Düzce University
Journal of Science & Technology

Research Article

A Comparative Analysis of Vision Transformers and Transfer Learning for Brain Tumor Classification

Ahmet SOLAK [a,*]

[a] *Department of Electrical and Electronics Engineering, Faculty of Engineering and Natural Sciences, Konya Technical University, Konya, TURKEY*
** asolak@ktun.edu.tr*
DOI: 10.29130/dubited.1521340

ABSTRACT

Accurate brain tumor classification is crucial in neuro-oncology for guiding treatment plans and improving patient outcomes. Leveraging the potential of Vision Transformers (ViTs), this study investigates their efficacy in binary classification of brain tumors using magnetic resonance (MR) images, comparing them to CNN-based models such as VGG16, VGG19, and ResNet50. Comprehensive evaluation using accuracy, precision, recall, and F1-score reveals ViTs' superior performance, achieving 92.59% accuracy, surpassing VGG16 (85.19%), VGG19 (74.04%), and ResNet50 (88.89%). These findings highlight ViTs as a transformative tool for clinical adoption, enhancing diagnostic accuracy and patient care in neuro-oncology.

Keywords: *Brain tumor classification, Vision transformers, Deep learning, Transfer learning, Medical imaging*

Beyin Tümörü Sınıflandırması için Görü Dönüştürücü ve Transfer Öğrenmenin Karşılaştırmalı Analizi

ÖZ

Beyin tümörlerinin doğru sınıflandırılması, nöro-onkolojide tedavi planlarını yönlendirmek ve hasta sonuçlarını iyileştirmek için kritik öneme sahiptir. Bu çalışmada, Manyetik Rezonans (MR) görüntüleri kullanılarak Vision Transformers (ViTs) yönteminin beyin tümörlerinin ikili sınıflandırmasındaki etkinliği araştırılmış ve VGG16, VGG19 ve ResNet50 gibi CNN tabanlı modellerle karşılaştırılmıştır. Doğruluk, kesinlik, duyarlılık ve F1-skoru gibi kapsamlı değerlendirme metrikleri, ViTs'in üstün performansını ortaya koymuştur; ViTs, %92,59 doğrulukla VGG16 (%85,19), VGG19 (%74,04) ve ResNet50'yi (%88,89) geride bırakmıştır. Bu bulgular, ViTs'in nöro-onkolojide tanısal doğruluğu artıran ve hasta bakımını iyileştiren dönüştürücü bir araç olarak klinik uygulamalara entegrasyonu için umut vadeden bir yöntem olduğunu göstermektedir.

Anahtar Kelimeler: *Beyin tümörü sınıflandırması, Görü dönüştürücü, Derin öğrenme, Transfer öğrenme, Tıbbi görüntüleme*

Received: 24/07/2024, Revised: 28/11/2024, Accepted: 09/12/2024

I. INTRODUCTION

Brain tumors represent a significant challenge in contemporary medicine, encompassing a heterogeneous group of neoplasms that arise within the central nervous system. These neoplasms are classified as either primary, originating from intrinsic brain tissue, or secondary, resulting from metastatic spread of extracerebral malignancies [1]. The clinical impact of brain tumors is profound and multifaceted, often manifesting as neurological deficits, cognitive impairment, and potentially life-threatening sequelae. The complex nature of these tumors, coupled with their location in critical neuroanatomical structures, presents unique diagnostic and therapeutic challenges that necessitate a multidisciplinary approach to patient care [2].

Timely and accurate diagnosis of brain tumors is of paramount importance in the fields of neurology and oncology, as it significantly influences treatment strategies, prognostic outcomes, and overall patient management. Advanced neuroimaging modalities, including high-resolution magnetic resonance imaging (MRI) with various sequences and multimodal computed tomography (CT), play a pivotal role in the initial detection and characterization of intracranial lesions. However, definitive diagnosis frequently requires histopathological examination of tumor tissue obtained through stereotactic biopsy or surgical resection. This diagnostic process is further augmented by molecular profiling and genetic analysis, which provide crucial information for tumor classification, prognostication, and the development of targeted therapeutic approaches. The integration of these diagnostic modalities not only guides clinical decision-making but also facilitates personalized treatment paradigms, potentially improving survival rates and preserving neurological function in patients with brain tumors.

In recent years, the application of artificial intelligence (AI) to brain tumor classification has emerged as a promising frontier in neuro-oncology, offering potential improvements in diagnostic accuracy, efficiency, and prognostic assessment. Machine learning algorithms, particularly deep learning neural networks, have demonstrated remarkable capabilities in analyzing complex medical imaging data, including MRI and CT scans [3-5]. These AI-driven systems can be trained on large datasets of annotated images to recognize subtle patterns and features that may elude human observers, potentially enhancing the detection and classification of brain tumors.

Convolutional Neural Networks (CNNs) have been widely adopted for brain tumor classification due to their capacity to extract hierarchical features from medical imaging data [6-8]. CNNs excel at capturing local spatial patterns through convolutional layers, making them particularly suitable for analyzing the intricate structural characteristics of brain tumors in MRI and CT scans. These networks have demonstrated high accuracy in tasks such as tumor detection, segmentation, and grading. However, CNNs have certain limitations in this context. They may struggle with capturing long-range dependencies within images, which can be crucial for understanding the global context of tumor appearance and spread. Additionally, CNNs typically require large, annotated datasets for training, which can be challenging to obtain in medical imaging due to privacy concerns and the scarcity of expert-labeled data. The fixed receptive field of CNNs may also limit their ability to adapt to variations in tumor size and shape across different patients.

Vision Transformers (ViTs) have emerged as a promising alternative to CNNs for brain tumor classification, offering several potential advantages [9]. Unlike CNNs, ViTs process images as sequences of patches, employing self-attention mechanisms to capture global relationships within the image. This approach allows ViTs to model long-range dependencies more effectively, potentially improving the understanding of tumor context and spatial relationships. ViTs have shown impressive performance on various computer vision tasks, often surpassing CNNs, and this success is beginning to translate to medical imaging applications. The self-attention mechanism in ViTs provides a more flexible approach to feature extraction, potentially adapting better to the heterogeneous nature of brain tumors. Furthermore, ViTs may require less extensive data augmentation compared to CNNs, which could be beneficial given the limited availability of medical imaging datasets. However, it is important to note that ViTs are computationally intensive and may require larger datasets for optimal performance.

The field of applying ViTs to brain tumor classification is still evolving, and further research is needed to fully elucidate their advantages and potential limitations in clinical settings.

This study presents a comprehensive investigation into the classification of brain tumors using MR imaging, employing both CNN-based transfer learning models and ViT. The research focuses on distinguishing between MR images with and without brain tumors, aiming to evaluate and compare the performance of these advanced deep learning architectures on the given dataset. The methodology encompasses the application of pre-trained CNN models, leveraging transfer learning techniques to adapt these networks to the specific task of brain tumor classification. Concurrently, the study implements ViT, a more recent approach in computer vision, to assess its efficacy in this medical imaging context. The performance of these models is rigorously examined, considering metrics such as accuracy, precision, recall and F1-score. This comparative analysis seeks to elucidate the strengths and limitations of each approach, potentially informing future directions in the application of artificial intelligence to neuroradiological diagnosis.

II. RELATED WORKS

Artificial intelligence-based applications for brain tumor classification and segmentation have proliferated in recent years. This study focuses specifically on binary brain tumor classification, limiting the scope of relevant literature review to studies addressing this aspect. Charfi et al. proposed a hybrid artificial intelligence system for the automatic detection of brain tumors using a dataset comprising 80 MR images, categorized into tumorous and non-tumorous samples. Their approach integrates computational methods including histogram-based thresholding, discrete wavelet transforms, principal component analysis, and backpropagation neural networks, achieving an accuracy rate of 90% [10]. Nazir et al. employed an artificial neural network for classifying tumors as malignant or benign. Their methodology encompasses three stages: preprocessing, feature extraction, and classification. The preprocessing stage utilizes filters for noise reduction, while the feature extraction phase extracts color moments from MR images, which are subsequently fed into a simple artificial neural network. The study, conducted on a dataset of 25 normal and 45 abnormal images, achieved an overall accuracy of 91.8% [11]. Vani et al. focused on the Support Vector Machine (SVM) algorithm for image classification. They developed a Simulink model for tumor classification, presenting a prototype for SVM-based object detection that classifies images and evaluates whether the classified image is cancerous. Their approach yielded an accuracy of 82% [12]. Gupta et al. evaluated the potential of machine learning in the accurate and rapid diagnosis of cerebral tumors using MRI. In a study involving 200 subjects, images were acquired using volumetric Fluid Attenuated Inversion Recovery (FLAIR) acquisition and normalized to 12 useful sections for classification. Discrete Wavelet Transform (DWT) was utilized for feature extraction, and Principal Component Analysis (PCA) for feature selection. Various classifiers (SVM, k-NN, CART, and Random Forest) were tested, with linear SVM yielding the best results: 84% sensitivity, 92% specificity, and 88% accuracy [13]. Asif et al. conducted a classification study using popular deep learning models including Xception[14], NasNet Large [15], DenseNet121 [16], and InceptionResNetV2 [17] on two different brain MR datasets, one binary and one multiclass. The images in the dataset were cropped, preprocessed, and augmented. Transfer learning models were trained and tested using ADAM [18], SGD [19], and RMSprop optimization algorithms. Performance was evaluated based on criteria such as accuracy, sensitivity, precision, specificity, and F1-score. Experimental results indicated that the proposed CNN model, the Xception architecture, outperformed other models. On the binary dataset, the Xception model demonstrated an accuracy of 91.94%, precision of 87.50%, recall of 96.55%, and an F1 score of 91.8% [20]. Shilaskar et al. propose a system for brain tumor detection and classification using machine learning models on MRI data. Their approach employs Histogram of Oriented Gradients (HOG) for feature extraction and evaluates various classifiers, including SVM, Logistic Regression, Gradient Boost, K-Nearest Neighbors (KNN), and XGBoost. Among these, the XGBoost classifier achieves the highest accuracy of 92.02%, outperforming other models such as Logistic Regression (77.62%) and SVM (74.19%). This highlights XGBoost's potential for precise medical diagnostics, although limitations in dataset size are noted. The

authors emphasize future research on larger datasets and deep learning techniques for better generalization and accuracy [21]. Pilaoon et al. use pretrained deep convolutional neural networks, specifically GoogLeNet and AlexNet, to classify Glioblastoma Multiforme (GBM) brain tumors using MRI images from the REMBRANDT database. Their transfer learning approach eliminates manual segmentation and feature extraction, reducing human error. AlexNet achieves a higher accuracy of 93.62%, while GoogLeNet achieves 80.85%. The study highlights the potential of automated deep learning methods to support medical staff in diagnosing brain tumors and improving patient survival rates [22]. Dhaniya et al. propose a hybrid CNN-LSTM model for classifying brain tumors from MRI images. Their approach involves preprocessing with a Wiener filter, data augmentation (including rotation, cropping, and CLAHE), and segmentation using the Adaptive Particle Swarm Optimization (APSO) algorithm. The CNN extracts features, and LSTM handles the classification. The model achieves an accuracy of 92.03%, outperforming standalone methods such as SVM (87.06%) and LSTM (88.5%), demonstrating its effectiveness for automated tumor detection [23]. These studies collectively demonstrate the evolving landscape of AI-based approaches in brain tumor classification, highlighting the potential for improved diagnostic accuracy and efficiency in clinical settings.

III. MATERIALS AND METHODS

A. DATASET

In this study, a publicly available dataset comprising 253 MRI images, 155 of which were tumorous and 98 of which were non-tumorous, was examined [24]. Figure 1 illustrates representative examples of the images included in the dataset. Given the variability in image dimensions across the dataset, a standardization process was employed to ensure consistency. This involved resizing the images to a fixed size of 128x128 pixels. Subsequently, data augmentation was conducted due to the limited number of images in the dataset. In this study, a variety of data augmentation techniques, including random horizontal flipping, small rotations (±2%), and random zooming (up to 20%), were employed. These augmentations were applied dynamically during the training process, meaning that the dataset size remained 253 images (155 tumorous and 98 non-tumorous). However, the model was exposed to different variations of each image across epochs, effectively enhancing its ability to generalize to unseen data. After, the data were employed for the purpose of training the classification network. Figure 2 illustrates the framework of the study.

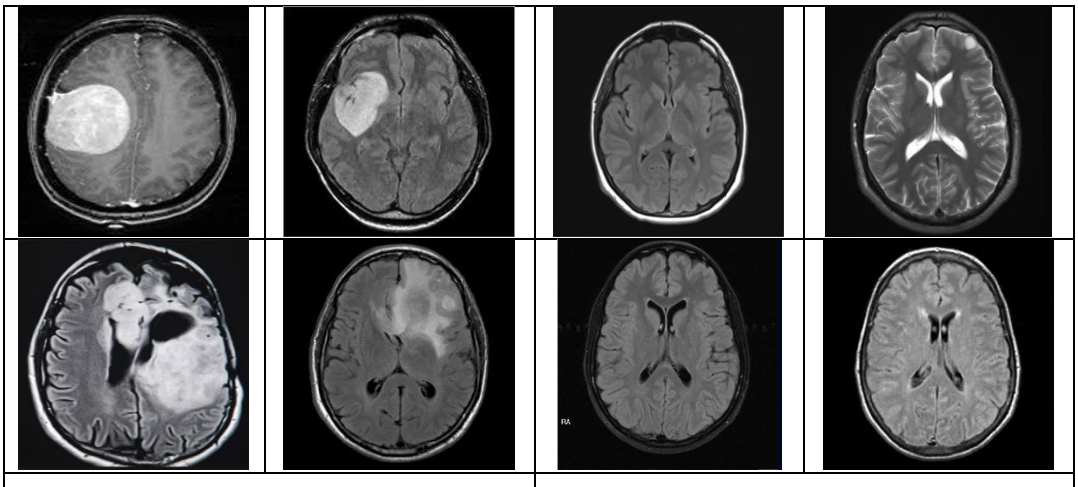


Figure 1. *Sample images from the data set.* ***(a)*** *MR images with tumor and* ***(b)*** *MR images without tumor*

Figure 2. *Flowchart of the study*

B. MODELS

B. 1. Transfer Learning Models

Transfer learning is a machine learning technique that employs the knowledge acquired from solving one problem to address a distinct but analogous problem. In the context of deep learning and computer vision, transfer learning typically involves the utilization of a neural network that has undergone pre-training on a substantial dataset (e.g., ImageNet [25]) as a foundation for a novel task with a comparatively limited dataset. This approach is particularly advantageous when the target task has limited labeled data, as it enables the model to leverage general features learned from the source domain. Transfer learning can be implemented through a variety of strategies, including feature extraction, in which the pre-trained network is utilized as a fixed feature extractor, and fine-tuning, in which some or all layers of the pre-trained network are further trained on the target dataset.

The VGG-16 architecture, introduced by Simonyan and Zisserman [26], is a convolutional neural network that is notable for its simplicity and depth. The network comprises 16 weight layers, including 13 convolutional layers and 3 fully connected layers. The convolutional layers utilize small 1-step 3x3 filters, interspersed with a 2x2 maximum pooling layer. The network maintains a uniform structure throughout, with the number of filters doubling after each maximum pooling layer, beginning with 64 and increasing to 512. The final layers comprise two fully connected layers, each comprising 4096 units, followed by a softmax output layer. The depth and uniform architecture of VGG-16 contributed to its success in image classification tasks, thereby establishing it as a popular choice for transfer learning in various computer vision applications.

VGG-19 represents an extension of the VGG-16 model, incorporating three additional convolutional layers, resulting in a total of 19 weight layers (16 convolutional and 3 fully connected) [26]. The overall structure is analogous to that of VGG-16, with additional depth provided by the incorporation of 3x3 convolutional layers in the subsequent stages of the network. Similarly to its predecessor, VGG-19 employs the use of small 3x3 filters throughout the network, in conjunction with a maximum of 2x2 pooling for the purpose of spatial minimization. The augmented depth of VGG-19 enables the acquisition of more intricate features, which may ultimately result in enhanced performance on specific tasks. However, the additional layers also result in increased computational requirements and a greater number of parameters, which may render the model more susceptible to overfitting in the context of smaller datasets.

ResNet-50, a component of the Residual Network family as outlined by He et al. [27], addresses the issue of degradation in exceedingly deep neural networks through the introduction of residual connections, representing a novel approach to network design. The network comprises 50 layers, including convolutional layers, batch normalization, rectified linear unit (ReLU) activations, and jump connections. The network is divided into stages, with each stage comprising multiple residual blocks.

These blocks facilitate the activation of shortcut connections that circumvent one or more layers, thereby enabling the network to learn residual functions with reference to layer inputs. This architectural design mitigates the vanishing gradient problem, thereby enabling the training of networks with considerably greater depth. ResNet-50 commences with a 7x7 convolutional layer, followed by a maximum pooling layer, and then progresses through four stages of residual blocks, concluding with a global average pooling layer and a fully connected layer. The incorporation of residual connections enables ResNet-50 to attain state-of-the-art performance on a diverse array of image classification tasks while maintaining a relatively modest number of parameters.

These architectures (VGG-16, VGG-19 and ResNet-50) have proven to be effective for transfer learning on many computer vision tasks, including medical image analysis. Their pre-trained weights on large-scale datasets such as ImageNet provide a valuable starting point for fine-tuning on specific tasks such as brain tumor classification, often leading to better performance and faster convergence compared to training from scratch.

B. 2. Vision Transformers

Advancements in artificial intelligence and deep learning have significantly enhanced researchers' capabilities in medical image classification. The literature abounds with studies leveraging CNNs, widely regarded as the foundation of deep learning, yielding remarkably successful outcomes. While CNNs excel in extracting local features such as color, texture, edges, and corners, they may exhibit limitations in capturing global features, including object shapes, sizes, and relationships. This distinction is particularly crucial in the analysis of medical images, where global context can be as important as local details. Consequently, this study endeavors to compare the performance of CNN-based deep learning networks with the more recently developed Vision Transformer (ViT) model [9], which offers potential advantages in capturing both local and global image characteristics.

The ViT represents a novel approach to image classification, drawing inspiration from transformer architectures that have demonstrated exceptional performance in natural language processing tasks. The general structure of the ViT model is illustrated in Figure 3. The ViT architecture comprises several key components:

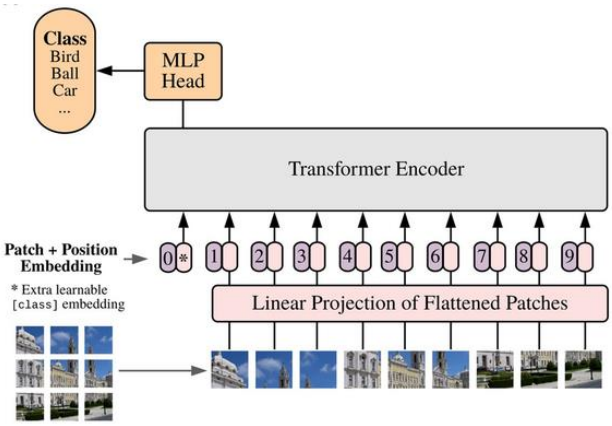


Figure 3. *ViT Architecture [9]*

B.2.1. Patch Embedding

The input image is segmented into fixed-size, non-overlapping patches. For an image of dimensions *HxW (Height x Width)*, divided into patches of size *PxP*, *N* patches are generated, where $N = \frac{HW}{P^2}$. Figure 4 illustrates the patch extraction process from a sample image in the dataset. Subsequently, each patch is flattened into a one-dimensional vector, as demonstrated in Figure 5. These flattened patches are then projected onto a higher-dimensional space through trainable linear transformations.

B.2.2. Position Embedding

To preserve the spatial information of the patches, positional embeddings are added to the patch embeddings. This process enables the model to maintain awareness of the original spatial arrangement of image fragments. Additionally, a specialized class token is prepended to the sequence of patch embeddings to facilitate image classification.

B.2.3. Transformer Encoder

This component forms the core of the ViT architecture and consists of two primary layers:

a) Multi-Head Self-Attention: This mechanism allows the model to weigh the importance of each patch relative to all others, enabling the capture of long-range dependencies and global context within the image.

b) Feed-Forward Neural Network (FFN) or Multi-Layer Perceptron (MLP): Each sub-layer is followed by Layer Normalization (LN) and residual connections. The FFN is applied uniformly to each position, processing the output of the self-attention mechanism.

Layer Normalization is applied before and after both the multi-head self-attention and FFN layers to stabilize training. Residual connections are incorporated around each sub-layer to facilitate gradient flow during the training process.

B.2.4. MLP Head

The final classification layer comprises one or more fully connected layers followed by a softmax activation. This component processes the output from the transformer encoder to determine the image's class membership.

This architecture enables the ViT model to potentially capture both fine-grained local features and broader global contexts, making it a promising candidate for medical image analysis tasks such as brain tumor classification.

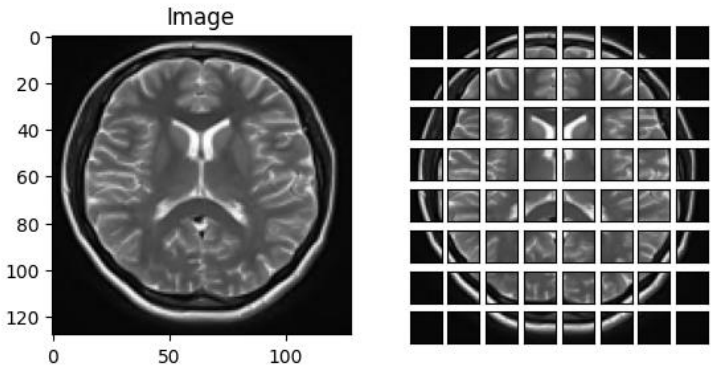


Figure 4. *Sample of Image Patches*

Figure 5. *Flattened Patches*

The selection of hyperparameters for the ViT model was a meticulous process designed to achieve a balance between performance and computational efficiency for brain tumor classification, particularly when dealing with potentially larger images. To accommodate the increased image size of 128x128 (*Height (H)* × *Width (W)*) pixels, the batch size was prudently reduced to 32. Furthermore, the patch size was correspondingly increased to 16x16 (*HxW*) pixels, resulting in 64 patches per image. To ensure training stability with these larger images, the learning rate was modestly adjusted to 0.0003, and weight

decay was set to 0.00003. In an effort to capture potentially more intricate patterns within the data, the number of training epochs was extended to 100.

Beyond hyperparameter selection, the architectural parameters were also fine-tuned to enhance the model's capacity for brain tumor classification. The number of Transformer encoder layers was strategically increased to 12, facilitating a deeper exploration of image features. The projection dimension was set to 128 to create a higher-dimensional representation of the image data. To enable a greater number of parallel attention mechanisms within the model, the number of attention heads was augmented to eight. The Transformer hidden units were meticulously adjusted to correspond with the new projection size, while the MLP head units were maintained for optimal performance. Finally, a dropout rate of 0.1 was implemented in both the attention and MLP layers to mitigate overfitting during training. In conclusion, these deliberate hyperparameter and architectural choices aimed to achieve a well-balanced ViT model, capable of efficient and accurate brain tumor classification.

C. EVALUATION METRICS

The assessment of a machine learning model's performance holds significant importance in understanding its effectiveness and facilitating comparisons across different methodologies. This section provides a comprehensive overview of the metrics utilized to evaluate model performance specifically within the domain of brain tumor classification.
In the realm of image classification, "accuracy" refers to the percentage of correctly classified images by a given model. It is computed as the ratio of correctly classified images to the total number of images, as defined in 1. Here, TP denotes true positives, TN represents true negatives, FP stands for false positives, and FN indicates false negatives. While achieving high accuracy is typically desirable, its interpretation can be misleading in scenarios where dataset classes are imbalanced, with one class (e.g., healthy images) significantly outnumbering others (e.g., tumor images).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Precision, a critical metric in classification tasks, quantifies the accuracy of positive predictions made by the model. It measures the proportion of true positives (correctly identified tumors) among all instances predicted as positive. Mathematically, precision is expressed as 2. A high precision value signifies that the model effectively identifies tumors while minimizing false positives. This metric is essential in evaluating the reliability of the model's positive predictions in clinical contexts and other applications where accurate identification is paramount.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

The recall metric quantifies the proportion of correctly identified true positive instances (tumors) out of all actual positive cases in the dataset. It is calculated using 3. A high recall value signifies that the model effectively detects a significant portion of the true tumor instances present in the dataset, indicating its ability to minimize false negatives and capture most actual positives.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

The F1 score represents a composite metric that integrates precision and recall, offering a balanced assessment of the model's effectiveness. It is computed using the harmonic mean of precision and recall, as defined by 4.

$$F1 - Score = \frac{2*Precision*Recall}{Precision+Recall} \quad (4)$$

The confusion matrix provides a detailed and structured overview of the model's predictive performance by displaying the counts of TP, TN, FP, and FN across each class. This matrix serves as a visual aid that effectively illustrates how well the model distinguishes between classes and where potential errors arise. In a binary classification scenario, the structure of the confusion matrix is delineated as Table 1.

Table 1. *Confusion Matrix*

Actual/Predicted	Positive	Negative
Positive	TP	FN
Negative	FP	TN

IV. EXPERIMENTATION AND RESULTS

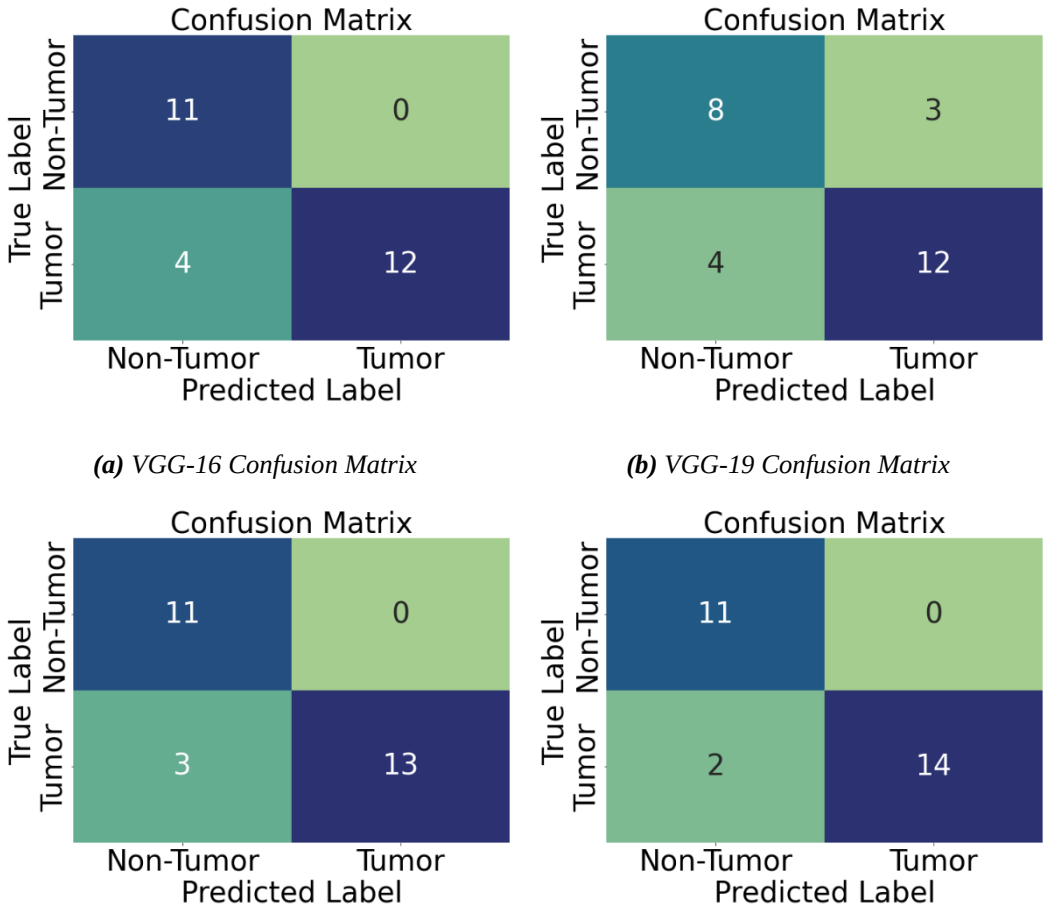
The training environment for this study was configured using Google Colab, which provides access to the Tesla T4 GPU and utilizes TensorFlow 2.15 as the primary deep learning framework. The Tesla T4 GPU, comprising 320 Turing Tensor Cores and 16 GB of GDDR6 VRAM, provided substantial computational capability, facilitating effective model training and inference. TensorFlow 2.15's extensive capabilities in deep learning enabled the implementation of complex neural network architectures and optimization algorithms that were necessary for achieving the study's objectives. This configuration not only guaranteed robust performance but also facilitated reproducibility and collaboration by Colab's cloud-based infrastructure.

The study employed a data partitioning strategy of 80:10:10 for training, validation, and testing sets, respectively, across all models. To enhance model generalization, data augmentation techniques were applied to the training set. Each model underwent independent training sessions, with subsequent performance evaluations conducted on the designated test data.

The training protocol adhered to the following parameters: input images were standardized to 128x128x3 dimensions, with a batch size of 32. The AdamW optimizer [28] was employed, utilizing a modestly adjusted learning rate of 0.0003 and a weight decay of 0.00003. Training proceeded for 100 epochs, incorporating a dropout rate of 0.1 to mitigate overfitting. Binary cross-entropy served as the loss function for all models.

The training processes for all models were executed sequentially. Figure 7 presents confusion matrices illustrating the classification performance of each model on the test set. In Figure 7 (a), the VGG-16 model erroneously classified 4 out of 27 tumor-positive images as tumor-free. Figure 7 (b) depicts the VGG-19 model's performance, which misclassified a total of 7 images: 4 tumor-positive images were incorrectly labeled as tumor-free, while 3 tumor-free images were mistakenly classified as tumor-positive. Figure 7 (c) shows the ResNet50 model's results, where 3 tumor-positive images were misclassified as tumor-free. Finally, Figure 7 (d) illustrates the ViT model's performance, which misclassified only 2 tumor-positive images as tumor-free.

Comparative analysis of model performance reveals that the ViT model achieved superior results, while the VGG-19 model demonstrated the least accurate performance. It is noteworthy that in the context of brain tumor classification, where even a single misclassification can have profound implications for patient care, the ViT model's superior performance is particularly significant. This underscores the potential of transformer-based architectures in medical image analysis, especially in applications where accuracy is paramount for patient outcomes.



(a) *VGG-16 Confusion Matrix*

(b) *VGG-19 Confusion Matrix*

(c) *ResNet50 Confusion Matrix*

(d) *ViT Confusion Matrix*

Figure 7. *Confusion matrices for all models*

Comparing the performance of different models using evaluation metrics is essential for enhancing the reliability and robustness of the study. Table 2 presents the accuracy, precision, recall, and F1-score values for each trained model on the test images. According to the table, the VGG-16 model achieved 85.19% accuracy, 86.67% precision, 87.50% recall, and an 85.16% F1-score. Conversely, the VGG-19 model attained 74.04% accuracy, 73.33% precision, 73.86% recall, and a 73.49% F1-score. The ResNet50 model demonstrated superior performance with 88.89% accuracy, 89.29% precision, 90.62% recall, and an 88.83% F1-score. Lastly, the ViT model outperformed the others, achieving 92.59% accuracy, 92.30% precision, 93.75% recall, and a 93.01% F1-score. Upon examining the evaluation metrics and performances, it is evident that the ViT model delivers the most successful results.

Table 2. *Test Results on Evaluation Metrics*

Models	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Training Time (mins)**	**Test Time (secs)**
VGG16	% 85.19	% 86.67	% 87.50	% 85.16	52.52	1.09
VGG19	% 74.04	% 73.33	% 73.86	% 73.49	51.51	0.91
ResNet50	% 88.89	% 89.29	% 90.62	% 88.83	20.89	0.38
ViT	% 92.59	% 92.30	% 93.75	% 93.01	16.49	0.52

In the evaluations conducted on the test images, the ViT model demonstrated the most successful results. To illustrate the model's performance on the test images, Figure 8 presents examples of the images and the corresponding predictions made by the model. As noted from the confusion matrix, only two images

were misclassified by the ViT model. Figure 8 includes an example of one such misclassification, providing further insight into the model's predictive behavior.

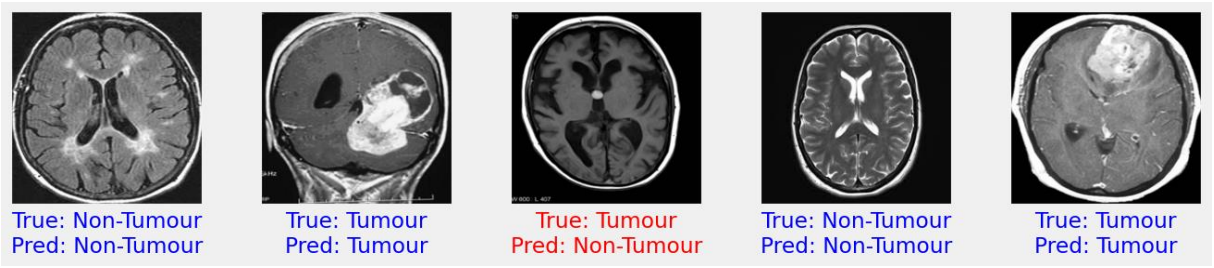


Figure 8. *Sample Test Images and Predictions of the ViT Model*

V. DISCUSSION

This study investigated the effectiveness of the ViT model for brain tumor classification using MR images. The results suggest that ViT may outperform traditional CNN-based transfer learning approaches, including VGG-16, VGG-19, and ResNet50. ViT achieved superior accuracy and performance across multiple image evaluation metrics, such as precision, recall, and F1-score.

The superior performance of the ViT model can potentially be attributed to its ability to capture long-range dependencies within the image data. This capability is particularly advantageous in the domain of medical imaging, where spatial relationships between image features hold significant importance for accurate diagnosis. Unlike CNNs that primarily rely on local feature extraction through convolutional filters, the ViT model leverages an attention mechanism. This mechanism allows the model to selectively focus on the most relevant regions of the image, potentially leading to enhanced classification accuracy in tasks like brain tumor classification.

Table 3. *Comparison with previous studies*

	Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
[10]	ANN	% 90	----------	% 82	----------
[11]	Normal densities based linear classifier	%73.03	----------	----------	----------
	Naive Bayes Classifier	% 76.92	----------	----------	----------
	SVM	% 88.46	----------	----------	----------
	ANN	% 91.8	----------	----------	----------
[12]	SVM	% 82	----------	----------	----------
[13]	Linear SVM	% 88	----------	% 84	----------
	k-NN	% 82	----------	% 80	----------
	Random Forest	% 86	----------	% 88	----------
	CART	% 76	----------	% 64	----------
[20]	Xception	% 91.94	% 87.50	% 96.55	% 91.80
	NasNet Large	% 91.74	% 90.00	% 93.10	% 91.53
	DenseNet121	% 90.32	% 89.66	% 89.66	% 89.86
	InceptionResNetV2	% 90.32	% 96.00	% 82.76	% 88.89
[21]	Gradient Boost	% 67.53	% 70.64	% 66.97	% 67.51
	KNN	% 68.52	% 95.15	% 68.19	% 77.85
	SVM	% 74.19	% 77.08	% 75.40	% 75.83
	Logistic Regression	% 77.62	% 94.99	% 76.40	% 84.24
	XG Boost	% 92.02	% 92.07	% 91.82	% 91.85

Table 3 (cont). *Comparison with previous studies*

[22]	Google Net	% 80.85	----------	% 93.94	----------
	Alex Net	% 93.52	----------	%96.97	----------
[23]	ANN	% 85.7	%86.7	% 87.8	% 88.6
	SVM	% 87.06	% 88.06	% 89.06	% 89.4
	LSTM	% 88.5	% 89.5	% 89.39	% 90.2
	CNN-LSTM	% 92.03	% 92.93	% 92.36	% 94.3
This Study	VGG16	% 85.19	% 86.67	% 87.50	% 85.16
	VGG19	% 74.04	% 73.33	% 73.86	% 73.49
	ResNet50	% 88.89	% 89.29	% 90.62	% 88.83
	ViT	**% 92.59**	**% 92.30**	**% 93.75**	**% 93.01**

It is also of great importance to compare the study with previous studies in the literature in terms of its reliability and robustness. In this regard, a comparison of this study with different studies is included in Table 3. Although the data sets may not be common to all studies, all compared studies belong to binary brain tumor classification studies. On the other hand, the study numbered [20] used the same data set as in this study in terms of the data set used. This is especially important in terms of comparison. As a result of the comparison with both machine learning-based models and deep learning-based models, it is clearly seen in Table 3 that the study conducted with ViT gives the most successful results.

The integration of ViT models into clinical workflows holds promise for significantly improving the accuracy and efficacy of brain tumor diagnosis. This advancement could translate to better patient outcomes. ViT-powered systems could potentially offer radiologists a valuable tool for automated second opinions, potentially leading to a reduction in misdiagnoses and more optimized treatment plans. The encouraging results obtained in this study warrant further investigation while acknowledging inherent limitations. The employed dataset, while valuable, may not comprehensively capture the entire spectrum of brain tumor presentations. This limitation highlights the need for future research to utilize even larger and more diverse datasets for model validation. Additionally, the computational resources required for training the ViT models are substantial, potentially hindering their real-world implementation in resource-constrained settings. Future research efforts should prioritize the optimization of these models for efficient training with lower computational resources. This would facilitate broader adoption and enhance the clinical applicability of ViT-based brain tumor classification techniques.

This study evaluates the performance of Vision Transformers (ViTs) in comparison to CNN-based transfer learning models (VGG-16, VGG-19, and ResNet50) for brain tumor classification. The ViT model achieved the highest accuracy (92.59%), precision (92.30%), recall (93.75%), and F1-score (93.01%), outperforming its counterparts.

Specifically, the ViT model misclassified only 2 tumor-positive images as tumor-free, demonstrating superior robustness compared to the VGG-16 (4 misclassifications) and VGG-19 (7 misclassifications). ResNet50, while showing competitive performance (88.89% accuracy), lagged slightly behind the ViT in precision and recall, underscoring the advantages of the transformer-based architecture in capturing global image features.

Remarks:

- **Superior Modeling of Global Features**: The ViT model's ability to process images as sequences of patches and leverage self-attention mechanisms likely contributed to its superior

classification accuracy. This highlights its potential for analyzing complex medical images where global spatial relationships are critical.

- **Clinical Implications**: The minimal misclassification rate of the ViT model emphasizes its potential for reducing diagnostic errors in clinical settings, particularly in tasks where false negatives (tumor-free classifications of tumor-positive images) could have serious consequences for patient care.
- **Model Efficiency**: While the ViT model demonstrates superior performance, its computational intensity remains a limitation. Future efforts could focus on optimizing transformer architectures for resource-constrained environments.

VI. CONCLUSION

This study presents a comprehensive investigation into the application of deep learning methodologies for the classification of MR images, specifically distinguishing between tumorous and non-tumorous brain scans. Our research focused on evaluating the performance of the Vision Transformer (ViT), a relatively novel architecture, in comparison to established CNN-based transfer learning models. The results unequivocally demonstrate the superiority of the ViT model in this classification task, consistently outperforming its CNN-based counterparts across multiple evaluation metrics.

Furthermore, our findings were benchmarked against previous studies in the literature, revealing that our ViT implementation achieved state-of-the-art performance in brain tumor classification from MR images. This outcome not only validates the efficacy of transformer-based architectures in medical image analysis but also underscores the potential for significant advancements in automated diagnostic tools for neuro-oncology.

Future Work

While the ViT model has shown remarkable promise, further research is warranted to enhance its robustness and clinical applicability. Future work could include:

- **Utilizing larger and more diverse datasets**: Expanding the dataset to include varied tumor types, multi-class classification tasks, and data from multiple imaging modalities will improve the generalizability of the model.
- **Exploring multimodal approaches**: Combining MR imaging with additional clinical data, such as patient history or genomic information, to develop a more holistic diagnostic model.
- **Optimizing model efficiency**: Reducing computational requirements of ViT models through techniques like pruning, quantization, or lightweight transformer architectures to facilitate deployment in resource-constrained clinical environments.
- **Validation in real-world settings**: Collaborating with healthcare providers to test the ViT model in live clinical workflows and gather feedback on its performance and usability.
- **Extending ViT applications**: Investigating the potential of ViTs for other medical imaging tasks, such as early detection of other cancers or organ-specific imaging challenges.

In conclusion, this study contributes valuable insights to the field of medical image analysis, particularly in brain tumor classification. The demonstrated efficacy of the ViT model suggests a promising direction for the development of more accurate and reliable diagnostic support systems in clinical settings. By addressing these future directions, the ViT framework can continue to evolve, offering significant advancements in early detection and diagnosis, ultimately leading to improved patient outcomes in neuro-oncology.

VII. REFERENCES

[1] L. M. DeAngelis, "Brain tumors," *New England journal of medicine,* vol. 344, no. 2, pp. 114-123, 2001.

[2] J. H. Sampson, M. D. Gunn, P. E. Fecci, and D. M. Ashley, "Brain immunology and immunotherapy in brain tumours," *Nature Reviews Cancer,* vol. 20, no. 1, pp. 12-25, 2020.

[3] G. S. Tandel, A. Balestrieri, T. Jujaray, N. N. Khanna, L. Saba, and J. S. Suri, "Multiclass magnetic resonance imaging brain tumor classification using artificial intelligence paradigm," *Computers in Biology and Medicine,* vol. 122, p. 103804, 2020.

[4] R. Mehrotra, M. Ansari, R. Agrawal, and R. Anand, "A transfer learning approach for AI-based classification of brain tumors," *Machine Learning with Applications,* vol. 2, p. 100003, 2020.

[5] R. Ranjbarzadeh, A. Caputo, E. B. Tirkolaee, S. J. Ghoushchi, and M. Bendechache, "Brain tumor segmentation of MRI images: A comprehensive review on the application of artificial intelligence tools," *Computers in biology and medicine,* vol. 152, p. 106405, 2023.

[6] W. Ayadi, W. Elhamzi, I. Charfi, and M. Atri, "Deep CNN for brain tumor classification," *Neural processing letters,* vol. 53, pp. 671-700, 2021.

[7] Ş. Öztürk and U. Özkaya, "Skin lesion segmentation with improved convolutional neural network," *Journal of digital imaging,* vol. 33, pp. 958-970, 2020.

[8] O. Dikmen, "Deep Learning Models for the Detection and Classification of COVID-19 and Associated Lung Diseases Using X-Ray Images," *Artificial Intelligence Theory and Applications,* vol. 4, no. 2, pp. 121-142, 2024.

[9] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929,* 2020.

[10] S. Charfi, R. Lahmyed, and L. Rangarajan, "A novel approach for brain tumor detection using neural network," *International Journal of Research in Engineering and Technology,* vol. 2, no. 7, pp. 93-104, 2014.

[11] M. Nazir, F. Wahid, and S. Ali Khan, "A simple and intelligent approach for brain MRI classification," *Journal of Intelligent & Fuzzy Systems,* vol. 28, no. 3, pp. 1127-1135, 2015.

[12] N. Vani, A. Sowmya, and N. Jayamma, "Brain tumor classification using support vector machine," *International Research Journal of Engineering and Technology (IRJET),* vol. 4, no. 7, pp. 792-796, 2017.

[13] T. Gupta, T. K. Gandhi, R. Gupta, and B. K. Panigrahi, "Classification of patients with tumor using MR FLAIR images," *Pattern Recognition Letters,* vol. 139, pp. 112-117, 2020.

[14] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251-1258.

[15] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8697-8710.

[16] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700-4708.

[17] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[19] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[20] S. Asif, W. Yi, Q. U. Ain, J. Hou, T. Yi, and J. Si, "Improving effectiveness of different deep transfer learning-based models for detecting brain tumors from MR images," *IEEE Access*, vol. 10, pp. 34716-34730, 2022.

[21] S. Shilaskar, T. Mahajan, S. Bhatlawande, S. Chaudhari, R. Mahajan, and K. Junnare, "Machine Learning Based Brain Tumor Detection and Classification using HOG Feature Descriptor," in *2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, 2023, pp. 67-75: IEEE.

[22] P. Pilaoon, N. Maneerat, A. Nakthewan, R. Varakulsiripunth, and K. Hamamoto, "Brain Tumor Classification using Pretrained Deep Convolutional Neural Network," in *2023 9th International Conference on Engineering, Applied Sciences, and Technology (ICEAST)*, 2023, pp. 84-88: IEEE.

[23] R. Dhaniya and K. Umamaheswari, "CNN-LSTM: A Novel Hybrid Deep Neural Network Model for Brain Tumor Classification," *Intelligent Automation & Soft Computing*, vol. 37, no. 1, 2023.

[24] N. Chakrabarty. (2019). *Brain MRI Images Dataset for Brain Tumor Detection* [Online]. Available: https://www.kaggle.com/datasets/navoneel/brain-mri-images-for-brain-tumor-detection

[25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, 2009, pp. 248-255: Ieee.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[28] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.