Self-Distilled Vision Transformer (SD-ViT) to Classify Brain Tumors using MRI images

Tanjina Ahmed Tuly*[1], Tanjid Ahammed Shafin[1], Jahangir Alam Tamal[1], Jamil Hasan[1], Md Zahid Hasan[1], Md Mashruf Hasan[1]

[1] Department of Computer Science and Engineering,
Daffodil International University, Dhaka 1216, Bangladesh

Emails:

tuly15-4902@diu.edu.bd, shafin15-4903@diu.edu.bd,
tamal15-5322@diu.edu.bd, hasan15-5222@diu.edu.bd,
zahid.cse@diu.edu.bd, 252-15-057@diu.edu.bd

*Corresponding author: tuly15-4902@diu.edu.bd

Abstract. Medical imaging is essential in the identification of brain tumors early and accurately in order to plan diagnosis and treatment. The latest progress in Vision Transformers (ViTs) has shown good promise in the medical image classification tasks. Nonetheless, typical ViT models typically need big data and many computational resources to have the best performance, which can be restrictive in medical imaging applications where the data are small and diverse. This study suggests a Self-Distilled Vision Transformer (SD-ViT) model to classify binary brain tumors to address this problem. The suggested model takes advantage of self-distillation concept, allowing the student branch of the network to learn the intermediate representations of the teacher branch of the same architecture, thus improving feature generalization and learning representations without any external supervision. The publicly available Brain Tumor MRI data was used in experiments and included images of tumor and non-tumor brain scans. The baseline Vision Transformer had an accuracy of 88% whereas a standard ResNet-50 model had a 75% accuracy with the same experimental conditions. Conversely, the proposed SD-ViT model achieved a high classification accuracy of 91% indicating that both the precision and robustness have been improved significantly. The findings imply that self-distillation improves the ability of the Vision Transformer to learn more discriminative tumor features, and reduce overfitting, especially on medical image classification problems with small amounts of data. In general, the suggested SD-ViT offers a simple but a powerful framework to detect brain tumors automatically, which can open the way to better clinical decision support systems with the help of radiological diagnosis.

Keywords: Brain Tumor Classification, Vision Transformer, Self-Distillation, Deep Learning, Medical Imaging, MRI.

1 Introduction

Brain tumor is one of the most devastating and threatening neurological diseases and is associated with cell proliferation abnormality in the brain tissue that may disrupt the normal brain functions including motor skills, visual perception, and memory. Early diagnosis of brain tumors is also important because the delay in diagnosis may cause lethal consequences or irreversible neurologic disability [1]. The Magnetic Resonance Imaging (MRI) is generally considered to be the most accurate imaging modality in identification of tumors because it has a high contrast and is non-invasive [2]. Nonetheless, manual MRI interpretation is tedious and there is a likelihood of inter-observer variations, which has led to the automated computer-aided diagnosis (CAD) systems in order to support radiologists in making clinical decisions [3].

Deep learning (DL) methods have been shown to have an outstanding potential in medical image classification in recent years. VGGNet, ResNet, and DenseNet models are CNN-based models that made numerous steps forward in the field of brain tumor detection and segmentation [4], [5]. In the meantime, self-attention-based models called Vision Transformers (ViTs) have become an exciting substitute to CNNs with excellent global context modeling capability [6]. A number of works have either used CNNs as part of hybrid models or Transformers as part of hybrid models, or both with attention mechanisms to improve the performance of tumor localization and classification [7], [8]. These developments have brought about significant gains in the automated diagnosis but there are still issues in getting reliable generalization between datasets that have limited data diversity [9]. In spite of these developments, current research studies are characterized by three significant limitations. First of all, deep networks are prone to overfitting because of inaccessibility of annotated MRI data. Secondly, the inter-layer knowledge transfer of traditional architecture is frequently underutilized because the general training procedures fail to facilitate deeper levels to be informed by the guidance of higher levels feature representations. Lastly, transformer models, despite their general vision capabilities, are more likely to be large data intensive to achieve the best results and as a result, they are less effective in medical imaging where such data sets are small [10]. This points to the fact that architectures able to provide superior generalization of features with little or no extra data or computational overhead are required.

To overcome these issues, we introduce a Self-Distilled Vision Transformer (SD-ViT) model that introduces self-distillation into the ViT framework in a seamless manner. Self-distillation allows the student branch of the model to learn using the intermediate representations of the student branch of the model, thus trying to transfer knowledge internally without the need to have an external teacher model. This procedure is useful to make the model generalize in limited data condition by imposing consistency among the various hierarchical representations [12]. Using SD-ViT on binary brain tumor classification, we will strengthen the features and reduce overfitting, as well as improve classification accuracy compared to baseline models. We summarize our contribution in this paper as the following:

- Proposed Self-distilled Vision Transformer (SD-ViT) architecture, which benefits internal knowledge transfer between teacher and student networks.

- Validated SD-ViT outperforms baseline ViT and ResNet in brain tumor classification.
- Ablation analysis and visualization to examine effects of self-distillation that resulted in better model robustness.

2 Literature Review

Medical image analysis, especially the detection and classification of brain tumors, has developed greatly through deep learning. The first methods mainly used convolutional neural networks (CNNs) to extract hierarchical features of MRI scans with high success on multiple data sets. Since the appearance of Vision Transformers (ViTs) and hybrid CNN-Transformer models, scientists have investigated how they can be used to capture long-range dependencies and provide better feature representations. Moreover, self-distillation methods have also been suggested to enhance model generalization and efficiency particularly where there is a lack of medical imaging data.

Swati et al. [4] trained a finely tuned GoogLeNet on brain tumor classification and attained the final accuracy scores of 98 percent on a small dataset of MRI. Equally, Sultan et al. [5] used a multi-layer deep CNN to identify various types of brain tumors with an accuracy of 96.13%. Rehman et al. [3] used a self-trained CNN model on MRI slices and achieved a 94.39 percent accuracy. Although these CNN-based architectures showed good results, the use of local receptive fields restricted these architectures in terms of long-range spatial dependencies which would be important in medical image contexts. As transformers-based architectures have emerged, scholars started considering the use of Vision Transformers (ViT) in medical image tasks. The ViT framework was initially proposed by Dosovitskiy et al. [9] and became competitive on ImageNet and then inspired other biomedical applications. A comparative study was carried out by Kumar et al. [7] between CNNs and ViTs, and indicates that ViTs are superior in feature representation when trained using adequate data but might not be effective in small data sets. Srinivasan et al. [8] suggested a hybrid CNN-Transformer model that included CNN feature extractors and self-attention modules and enhanced accuracy and interpretability. Their proposed model was able to achieve 98.56% accuracy on a dataset of brain MRI scans showing the potential of the transformer to improve the localization and classification of tumors.

The techniques of knowledge distillation (KD) and self-distillation have also been studied as recent innovations to enhance the generalization of deep models in low-data settings. Touvron et al. [10] suggested Data-efficient Image Transformers (DeiT), which introduces distillation using attention to learn on a teacher network and apply the results to a transformer model to obtain near-state-of-the-art results with less and less samples on ImageNet. The hybrid CNN-Transformer model developed by Liu et al. [11] included pyramidal convolution and MLP modules to be used in medical image segmentation. Although feature representation quality improved significantly with the model, their main role was that of segmentation and not classification, even though they showed the usefulness of convolutional locality and transformer global reasoning fusion. Saraei et al. [12] examined attention-based deep learning methods such as transformer and attention-based CNNs to analyze brain tumors. Nevertheless, with these developments,

few studies have used self-distillation in ViT models specifically to medical imaging-particularly in binary brain tumor recognition problems.

To conclude, CNN-based models [3-5] have a high accuracy and no global contextual knowledge, while ViT models, and hybrid models [6-8] possess better global representation but demand huge datasets and fail in medical imaging benchmarks with small samples. There are also promising transformers based on distillation [10-12], which have not been exhausted in brain MRI classification. The gap is what is behind our proposed Self-Distilled Vision Transformer (SD-ViT) which utilizes intra-model self-distillation as a ViT architecture to increase the learning efficiency, feature generalization, and classification accuracy on small MRI datasets. Table 1 shows the comparison among existing works

Table 1. Comparison of Existing Works.

Author (Year)	Accuracy (%)	Model / Method	Contribution
Rehman et al. [3]	94.39	Custom CNN	Basic CNN for MRI classification.
Swati et al. [4]	98.00	Fine-tuned GoogLeNet	Transfer learning with ImageNet weights.
Sultan et al. [5]	96.13	Deep CNN	Multi-layer CNN for tumor detection.
Dosovitskiy et al. [9]	88.55	Vision Transformer (ViT)	Introduced ViT for image tasks.
Raghu et al. [10]	—	CNN vs. ViT	Compared CNN and ViT performance.
Touvron et al. [13]	83.10	DeiT	Data-efficient ViT via distillation.
Zhang et al. [15]	92.00	Self-Distillation	Intra-model distillation for efficiency.
Proposed (SD-ViT)	91.00	Self-Distilled ViT	Self-distilled ViT for better generalization.

3 Methodology

The suggested architecture, Self-Distilled Vision Transformer (SD-ViT), is designed to improve the performance of tumor classification of MRI scans by means of internal knowledge distillation in the transformer itself. The whole pipeline involves dataset preparation, data preprocessing and augmentation, Vision Transformer backbone design which is a combination of a self-distillation training framework and explainability mechanisms. Fig. 1 provides the overview of the entire proposed workflow.

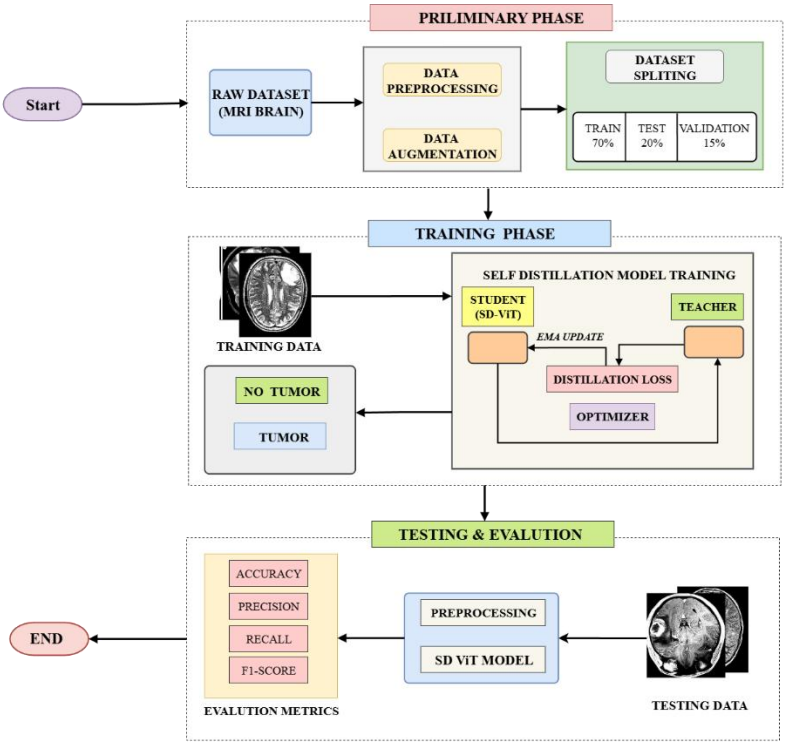

Fig. 1. Proposed Methodology

3.1 Dataset Description

The Brain Tumor MRI Dataset that can be found on Kaggle was used in the experiment [18]. The dataset has 3000 MRI images which are divided into two groups (Yes, tumor and No, non-tumor). All the images are T1-weighted axial brain MRI scan slices. The scanning conditions show different images that have varying resolutions, intensity and contrast. Fig. 2 provides the visualization of sample images for both classes. Table 2 shows the summary of the dataset.

(a) (b)

Fig. 2. Sample Image (a) Yes [Tumor] (b) No [No-Tumor].

Table 2. Original Dataset Image and Class Description

Class Name	No of Images
Yes (Tumor)	1500
No (No-Tumor)	1500

3.2 Data Preprocessing

Medical imaging data, especially MRI scans, can also be affected by illumination variation, resolution, orientation, and scanner intensity, which adversely affect deep learning model performance. Thus, to achieve consistency, better generalization of the model, and effective learning of the proposed Self-Distilled Vision Transformer (SD-ViT), a systematic and standardized preprocessing pipeline was used. Fig. 3 shows the comparison of original image and preprocessed image.

(a) (b)

Fig. 3. (a) Original Image (b) Preprocessed Image.

The preprocessing process was divided into the following main steps:

Image Resizing and Normalization: As the architecture of the Vision Transformer algorithm needs a fixed image size, to ensure consistency and compatibility of all images with the model input layer, all the images were resized to 224 x 224 pixels. The values of the pixel intensities were scaled to the interval [0, 1] by dividing with 255 to make calculations less complex and gradient updates more stable with training. Normalization reduces also the effect of a change in intensity between MRI scans of different subjects or acquisition devices.

$$I_{norm} = \frac{I - I_{min}}{I_{max} - I_{min}} \quad (1)$$

Here I represent the original pixel intensity and I_{norm} denotes the normalized pixel value.

Noise Removal and Smoothing: Certain noise is usually inevitable in MRI images because of artifacts in the process of acquisition and distortion of the magnetic field. To counter this, a Gaussian smoothing filter has been applied which is useful in keeping the important structures of the body and also diminishing undesired noise. This preprocessing improves the effect of the model in concentrating on structural tumor boundaries and not on background inconsistencies.

Data Augmentation: To help avoid overfitting and increase model strength, massive on-the-fly augmentation was undertaken with the TensorFlow image augmentation pipeline. The transformations applied are illustrated in Table 3.

Table 3. Data augmentation techniques and their applied values/ranges.

Augmentation Type	Value/Range Applied
Flipping	Horizontal and Vertical
Rotation	±8%
Shifting (Translation)	±6% in x or y direction
Zoom Augmentation	(0.08-0.12)

These changes are a simulation of real-world variations (including variations in head position and imaging perspective) that relieve the training samples of the burden of additional data collection. The augmented image does not distort medically relevant features of the brain MRI because each of the augmented images retains the structural and spatial features of the brain MRI.

Data Splicing: To ensure fair assessment the dataset was separated into training (70%) and validation (10%) and testing (20%) sets through stratified sampling to ensure that classes were balanced. Table 3 shows the manner in which the data are distributed for both the classes.

Table 3. Distributed Dataset

Class	**Train**	**Test**	**Validation**
Yes (Tumor)	1050	300	150
No (No-Tumor)	1050	300	150

3.3 Proposed SD-ViT Model Architecture

The proposed architecture combines the principles of the Vision Transformer (ViT) and the Self-Distillation that would provide better feature representation.

Patch Embedding Layer: ViT divides an image $x \in R^{H \times W \times C}$ into regular-sized patches, all flattened into series of embeddings:

$$z_0 = [x_1E, x_2E, x_3E; \cdots; x_NE] + E_{pos} \quad (2)$$

where E represents the learnable linear projection, E_{pos} represents positional embeddings and $N = \frac{HW}{P^2}$ is the thickness of the patches.

Transformer Encoder Block: Every embedding undergoes several Transformer Encoder layers, which consist of Multi-Head Self-Attention (MHSA) and Feed-Forward Networks (FFN). MHSA enables the model to acquire contextual dependency

between remote spatial areas within the image. The fundamental attention process is specified as:

$$A(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{D_K}}\right)V \quad (3)$$

Q, K, V are query, key and value matrices which are the result of the input features. Using several layers of such transformers, the model can solve the local tissue structures in addition to long distances that are essential in the accurate localization of tumors.

Self-distillation Module: The self-distillation mechanism is the main novelty of the proposed model, and it allows transferring knowledge between the intermediate layers of the same model. The deeper layers of the transformer serve as the teacher, and the previous layers as the student as opposed to utilizing a stand-alone teacher net. In order to train efficiently and avoid overfitting with few medical data points, we used Distillation Through Attention - teacher model instructs a student model by transferring knowledge in the form of attention as suggested by Touvron et al. [10]. Parameters Exponential Moving Average (EMA) is used to update the weights of the teacher with the student parameters:

$$\theta_t^{(teacher)} = \propto \theta_{t-1}^{(teacher)} + (1-\propto)\theta_t^{(student)} \quad (4)$$

where α EMA decay factor is denoted by α (=0.999, in this paper).

The overall loss is a sum of supervised binary cross-entropy loss and distillation loss:

$$\mathcal{L}_{total} = \lambda_{sup}\mathcal{L}_{BCE(y,\hat{y})} + \lambda_{distill}\mathcal{L}_{KD(p_t,p_s)} \quad (5)$$

where λ_{sup} and $\lambda_{distill}$ are weighting factors, p_t and p_s represent the logits of teacher and student respectively. This hybrid training will enable the student ViT to take advantage of the soft-target distributions of the teacher, enhancing the stability and generalization. where $\mathcal{L}_{BCE}$ represents Binary Cross Entropy, $\mathcal{L}_{KD}$ denotes Kullback–Leibler Divergence between teacher–student outputs [13–15].

Classification Head: The completed all blocks of transformer encoder processing pass the [CLS] token of the whole image to a fully connected layer and a Softmax activation function to classify to binary. The last stage is the one that outputs the possibility that the input image is either of the Tumor or No Tumor classes.

Training Setup: The model was trained on a GPU environment with the following hyperparameters TensorFlow 2.19.0. Early stopping and learning rate schedule were used to avoid overfitting, which was done depending on the accuracy of validation. Table 4 represents the training setup used for our proposed methodology.

Table 4. Training Parameter and Value.

Parameter	Value
Learning Rate	1e-4 (with cosine decay)
Batch Size	32
Epochs	50
Loss Function	Cross-Entropy + Self-Distillation Loss
Dropout	0.1

4 Experimental Result Analysis

4.1 Evaluation metrics

Several common classification metrics were used to fully analyze the performance of the proposed Self-Distilled Vision Transformer (SD-ViT) model, based on the confusion matrix of each model. These metrics are: Accuracy, Precision, Recall (Sensitivity) and F1-Score which are defined as follows:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{6}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{7}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{8}$$

$$\text{F1 Score} = 2 * \frac{\text{Precision}*\text{Recall}}{\text{Precision+Recall}} \tag{9}$$

TP, TN, FP and FN are True Positives, True Negatives, False Positives, and False Negatives, respectively.
These measures were calculated in both classes (Tumor and No Tumor) to give a balanced analysis of the discriminative ability of the models.

4.2 Performance Evaluation

The proposed study was carried out as an experiment to determine the performance of three models, ResNet50, Baseline Vision Transformer (ViT), and the Proposed Self-Distilled Vision Transformer (SD-ViT) in binary classification of brain tumors. Both models were trained and tested using the same data to be fairly compared. Several evaluation metrics such as accuracy, precision, recall, and F1-score were used to analyze the performance, and confusion matrices were used to gain an in-depth insight into the class-wise predictions of the models.
In general, the findings indicate a definite performance improvement of traditional CNN-based models in favor of transformer models, with the proposed SD-ViT showing the highest results in all of the metrics. The quantitative findings are presented in Table 5, and the confusion matrices are demonstrated in Fig. 4.

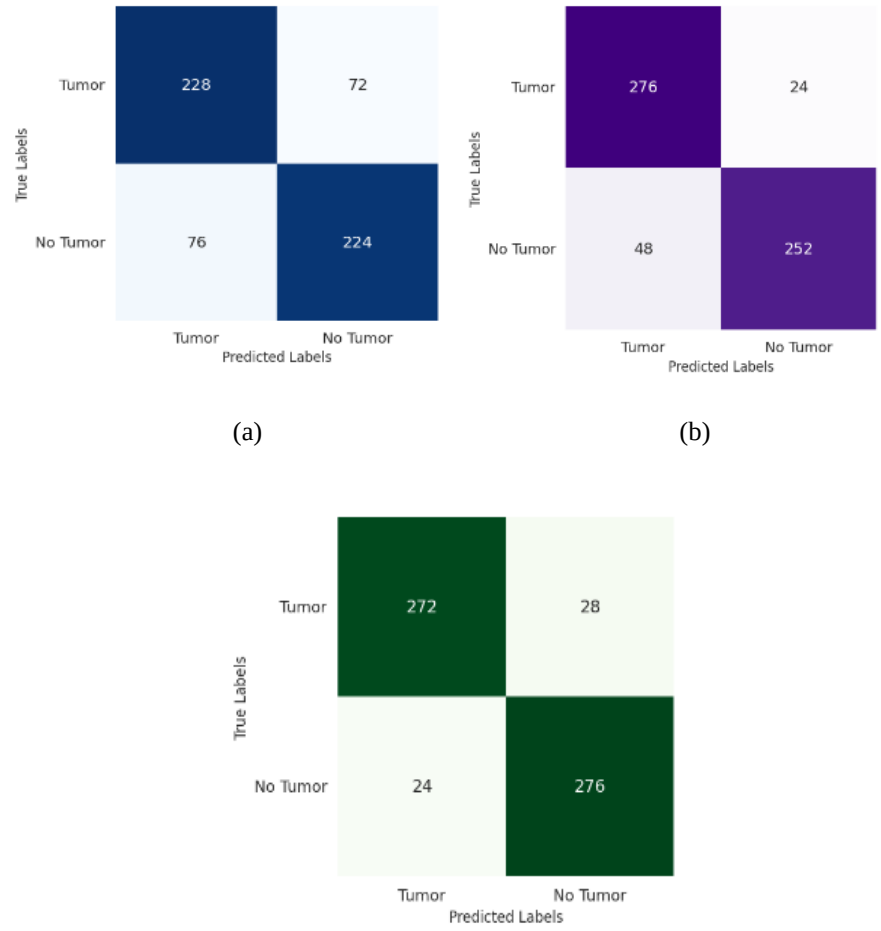

Fig.4. Confusion Matrix (a) ResNet50 (b) Baseline ViT (C) SD-ViT

Based on the findings, it is clear that the ResNet50 model can perform moderately with an accuracy of 75 and hence it has limited capacity of generating high-level contextual features required to differentiate between subtle tumor boundaries. The confusion matrix reveals that there was a lot of misclassifications between tumor and non-tumor cases, meaning that convolutional filters did not sufficiently provide the extracts of the global dependencies among the MRI slices.

The Baseline ViT which utilizes the attention mechanisms, greatly increased the classification performance up to 88% accuracy. The results of its greater recall and F1-score indicate that the transformer-based representation learning was more efficient in the sense that it was able to capture spatial relationships in medical images. Nonetheless, certain mis-classifications were still observed especially when a region of

the tumor had irregular shapes or the intensity overlapped with normal tissue which means that the internal representation of the model can still be improved.

Table 5. Performance Metrics Comparison.

Model	Precision	Recall	F1Score	Accuracy
ResNet50	0.75	0.75	0.75	0.75
Baseline ViT	0.88	0.88	0.88	0.88
Proposed SD-ViT	**0.91**	**0.91**	**0.91**	**0.91**

The Proposed SD-ViT performed better than the previous two models with an accuracy of 91, with the same amount of accuracy, recall, and F1-score values of both classes. This is because it has been enhanced by the incorporation of self-distillation in the ViT framework, which enables the model to self-transfer knowledge in lower layers using the information at the higher levels. A consequence of this is that SD-ViT features better inter-layer feature coherence and higher discriminative power. The confusion matrix also substantiates the fact it is more able to correctly classify the tumor and non-tumor cases with only a few misclassifications.

These results support the idea that the proposed SD-ViT is able to overcome the weaknesses of previous models in that it improves the representation of features, the distance between the classes, and the learning dynamics. The good sensitivity/specificity ratio in the model demonstrates that it is a reliable tool in supporting clinical diagnosis and is a more interpretable and effective model than traditional CNN and transformer architectures in brain tumor classification.

5 Conclusion & Discussion

The Self-Distilled Vision Transformer (SD-ViT) is a new proposal that addresses the problem of brain tumor classification by providing a self-distillation mechanism that is included in the transformer. This natural self-distillation increases feature refinement and brings about better generalization across complicated tumor pattern. This model had an accuracy of 91% which was higher than ResNet50 (75%) and the baseline ViT (88%) which showed that the model was better at detecting subtle differences between tumor and non-tumor regions. This progress brings out the novelty of incorporating self-distillation into transformer layers, which have created a more robust, efficient and interpretable medical image analysis architecture.

In the future, SD-ViT can be implemented in multi-class tumor classification to be able to sort different types of tumors like glioma, meningioma, and pituitary tumors. MRI sequences to incorporate multi-modes (e.g., T1, T2, and FLAIR) could be used to complementary images to further increase diagnostic accuracy. It can also be adapted as semi-supervised or self-supervised learning to utilize a large amount of unlabeled medical data. In addition, the need to include explainability tools and apply SD-ViT to multi-center clinical datasets will make it more interpretable, scalable, and applicable in real-life scenarios.

Although the proposed SD-ViT has a good performance, it has various limitations. The present study is interested in binary classification, which makes the real clinical conditions simpler. The sample size is quite small and balanced and it might not reflect the diversity of MRI data in actual conditions. Also, the model may be sensitive to variation in the MRI acquisition conditions and image noise. Even though self-distillation enhances internal representation learning, it does not directly deal with domain adaptation between scanners or institutions. By tackling such limitations in future studies, the model can be made more robust, have a higher degree of generalizability, and increase clinical reliability.

References

1. Bauer, S., Wiest, R., Nolte, L.-P., Reyes, M.: A survey of MRI-based medical image analysis for brain tumor studies. *Phys. Med. Biol.* 58(13), R97–R129 (2013). https://doi.org/10.1088/0031-9155/58/13/R97
2. Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C.: N4ITK: improved N3 bias correction. *IEEE Trans. Med. Imaging* 29(6), 1310–1320 (2010). https://doi.org/10.1109/TMI.2010.2046908
3. Rehman, A., Naz, S., Razzak, M.I., Akram, F., Imran, M.: A deep learning-based framework for automatic brain tumor classification using transfer learning. *Circuits Syst. Signal Process.* 39(2), 757–775 (2019). https://doi.org/10.1007/s00034-019-01246-3
4. Swati, Z.N.K., Zhao, Q., Kabir, M.H., Ali, F., Feng, Q., El-Baz, A.: Brain tumor classification for MR images using transfer learning and fine-tuning. *Comput. Med. Imaging Graph.* 75, 34–46 (2019). https://doi.org/10.1016/j.compmedimag.2019.05.001
5. Sultan, H.H., Salem, N.M., Al-Atabany, W.: Multi-classification of brain tumor images using deep neural network. *IEEE Access* 7, 69215–69225 (2019). https://doi.org/10.1109/ACCESS.2019.2919122
6. Jamshidi, B., Rostamy-Malkhalifeh, M.: Using VGG19 transfer learning to diagnose and classify brain tumors based on CNN, and predict with deep learning-based ANN transfer learning via MR images. *SSRN Electron. J.* (2023). https://doi.org/10.2139/ssrn.4624936
7. Kumar, S., Choudhary, S., Jain, A., Singh, K., Ahmadian, A., Bajuri, M.Y.: Brain tumor classification using deep neural network and transfer learning. *Brain Topogr.* (2023). https://doi.org/10.1007/s10548-023-00953-0
8. Srinivasan, S., Francis, D., Mathivanan, S.K., Rajadurai, H., Shivahare, B.D., Shah, M.A.: A hybrid deep CNN model for brain tumor image multi-classification. *BMC Med. Imaging* 24(1), 1–18 (2024). https://doi.org/10.1186/s12880-024-01195-7
9. Dosovitskiy, A. et al.: An image is worth 16×16 words: transformers for image recognition at scale. *arXiv preprint* arXiv:2010.11929 (2020).
10. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A.: Do vision transformers see like convolutional neural networks? *arXiv preprint* arXiv:2108.08810 (2021).
11. Liu, X., Hu, Y., Chen, J.: Hybrid CNN–Transformer model for medical image segmentation with pyramid convolution and multi-layer perceptron. *Biomed. Signal Process. Control* 86, 105331 (2023). https://doi.org/10.1016/j.bspc.2023.105331
12. Saraei, M., Liu, S.: Attention-based deep learning approaches in brain tumor image analysis: a mini review. *Front. Health Inform.* 12, 164 (2023). https://doi.org/10.30699/fhi.v12i0.493
13. Touvron, H. et al.: Training data-efficient image transformers & distillation through attention. *arXiv preprint* arXiv:2012.12877 (2020).

14. Dai, Y., Gao, Y.: TransMed: transformers advance multi-modal medical image classification. *arXiv preprint* arXiv:2103.05940 (2021).
15. Zhang, L., Bao, C., Ma, K.: Self-distillation: towards efficient and compact neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.* (2021). https://doi.org/10.1109/TPAMI.2021.3067100
16. Selvaraju, R.R. et al.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: *ICCV*, 618–626 (2017). https://doi.org/10.1109/ICCV.2017.74
17. Pacal, I.: A novel Swin transformer approach utilizing residual multi-layer perceptron for diagnosing brain tumors in MRI images. *Int. J. Mach. Learn. Cybern.* (2024). https://doi.org/10.1007/s13042-024-02110-w
18. Ahmed Hamada: Br35H – Brain tumor detection (2020). Kaggle. https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection