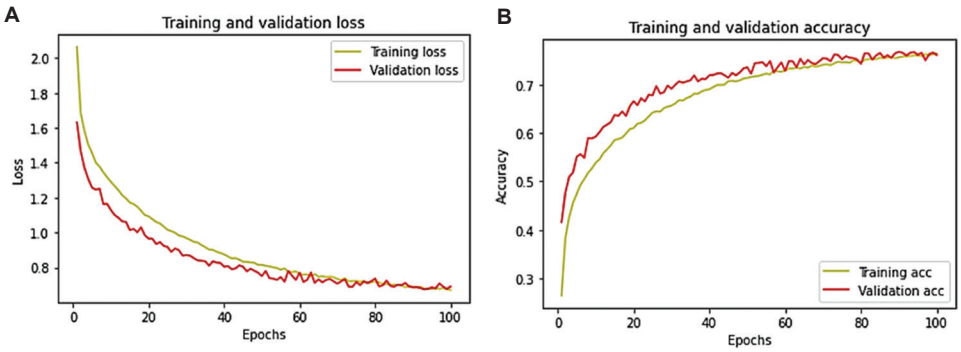


Figure 3. Vision transformer model performance during pretraining. (A) Training and validation losses. (B) Training and validation accuracy.

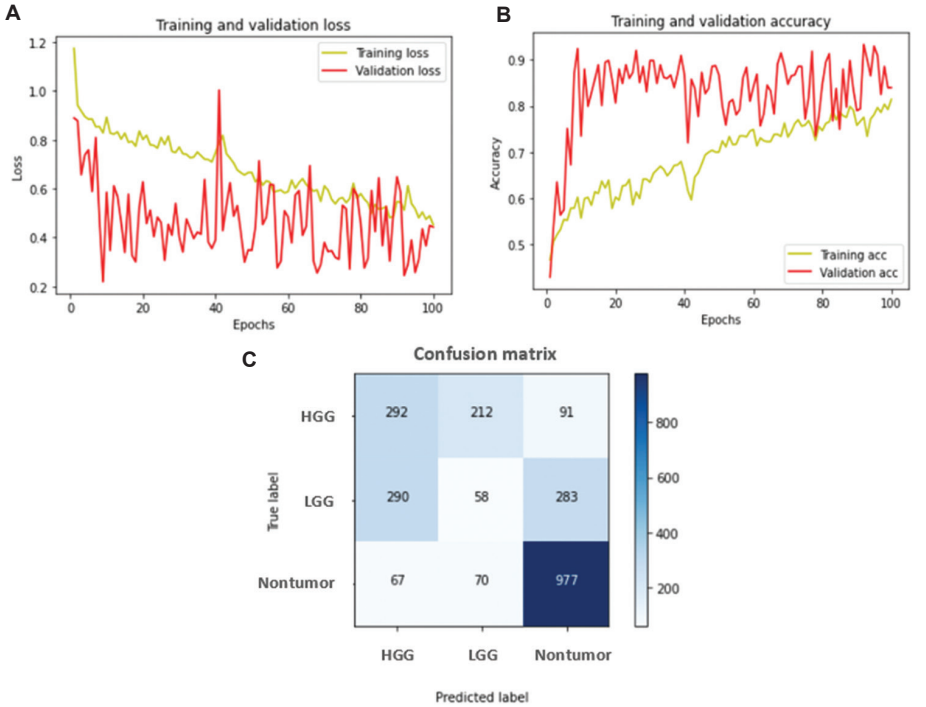


Figure 4. Performance of model fine-tuning using 16 × 16 patches. (A) Variation of model loss versus epoch. (B) Variation of model accuracy versus epoch. (C) Classification performance of the model presented using the confusion matrix.
Abbreviations: HGG: High-grade glioma; LGG: Low-grade glioma.

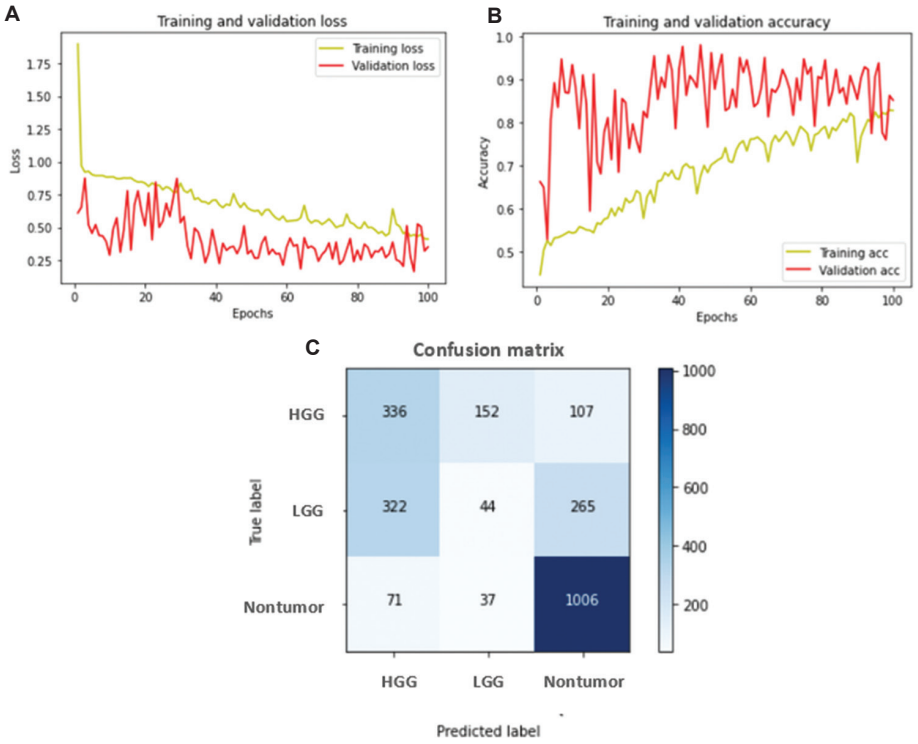


Figure 5. Performance of model fine-tuning using 8 × 8 patches. (A) Variation of model loss versus epoch. (B) Variation of model accuracy versus epoch. (C) Classification performance of the model presented using the confusion matrix.
Abbreviations: HGG: High-grade glioma; LGG: Low-grade glioma.

with 4 × 4 patch resolution. This CNN model has four convolution layers with only 8 million trainable parameters and was trained using the same dataset as ViT to classify the brain tumors.

Figure 7A shows the training and validation accuracy for the CNN model while Figure 7B shows that of the ViT model for brain tumor classification. Both models were trained using the same dataset under optimized settings. Although the CNN model showed nearly prefect accuracy with training, it underperformed during validation and indicated model overfitting. By contrast, the ViT model exhibited better performance in training and validation settings. As shown in Figure 7B, the ViT model exhibited a considerable level of instability. To stabilize the ViT model, it needs to be further trained using a large dataset. However, one of the critical factors in medical image analysis is the limitations in dataset; therefore, stabilizing the ViT model with small datasets is challenging.

4.3. Model performance under two-class problem

Furthermore, the accuracy of the ViT model with 4 × 4 patch size was analyzed for the task of classifying MRI images as with tumor or without tumor. According to the confusion matrix shown in Figure 8, the overall accuracy of classification for a three-class problem was 63.2% (Figure 8A), whereas that of two-class problem was 81.8% (Figure 8B), i.e., the trained and fine-tuned ViT model could detect the presence and absence of tumors with higher accuracy than classifying the different grades of tumors. The main reason behind this observation is the restriction in the number of images belonging to each class. For the three-class problem, the dataset showed a class imbalance, whereas it was balanced for the two-class problem. This observation indicated that the dataset used was suitable for tumor identification with two classes: with tumor and without tumor.

5. Discussion

CNN-based approaches are a popular choice for brain tumor classification using MRI images. They are highly effective in processing and analyzing medical data owing to their ability to automate feature extraction, capture hierarchical features, perform end-to-end learning, and yield high-accuracy output. However, transformers are emerging as leading contenders for this task, mainly because of their global context modeling features. In particular, their capacity to capture long-range dependencies and ability to focus on relevant parts of the input images are noteworthy.

CNN-based architectures perform weakly, particularly with datasets that show large variation in terms of texture,

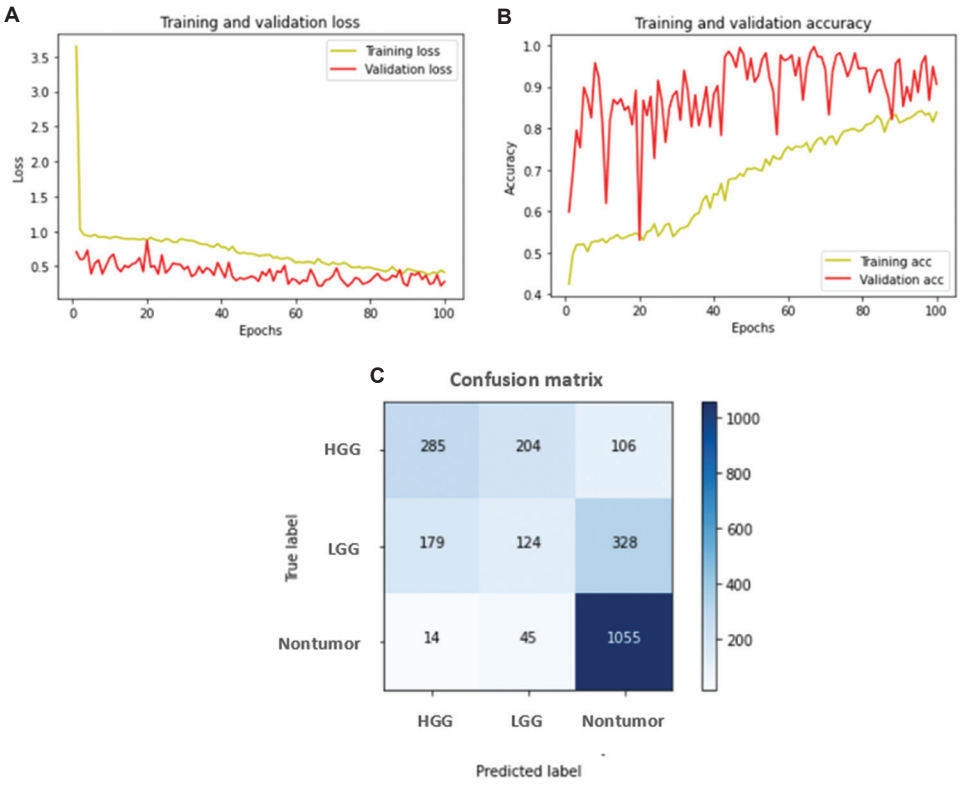


Figure 6. Performance of model fine-tuning using 4 × 4 patches. (A) Variation of model loss versus epoch. (B) Variation of model accuracy versus epoch. (C) Classification performance of the model presented using the confusion matrix.
Abbreviations: HGG: High-grade glioma; LGG: Low-grade glioma

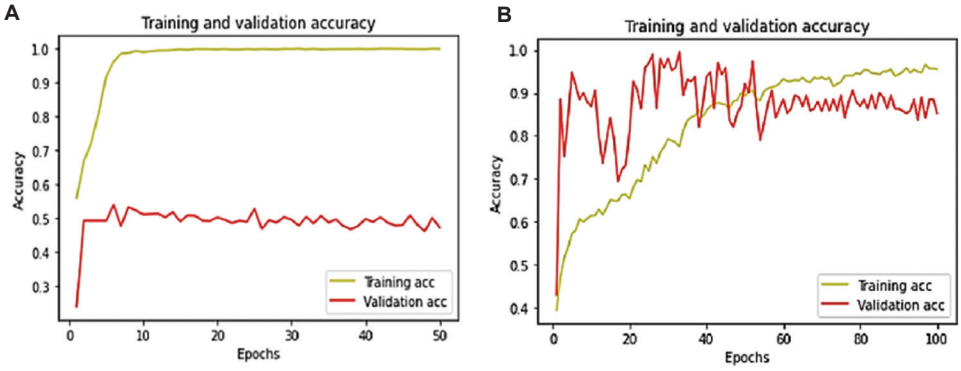


Figure 7. Model performance comparison between CNN and ViT models. (A) Performance accuracy versus epochs for CNN model. (B) Performance accuracy versus epochs for the ViT model.
Abbreviations: CNN: Convolutional neural network; ViT: Vision transformer.

shape, and size. The newly emerged transformer-based DL architectures, especially ViTs, show promising capacity to overcome these limitations. Although ViTs are a new concept for medical imaging, the accuracy of medical image classification can be improved using self-attention. For instance, the model can be trained to focus on abnormal cells in MRI by dynamically adjusting the weight assigned to these areas using attention mechanisms. This eventually improves the overall model performance. Moreover, the model can capture the relationship between tumors that are far apart owing to the inherent long-range dependency of ViTs. This introduces a provision for the model to learn dependencies between different slices of different planes of MRI images. Figure 7 shows the performance improvements achieved owing to these inherent characteristics of the ViT model in comparison with those of simple CNN.

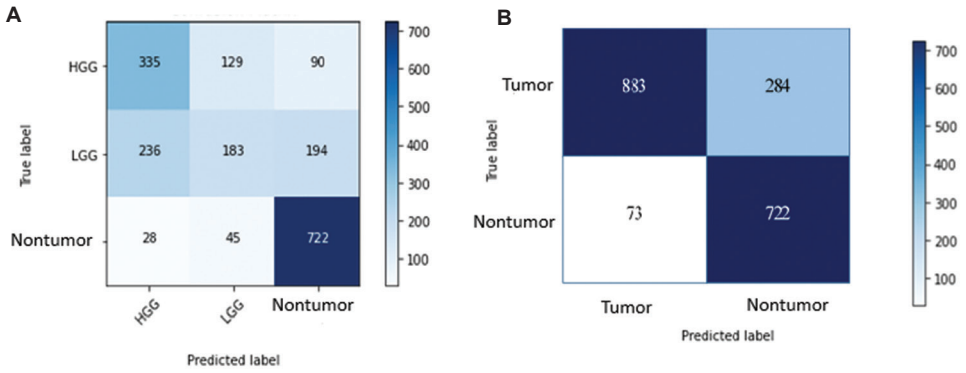


Figure 8. Comparison between three- and two-class classification problem. (A) Confusion matrix for a three-class problem with HGG, LGG, and nontumor. (B) Confusion matrix for a two-class problem with tumor (HGG and LGG tumors) and nontumor.
Abbreviations: HGG: High-grade glioma; LGG: Low-grade glioma.

Table 2. Comparison of the model performance for different patch sizes with learning rate=0.001 and weight decay=0.0001 and Adam optimizer

Patch size	Number of patches	Overall accuracy	Time taken to process (s)
16×16	4	56.70%	700
8×8	16	59.23%	1,900
4×4	64	62.56%	8,600
2×2	256	-	5,6100 (Estimated)

To combat the negative performance of ViTs owing to data scarcity, a pretraining approach coupled with transfer learning is presented herein. Moreover, the effects of the patch resolution on the overall performance accuracy and the loss curve behavior are discussed. With a 4 × 4 patch resolution, the stability of the model increased at the expense of inference time. Experimental results showed that the model performed better on the two-class problem of tumor and nontumor detection than on the three-class problem of HGG, LGG, and nontumor detection owing to class imbalance present in the BraTS 2015 dataset.

Moreover, the proposed model achieved an average classification accuracy of 81.8% for the BraTS 2015 dataset for the two-class problem. The confusion matrix in Figure 8 shows a model accuracy of 75.6% in detecting tumors and 90.8% in detecting nontumors. These results agreed well with previous studies using the BraTS 2015 data. For instance, the DL ensemble model that concatenates the weighted outputs of the cascaded anisotropic CNN (CA-CNN), DFKZ Net, and 3D U-Net achieved a classification accuracy of 46.4% during validation and 61% during testing with the BraTS 2018/2015 dataset.[29] The multiclass glioma tumor classification architecture presented in a previous study[30] achieved a 96.3% classification accuracy on a custom-built dataset that mainly used the BraTS 2015 dataset along with the other MRI images collected from different sources. The same custom-built dataset achieved a classification accuracy of 80.85% using 10 statistical features along with random forest[31] and 84.9% with dual-path residual CNNs.[32] The classification algorithm presented by Amin *et al.*[33] used discrete wavelet transform (DWT) to fuse MRI image sequences during preprocessing. The fused images followed the pipeline of denoising with a partial differential diffusion filter, segmentation using a global thresholding method, and classification of the segmented output into glioma, meningioma, and sarcoma using a CNN. This algorithm yielded a very high accuracy of nearly 100% in image fusion of all four MRI sequences, 89% in Flair + T1 fused images, and 78% in T1 images used herein. However, this algorithm first segmented tumor regions and then applied classification on the segmented region. Therefore, the results do not clearly present the detection accuracy on the initial dataset before segmentation.

Moreover, the BraTS datasets yielded better model performance. For instance, B. Maram and P. Rana achieved a quick and accurate image classification with a training accuracy of 98.485% using a U-Net architecture and BraTS 2020 dataset.[34] The novel linear-complexity data-efficient image transformer[35] achieved a classification accuracy of 97.86% with BraTS 2021 dataset. The ViT model discussed herein achieved a substantial level of classification accuracy using the BraTS 2015 dataset compared with those reported in the literature. However, if the input was preprocessed[33] or tested on an improved dataset such as BraTS 2021,[35] the performance accuracy of ViTs may increase compared with the current classification accuracy of 81.8%. Thus, the ViT model will be tested using the BraTS 2021 dataset and image preprocessing will be performed to facilitate better comparison and understanding on the performance of transformers for brain tumor classification.

6. Conclusion

Herein, the ViT architecture was studied for MRI image classification, focusing on glioma. To address the issues of data scarcity and class imbalance, ViT was pretrained using the CIFAR-10 dataset and fine-tuned using the BraTS 2015 dataset. The fine-tuned ViT could accurately and effectively identify glioma compared with the popular CNN architecture. Moreover, the effects of the patch resolution on the overall performance accuracy and the behavior of the loss curve were discussed. Overall, this study proposed a feasible and resource-optimized solution for the early detection and better prognosis of brain tumors. Further research is required to improve the predictions of the model while making the results understandable with explainable AI techniques for the advancement of automated systems for brain tumor detection and diagnosis.

Acknowledgments

None.

Funding

None.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: Maheshi B. Dissanayake
Formal analysis: All authors
Investigation: All authors
Methodology: All authors
Writing – original draft: All authors
Writing – review & editing: Maheshi B. Dissanayake

Ethics approval and consent to participate

The data collection was not part of this research. We use publicly available BRATS dataset. Ethical clearance had already been obtained before the upload of the medical dataset BRATS onto the public domain by Menze BH *et. al.*[7]

Consent for publication

Not applicable.

Availability of data

The data utilized in this research are publicly available. The authors have released the code on the GitHub page (https://github.com/SaneruW/Vision-transformers-for-glioma-classifications-using-T1-magnetic-resonance-images). Regarding materials or details related to the implementation, please contact Mr. Saneru Wickramasinghe (saneruw@gmail.com).

References

1. Pal A, Chaturvedi A, Garain U, Chandra A, Chatterjee R. *Severity Grading of Psoriatic Plaques using Deep CNN Based Multi-task Learning*. Mexico: ICPR; 2016.

   doi: 10.1109/ICPR.2016.7899846

2. Wang G. A perspective on deep imaging. *IEEE Access*. 2016;4:8914-8924.

   doi: 10.1109/ACCESS.2016.2624938

3. Ker J, Wang L, Rao J, Lim T. Deep learning applications in medical image analysis. *IEEE Access*. 2018;6:9375-9389.

   doi: 10.1109/ACCESS.2017.2788044

4. Kabir Anaraki A, Ayati M, Kazemi F. Magnetic resonance imaging-based brain tumor grades classification and grading via convolutional neural networks and genetic algorithms. *Biocybern. Biomed. Eng*. 2019;39(1):63-74.

   doi: 10.1016/j.bbe.2018.10.004

5. Kaldera HNTK, Gunasekara SR, Dissanayake MB. Brain Tumor Classification and Segmentation Using Faster R-CNN. In: *Proceedings ASET*. United States: IEEE; 2019.

   doi: 10.1109/ICASET.2019.8714263

6. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *In: NIPS'17*: *Proceedings of the 31st International Conference on Neural Information Processing Systems*; 2017:6000-6010.

7. Menze BH, Jakab A, Bauer S, *et al*. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans Med Imaging*. 2015;34(10):1993-2024.

   doi: 10.1109/tmi.2014.2377694

8. Alsaif H, Guesmi R, Alshammari BM, *et al*. A novel data augmentation-based brain tumor detection using convolutional neural network. *Appl Sci*. 2022;12(8):3773.

   doi: 10.3390/app12083773

9. Pan X, Ge C, Lu R, Song S, Chen G, Huang Z, *et al*. On the Integration of Self-Attention and Convolution. *In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2022:805-815.

   doi: 10.1109/cvpr52688.2022.00089

10. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Published online 2018.

    doi: 10.48550/ARXIV.1810.04805

11. Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Published online 2020.

doi: 10.48550/ARXIV.2010.11929

12. Parmar N, Vaswani A, Uszkoreit J, *et al*. Image Transformer. *In: JMLR Workshop and Conference Proceedings*; 2018:4055-4064.

13. Zheng S, Lu J, Zhao H, *et al*. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. Published online 2020.

doi: 10.48550/ARXIV.2012.15840

14. Child R, Gray S, Radford A, Sutskever I. Generating Long Sequences with Sparse Transformers. Published online 2019.

doi: 10.48550/ARXIV.1904.10509

15. Wu H, Xiao B, Codella N, *et al*. CvT: Introducing Convolutions to Vision Transformers. *In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE; 2021:22-31.

doi: 10.1109/iccv48922.2021.00009

16. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-End Object Detection with Transformers. *In: Computer Vision – ECCV 2020 (Lecture Notes in Computer Science)*. Springer International Publishing; 2020:213-229.

doi: 10.1007/978-3-030-58452-8_13

17. Aloraini M, Khan A, Aladhadh S, Habib S, Alsharekh MF, Islam M. Ombining the transformer and convolution for effective brain tumor classification using MRI Images. *Appl Sci*. 2023;13:3680.

doi: 10.3390/app13063680

18. Mehta S, Lu X, Weaver D, Elmore JG, Hajishirzi H, Shapiro L. HATNet: An End-to-End Holistic Attention Network for Diagnosis of Breast Biopsy Images. arXiv. Preprint posted online 2020.

doi: 10.48550/arXiv.2007.13007

19. Lan Y, Zou S, Qin B, Zhu X. Potential roles of transformers in brain tumor diagnosis and treatment. Brain-X. 2023;1(2):ae23.

doi: 10.1002/brx2.23

20. Courant R, Edberg M, Dufour N, Kalogeiton V. Transformers and visual transformers. In: Colliot O, editors. *Machine Learning for Brain Disorders. Neuromethods*. vol. 197. United States: Humana; 2023.

doi: 10.1007/978-1-0716-3195-9_6

21. Zunair H, Ben Hamza A. Sharp U-Net: Depthwise convolutional network for biomedical image segmentation. *Comput Biol Med*. 2021;136:104699.

doi: 10.1016/j.compbiomed.2021.104699

22. Dasanayaka C, Dharmasena B, Bandara WR, Dissanayake MB, Jayasinghe R. Segmentation of Mental Foramen in Dental Panoramic Tomography Using Deep Learning. In: *2019 IEEE 14th Conference on Industrial and Information Systems (ICIIS)*. IEEE; 2019:81-84.

doi: 10.1109/ICIIS47346.2019.9063312

23. Wang P, Yang Q, He Z, Yuan Y. Vision transformers in multi-modal brain tumor MRI segmentation: A review. *Meta Radiol*. 2023;1:100004.

doi: 10.1016/j.metrad.2023.100004

24. Marathe A, Kadam V, Chaumal A, Kodilkar S, Joshi A, Sawant S. Performance analysis of memory-efficient vision transformers in brain tumor segmentation. In: *Artificial Intelligence-Based Healthcare Systems*. Cham: Springer Nature Switzerland; 2023:125-133.

doi: 10.1007/978-3-031-41925-6_9

25. Asiri AA, Shaf A, Ali T, *et al.* Exploring the power of deep learning: Fine-tuned vision transformer for accurate and efficient brain tumor detection in MRI Scans. *Diagnostics*. 2023;13(12):2094.

doi: 10.3390/diagnostics13122094

26. Salama K. *Image Classification with Vision Transformer*; 2022. Available: https://keras.io/examples/vision/image_classification_with_vision_transformer [Last accessed on 2022 Oct 10].

27. Mabu S, Atsumo A, Kido S, Kuremoto T, Hirano Y. Investigating the effects of transfer learning on ROI-based classification of chest CT images: A case study on diffuse lung diseases. *J Signal Process Syst*. 2020;92:307-313.

doi: 10.1007/s11265-019-01499-w

28. Kanesamoorthy K, Dissanayake MB. Prediction of treatment failure of tuberculosis using support vector machine with genetic algorithm. *Int J Mycobacteriol*. 2021;10(3):279-284.

doi: 10.4103/ijmy.ijmy_130_21

29. Sun L, Zhang S, Chen H, Luo L. Brain tumor segmentation and survival prediction using multimodal MRI scans with deep learning. *Front Neurosci*. 2019;13:810.

doi: 10.3389/fnins.2019.00810

30. Latif G. DeepTumor: Framework for brain MR image classification, segmentation and tumor detection. *Diagnostics* (*Basel*). 2022;12(11):2888.

doi: 10.3390/diagnostics12112888

31. El-Melegy MT, El-Magd KMA. A Multiple Classifiers System for Automatic Multimodal Brain Tumor Segmentation. In: *Proceedings of the 2019 15th International Computer Engineering Conference (ICENCO), Giza, Egypt. 29-30 December 2019*. New York, NY, USA: IEEE; 2019.

doi: 10.1109/ICENCO48310.2019.9027389

32. Xue Y, Yang Y, Farhat FG, *et al*. Brain tumor classification with tumor segmentations and a dual path residual

convolutional neural network from MRI and pathology images. In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Germany: Springer; 2020. p. 360-367.

doi: 10.1007/978-3-030-46643-5_36

33. Amin J, Sharif M, Gul N, Yasmin M, Shad SA. Brain tumor classification based on DWT fusion of MRI sequences using convolutional neural network. *Pattern Recognit Lett*. 2020;129:115-122.

doi: 10.1016/j.patrec.2019.11.016

34. Maram B, Rana P. Brain Tumour Detection on BraTS 2020 using U-Net. In: *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, Noida, India; 2021. p. 1-5.

doi: 10.1109/ICRITO51393.2021.9596530

35. Ferdous GJ, Sathi KA, Hossain MA, Hoque MM, Dewan MAA. LCDEiT: A linear complexity data-efficient image transformer for MRI brain tumor classification. *IEEE Access*. 2023;11:20337-20350.

doi: 10.1109/ACCESS.2023.3244228