Explainable AI Solutions for MRI-Based Brain Tumor Diagnosis

Zhicheng Zhen Mohammad Ahsan Siddiqui Keaton Banik

***Abstract*—The use of artificial intelligence (AI) for medical image classification has been an active field of research. As neural network architectures become increasingly powerful and efficient, their performance in tackling image classification problems has also improved drastically. However, a major challenge in the way of AI-based solutions' adoption in the real-world settings is their uninterpretable decision-making process, due to which they fail to earn the trust of their potential adopters. Our study involves leveraging two powerful pre-trained convolutional neural networks and a pre-trained vision transformer model to classify MRI-based human brain tumor images, and apply explainable AI methods to discover the type of critical information each algorithm is able to detect from the images.**

I. Introduction

Over 300,000 cases of brain tumors are reported worldwide every year, which makes it a pressing concern for the international medical community [1]. Manual inspection of brain MRI images by radiologists is the traditional approach to detect brain tumors. An identified brain tumor can belong to one of the following categories: glioma, meningioma, and pituitary. The traditional brain tumor detection method, however, is subjective, time-consuming, and prone to inter-observer variability.

Convolutional neural networks (CNNs) are powerful algorithms that can perform vision tasks with accuracy that was unachievable a decade ago. Neural Architecture Search (NAS) is a revolutionary algorithm that automates the design process of neural architectures for a specific task, and has outperformed many top-performing human-designed neural architectures on many tasks [2]. EfficientNet is a family of CNNs that can achieve high performance with fewer computational resources compared to previous architectures [3]. Each EfficientNet family member was designed using NAS.

Vision transformers (ViT) [4] are another category of neural networks that have revolutionized the field of image classification. In essence, a ViT is a derivative of a transformer architecture [5] used in sequence-to-sequence modeling; popular examples of their use being large language models such as GPT-4.

CNNs are known to have strong inductive biases. Firstly they assume that a pixel is more related to another pixel that is located near to it than a pixel that is located farther in the image. Secondly, they assume that edges and textures have the same appearance in all images. Therefore, CNNs are faster to train and require less data set for training as compared to ViTs, due to the pre-assumptions they have about spatial locality and translation variance. ViTs, on the other hand, have less inductive bias and use self-attention mechanisms to learn relationships between patches of an image. As a result, they can model long-range dependencies in an image, unlike CNNs that assume spatial locality matters. However, due to their lack of inductive biases, transformers not only need larger data sets than CNNs to generalize for an image classification problem, they also require more time for training.

In this study, we fine-tuned two pre-trained CNN models, EfficientNet-B0 and ResNet, and a pretrained vision transformer model, ViT. Our goal was how do the results obtained from our models compare based on their architectural differences. Additionally, we applied Grad-CAM [6] to generate heatmaps of our classified images for each model to explore the extent to which global context captured by ViT is useful in the accurate classification of brain tumor MRI images.

II. Related Work

Liu et Wang explored four pre-trained CNNs, MobileNet, EfficientNet-B0, ResNet-18, and VGG16, for brain tumor image classification, and obtained insights from their performance analysis to propose a CNN model of their own named MobileNet-BT [7]. MobileNet-BT utilized the pre-trained weights of MobileNetV2. All the layers of MobileNetV2 were then unfrozen to allow the model to learn more specific characteristics from the data. The final layer was then replaced by a custom classifier that enabled the model to learn different levels of abstraction. MobileNet-BT demonstrated a test accuracy of 99% on the brain tumor MRI image dataset. The work is entirely focused on maximizing classification accuracy with a trained classifier, and proves the power of pre-trained CNN models.

A work with a similar focus on maximizing classification accuracy was by Oh et al [8]. The result of the study was a modified version of ViT. The modifications included feature calibration mechanism (FCM) and selective cross-attention (SCM) to improve the performance of the cross-attention fusion module in ViT. The FCM makes features from different branches more compatible using calibration, and SCA selectively attends to the most useful features.

There has been a lot of work focused on boosting the classification performance of CNNs and vision transformers on brain tumor MRI images. However, the interpretability of various models' decision-making process, which also impacts their performance, is unexplored, especially for human brain tumor MRI images dataset.

III. Methods

Dataset

We used publically available brain tumour MRI data to train our three CNN and ViT models [9]. The dataset contained 7023 human brain MRI images labeled into four classes: glioma, meningioma, pituitary and no tumour (Table 1). We observed a roughly 80/20 split between the training and testing data. Validation was done on 20% of the training data during model refinement. Each image was resized to 224 pixels by 224 pixels and normalized.

Table 1: Brain MRI image distribution for the training and testing dataset. The dataset is split into four categories: glioma, meningioma, pituitary and no tumour

	Training	Testing
Glioma	1321	300
Meningioma	1339	306
Pituitary	1457	300
No Tumour	1595	405

Fig. 1. Brain MRI image distribution for the training and testing dataset. The dataset is split into four categories: glioma, meningioma, pituitary and no tumour

Model Fine-Tuning

We employed three pre-trained models for this study: EfficientNet-B0, ResNet and ViT-B16.

EfficientNet-B0 uses a method called compound scaling, which balances three key aspects of a convolutional neural network (CNN): depth (number of layers), width (number of channels per layer), and input resolution (image size) [10]. Unlike traditional scaling, where these parameters are adjusted independently—often resulting in inefficiencies—compound scaling systematically scales all three dimensions in a coordinated manner. This approach enables EfficientNet-B0 to achieve optimal performance with minimal computational overhead.

Unlike traditional approaches that adjust these dimensions independently, often causing inefficiencies, compound scaling adjusts them together in a coordinated way. This ensures the network achieves high accuracy while using fewer computational resources.

The architecture includes Mobile Inverted Bottleneck Convolutions (MBConv), which process images efficiently by expanding the features, applying lightweight depthwise convolutions to each feature, and then compressing the output back to its original size. This reduces the computational load without losing important details. Additionally, Squeeze-and-Excitation (SE) blocks help the model focus on the most important features in an image by emphasizing relevant information and reducing noise.

These innovations allow EfficientNet-B0 to deliver high accuracy with fewer parameters and lower computational costs compared to traditional CNNs. However, its focus on local features can limit its ability to recognize broader patterns in an image, which may be important for tasks like medical image analysis. Explainability is crucial for understanding how this model balances efficiency and its ability to process important global features.

ResNet, or Residual Network, uses a unique approach called residual connections, which allows the model to focus on learning small adjustments (residuals) instead of trying to learn the full mapping from input to output. This makes training more efficient and reduces the risk of computational issues like vanishing gradients.

ResNet's architecture is built from residual blocks, which include shortcut connections that skip certain layers. These connections ensure that information flows smoothly through the network, allowing it to perform well even when the model has many layers. While ResNet is highly stable and generalizes well in deep architectures, it is limited in its ability to capture global relationships within an image, as it primarily focuses on local features.

ViT-B16 applies transformer architecture, originally designed for natural language processing, to image classification. The model divides each input image into fixed-size patches, which capture local details. These patches are flattened into vectors and embedded with positional information to retain their spatial relationships within the image.

ViT-B16 uses self-attention mechanisms, which allow it to analyze how different patches relate to one another, capturing both local and global patterns. This flexibility makes it particularly powerful for identifying complex relationships in images. However, these models are computationally expensive and require more resources compared to CNNs. Despite this, they can provide more robust and comprehensive conclusions by modelling relationships across the entire image.

Each model was initialized with pre-trained weights in a Python environment, and the classification layers were modified to classify three brain tumor types: glioma, meningioma, and pituitary tumors. The models were fine-tuned for 20 epochs with a batch size of 32. Training used the Adam optimizer with an initial learning rate of 0.001, which was dynamically adjusted between epochs based on validation loss to ensure optimal convergence.

Grad-CAM

Grad-CAM was employed to create activation heatmaps for the CNN models, providing a visual representation of how each model made its decisions (Figure 1). These heatmaps were generated by extracting gradients from the final layers of the networks, highlighting the contributions of each region (or pixel) in the input image to the model's output. For ResNet, the heatmaps were based on the third convolutional layer

in block 4.2, while for EfficientNet-B0, they were generated using Feature Layer 8.

The resulting heatmaps were superimposed onto the original MRI images, making it possible to identify which regions of the image were most influential in the model's decision-making process. However, ViT-B16 was unable to produce comparable decision heatmaps due to its architectural differences, which limited observations regarding its explainability.

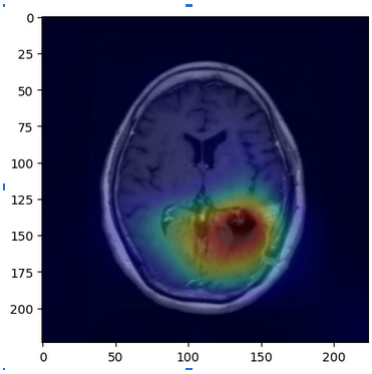

Fig. 2. Grad-Cam generated heatmap from the EfficientNet B0 model. Heatmaps were generated using Grad-Cam from the gradients in the last layer (Features 8.0) of the CNN. Higher decision weight is depicted in red and cascades down to little or no decision weight being depicted in blue. These heatmaps are superimposed onto the input brain MRI scan to visualize the regions of importance.

Model Performance and Explainability

Each model was evaluated using performance metrics such as precision, recall, and F1-score, providing a quantitative measure of their classification accuracy in identifying MRI images as glioma, meningioma, pituitary tumor, or no tumor. In addition to these metrics, Grad-CAM (Gradient-weighted Class Activation Mapping) was employed to generate heat maps for individual predictions, enabling a visual analysis of the regions in each MRI image that influenced the model's decision-making process. These heat maps, when overlaid onto the original MRI images, allowed for a comparison of how the different models interpreted image features and highlighted areas of interest. This approach was instrumental in evaluating the explainability of each model, revealing whether their focus aligned with medically relevant regions, such as tumor boundaries, or if they relied on extraneous and irrelevant image features.

To further investigate model performance and explainability, particular attention was given to misclassified images. These included images with tumors that were incorrectly detected as tumor-free. Grad-CAM heat maps for these cases were carefully analyzed to understand the underlying reasons for the misclassifications. For instance, the analysis focused on whether the models were overly sensitive to noise or benign image artifacts that mimicked tumor-like features.

By comparing the heat maps generated by the different models, we assessed the consistency and relevance of their decision-making processes. The analysis revealed that ViT-B16, with its global attention mechanism, often highlighted broader regions of the image, while CNN-based models like EfficientNet-B0 and ResNet focused on localized features. This distinction underscored the influence of architectural differences on the explainability and interpretability of the models' predictions. Overall, this dual evaluation—using performance metrics and Grad-CAM heat maps—allowed for a comprehensive assessment of the models, balancing their technical effectiveness with their practical applicability in clinical decision-making.

IV. Results

A. Classification Report

The classification reports of each model can be found as follows:

Classification Report:

	precision	recall	f1-score	support
glioma	0.90	0.84	0.87	300
meningioma	0.81	0.71	0.75	306
notumor	0.90	0.96	0.93	405
pituitary	0.88	0.98	0.93	300
accuracy			0.88	1311
macro avg	0.87	0.87	0.87	1311
weighted avg	0.87	0.88	0.87	1311

Fig. 3. Classification Report of the ResNet CNN Model

Classification Report:

	precision	recall	f1-score	support
glioma	0.97	0.82	0.89	300
meningioma	0.81	0.86	0.83	306
notumor	0.94	0.99	0.97	405
pituitary	0.94	0.96	0.95	300
accuracy			0.91	1311
macro avg	0.91	0.91	0.91	1311
weighted avg	0.92	0.91	0.91	1311

Fig. 4. Classification Report of the EfficientNet CNN Model

Classification Report:

	precision	recall	f1-score	support
glioma	0.98	0.82	0.89	300
meningioma	0.84	0.92	0.88	306
notumor	0.98	1.00	0.99	405
pituitary	0.94	0.98	0.96	300
accuracy			0.93	1311
macro avg	0.93	0.93	0.93	1311
weighted avg	0.94	0.93	0.93	1311

Fig. 5. Classification Report of the ViT Model

Fig. 6. Training Loss Comparisons

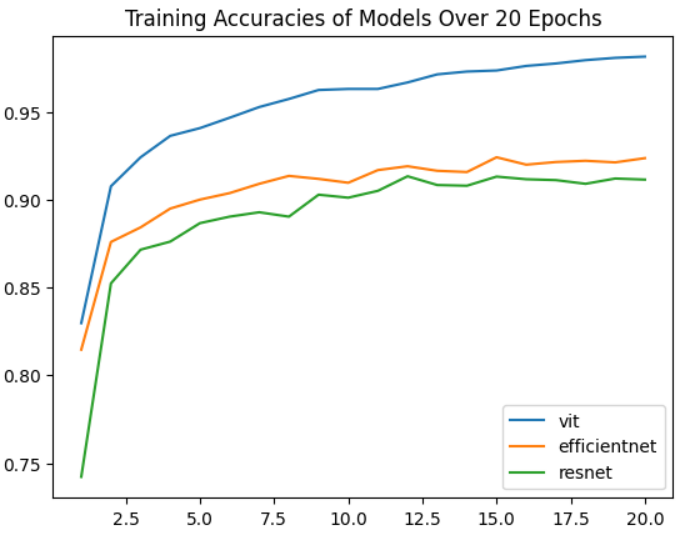

Fig. 7. Training Accuracy Comparisons

The Vision Transformer (ViT-B16) model demonstrated superior performance in classifying brain tumor MRI images, achieving the highest training accuracy of 98.14% and testing accuracy of 93.36%. This surpasses the results of both CNN-based models. EfficientNet-B0 ranked second, with a training accuracy of 92.36% and a testing accuracy of 91.38%, followed by ResNet, which achieved a training accuracy of 91.14% and a testing accuracy of 87.64%. These results suggest that the ViT model's ability to model global dependencies and extract features across the entire image makes it more suitable for this task compared to the localized feature extraction of CNN-based architectures. The training accuracies for all models were tracked over 20 epochs, revealing notable trends. The ViT model exhibited a steady increase in accuracy throughout the fine-tuning process, consistently improving without noticeable plateaus or declines. In contrast, ResNet experienced a slight decrease in precision after reaching its highest accuracy of 91.33% around the 12th epoch. Similarly, EfficientNet displayed a minor drop in accuracy between epochs 15 and 20, suggesting that both CNN-based models may be prone to overfitting during extended training. These trends highlight the robustness of the ViT model in maintaining stable and consistent learning throughout the training process.

To further evaluate model performance and interpretability, Grad-CAM heatmaps were analyzed for both correct and incorrect classifications. Figures 8–12 illustrate key examples:

Figure 8 presents the heatmaps for a correct glioma classification by both ResNet and EfficientNet. While both models reached the correct conclusion, the heatmaps display noticeable variation in the regions of focus, reflecting differences in their feature extraction mechanisms.

Figure 9 highlights the classification of a meningioma tumor, correctly identified by both ResNet and EfficientNet. Compared to glioma classification, the heatmaps show less variation, suggesting greater consistency in identifying features characteristic of meningiomas.

Figure 10 demonstrates the correct classification of a pituitary tumor by both models. The heatmaps for pituitary

tumors exhibit patterns similar to those observed in glioma classifications, with moderate variation in the regions of focus.

Figure 11 depicts a case where both models correctly classified an MRI as ”no tumor.” These heatmaps showed minimal variation, indicating strong agreement in how ResNet and EfficientNet interpret non-tumorous images.

Figure 12 displays a heatmap where ResNet misclassified a tumor as ”no tumor,” whereas EfficientNet correctly identified it. The heatmap reveals that ResNet focused on irrelevant regions, leading to its incorrect classification.

Figure 13 shows a case where EfficientNet misclassified a tumor, while ResNet accurately detected its presence. The heatmap analysis indicates that EfficientNet’s attention was distributed over less relevant features, contributing to its error.

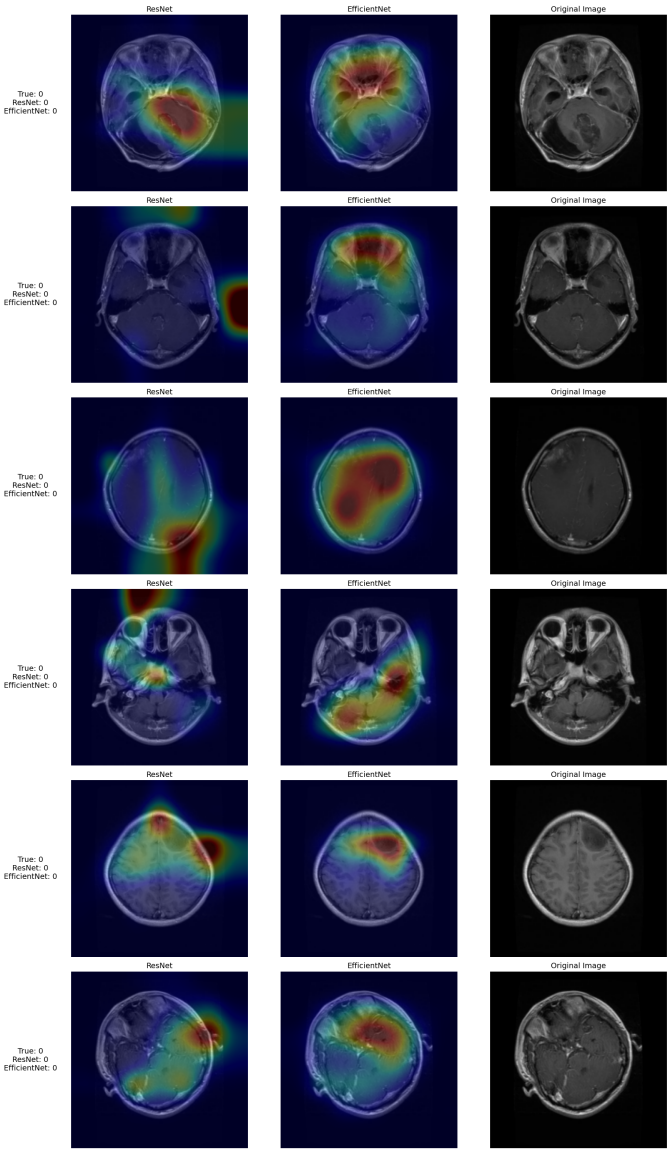

Fig. 8. Class 0 Glioma - Correct Classification

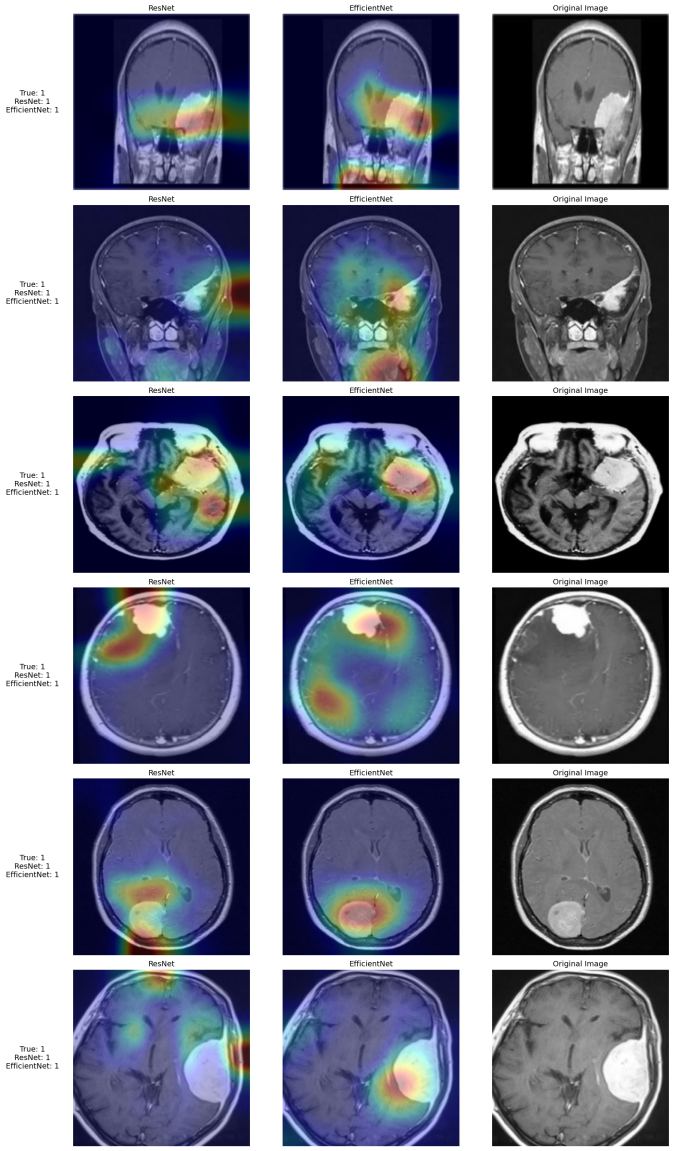

Fig. 9. Class 1 Meningioma - Correct Classification

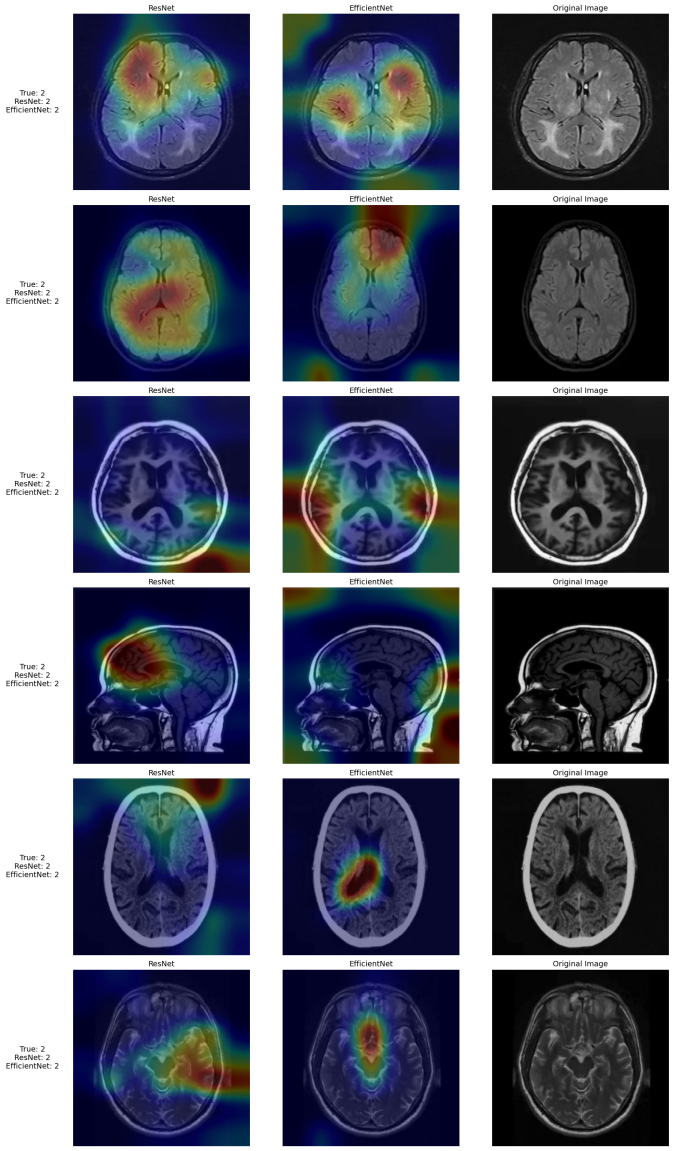

Fig. 10. Class 2 Pituitary - Correct Classification

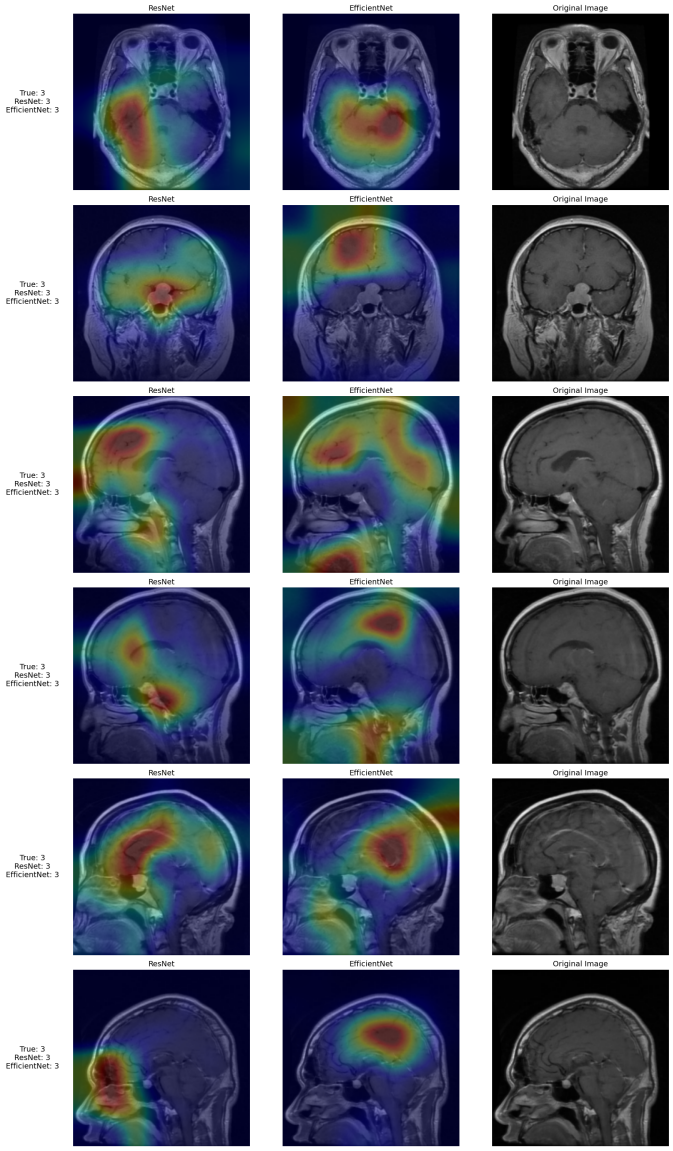

Fig. 11. Class 3 No Tumor - Correct Classification

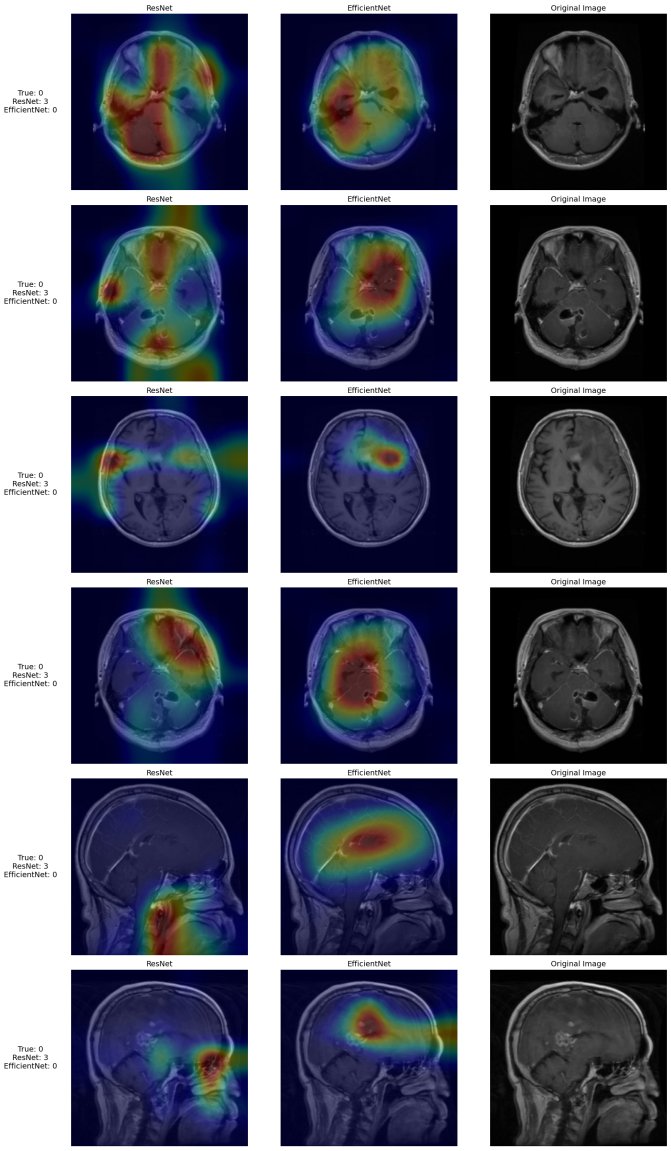

Fig. 12. Resnet - Failed Identification of Tumor

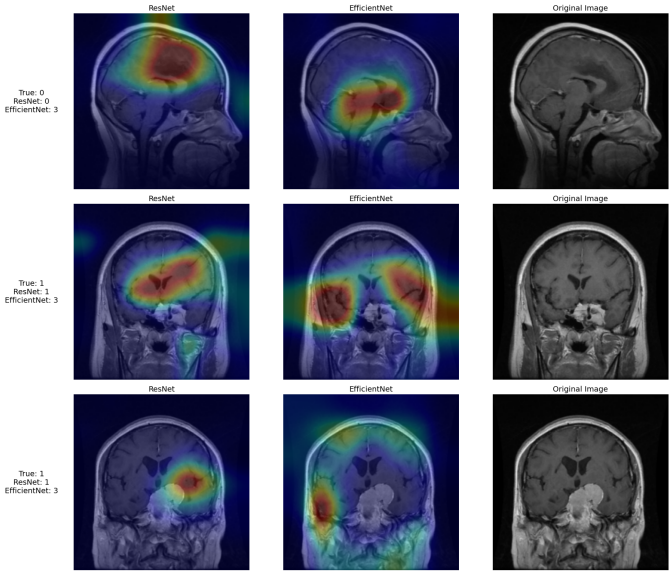

Fig. 13. EfficientNet - Failed Identification of Tumor

These analyses of Grad-CAM heatmaps highlight the strengths and weaknesses of each model's decision-making process. The ViT model's superior accuracy and ability to maintain stable performance suggest its potential for clinical applications, while the complementary strengths of ResNet and EfficientNet demonstrate the utility of CNN-based models in identifying specific tumor types. The variation in heatmaps for tumor classifications also emphasizes the importance of explainability tools for understanding and validating model predictions in medical image analysis.

DISCUSSION

The results of this study underscore the importance of model architecture in determining the accuracy, robustness, and explainability of artificial intelligence (AI) systems for medical image classification. Among the three models tested, the Vision Transformer (ViT-B16) demonstrated the highest performance, with superior training and testing accuracies compared to the CNN-based models, EfficientNet-B0 and ResNet. The ViT's ability to model global dependencies using self-attention mechanisms likely contributed to its superior classification accuracy, particularly for challenging cases where spatial relationships across the entire image were crucial. This suggests that transformers, though computationally expensive, hold significant promise for medical imaging tasks that require both precision and holistic analysis.

The CNN-based models, while trailing ViT in overall accuracy, demonstrated complementary strengths. EfficientNet-B0, for instance, achieved high testing accuracy with relatively low computational requirements, highlighting its efficiency for resource-constrained environments. ResNet, on the other hand, showed stability in deeper architectures, albeit with limitations in capturing global image features. Both CNN models displayed tendencies toward overfitting in later epochs, as

evidenced by slight declines in accuracy during extended training. These findings emphasize the importance of balancing model complexity with training dynamics to avoid overfitting, particularly in medical datasets with limited diversity. ViTs may be able to solve this issue at the cost of computational power.

The analysis of Grad-CAM heatmaps provided critical insights into the explainability of the CNN-based models, ResNet and EfficientNet-B0, as they allowed us to visually interpret the regions most influential in the models' decision-making processes. However, we were unable to extract usable gradient information from the attention blocks of the ViT model due to its architectural differences, which made it incompatible with the Grad-CAM method. As a result, our comparative analysis of explainability was limited to ResNet and EfficientNet-B0. The CNN-based models demonstrated distinct strengths and weaknesses, with ResNet and EfficientNet focusing on more localized image details, which occasionally led to false positives and negatives. Consistent classifications, such as no-tumor cases, displayed strong agreement in focus regions across the CNN models.

In future work, ViT models could be made more interpretable by developing methods to extract attention weights in a manner analogous to gradient-based heatmaps used for CNNs [11]. By making attention weights accessible and interpretable, similar to Grad-CAM outputs for CNNs, researchers could gain a deeper understanding of how ViT models arrive at their decisions, potentially bridging the current gap in explainability between these architectures. This advancement would also facilitate a more holistic comparison of interpretability between transformer-based and CNN-based approaches, further driving the adoption of explainable AI in medical imaging.

Conclusion

This study highlights the trade-offs between accuracy, computational efficiency, and explainability in different AI architectures for brain tumour classification. While ViT demonstrated superior accuracy, the complementary strengths of CNN-based models like EfficientNet-B0 and ResNet point to the potential for hybrid approaches that combine the benefits of both architectures. These findings underscore the need for further exploration of AI models that balance technical performance with practical considerations, paving the way for more reliable and interpretable solutions in medical diagnostics.

References

[1] A. Cohen-Gadol, "Brain tumor statistics," Aaron Cohen-Gadol, [Online]. Available: https://www.aaroncohen-gadol.com/en/patients/brain-tumor/types/statistics: :text=With

[2] "Investigating Large Language Models' Understanding of General Machine Knowledge," arXiv, vol. 2301.08727, 2023. [Online]. Available: https://arxiv.org/abs/2301.08727. [Accessed: Dec. 20, 2024].

[3] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," arXiv, vol. 1905.11946, 2019. [Online]. Available: https://arxiv.org/abs/1905.11946. [Accessed: Dec. 20, 2024].

[4] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv, vol. 2010.11929, 2020. [Online]. Available: https://arxiv.org/abs/2010.11929. [Accessed: Dec. 20, 2024].

[5] A. Vaswani et al., "Attention Is All You Need," arXiv, vol. 1706.03762, 2017. [Online]. Available: https://arxiv.org/abs/1706.03762. [Accessed: Dec. 20, 2024].

[6] S. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," arXiv, vol. 1610.02391, 2016. [Online]. Available: https://arxiv.org/abs/1610.02391. [Accessed: Dec. 20, 2024].

[7] "Title of the Paper," arXiv, vol. 2408.00636, 2024. [Online]. Available: https://arxiv.org/pdf/2408.00636. [Accessed: Dec. 20, 2024].

[8] S. Liu, "Wi-Fi Energy Detection Testbed (12MTC)," 2023, gitHub repository. [Online]. Available: https://github.com/liustone99/Wi-Fi-Energy-Detection-Testbed-12MTC

[9] "Title of the Paper," arXiv, vol. 2410.12692, 2024. [Online]. Available: https://arxiv.org/pdf/2410.12692. [Accessed: Dec. 20, 2024].

[10] M. Nickparvar, "Brain tumor MRI dataset," Kaggle, [Online]. Available: https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset. [Accessed: Dec. 20, 2024].

[11] "Title of the Paper," arXiv, vol. 2311.06786, 2023. [Online]. Available: https://arxiv.org/abs/2311.06786. [Accessed: Dec. 20, 2024].