Adıyaman University Journal of Science
https://dergipark.org.tr/en/pub/adyujsci
DergiPark AKADEMİK
ISSN 2147-1630
e-ISSN 2146-586X

Tumor Detection by Classification of Brain MRI Images Using the Vision Transformers

Uğur DEMİROĞLU[1,*]

[1]*Kahramanmaraş İstiklal University, Faculty of Engineering, Architecture and Design, Department of Software Engineering, Kahramanmaraş, Türkiye*
ugur.demiroglu@istiklal.edu.tr, ORCID: 0000-0002-0000-8411

Received: 23.10.2024 **Accepted:** 19.12.2024 **Published:** 31.12.2024

Abstract

The interplay between applied mathematics and artificial intelligence is pivotal for advancing both fields. AI fundamentally relies on statistical and mathematical techniques to derive models from data, thus enabling computers to improve their performance over time. Classification of brain MRI images for tumor detection has improved significantly with the advent of machine learning and deep learning techniques. Classical classifiers such as Support Vector Machines (SVM), Tree, and k-Nearest Neighbors (k-NN) have been widely used in conjunction with feature extraction methods to improve the accuracy of tumor detection in MRI scans. Recent studies have shown that classical classifiers can effectively analyze features extracted from MRI images, which can lead to improved diagnostic capabilities. Feature extraction is a critical step in the classification process. Classification of brain MRI images using Vision Transformers (ViTs) represents a significant advancement in medical imaging and tumor detection. ViTs leverage the transformer architecture, which is highly successful in natural language processing, to effectively process visual data. This approach allows for capturing long-range dependencies within images and enhances the ability of the model to distinguish complex patterns associated with brain

* Corresponding Author DOI: 10.37094/adyujsci.1572289

tumors. Recent studies have demonstrated the effectiveness of ViTs in various classification tasks, including medical imaging. In our study, the classification accuracy of the dataset from the ViTs network was 78.26%. In order to increase tumor detection performance, features of the ViTs network were extracted and given to classical classifiers, and 81.9% accuracy was achieved in Tree classifier. As a result, classification of brain MRI images using ViTs represents a new approach with the strengths of deep learning and traditional machine learning methods, namely feature extraction and classification in classical classifiers.

Keywords: Brain MRI; Tumor detection; Classification; Vision transformers; Applied mathematics.

Vision Transformers Kullanılarak Beyin MRI Görüntülerinin Sınıflandırılmasıyla Tümör Tespiti

Öz

Uygulamalı matematik ve yapay zeka arasındaki etkileşim, her iki alanın da ilerlemesi için çok önemlidir. Yapay zeka, verilerden modeller türetmek için temelde istatistiksel ve matematiksel tekniklere güvenir ve böylece bilgisayarların zamanla performanslarını iyileştirmelerini sağlar. Beyin MRI görüntülerinin tümör tespiti için sınıflandırılması, makine öğrenimi ve derin öğrenme tekniklerinin ortaya çıkmasıyla önemli ölçüde iyileşmiştir. Destek Vektör Makineleri (SVM), Ağaç ve k-En Yakın Komşular (k-NN) gibi klasik sınıflandırıcılar, MRI taramalarında tümör tespitinin doğruluğunu artırmak için özellik çıkarma yöntemleriyle birlikte yaygın olarak kullanılmıştır. Son çalışmalar, klasik sınıflandırıcıların MRI görüntülerinden çıkarılan özellikleri etkili bir şekilde analiz edebileceğini ve bunun da gelişmiş tanı yeteneklerine yol açabileceğini göstermiştir. Özellik çıkarma, sınıflandırma sürecinde kritik bir adımdır. Görme Dönüştürücüleri (ViT) kullanılarak beyin MRI görüntülerinin sınıflandırılması, tıbbi görüntüleme ve tümör tespitinde önemli bir ilerlemeyi temsil etmektedir. ViT, görsel verileri etkili bir şekilde işlemek için doğal dil işlemede oldukça başarılı olan dönüştürücü mimarisinden yararlanır. Bu yaklaşım, görüntüler içindeki uzun menzilli bağımlılıkları yakalamaya olanak tanır ve modelin beyin tümörleriyle ilişkili karmaşık örüntüleri ayırt etme yeteneğini artırır. Son çalışmalar, tıbbi görüntüleme dahil olmak üzere çeşitli sınıflandırma görevlerinde ViT'in etkinliğini göstermiştir. Çalışmamızda, ViT ağından gelen veri setinin sınıflandırma doğruluğu %78,26 idi. Tümör tespit performansını artırmak için ViT ağının özellikleri çıkarılıp klasik sınıflandırıcılara verildi ve Ağaç sınıflandırıcısında %81,9 doğruluk

elde edildi. Sonuç olarak, Görme Dönüştürücüleri kullanılarak beyin MRI görüntülerinin sınıflandırılması, klasik sınıflandırıcılarda özellik çıkarma ve sınıflandırma olmak üzere derin öğrenme ve geleneksel makine öğrenme yöntemlerinin güçlü yönlerine sahip yeni bir yaklaşımı temsil etmektedir.

Anahtar Kelimeler: Beyin MRI; Tümör tespiti; Sınıflandırma; Görüntü transformatörleri, Uygulamalı matematik.

1. Introduction

Brain tumors are abnormal growths that develop inside the brain or its surrounding tissues. These tumors can be benign (non-cancerous) or malignant (cancerous), with symptoms varying greatly depending on their location, size, and form [1]. Common symptoms include migraines, seizures, vision problems, balance challenges, personality or behavioral changes, and trouble speaking or swallowing. Imaging examinations, such as magnetic resonance imaging (MRI) or computed tomography (CT) scans, are commonly used to diagnose tumors. A biopsy, which includes removing a sample of tumor tissue for examination under a microscope, is frequently required to diagnose the type of tumor. Treatment choices for brain tumors are determined by the tumor's features, such as type, size, location, and the patient's overall health. Surgery is commonly used to remove benign and certain malignant tumors, whereas radiation treatment and chemotherapy are used to kill cancer cells. Targeted therapy, which utilizes medications to specifically target cancer cells, is also becoming more essential in the treatment of brain tumors. Some brain tumors are curable, but others are more difficult to treat. The prognosis of brain tumors varies greatly depending on these characteristics. Ongoing research aims to produce more effective and targeted medicines for brain tumors, with the ultimate goal of improving patient outcomes. Even if the tumor is treatable, early diagnosis of the disease is life-saving. Thus, developments on this issue are significantly important.

As mentioned in the previous paragraph, MRI and CT have been frequently used to diagnose tumors. MRI is a non-invasive diagnostic technique that uses a powerful magnet and radio waves to create detailed images of the body's internal structures. Unlike X-rays or CT scans, MRI does not use ionizing radiation. Instead, it aligns the hydrogen atoms in the body with the magnetic field and then disturbs them using radio waves. As the atoms realign, they emit signals that are detected by the MRI machine. These signals are processed by a computer to create images that can show organs, bones, muscles, and blood vessels in great detail. MRI is particularly useful for examining soft tissues and is often used to diagnose conditions such as brain tumors, spinal cord injuries, and joint problems. Specifically, the study in this paper focuses on MRI to detect

brain tumors. Brain MRI is a technique to obtain detailed images of the brain and surrounding structures [2]. The patient lies inside a cylindrical magnet, and radiofrequency pulses are applied to the brain, causing the hydrogen atoms in the body to temporarily shift their alignment. As these atoms realign, they emit signals that are detected by the MRI machine. These signals are processed by a computer to create images that can show the brain's anatomy, blood flow, and metabolism. The advantage of Brain MRI lies in its ability to provide high-resolution images without the use of ionizing radiation, making it a safer option compared to CT scans or X-rays. MRI can also be used to detect subtle changes in brain tissue, making it valuable for diagnosing conditions such as tumors, strokes, and multiple sclerosis. Additionally, MRI can be used to assess brain function and monitor treatment response. Although it has a higher cost compared to other imaging techniques, the longer scan time, and the potential for discomfort or claustrophobia in some patients, MRI remains one of the most reliable imaging techniques.

Classification techniques for medical images play a crucial role in the field of healthcare by enabling accurate and efficient diagnosis and treatment. These techniques involve the use of algorithms to categorize medical images into different classes based on their visual characteristics. For instance, they can be used to differentiate between benign and malignant tumors, identify various types of diseases, or analyze the progression of a disease over time. By automating the classification process, medical professionals can save time and enhance diagnostic accuracy, leading to improved patient outcomes. Furthermore, classification techniques can be used to assist in treatment planning and monitoring, ensuring that patients receive appropriate care. This paper implements the ViTs method for classification [3]. ViTs have significantly impacted the field of computer vision, particularly in image classification. Unlike traditional convolutional neural networks (CNNs), ViTs utilize transformers, a sequence-to-sequence modeling architecture originally developed for natural language processing. ViTs divide images into patches, flatten them into vectors, and then feed them into a transformer encoder. This approach allows ViTs to handle images of various sizes and resolutions efficiently. Additionally, ViTs have demonstrated competitive performance with CNNs, especially on large-scale datasets. Their scalability, flexibility, and strong theoretical foundation make them a promising choice for various computer vision tasks, including medical image analysis and remote sensing. More information about the method can be found in the further sections.

Based on the above information, computerized techniques have significant importance in early diagnosis of tumors. This idea has attracted the attention of researchers and has led to valuable studies on the subject. For instance, Elbedoui et al. studied deep learning approaches for dermoscopic image-based skin cancer diagnosis in [4]. They concentrated on utilizing deep

learning methodologies for the detection of skin cancer through dermoscopic images, demonstrating the efficacy of neural networks in identifying malignant patterns. Mejri et al. implemented the Visual Geometry Group and ResNet-50 in their study in [5]. In the study, they employed Visual Geometry Group (VGG) networks and ResNet-50, two powerful convolutional neural network (CNN) architectures, to proficiently categorize skin cancer photos. The ViT formed the basis of the study by Hameed et al. [6]. In this study, Hameed et al. investigated the Vision Transformer (ViT) for skin cancer classification, highlighting its capacity to grasp intricate connections among picture components. Similarly, the ViT is combined with MobileNetV2 for skin cancer classification in [7]. The integration of ViT with MobileNetV2 exemplifies a hybrid strategy that optimizes performance while ensuring computational economy, rendering it appropriate for use in resource-limited settings. Specialized in brain tumor detection, Karthik et al. used a fusion of advanced methodologies for this purpose in [8]. The study utilized a combination of modern technologies in brain tumor detection to improve diagnosis accuracy. Subba and Sunaniya implemented an attention based GoogLeNet-style CNN to optimize brain tumor classification in [9]. They developed an attention-based GoogLeNet-style CNN that enhances categorization by concentrating on the most pertinent areas of the picture. A study of Sathya et al. employed Xception CNN through high-precision MRI analysis for brain tumor diagnosis in [10]. The Xception model was selected for its depthwise separable convolutions, which improve computing efficiency and precision. The methodology prioritizes the mitigation of demographic biases through the integration of various data and the establishment of bias detection systems. The ConvNext architecture is used to classify brain tumor grade in [11]. The study, utilized the ConvNext architecture to identify brain tumor grades, hence satisfying the essential requirement for accurate tumor grading in treatment planning. A ViT named as ViT-BT has been designed for classifying brain tumors in [12] and a mobile ViT model is presented in [13]. Similar to these studies, more studies can be found related to classification of medical images and also ViT based classification problems. In these studies, the Vision Transformer was specifically modified for brain tumor categorization, resulting in the creation of a dedicated ViT-BT model to enhance its efficacy for this purpose. A mobile-compatible ViT model was developed, integrating the sophisticated feature extraction abilities of transformers with lightweight, efficient processing for use in portable or resource-constrained environments. These publications together highlight the progression of deep learning in medical imaging, namely the transition from conventional CNNs to sophisticated architectures such as ViTs and their hybrid forms. They emphasize the use of attention processes and streamlined models, which improve accuracy while maintaining application in various clinical and real-world settings.

This paper introduces an innovative method for brain tumor diagnosis by combining ViTs with conventional machine learning classifiers. This research utilizes ViTs for feature extraction, distinguishing it from traditional approaches that depend exclusively on classical classifiers or CNNs, as ViT adeptly captures long-range correlations in pictures. It subsequently improves classification accuracy by integrating ViT-derived features with conventional classifiers like Decision Trees. The research demonstrated a significant increase in accuracy, increasing from 78.26% with ViT alone to 81.9% with a Tree classifier, illustrating the collaboration between sophisticated deep learning and traditional techniques. This hybrid method represents a substantial improvement in medical imaging, delivering a more reliable and precise diagnostic instrument for brain tumor identification. This study's originality stems from its revolutionary integration of ViTs with traditional machine learning classifiers to enhance the precision of brain tumor identification in MRI images. ViTs, originally developed for natural language processing, have lately been repurposed for computer vision applications, including medical imaging. Nonetheless, its independent use in tumor diagnosis frequently encounters difficulties in attaining maximum accuracy due to the restricted quantity and intricacy of medical datasets. This study mitigates these limitations by employing ViT for its robust capacity to capture long-range dependencies and complex visual patterns, while also extracting features from the ViT model to input into traditional classifiers such as Decision Trees, Naïve Bayes, and k-Nearest Neighbors. This dual-stage technique is innovative since it combines the representational capabilities of deep learning with the interpretability and simplicity of conventional machine learning models. The study indicates that ViT attains a baseline accuracy of 78.26%, while the incorporation of ViT-derived features with a Tree classifier enhances performance to 81.9%, resulting in a 3.64% improvement. This technique leverages the advantages of both paradigms—ViT for sophisticated feature extraction and traditional classifiers for effective and precise decision-making. By integrating these techniques, the study establishes a comprehensive and scalable framework for enhancing diagnostic precision, establishing a new standard for hybrid models in medical imaging. This concept improves tumor detection efficacy and demonstrates the possibility of combining contemporary and classic methods to address intricate challenges in healthcare.

This paper is organized in the following way. The materials and methods used in the study are presented in Section 2. Section 3 gives the case study with illustrations and the last section has the concluding remarks.

2. Materials and Methods

2.1. The Dataset

A malignant brain tumor is a potentially fatal disorder. Glioblastoma is the most common type of brain cancer in adults and has the worst prognosis, with a median survival of less than a year. The presence of a specific genetic sequence in the tumor known as MGMT promoter methylation is a good prognostic indicator and a strong predictor of treatment response. Currently, genetic analysis of cancer needs surgery to get a tissue sample. It may take many weeks to discover the genetic characterization of the tumor. Depending on the findings and the type of initial therapy chosen, more surgery may be required. If an accurate approach for predicting cancer genetics only by imaging (i.e., radiogenomics) could be developed, it could reduce the number of surgeries and modify the type of therapy required.

The dataset used is "Brain MRI Scan Images" and was downloaded from the Kaggle website [14]. The scanned images are divided into 2 subclasses to be used in tumor detection. These images are images exported from the RSNA-MICCAI Brain Tumor Competition. The dataset size is 7MB and consists of 2 subfolders as negative and positive. While there are 98 negative images, there are 129 positive images, a total of 227 Brain MRI images. The images are in 96 DPI resolution, 24-bit depth, the images are at least 200 pixels in width and height and jpg image format. The image format is jpg Positive/Negative MGMT status, used to label the tumor types. The dataset used in the study has a public license and is frequently used in the fields of medicine, cancer, computer vision free of charge and provides uninterrupted access and download. Two sample images from the dataset are shown in Fig. 1.

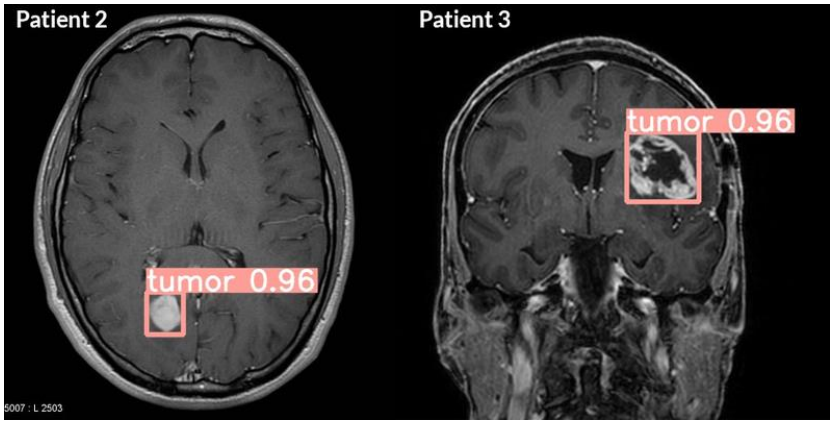

Figure 1: Sample images from the dataset.

2.2. The Vision Transformers

In 2022, the ViT emerged as a competitive alternative to CNNs, which are currently state-of-the-art in computer vision and thus widely used for various image recognition tasks. ViT models outperform the latest CNN technology by nearly four times in terms of computational efficiency and accuracy. Unlike traditional CNNs, which eliminate the need for manually crafted features, the ViT distinguishes itself by leveraging a self-attention mechanism to gather global contextual information from the entire image. This innovative approach involves dividing an input image into fixed-size patches, subjecting each patch to linear embeddings that transform them into high-dimensional vectors, and then processing these vectors through a transformer encoder. This methodology empowers ViT with the ability to skillfully capture complex long-range dependencies and subtle relationships between different regions of the image.

In this study, a pre-trained ViT neural network will be used for the classification of brain MRI images. The model utilizes a transformer architecture to encode image inputs into feature vectors. The network consists of two main components: the backbone and the head. The backbone is responsible for the encoding step, where it takes input images and extracts feature vectors. The head is responsible for making predictions by mapping the encoded feature vectors to prediction scores. By employing transfer learning, the model can be fine-tuned for better performance on specific tasks. The block diagram of the ViT network is illustrated in Fig. 2 [15].

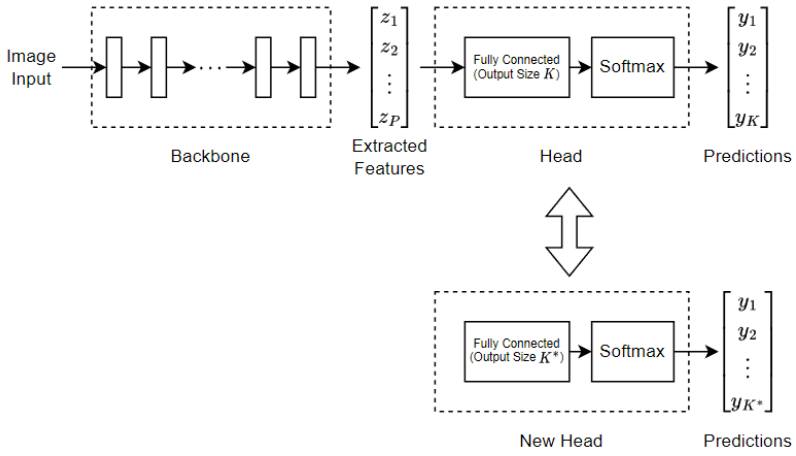

Figure 2: The block diagram of the ViT network.

This diagram outlines the architecture of a ViT network that makes predictions for K classes and illustrates how the network is structured to enable transfer learning for a new dataset with K classes. ViTs represent a revolutionary architecture in computer vision, using the ideas of the Transformer model—originally designed for natural language processing (NLP)—to interpret visual information. ViTs analyze pictures by segmenting them into fixed-size patches, typically

16x16 pixels, and using each patch as a token, similar to words in natural language processing tasks. The patches are flattened and linearly inserted into a fixed-dimensional space, with positional encodings used to preserve spatial information. The Transformer encoder, central to ViTs, employs self-attention processes to represent global interactions across patches, enabling the network to successfully capture long-range dependencies and contextual characteristics. In contrast to CNNs, which depend on localized operations, ViTs are proficient at collecting global context, especially when trained on extensive datasets. Nonetheless, their quadratic complexity in self-attention renders them computationally demanding for high-resolution photos. To address this, versions like as Swin Transformers implement hierarchical structures and localized self-attention, enhancing computing efficiency. Hybrid models integrate the advantages of CNNs and ViTs by employing convolutional layers for preliminary feature extraction before transmitting data to Transformer layers. Mobile ViTs tackle the issue of resource use, facilitating implementation on edge devices. Notwithstanding their benefits, ViTs are data-intensive, sometimes necessitating considerable pretraining on large datasets like ImageNet or JFT-300M to achieve optimal performance. In the absence of such resources, their performance may be inferior to that of CNNs, which are intrinsically more efficient with smaller datasets. Nevertheless, ViTs exhibit exceptional scalability and adaptability, rendering them suitable for a range of vision applications, such as image classification, object identification, medical image analysis, and video comprehension. Their applicability is expanding, with research aimed at enhancing computing efficiency and versatility.

It would be comprehensive to include information about how the components of the ViTs model contribute to the model performance. The Backbone in a deep learning model refers to the feature extraction component. This module processes input data, such as an image, and extracts both low- and high-level features that are subsequently used by other parts of the model. In ViTs, the backbone is built on the Transformer architecture. Unlike traditional convolutional neural networks (CNNs), ViTs divide visual input into patches and process each patch as a vector. These vectors are then analyzed using the Transformer model's attention mechanism, making it a unique method for feature extraction. In detail:

- **Input Images and Patches**: ViT segments an image into fixed-size patches. For example, an image of size 224x224 can be divided into 16x16 patches, with each patch converted into a vector.
- **Patch Embeddings**: Each patch is transformed into a vector and serves as input for the Transformer model.

- **Transformer Encoder**: These patch vectors are processed by the encoder in ViT, which leverages the attention mechanism to understand the context of each patch, thereby extracting higher-level features.

In summary, the backbone provides the foundational structure for extracting meaningful features from visual data. The Head module represents the inference component of the model. It transforms the features extracted by the backbone into a final output. In ViT, the head module typically performs the following steps:

- **Pooling and Classification**: Features from the backbone are aggregated and formatted for final classification, often using global average pooling or a Multi-Layer Perceptron (MLP).
- **Result Generation**: The output is tailored for specific tasks, such as classification, regression, or other objectives. For instance, in image classification, the head module predicts the class of the input image.

ViT commonly uses a single "class token" in its head module, which combines the representations of all patches for classification. This process is carried out using the attention mechanism typical of Transformer architectures.

Functions of the Backbone and Head Modules

- **Backbone**: Extracts and transforms features from the image into meaningful high-level representations.
- **Head**: Utilizes these features to perform specific tasks, such as classification.

In summary:

- The **Backbone** processes visual data and extracts features in a format that the model can understand.
- The **Head** uses these features to produce the final output, such as a class prediction.

Together, these two components form the core structure of Google's Vision Transformer, playing a crucial role in processing visual data and generating accurate predictions. In traditional artificial intelligence models, performance augmentation techniques are often employed in the literature by extracting features from the model post-training, prior to the classification layer, and utilizing alternative classifiers instead of the network's classifier. This study utilized the

characteristics from the head layer of the ViT network post-training, employing classical classifiers for classification.

The ViTs represents a notable advancement in deep learning and image processing, presenting distinct benefits while encountering certain limits. In contrast to conventional CNNs, ViT obviates the necessity for convolutional layers by employing the Transformer architecture, hence offering a more adaptable and generalizable framework for visual applications. By modifying the Transformer, first created for natural language processing, for visual input, ViT acquires a more profound comprehension of contextual relationships inside pictures. The attention mechanism proficiently captures long-range dependencies, accurately representing the links between distant components in a picture. Moreover, ViT provides versatility in data preprocessing and network architecture, enabling the modification of patch dimensions and attention techniques. It excels on extensive datasets, such as ImageNet, frequently attaining elevated accuracy rates. Nonetheless, ViT possesses some restrictions. Its dependence on attention processes and a substantial number of parameters renders it computationally demanding, necessitating robust hardware such as GPUs or TPUs and extended training durations. The large number of parameters leads to prolonged and more resource-demanding training. Another difficulty is its inferior performance on tiny datasets, as CNNs frequently surpass ViT. To get optimal outcomes with constrained data, ViT generally necessitates methodologies such as pre-trained models or data augmentation. In our work utilizing a dataset of 227 brain MRI images, we employed transfer learning to address these problems. This method increased the model's accuracy to 81.9%, illustrating the efficacy of transfer learning in augmenting ViT's performance with limited datasets. In conclusion, although ViT presents revolutionary advancements in image processing, its efficacy is contingent upon the accessibility of computing resources and extensive datasets. However, methods like transfer learning can alleviate its shortcomings, rendering it an effective instrument for visual analytic tasks.

3. Results

80% of the dataset is used for training, while 20% is strictly reserved for test data that is not involved in the training process. The scanned images in the dataset are scaled and normalized to a uniform size of 384x384x3 and treated as colored images during both the training and testing phases. The training parameters are determined as Mini Batch Size = 16, Max Epochs = 5, Iterations Per Epoch = 11 and Validation Frequency = 3. This study utilized hyperparameter values commonly seen in the literature. The MiniBatchSize was set at 16, and the MaxEpochs was established at 5. These options are typically used due to their efficacy in training machine learning models. The IterationsPerEpoch value was computed as the estimated ratio of the entire

dataset size utilized for training divided by the MiniBatchSize. The ValidationFrequency values were determined by dividing the IterationsPerEpoch by the MaxEpochs, resulting in a framework where validation steps are evenly distributed throughout the training phase. The model's performance was consistently assessed throughout and following the training process. The selection of these hyperparameters seeks to enhance the efficiency of the training process as well as the validation and generalization performance of the model.

The training of the network utilized Stochastic Gradient Descent with Momentum (SGDM) as the optimizer, employing a stochastic solver. Parallel computing was leveraged on a graphics card, with 16 parallel workers running simultaneously to accelerate the training process. The remaining training parameters are InitialLearnRate = 1e-4, Shuffle = every-epoch and ExecutionEnvironment = parallel. The specifications of the computer used in the experiment are listed in Table 1.

Table 1: Specifications of the computer used in the experiment.

Processor	12th Gen Intel(R) Core(TM) i9-12900F 2.40 GHz
Cores, Processors	16, 24
Installed RAM	64.0 GB (63.7 GB usable)
GPU	NVIDIA RTX A4000
DirectX version	12 (FL 12.1)
GPU Memory	47.9 GB (16.0 GB Dedicated, 31.9 GB Shared)
SSD Capacity	477 GB

The training process was completed in 26 minutes and 3 seconds. Training accuracy was achieved as 0.7826. Fig. 3 gives the confusion matrix obtained.

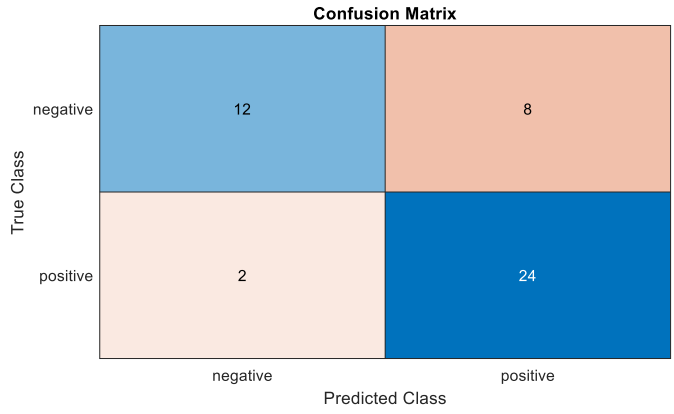

Figure 3: The confusion matrix.

Upon examining the Confusion Matrix, it is observed that out of 46 test samples, 20 are negative and 26 are positive images. Among the negative images, 12 were correctly predicted, while 8 were misclassified. Similarly, out of 26 positive images, 24 were accurately predicted.

The progress of training iterations, the duration of each iteration, mini-batch performance, test performance, and errors are presented in Table 2.

Table 2: Training iterations.

Epoch	Iteration	Time Elapsed (hh:mm:ss)	Mini-batch Accuracy	Validation Accuracy	Mini-batch Loss	Validation Loss
1	1	00:00:36	31.25%	47.83%	2.6602	1.1102
1	3	00:01:36	43.75%	56.52%	2.1689	1.1805
1	6	00:03:07	75.00%	56.52%	0.7524	0.9590
1	9	00:04:32	50.00%	63.04%	2.2467	1.0096
2	12	00:06:00	50.00%	63.04%	1.3735	0.6843
2	15	00:07:23	37.50%	58.70%	2.3455	0.9527
2	18	00:08:45	81.25%	76.09%	0.3105	0.5432
2	21	00:10:15	75.00%	58.70%	1.1402	1.4605
3	24	00:11:40	87.50%	71.74%	0.4229	0.6494
3	27	00:13:02	62.50%	60.87%	1.8355	0.9578
3	30	00:14:25	68.75%	89.13%	1.8386	0.4378
3	33	00:15:51	62.50%	89.13%	0.9929	0.4451
4	36	00:17:12	81.25%	65.22%	0.5713	0.8979
4	39	00:18:34	87.50%	80.43%	0.6988	0.4610
4	42	00:19:59	56.25%	89.13%	1.3256	0.4217
5	45	00:21:21	81.25%	60.87%	0.6538	1.0530
5	48	00:22:45	75.00%	71.74%	0.5961	0.5908
5	50	00:23:38	62.50%		2.0322	
5	51	00:24:08	93.75%	80.43%	0.1706	0.4489
5	54	00:25:31	93.75%	73.91%	0.1351	0.6625
5	55	00:26:01	68.75%	78.26%	0.6194	0.5166

It would be useful to enlighten the meanings of the above training parameters. An Epoch signifies a full traversal of the whole training dataset. This is a quantitative measure of the number of instances the model has encountered the complete dataset throughout the training process. During each epoch, the dataset is partitioned into smaller segments known as mini-batches, and the model is modified iteratively for each mini-batch. These updates are termed Iterations, and the quantity of iterations per epoch is contingent upon the batch size and the dataset size. The Time Elapsed (hh:mm:ss) captures the total duration since the commencement of training, facilitating the assessment of training efficiency and progress over time. It is very beneficial for predicting the completion time of the instruction. Mini-batch Accuracy evaluates the model's performance on the current mini-batch during training, offering a rapid yet localized assessment of its predictive capability. Validation Accuracy assesses the model's performance on a distinct validation set post-epoch, indicating its capacity to generalize to novel data. Loss values are also essential. Mini-batch Loss measures the model's mistake on the current mini-batch during training, informing the necessary adjustments to the model's weights. The validation loss, computed post-epoch, reflects the model's performance on the validation dataset. If the validation loss ceases to decline or starts to rise while the training loss continues to fall, it may indicate

overfitting. Monitoring these variables collectively offers an extensive perspective on the training process, facilitating the prompt identification of problems such as overfitting, underfitting, or ineffective training.

The training and the error visualities are given in Fig. 4. As seen in the figure, the test accuracy of the training process using the ViT network reached 78.26%. Before passing through the classification layer, the dataset's training features were extracted and classified using classical classifiers such as Tree, Discriminant Analysis, SVM, KNN, and others.

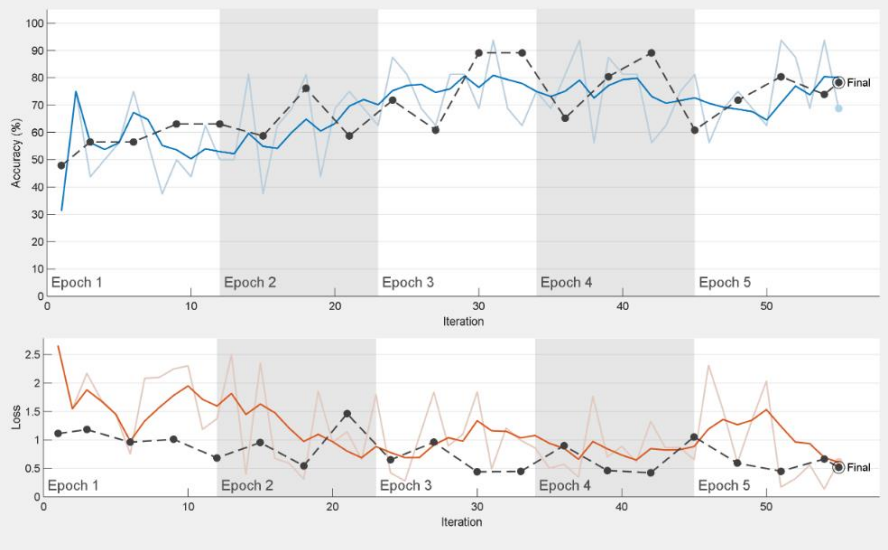

Figure 4: The training and the error graphs.

The classification results are shown in Table 3. Upon reviewing the results, the accuracy increased to 81.9%, representing a 3.64% improvement compared to the standard training performance.

Table 3: Top 12 accuracies of classical classifiers.

No	Models	Sub Models	Accuracy (%)
1	Tree	Coarse Tree	81.9%
2	Quadratic Discriminant	Quadratic Discriminant	81.9%
3	Navie Bayes	Gaussian Naïve Bayes	81.9%
4	Navie Bayes	Kernek Navie Bayes	81.9%
5	KNN	Medium KNN	81.5%
6	Ensemble	Subspace Discriminant	81.5%
7	Binary GLM Logistic Regression	Binary GLM Logistic Regression	81.1%
8	Efficient Linear SVM	Efficient Linear SVM	81.1%
9	SVM	Linear SVM	81.1%
10	SVM	Medium Gaussian SVM	81.1%
11	Linear Discriminant	Linear Discriminant	80.6%
12	SVM	Quadric SVM	80.6%

Similarly, when examining the confusion matrix for the classical classifier that achieved the highest performance given in Fig. 6, it is observed that out of 227 training and test samples—the entire dataset—98 are negative and 129 are positive images. Among the negative images, 76 were correctly predicted, while 22 were misclassified. Similarly, out of 129 positive images, 110 were accurately predicted.

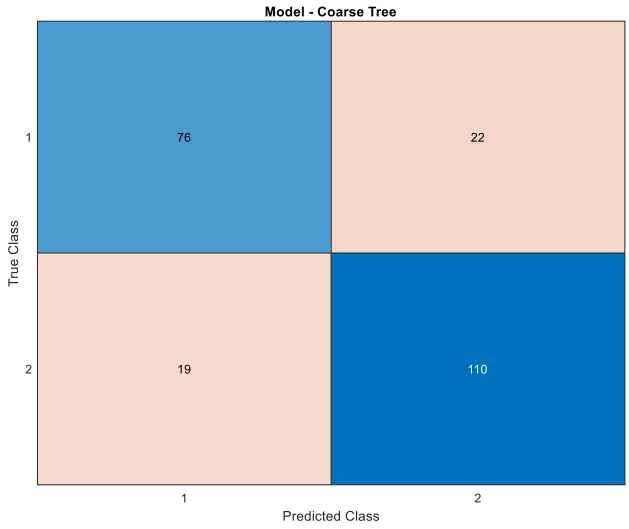

Figure 5: Coarse tree confusion matrix.

Figure 6 shows the prediction distribution of the classical classifier that achieved the highest performance. This distribution illustrates how the model predicted across different classes, highlighting the frequency of correct and incorrect predictions for both the positive and negative samples. The visualization provides insights into the classifier's overall accuracy and the balance between true positives, false positives, true negatives, and false negatives.

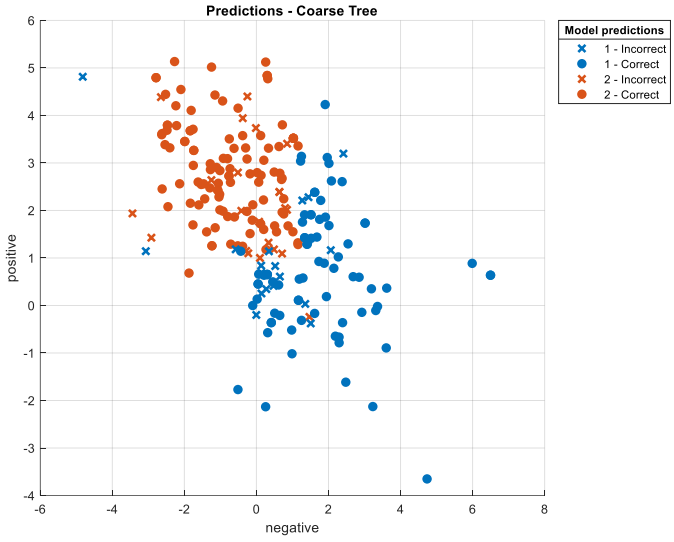

Figure 6: Coarse tree predictions.

Hence, the success of the proposed method is shown.

4. Conclusion

The field of medical imaging has witnessed substantial progress in brain tumor detection, primarily driven by advancements in machine learning and deep learning algorithms. Traditional classifiers like Support Vector Machines, Decision Trees, and k-Nearest Neighbors, when combined with effective feature extraction techniques, have demonstrated promising results in identifying tumors from MRI scans. Recent research has consistently highlighted the efficacy of these classical classifiers in analyzing extracted features from MRI images, leading to enhanced diagnostic capabilities. Feature extraction, a pivotal step in the classification process, plays a crucial role in optimizing the performance of these models.

ViT, a groundbreaking development in medical imaging and tumor detection, has emerged as a powerful tool. Leveraging the transformer architecture, which has proven highly effective in natural language processing, ViT enables the efficient processing of visual data. This approach facilitates the capture of long-range dependencies within images, empowering the model to discern intricate patterns associated with brain tumors. Numerous studies have validated the effectiveness of ViT in various classification tasks, including medical imaging. In our research, the ViT network achieved a classification accuracy of 78.26% on the given dataset. To further enhance tumor detection performance, we extracted features from the ViT network and fed them to classical classifiers. Notably, the Decision Tree classifier exhibited an impressive accuracy of 81.9% when utilizing these extracted features.

In conclusion, the classification of brain MRI images using ViT presents a novel approach that seamlessly integrates the strengths of deep learning and traditional machine learning methods. By combining the powerful feature extraction capabilities of deep learning models with the effective classification techniques of classical classifiers, this approach offers a promising avenue for improving the accuracy and reliability of brain tumor detection.

References

[1] Amin, J., Sharif, M., Haldorai, A., Yasmin, M., Nayak, R. S., *Brain tumor detection and classification using machine learning: a comprehensive survey*, Complex & intelligent systems, 8(4), 3161-3183, 2022.

[2] Ali, H., Biswas, M. R., Mohsen, F., Shah, U., Alamgir, A., Mousa, O., Shah, Z., *The role of generative adversarial networks in brain MRI: a scoping review*, Insights into imaging, 13(1), 98, 2022.

[3] Khan, A., Rauf, Z., Sohail, A., Khan, A. R., Asif, H., Asif, A., Farooq, U., *A survey of the vision transformers and their CNN-transformer based variants*, Artificial Intelligence Review, 56 (Suppl 3), 2917-2970, 2023.

[4] Elbedoui, K., Mzoughi, H., Slima, M. B., *Deep Learning Approaches for Dermoscopic Image-Based Skin Cancer Diagnosis*, In 2024 IEEE 7th International Conference on Advanced Technologies, Signal and Image Processing (ATSIP) 1, 1-7, 2024.

[5] Mejri, S., Oueslati, A. E., *Dermoscopic Images Classification Using Pretrained VGG-16 and ResNet-50 Models*, IEEE 7th International Conference on Advanced Technologies, Signal and Image Processing, 1, 342-347, 2024.

[6] Hameed, M., Zameer, A., Raja, M. A. Z., *A Comprehensive Systematic Review: Advancements in Skin Cancer Classification and Segmentation Using the ISIC Dataset*, Computer Modeling in Engineering & Sciences, 140(3), 2024

[7] Kumar, M. R., Priyanga, S., Anusha, J. S., Chatiyode, V., Santiago, J., Revathi, P., *Synergistic Skin Cancer Classification: Vision Transformer alongside MobileNetV2*, 4th International Conference on Intelligent Technologies, 1-7, 2024.

[8] Karthik, A., Sahoo, S. K., Kumar, A., Patel, N., Chinnaraj, P., Maguluri, L. P., Rajaram, A., *Unified approach for accurate brain tumor Multi-Classification and segmentation through fusion of advanced methodologies*, Biomedical Signal Processing and Control, 100, 106872, 2025.

[9] Subba, A. B., Sunaniya, A. K., *Computationally optimized brain tumor classification using attention based GoogLeNet-style CNN*, Expert Systems with Applications, 125443, 2024.

[10] Sathya, R., TR, M., Bhatia Khan, S., Malibari, A. A., Asiri, F., *Employing Xception Convolutional Neural Network through High-Precision MRI Analysis for Brain Tumor Diagnosis*, Frontiers in Medicine, 11, 1487713, 2024.

[11] Mehmood, Y., Bajwa, U. I., *Brain tumor grade classification using the ConvNext architecture*, Digital Health, 10, 20552076241284920, 2024.

[12] Ali Al-Hamza, K., *ViT-BT: Improving MRI Brain Tumor Classification Using Vision Transformer with Transfer Learning*. Available at SSRN, http://dx.doi.org/10.2139/ssrn.4959261, 2024.

[13] Odusami, M., Damasevicius, R., Milieskaite-Belousoviene, E., Maskeliunas, R., *Multimodal Neuroimaging Fusion for Alzheimer's Disease: An Image Colorization Approach with Mobile Vision Transformer*, International Journal of Imaging Systems and Technology, 34(5), e23158, 2024.

[14] https://www.kaggle.com/datasets/volodymyrpivoshenko/brain-mri-scan-images-tumor-detection, Last Accesed in 10.10.2024.

[15] https://www.mathworks.com/help/deeplearning/ug/train-vision-transformer-network-for-image -classification.html, Last Accessed in 10.10.2024.