2024 27th International Conference on Computer and Information Technology (ICCIT),
20-22 December 2024, Cox's Bazar, Bangladesh.

An Exploratory Approach Towards Investigating and Explaining Vision Transformer and Transfer Learning for Brain Disease Detection

Shuvashis Sarker*, Shamim Rahim Refat†, Faika Fairuj Preotee‡,
Shifat Islam§, Tashreef Muhammad¶, Mohammad Ashraful Hoque‖,
*†‡§¶‖Department of Computer Science and Engineering,
*†‡Ahsanullah University of Science and Technology,
§Bangladesh University of Engineering and Technology, ¶‖Southeast University,
Email:{*shuvashisofficial, †n.a.refat2000, ‡faikafairuj2001, §shifat.islam.buet, ¶tashreef.muhammad}@gmail.com
‖ashraful@seu.edu.bd.

***Abstract*—The brain is a highly complex organ that manages many important tasks, including movement, memory and thinking. Brain-related conditions, like tumors and degenerative disorders, can be hard to diagnose and treat. Magnetic Resonance Imaging (MRI) serves as a key tool for identifying these conditions, offering high-resolution images of brain structures. Despite this, interpreting MRI scans can be complicated. This study tackles this challenge by conducting a comparative analysis of Vision Transformer (ViT) and Transfer Learning (TL) models such as VGG16, VGG19, Resnet50V2, MobilenetV2 for classifying brain diseases using MRI data from Bangladesh based dataset. ViT, known for their ability to capture global relationships in images, are particularly effective for medical imaging tasks. Transfer learning helps to mitigate data constraints by fine-tuning pre-trained models. Furthermore, Explainable AI (XAI) methods such as GradCAM, GradCAM++, LayerCAM, ScoreCAM, and Faster-ScoreCAM are employed to interpret model predictions. The results demonstrate that ViT surpasses transfer learning models, achieving a classification accuracy of 94.39%. The integration of XAI methods enhances model transparency, offering crucial insights to aid medical professionals in diagnosing brain diseases with greater precision.**

***Index Terms*—Vision Transformer (ViT), Transfer Learning (TL), Explainable AI (XAI), Magnetic Resonance Imaging (MRI), State-of-the-Art (SOTA).**

I. Introduction

The human brain serves as the central nervous system, directing and coordinating all physiological processes and functions. Brain diseases, including tumors and neurodegenerative conditions, pose substantial hurdles in terms of both diagnosis and treatment. Magnetic Resonance Imaging (MRI) has emerged as a crucial diagnostic tool for identifying brain disorders, providing highly detailed images of soft tissues and enabling specialists to detect abnormalities with remarkable precision. However, the complex nature of brain structures makes it challenging to interpret MRI data accurately.

In recent years, advancements in Artificial Intelligence (AI), particularly deep learning techniques, have revolutionized medical imaging. Vision Transformers (ViT) outperform in medical image analysis by leveraging self-attention processes to capture global dependencies, rendering them suitable for applications like as brain tumor classification. However, ViT requires extensive datasets, posing a challenge in areas with constrained medical imaging data, such as Bangladesh [1]. Transfer learning (TL) mitigates this issue by refining pre-trained models on limited medical datasets, thereby strengthening brain disease identification and decreasing the necessity for considerable expert involvement, which ultimately improves clinical decision-making.

This study employs Vision Transformers (ViT) to diagnose brain disorders using MRI data from a Bangladeshi population. Transfer learning is utilized to address data shortages by refining pre-trained models, hence enhancing performance with less training data. Explainable AI (XAI) techniques are employed to identify critical MRI regions that affect predictions, hence improving model transparency. The research compares the efficacy of ViT and transfer learning models to determine their robustness and generalizability in medical imaging. The primary objectives of this study are:

i. To develop a brain MRI classification framework using a localized Bangladeshi dataset collected from the National Institute of Neurosciences & Hospital (NINS).
ii. To improve classification accuracy by incorporating the Vision Transformer (ViT) model, which has not been previously applied to this dataset, leveraging its self-attention processes for enhanced feature extraction.
iii. To compare the performance of ViT with other transfer learning models to ensure robustness and evaluate their effectiveness on medical imaging.
iv. To apply multiple Explainable AI (XAI) techniques to enhance the interpretability of model predictions, facilitating clearer insights into the classification process.

The remainder of this paper is structured as follows: Section II discusses the related work, while Section III outlines the proposed methodology, and Section IV presents the analysis of the results. Techniques related to Gradient Based Explainable AI are detailed in Section V, Limitations and Future Works are described in Section VI and concluding remarks are provided

979-8-3315-1909-4/24/$31.00 ©2024 IEEE

in Section VII.

II. Related Works

A. Vision Transformer(ViT)

In recent years, Vision Transformers (ViTs) have gained significant attention for medical image classification tasks. Van Dongen [2] explored the use of Vision Transformers (ViTs) for brain tumor classification using MRI images. It tested four ViT models (ViT-B/16, ViT-B/32, ViT-L/16, and ViT-L/32) and introduced a "model soup" approach, averaging the weights of multiple fine-tuned models. The best individual model, ViT-L/32, achieved 95.31% accuracy, while the Combi Soup of ViT-L/16 models outperformed it with 96.12% accuracy in validation. A new way to classify brain tumors was created by Jaffar [1] that combines iResNet and Vision Transformers (ViTs). The model used iResNet to get local feature extraction and ViTs to get global environmental information. An attention-based feature merge module optimized the accuracy of classification. The suggested model was 99.2% accurate, which was much better than other models like InceptionV3, ResNet, and DenseNet. Similarly, Aloraini et al. [3] introduced a hybrid Transformer-Enhanced Convolutional Neural Network (TECNN) for brain tumor classification using MRI images. The model combines CNN for local feature extraction and transformers for global context. It outperformed several state-of-the-art methods, achieving 96.75% accuracy on the "*BraTS 2018*" dataset and 99.10% accuracy on the "*Figshare*" dataset. Krishnan et al. [4] presented a Rotation Invariant Vision Transformer (RViT) for the classification of brain tumors utilizing MRI data. The model adeptly managed changes in image orientation by including rotational patch embeddings. The RViT surpassed baseline ViTs and other leading methodologies, with an accuracy of 98.6% with flawless sensitivity and elevated specificity.

B. Transfer Learning with XAI

Transfer learning utilizing Explainable AI(XAI) integrates the efficacy of pre-trained models and fine-tuning with interpretative strategies. Zeineldin et al. [5] introduced the NeuroXAI framework to enhance the explainability of deep learning models for brain tumor analysis. Using XAI techniques like Grad-CAM and SmoothGrad, they provided visual explanations for model predictions. Their ResNet-50 model achieved 98.62% accuracy, while the segmentation model demonstrated strong performance with a 92% Dice score for tumor segmentation. Moreover, Hasan et al. [6] created a brain tumor classification system utilizing transfer learning models on the BrainTumorInSight dataset. The ResNet50 model attained an accuracy of 96.3%, utilizing LIME to offer visual elucidation of critical MRI regions that affect the model's determinations. Another researcher Shad et al. [7] employed a pipeline of ResNet-50, VGG16, and Inception V3 to categorize phases of Alzheimer's disease with T1-weighted MRI images from a hybrid Kaggle dataset. The maximum categorical accuracy of 86.82% was attained with VGG16, with XAI techniques such as LIME applied for interpretability. Furthermore, Mahmud et al. [8] developed a model for Alzheimer's disease diagnosis using the OASIS-2 MRI dataset and transfer learning with VGG16 and DenseNet169. The model achieved 96% accuracy, leveraging saliency maps and Grad-CAM for enhanced interpretability and clinical application.

III. Methodology

A. Dataset

The Dataset which we have utilized in this study was jointly developed by the National Institute of Neuroscience (NINS), Bangladesh, and the Computer Science and Engineering department of the University of Dhaka [9]. This Dataset is publicly available at Figshare [10] for further study. The dataset consists of 5,285 T1-weighted contrast-enhanced brain MRI images, with various pixel resolution. This dataset is

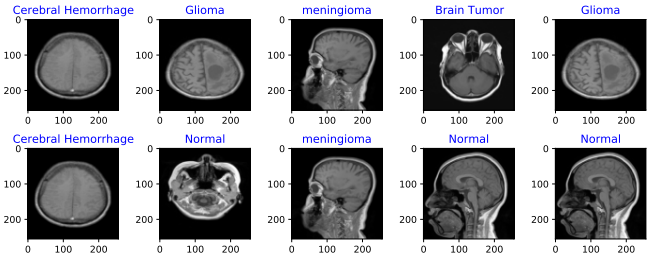


Fig. 1: Sample Images of the dataset.

visually represented in Figure 1, which showcases a selection of random images from the dataset. These images are classified into 37 distinct categories. The extensive variety of categories in the dataset facilitates the creation of sophisticated diagnostic algorithms. This invaluable resource plays a crucial role in the training and validation of machine learning models that utilize computational techniques in medical imaging to enhance the precision and effectiveness of neurological diagnostics.

Figure 2 depicts a bar plot that visually illustrates the dataset comprising 37 distinct classes, effectively depicting the quantity of MRI pictures for each class.

B. Data Preprocessing

In the data preprocessing stage, we have at first selected brain MRI images from the dataset to ensure consistency in how the model processes the images during training and testing. To rectify class imbalance, we have employed SMOTE [11], whereby the minority classes are augmented to align with the mean sample count across all classes, determined by averaging the complete dataset. Figure 3 provides a comprehensive overview of the data preprocessing steps which are involved in this process.

Next, we have implemented a data augmentation pipeline using TensorFlow to introduce variability in the training set. This process includes normalization, resizing the images to 128x128, random horizontal flipping, random rotations with a reduced factor of 0.01, and random zooming with height and width factors of 0.05. These augmentation techniques help the

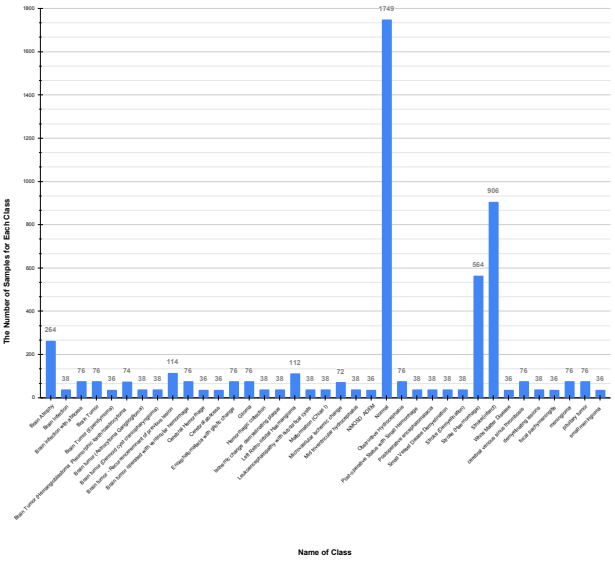


Fig. 2: Distribution of Classes in the MRI Dataset

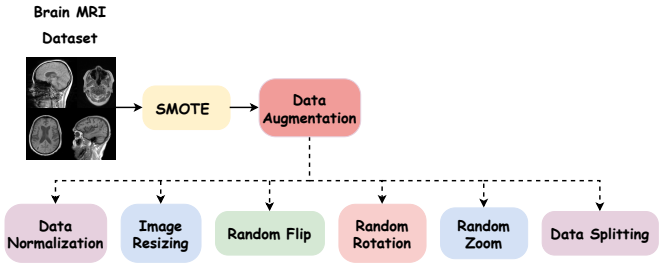


Fig. 3: Overview of the data preprocessing steps

model generalize better by exposing it to diverse versions of the images. Finally, the dataset is split into training, validation, and test sets with an 80:10:10 ratio to ensure balanced and unbiased evaluation of the model's performance.

C. Model Architecture

Figure 4 illustrates the process of the suggested methodology. This demonstrates a comparative analysis between TL models and ViT, with the application of Explainable AI (XAI) techniques on the transfer learning models.

This study utilizes a range of pre-trained models grounded in CNN architectures, such as VGG16, VGG19, ResNet50V2 and MobileNetV2 to enhance classification performance. These models, initially trained on ImageNet, are subsequently fine-tuned for the diagnosis of brain diseases utilizing MRI data. Each model has unique benefits: VGG models employ 3x3 filters for detailed feature extraction, ResNet50V2 integrates residual connections to mitigate vanishing gradients, captures multi-scale features, improves efficiency through depthwise separable convolutions, and MobileNetV2 is designed for optimal performance on resource-constrained devices [12]. The task is effectively modified for brain disease classification by fine-tuning pre-trained models, thereby retaining features acquired from large datasets. The architecture includes a flattening layer, dense layers with 1024 and 512 neurons, dropout layers to reduce overfitting and a softmax output for classification. Fine-tuning enhanced model efficacy and diminished training duration by utilizing pretrained ImageNet weights. This method has competently addressed the complexity of brain MRI categorization with enhanced efficiency.

The Vision Transformer (ViT) [13], analyzes images by segmenting them into smaller patches, similarly to how the original transformer processes words in natural language processing . In contrast to conventional transformer models that employ both an encoder and a decoder, ViT exclusively utilizes the encoder. An input image I is represented as $\mathbb{R}^{H\times W\times C}$, where H, W, and C correspond to the height, width, and number of channels of the image, respectively. The image is divided into N patches, each of size $P\times P\times C$, where:$N = \frac{HW}{P^2}$. These patches are then flattened and embedded with positional information to preserve their spatial arrangement. Figure 5 shows the architechture of the Vision Transformer(ViT). A trainable classification token is incorporated into the patches, enabling final classification by a Multilayer Perceptron (MLP) head. The combined patches and positional embeddings are fed into the transformer encoder, which alternates between Multi-head Self-Attention (MSA) layers and MLP blocks. The MSA enables the model to discern relationships among patches, whereas the MLP has fully connected layers activated by the Gaussian Error Linear Unit (GELU). The MSA outputs are derived by concatenating several attention heads, followed by processing through a final linear layer. The ViT's capacity to collect both local and global information in images renders it appropriate for applications such as medical imaging and object recognition, where comprehension of intricate details and long-range connections is essential. This method renders ViT a viable substitute for convolutional neural networks (CNNs) in diverse picture categorization tasks [13].

This study has employed a state-of-the-art ViT model with critical hyperparameters optimized for peak performance. The batch size of 256 is used here. Input images (128x128 pixels) are segmented into 64x64 patches and mapped into a 64-dimensional space. The model comprises eight transformer layers, each with four attention heads, while the MLP head incorporates two dense layers with 2048 and 1024 units respectively. To prevent overfitting, we have implemented a dropout rate of 0.1 in the transformer layers and 0.5 prior to the final classification layer.

IV. Experimental Setup & Result Analysis

A. Experimental Setup

The experimental configuration includes a machine featuring an Intel Core i5 13400F processor, an NVIDIA GeForce RTX 3060 12GB GPU, and 16GB of DDR4 RAM. All models are executed with Keras [14], a high-level API for neural networks, with TensorFlow [15] serving as the backend. To guarantee robustness, each model is executed five times, with

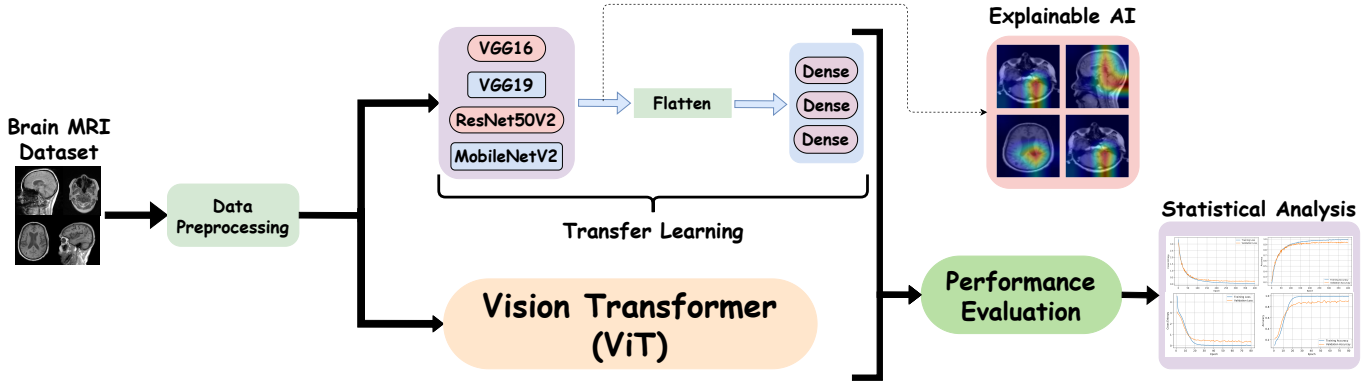


Fig. 4: Proposed Model Architecture

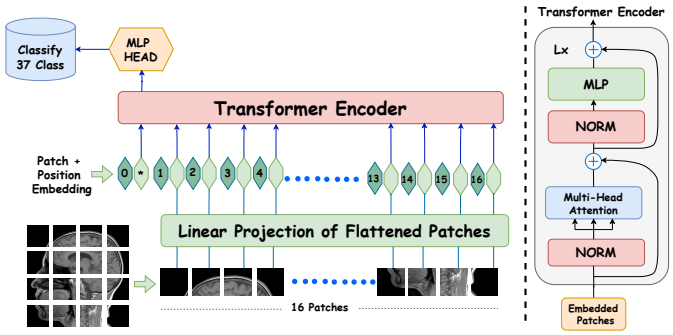


Fig. 5: Vision Transformer Architecture

TABLE I: Performance Metrics for Different Models

Model	Accuracy	Precision	Recall	F1 Score
VGG16	0.8844	0.9103	0.8854	0.8947
VGG19	0.8905	0.8951	0.9108	0.8995
Resnet50V2	0.9141	0.9276	0.9389	0.9211
MobilenetV2	0.8856	0.8953	0.8938	0.8915
ViT	**0.9439**	**0.9651**	**0.9644**	**0.9638**

TABLE II: Comparison of Our Proposed Approach with Existing Methods

Authors	Approach	Accuracy	Use of XAI
Brima et al. [9]	Transfer Learning of ResNet50	0.87	NO
	Transfer Learning of ResNet50 with Augmented Data	0.81	
	Fine-tuning ResNet50 with Augmented Data	0.84	
Ahmed et al. [16]	Transfer Learning of MobileNetV3 with Augmented Data	0.92	NO
Our Proposed Method	**ViT**	**0.94**	**NO**
	ResNet50V2	**0.91**	**YES (Gradient-Based)**

the optimal results chosen for analysis. For the Vision Transformer, the learning rate is set to 1×10^{-4}, using the Adam optimizer, over 400 epochs, and the loss function applied was Sparse Categorical Crossentropy. In the TL models, a learning rate of 1×10^{-5} have been employed, with Adam as the optimizer and an identical loss function.

B. Result Analysis

The results from the study clearly demonstrate the superior performance of the Vision Transformer (ViT) model compared to traditional convolutional models like VGG16, VGG19, ResNet50V2 and MobileNetV2. Table I indicates that ViT has attained the maximum accuracy of 94.39%, surpassing all other models, with ResNet50V2 following at 91.41%. The transformer-based architecture of ViT enables the capturing of long-range dependencies in data, offering a considerable advantage in feature extraction and model generalization. This is especially evident compared to convolutional models that rely more on local feature detection. The superior accuracy of ViT highlights its ability to better understand complex patterns in the dataset, leading to more precise predictions.

The training and validation loss curves, shown in Figure 7, further emphasize the efficiency of ViT. ViT not only converges faster during training but also maintains a lower validation loss compared to ResNet50V2. This indicates that ViT generalizes better to unseen data, with a reduced risk of overfitting. In contrast, the higher validation loss of ResNet50V2 suggests that it struggles more with generalization. Table II presents a comparative analysis of our proposed approach with existing methods in the field.

Overall, the results confirm that the Vision Transformer (ViT) significantly outperformed traditional CNN models like VGG16, VGG19, ResNet50V2 and MobilenetV2 achieving the highest accuracy of 94.39%. ViT's transformer-based architecture enables better generalization and feature extraction, especially in complex image classification tasks, where it excels in capturing long-range dependencies. In contrast, while CNN models show solid performance, they have struggled more with generalization and are less effective in capturing global features across the dataset.

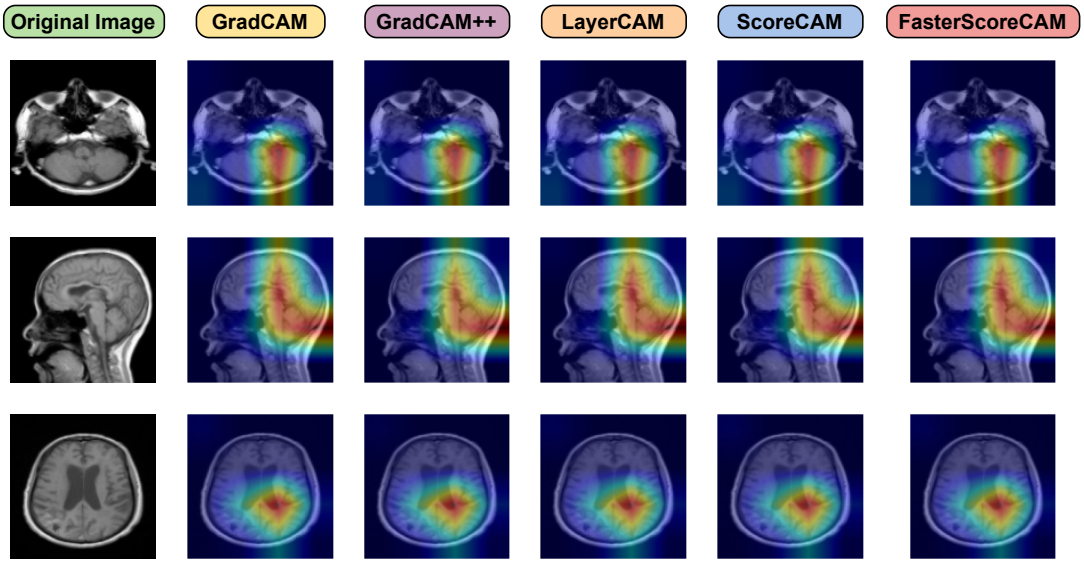
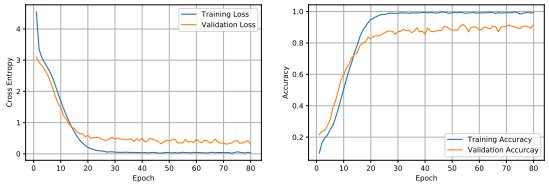
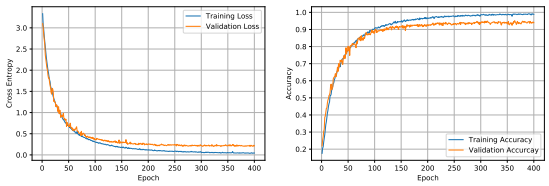


Fig. 6: Gradinat Based XAI Visualization of Resnet50V2



(a) Training and Validation Loss for ResNet50V2



(b) Training and Validation Loss for Vision Transformer (ViT)

Fig. 7: Comparison of Training and Validation Loss Curves between ResNet50V2 and Vision Transformer (ViT)

V. Gradient Based Explainable AI

XAI approaches aim to enhance the interpretability of AI models by revealing the mechanisms through which transfer learning models process and analyze data to generate predictions. **GradCAM**, a pivotal technology in XAI, offers visual elucidations for convolutional neural networks (CNNs) by employing target class gradients to pinpoint significant picture regions that affect predictions. It produces a heatmap to emphasize these regions, enhancing model transparency and interpretability without modifying the architecture, rendering it beneficial for tasks such as picture categorization, Visual Question Answering(VQA), and captioning [17]. **GradCAM++** enhances this by overcoming constraints in identifying multiple instances of the same category. It integrates second-order derivatives, yielding enhanced object boundaries and augmenting heatmap precision. Conversely, **ScoreCAM** employs a gradient-free methodology, generating class-specific maps by assessing the impact of alterations in the activation maps on the resulting score. This technique improves visual clarity and is particularly helpful in pinpointing essential areas for picture categorization jobs. **FasterScore-CAM** enhances the procedure by concentrating on the most active channels within feature maps. This diminishes processing requirements while preserving interpretability, proving it beneficial for extensive datasets and intricate models. **LayerCAM**, an advanced methodology, produces class activation maps using multiple layers of the CNN, rather than exclusively from the terminal convolutional layer. This enhances comprehension of spatial localization and object characteristics, consequently augmenting the accuracy of both object localization and classification [18].

Gradient-based XAI techniques encounter difficulties with ViT due to architectural imbalances, including the absence of convolutional layers. ViT processes images as patch sequences and rely on self-attention mechanisms, which makes it more difficult to apply XAI methods on spatial feature maps and gradients. ViT analyzes images in segments, complicating the application of traditional CAM techniques. Furthermore, interpretability may be hampered by attention scores in ViT that do not correspond to human-intuitive relevance [19]. For this reason, instead of ViT five different XAI approaches are used in ResNet50V2 model, which has outperformed all other TL models in evaluating brain MRI data.

Figure 6 highlights the consistent activation patterns seen across the ResNet50V2 model. This model demonstrates that GradCAM and GradCAM++ consistently highlight critical areas, whereas LayerCAM provides a broader yet still focused depiction.These methods employ feature maps, generally ob-

tained from the intermediate or final layers preceding the flattening stage. GradCAM consistently emphasizes significant activations in essential brain regions, whereas GradCAM++ provides more concentrated and accurate heatmaps. LayerCAM records extensive activations while maintaining focus on critical regions. ScoreCAM allocates attention more uniformly throughout the brain image. In contrast, FasterScoreCAM concentrates on particular regions exhibiting elevated confidence levels. These methods consistently indicate significant regions, demonstrating that the models reliably recognize alike areas of interest across several XAI methodologies. XAI dramatically improves the reliability and interpretability of AI models through improved visual explanations, which is essential in vital domains like brain disease detection.

VI. Limitations and Future Work

Despite the strength of the Vision Transformer (ViT) in classifying 37 distinct brain MRI classes, there are still areas for future work. One such area is use of Generative Adversarial Networks (GANs), which could be employed to generate synthetic MRI images and further test the model's robustness, ensuring it generalizes well to more varied scenarios. Although the dataset is diverse and comprehensive, incorporating GANs could help justify the model's true accuracy by challenging it with additional data variations. Practical implementation in clinical settings also presents challenges, such as computational resource demands and integration with existing hospital workflows. Future efforts will focus on incorporating GANs for data augmentation and validation, exploring hybrid architectures that blend the strengths of ViTs and CNNs, and optimizing the model for real-time deployment in healthcare environments.

VII. Conclusion

In conclusion, this study demonstrates the significant potential of Vision Transformers (ViTs) for brain MRI classification, particularly in capturing global dependencies and improving overall accuracy in medical imaging tasks. By applying transfer learning and utilizing explainable AI (XAI) techniques, ViTs have been shown to outperform traditional CNN-based models like VGG16, VGG19, and ResNet50V2, achieving superior accuracy and generalization capabilities. The integration of XAI methodologies provided crucial insights into model predictions, making the classification process more transparent and interpretable. These findings underscore the importance of transformer-based architectures in medical image analysis and highlight the effectiveness of explainable AI in ensuring the reliability and trustworthiness of deep learning models in critical applications like brain disease diagnosis.

References

[1] A. Y. Jaffar, "Combining local and global feature extraction for brain tumor classification: A vision transformer and iresnet hybrid model," *Engineering, Technology & Applied Science Research*, vol. 14, no. 5, pp. 17011–17018, 2024.

[2] I. VAN DONGEN, *COMPARISON OF INDIVIDUAL VISION TRANSFORMERS AND MODEL SOUPS FOR BRAIN TUMOR CLASSIFICATION ON MAGNETIC RESONANCE IMAGES*. PhD thesis, tilburg university.

[3] M. Aloraini, A. Khan, S. Aladhadh, S. Habib, M. F. Alsharekh, and M. Islam, "Combining the transformer and convolution for effective brain tumor classification using mri images," *Applied Sciences*, vol. 13, no. 6, p. 3680, 2023.

[4] P. T. Krishnan, P. Krishnadoss, M. Khandelwal, D. Gupta, A. Nihaal, and T. S. Kumar, "Enhancing brain tumor detection in mri with a rotation invariant vision transformer," *Frontiers in Neuroinformatics*, vol. 18, p. 1414925, 2024.

[5] R. A. Zeineldin, M. E. Karar, Z. Elshaer, ·. J. Coburger, C. R. Wirtz, O. Burgert, and F. Mathis-Ullrich, "Explainability of deep neural networks for mri analysis of brain tumors," *International journal of computer assisted radiology and surgery*, vol. 17, no. 9, pp. 1673–1683, 2022.

[6] S. Hasan, M. M. Nabila, R. B. Anis, and R. Rab, "Deep learning-based model with xai for brain tumor classification and segmentation using mri images," in *2023 IEEE 9th International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)*, pp. 1–6, IEEE, 2023.

[7] H. A. Shad, Q. A. Rahman, N. B. Asad, A. Z. Bakshi, S. F. Mursalin, M. T. Reza, and M. Z. Parvez, "Exploring alzheimer's disease prediction with xai in various neural network models," in *TENCON 2021-2021 IEEE Region 10 Conference (TENCON)*, pp. 720–725, IEEE, 2021.

[8] T. Mahmud, K. Barua, S. U. Habiba, N. Sharmen, M. S. Hossain, and K. Andersson, "An explainable ai paradigm for alzheimer's diagnosis using deep transfer learning," *Diagnostics*, vol. 14, no. 3, p. 345, 2024.

[9] Y. Brima, M. H. K. Tushar, U. Kabir, and T. Islam, "Deep transfer learning for brain magnetic resonance image multi-class classification," *arXiv preprint arXiv:2106.07333*, 2021.

[10] Y. Brima, M. H. K. Tushar, U. Kabir, and T. Islam, "Brain mri dataset." https://doi.org/10.6084/m9.figshare.14778750.v2, 2021.

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[12] A. W. Salehi, S. Khan, G. Gupta, B. I. Alabduallah, A. Almjally, H. Alsolai, T. Siddiqui, and A. Mellit, "A study of cnn and transfer learning in medical imaging: Advantages, challenges, future scope," *Sustainability*, vol. 15, no. 7, p. 5930, 2023.

[13] S. Tummala, S. Kadry, S. A. C. Bukhari, and H. T. Rauf, "Classification of brain tumor from magnetic resonance imaging using vision transformers ensembling," *Current Oncology*, vol. 29, no. 10, pp. 7498–7511, 2022.

[14] F. Chollet *et al.*, "Keras." https://keras.io, 2015.

[15] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software available from tensorflow.org.

[16] A. F. Majeed, P. Salehpour, L. Farzinvash, and S. Pashazadeh, "Multiclass brain lesion classification using deep transfer learning with mobilenetv3," *IEEE Access*, 2024.

[17] M. H. Rafi, S. Islam, S. H. I. Labib, S. S. Hasan, F. M. Shah, and S. Ahmed, "A deep learning-based bengali visual question answering system," in *2022 25th International Conference on Computer and Information Technology (ICCIT)*, pp. 114–119, IEEE, 2022.

[18] F. T. J. Faria, M. B. Moin, P. Debnath, A. I. Fahim, and F. M. Shah, "Explainable convolutional neural networks for retinal fundus classification and cutting-edge segmentation models for retinal blood vessels from fundus images," *arXiv preprint arXiv:2405.07338*, 2024.

[19] S. Stassin, V. Corduant, S. A. Mahmoudi, and X. Siebert, "Explainability and evaluation of vision transformers: An in-depth experimental study," *Electronics*, vol. 13, no. 1, p. 175, 2023.