Ensemble Deep Learning for Brain Tumor Classification Using MRI: A Comparative Study of CNN Architectures

Abdullah Al Mahmud Joy*, Md Faizer Islam†, Md. Mamun Hossain‡
Department of Computer Science and Engineering
Bangladesh University of Business and Technology, Dhaka, Bangladesh
abdullahalmahmudjoy39@gmail.com*, faizarislam@gmail.com†, mamunhasan.cs@gmail.com‡

***Abstract*—Early and accurate detection of brain tumors is critical for improving patient outcomes; however, it remains challenging due to the variability in tumor morphology. This study presents a comparative evaluation of four deep learning architectures: Custom CNN, MobileNetV2, VGG16, and EfficientNetB0, along with a soft-voting ensemble method for brain tumor classification using magnetic resonance imaging (MRI) scans. The data set included balanced images of four classes: glioma, meningioma, pituitary tumors, and no tumors. The preprocessing involved intensity normalization, spatial resizing, and label encoding, with training optimized through class weighting, adaptive learning rate scheduling, and early termination. The interpretability of the model was supported using Grad-CAM visualizations to highlight discriminatory regions. The experimental results show that the proposed ensemble model achieved an accuracy of 98.93%, outperforming EfficientNetB0 (97.62%), MobileNetV2 (97.38%), VGG16 (96.99%), and Custom CNN (96.79%). The ensemble also delivered superior precision, recall, and F1 scores while reducing interclass confusion, particularly between gliomas and meningiomas. These findings demonstrate the effectiveness of ensemble learning in combining complementary model strengths and highlight its potential as a scalable, interpretable, and clinically applicable solution for the automated diagnosis of brain tumors.**

Keywords: Brain tumor*, *CNN*, *VGG16*, *MobileNetV2*, *Grad-CAM*, *EfficientNetB0*, *Ensemble

I. Introduction

Despite being the 21st most common cancer worldwide, brain tumors are among the most deadly. They have few options for treatment and high mortality rates. Particularly in low- and middle-income nations, accurate data is hard to come by [1]. To find brain tumors, radiologists employ medical imaging methods [2]. Because it is non-invasive and radiation-free, MRI is the method of choice among those that are available [3]. In the conventional method, radiologists use MRI scan analysis to diagnose brain tumors. However, because of workload and differing levels of expertise, this approach is laborious and prone to mistakes. Accurate classification is made more difficult by the intricate anatomy of tumors [4]. A brain tumor that is misdiagnosed can have detrimental effects and reduce the patient's chances of survival. Researchers have investigated several classification techniques to assist radiologists, such as traditional machine learning, which depends on human feature extraction and prior knowledge. As medical data becomes more complex, traditional approaches frequently fail, but deep learning in particular, CNN has demonstrated impressive tumor detection capabilities. CNNs automatically extract features from images, eliminating the need for manual feature engineering or domain-specific expertise. Models like VGG16, MobileNet, and EfficientNet have recently shown impressive results in medical image classification tasks, each balancing accuracy and computational cost differently [5]. Based on this, we propose a reliable CNN-based approach to accurately classify brain tumors from MRI scans.

Unlike prior studies, which mainly used homogeneous CNNs, our study proposes a heterogeneous ensemble combining lightweight (MobileNetV2, Custom CNN) and deep (VGG16, EfficientNetB0) models via soft voting. This balances accuracy and efficiency while reducing inter-class confusion, making it practical for clinical use.

The main contributions of this study:

- Performed a comparative analysis of four CNN-based architectures using a balanced MRI brain tumor dataset.
- Proposed a soft-voting ensemble approach that achieved 98.93% accuracy and reduced inter-class confusion.
- We incorporated Grad-CAM visualizations to support clinical applicability and improve interpretability.

The paper is structured as follows: Section II background study on Brain Tumor Classification. Section III describes our dataset, preprocessing steps, and the model architectures. Section IV presents quantitative results, confusion matrices, and Grad-CAM visualizations. Finally, Section V discusses implications, limitations, and future research directions.

979-8-3315-5525-2/25/$31.00 ©2025 IEEE

II. Related Works

Using MRI images, researchers have put forth a plethora of techniques and algorithms for identifying various brain tumor types as well as other anomalies in the human brain. Veeranki et al. [6] classified brain MRI images into four groups using the VGG-16 model: pituitary tumor, glioma tumor, meningioma tumor, and no tumor. They obtained 89% accuracy, 81% precision, 89% recall, and 84% F1 Score using the publicly accessible Brain Tumor Dataset (BTD).

TABLE I
Comparative Analysis of Existing Studies

Study	**Technique**	**Results**
Veeranki et al. [6]	VGG-16, Transfer Learning	Accuracy: 89%, Precision: 81%, Recall: 89%, F1 Score: 84%
Remzan et al. [7]	Custom CNN Model	Accuracy: 89%
Lakshmi et al. [8]	Inception-V3, Transfer Learning	Accuracy: 89%
Hashan et al. [9]	Custom CNN Model	Accuracy: 90%, F1 Score: 89%
Kabir Anakari et al. [10]	Genetic Algorithm, CNN	Accuracy: 90.9%
Chitnis et al. [11]	LeaSE	Accuracy: 88.87%, Precision: 90.62%, Recall: 88.63%, F1 Score: 89.61%
Pashaei et al. [12]	Custom CNN Model, KELM	Accuracy: 93.68%, Precision: 94.6%, Recall: 91.43%, F1 Score: 93%

A novel CNN architechture was introduced by Remzan et al. [7] to categorize various types of brain tumors. The Brain Tumor Dataset (BTD) was used to assess the model. 89% accuracy was attained. An empirical study on a dataset of 3,064 T1-weighted brain MR images were carried out by Lakshmi et al. [8]. The authors assessed the VGG-16, ResNet50, and Inception-v3 pre-trained CNN models. With Inception-V3, the highest accuracy of 89% on validation data was attained. Using a dataset of 400 images, Hashan et al. [9] created a novel CNN model to separate through normal and abnormal brain MRI images. To improve the model, the authors chose ideal hyperparameters, attaining an F1 Score of 89% and 90% accuracy. To categorize glioma grades, Kabir Anakari et al. [10] proposed a novel approach that combines the genetic algorithm and CNN. The CNN architecture's hyperparameters were established by the authors using the genetic algorithm. On validation data, they obtained an accuracy of 90.9%. Chitnis et al. [11] introduced the 'Learning by Self Explanation' (LeaSE) method, which automatically searches for effective neural architectures and gain an F1 score of 89.61% and accuracy of 88.87% on brain tumor data. Similarly, Pashaei et al. [12] developed two methods: a CNN that achieved 81.09% accuracy and a hybrid CNN-KELM approach that improved the accuracy to 93.68% with an F1 score of 93%. The summary of previous research is given in Table I.

The efficiency of transformer-based and hybrid approaches for medical image analysis has been demonstrated in recent studies. Vision Transformers (ViT) have shown notable success in tumor classification by capturing global contextual features [13]. Hybrid CNN–Transformer architectures further enhance representation by combining the local feature extraction capability of CNNs with the long-range dependency modeling of transformers [14]. In addition, self-supervised learning techniques have proven valuable in reducing the reliance on large annotated datasets while still delivering competitive accuracy [15]. Despite these advances, such approaches often require substantial computational resources, whereas our proposed ensemble offers a more efficient and accessible solution without sacrificing performance.

III. Methodology

A. Dataset Analysis

The original Brain Tumor Dataset contained 10,287 MRI images divided into four categories: glioma (2,547), meningioma (2,582), pituitary (2,658), and no tumor (2,500), resulting in a slight class imbalance in Figure 1. To address this issue, the dataset was balanced by downsampling each class to 2,500 images, yielding a total of 10,000 MRI data images evenly divided through the four categories. The balanced dataset was then split into 8,000 training images (2,000 per class) and 2,000 testing images (500 per class) to ensure a fair model training and evaluation. A sample data is shown in Figure 5

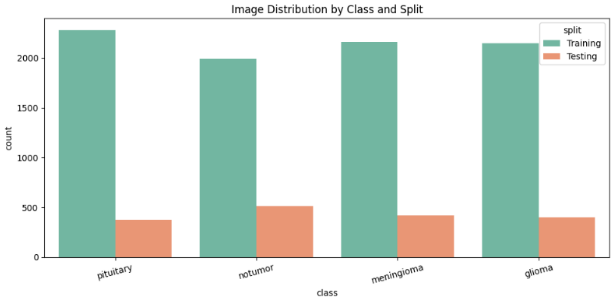

Fig. 1. Class Distribution of Brain Tumour Dataset

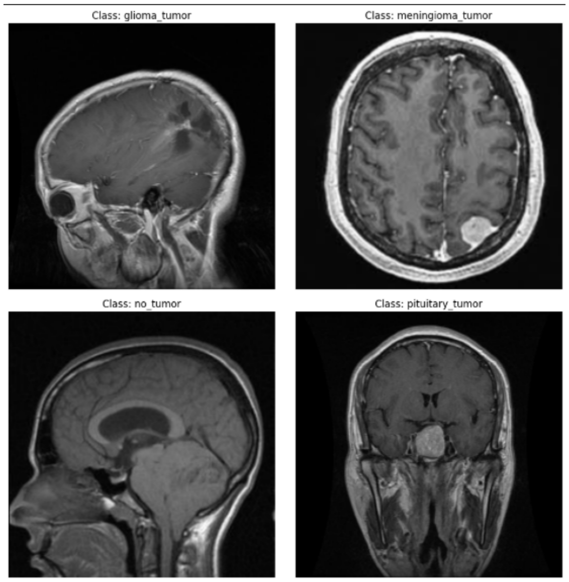

Fig. 2. Sample MRI Images of Class Distribution

B. Data Preprocessing

Several preprocessing steps were performed on the MRI images to maintain consistency and improve model performance. Specifically, all images were converted to 224×224 pixels to align the input shape expected by CNNs, and pixel intensities were normalized to the [0,1] range to facilitate faster convergence. The labels were one-hot encoded for multi-class classification, while data augmentation techniques such as flips, rotations, zooming, and brightness adjustments were used to increase variability and reduce overfitting. Although the dataset was relatively balanced, class weights were incorporated to preserve sensitivity across categories. Finally, the data was divided into training and testing sets to avoid leakage.

For reproducibility, the complete codebase, trained model weights, and preprocessing scripts will be made publicly available through an open-access repository, upon acceptance. This ensures transparency and enables other researchers to validate and extend our study.

C. Architecture of Deep Learning Models

For working with brain tumor analysis from MRI scans into four categories (glioma, meningioma, pituitary tumor, and no tumor), we utilized four CNN architectures: a Custom CNN, MobileNetV2, VGG16, and EfficientNetB0. These models integrate a baseline with powerful pre-trained networks, modified with a softmax output layer and a 0.5 dropout rate to mitigate overfitting. Where applicable, ImageNet pre-trained weights were fine-tuned to enhance performance in the medical imaging context. Further implementation details are presented in the subsequent sections.

1) Custom CNN

Our Custom CNN, designed from the ground up, comprised three convolutional blocks, each integrating convolution, batch normal., ReLU activation, and 2 by 2 max-pooling. The number of filters increased from 32 to 64 to 128, enabling the model to capture progressively complex features. The output was flattened and passed through two dense layers (256 and 128 neurons) with a 0.5 dropout, through a softmax layer for four-class tumor classification.

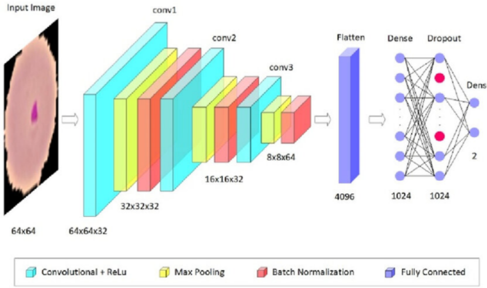

Fig. 3. CNN architecture [16]

2) MobileNetV2

MobileNetV2, introduced by Sandler et al. (2018), was chosen for its efficiency in resource-limited settings, such as medical imaging. It uses linear bottlenecks, inverted residual blocks, and depth-wise distinguishable convolutions to reduce the computational complexity without sacrificing accuracy. For our task, we froze the top 20 layers of the pre-trained model and added global average pooling, a 128-neuron dense layer, 0.5 dropout, and a softmax classifier to the model.

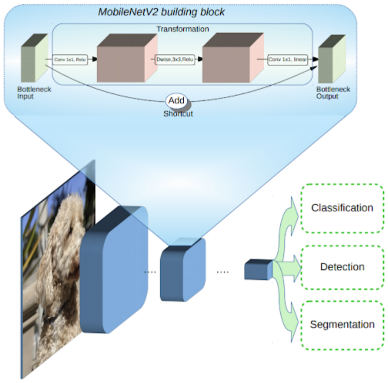

Fig. 4. MobileNetV2 architecture [17]

3) VGG16

VGG16, introduced by Simonyan and Zisserman (2014), has 13 convolutional layers in five blocks with max pooling and three fully joint layers. The filters were scaled from 64 to 512 to capture the hierarchical spatial features. For our task, we replaced the dense layers with a 256-neuron fully connected layer, global average pooling, and softmax output, adding dropout and batch normalization for stability. The last five convolutional blocks were unfrozen for domain-specific, fine-tuning.

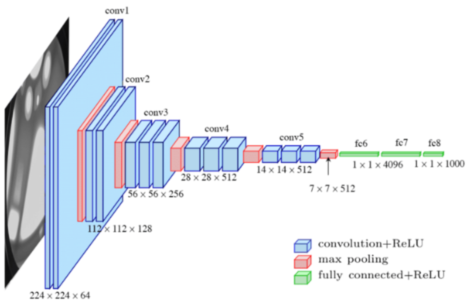

Fig. 5. VGG16 architecture [18]

4) EfficientNetB0

EfficientNetB0, proposed by Tan and Le (2019), applies compound scaling to balance the width, depth, and resolution. It begins with a 3×3 convolution (32 filters) followed by MBConv blocks with depthwise convolutions, a Squeeze-and-Excitation, and residual connections, scaling from 16 to 1280 channels. For our task, we added global average pooling, a 128-unit dense layer, 0.5 dropout, and a softmax classifier while fine-tuning the top 15% of layers with a reduced learning rate for MRI adaptation.

5) Ensemble Model

To improve the classification performance, a soft-voting ensemble was created by combining the Custom CNN, MobileNetV2, VGG16, and EfficientNetB0. Each model was trained separately, and their softmax outputs were averaged to make the final predictions. This method balanced the model strengths, reduced the variance, and boosted the robustness, achieving a high accuracy of 98.93%, outperforming all individual models.

6) Training Configuration

Adam (lr=1e-4, reduced ×0.1 on plateau) was used to train all models on an NVIDIA GPU. We employed early stopping (patience = 10), a maximum of 50 epochs, and a batch size of 32. Online data augmentation and class weights handled minor imbalances, ensuring stable convergence across the models.

D. Grad-CAM Visualization

We applied Grad-CAM to interpret the model decisions in brain tumor classification. Heatmaps revealed that EfficientNetB0 and VGG16 focused sharply on tumor regions, whereas MobileNetV2 and the Custom CNN had broader attention owing to their lighter architectures. The ensemble combines these strengths to produce localized heatmaps that enhance trust and provide visual support for radiologists.

IV. Results & Discussion

This study evaluated five deep neural models for brain tumor analysis of classification task from MRI scans. The Custom CNN achieved an accuracy of 96.79%, MobileNetV2 97.38%, VGG16 96.99%, and EfficientNetB0 97.62%. The ensemble model achieved the highest performance, reaching an accuracy of 98.93% along with superior precision, recall, and F1-scores.

TABLE II
Comparative Performance Metrics of the Models

Model	Accuracy	Precision	Recall	F1-Score
Ensemble	**0.98931**	**0.98935**	**0.989267**	**0.989238**
EfficientNetB0	0.97619	0.976157	0.976158	0.976078
MobileNetV2	0.973761	0.974009	0.973746	0.973699
VGG16	0.969874	0.970231	0.969843	0.969882
CustomCNN	0.96793	0.967829	0.967824	0.967704

The training and validation accuracies and loss curves in Figure 7 and 8 for the models demonstrated steady learning and good convergence. The Ensemble and EfficientNetB0 models showed the most stable curves with minimal overfitting, whereas the CustomCNN and VGG16 exhibited slight fluctuations in the validation performance compared to MobileNetV2.

The comparison of the test performance of the five models shown in Figure 6, was based on precision, recall, and f-score. The Ensemble Model achieved the best results across all metrics, followed closely by EfficientNetB0 and VGG16, indicating a strong and balanced classification performance.

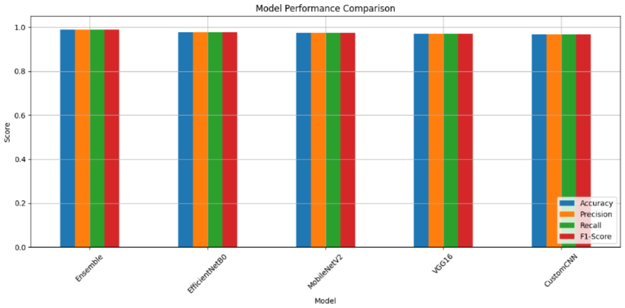

Fig. 6. Model Performance Comparison

The Custom CNN excelled at detecting pituitary tumors but struggled to distinguish gliomas from meningiomas. MobileNetV2 exhibited a balanced performance with minor confusion between these classes. VGG16 reliably classified meningiomas and pituitary tumors but had some difficulty with gliomas, indicating a feature

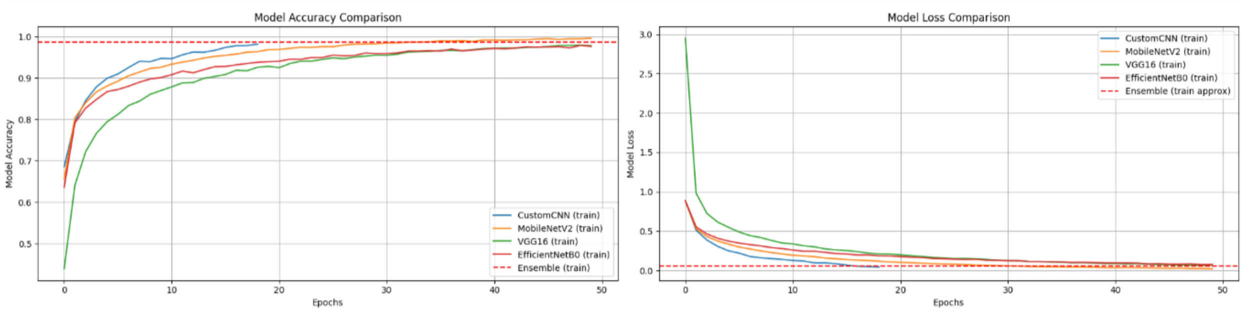

Fig. 7. Accuracy and Loss Curves During Training

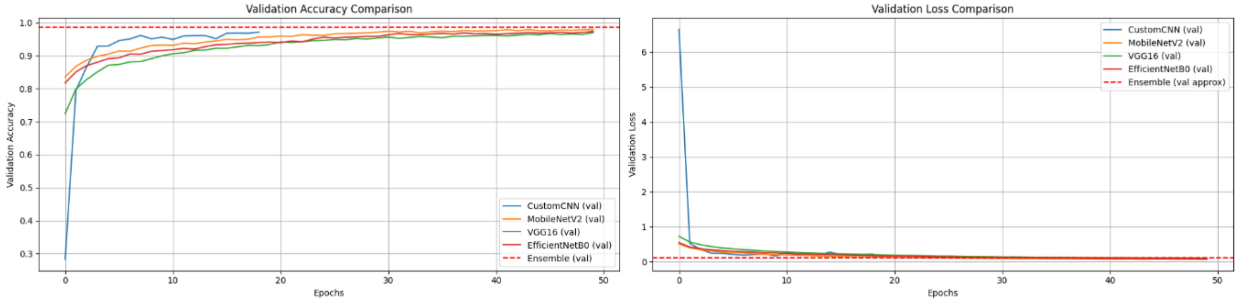

Fig. 8. Accuracy and Loss Curves During Validation

overlap. EfficientNetB0 achieved the highest accuracy, notably reducing the glioma–meningioma misclassifications.Overall, the Ensemble Model in Figure 9 achieved the best results, offering the highest accuracy and most balanced predictions with minimal misclassifications across all categories.

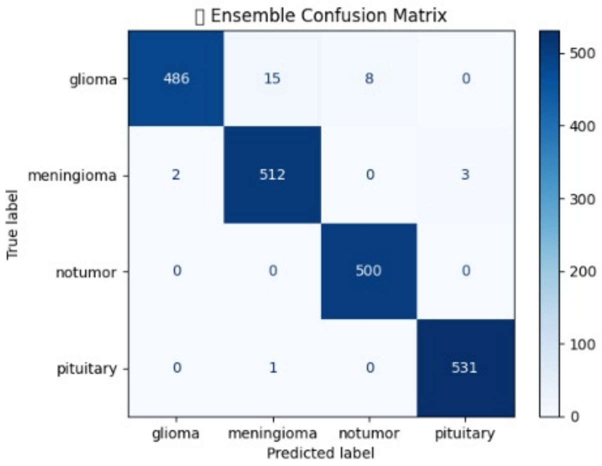

Fig. 9. Confusion Matrix in Ensemble Model

Grad-CAM visualizations Figure 10 provide interpretability by highlighting the regions in brain MRI images on which the model focuses during prediction. These heatmaps confirm that the model attends to relevant tumor regions, enhancing trust and understanding of the decision-making process.

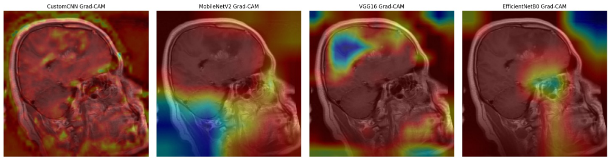

Fig. 10. GradCAM Visualization

V. Conclusion & Future Directions

We presented a comparative study of CNN architectures and introduced a heterogeneous ensemble that consistently outperformed individual models in brain tumor classification. Grad-CAM visualizations were incorporated to improve interpretability, helping build clinical trust in the predictions. While the results are promising, the work was limited to a single dataset without external or k-fold validation, and performance may vary with lower-quality MRI scans, highlighting the need for broader validation before clinical use.

Future research should aim to validate the model with external and multi-center datasets, in addition to applying cross-validation, to enhance its robustness. Incorporating hybrid CNN–Transformer architectures and self-supervised learning could further boost performance while minimizing reliance on large annotated datasets.

Ultimately, real-world deployment and evaluation in hospital workflows will be essential to determine the ensemble's practicality, scalability, and clinical value.

References

[1] F. Bray, J. Ferlay, I. Soerjomataram *et al.*, "A global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide," *CA Cancer Jour. Clin.*, vol. 68, no. 6, pp. 394–424, 2018.

[2] M. Kim and H. S. Kim, "Emerging techniques in brain tumor imaging: what radiologists need to know," *Korean Jour. Radi.*, vol. 17, no. 5, p. 598, 2016.

[3] Y. Guo, H. Chai, and Y. Wang, "A global approach for medical image denoising via sparse representation," *Int. Jour. of Biosci.*, vol. 5, no. 1, pp. 26–35, 2015.

[4] N. Arunkumar, M. AbedMohammed, S. A Mostafa *et al.*, "Fully automatic model-based segmentation and classification approach for mri brain tumor using artificial neural networks," *Concurrency Computa.: Pract. Exper.*, vol. 32, no. 1, pp. 1–9, 2018.

[5] M. A. Rahman, M. M. Hossain, S. P. Singh, and N. Sharmin, "Predicting early asd traits of adults and toddlers using machine learning and deep learning with explainable ai and optimization," *Neural Computi. Appli.*, pp. 1–28, 2025.

[6] P. S. L. Veeranki, G. L. Banavath *et al.*, "Detection and classification of brain tumors using convolutional neural network," in *Proce. 7th Int. Conf. Trends Electr. Informatics (ICOEI).* IEEE, 2023, pp. 780–786.

[7] N. Remzan, K. Tahiry, and A. Farchi, "Deep learning approach for brain tumor classification implemented in raspberry pi," in *Proce. Int. Conf. Advan. Intelli. Sys. Sust. Develop.* Springer, 2022, pp. 136–147.

[8] M. J. Lakshmi and S. Nagaraja Rao, "Retracted article: Brain tumor magnetic resonance image classification: a deep learning approach," *Soft Computi.*, vol. 26, no. 13, pp. 6245–6253, 2022.

[9] A. M. Hashan, E. Agbozo, A. A. K. Al-Saeedi *et al.*, "Brain tumor detection in mri images using image processing techniques," in *Proce. 4th Int. Sympo. Agents, Multi-Agent Sys. Robo. (ISAMSR).* IEEE, 2021, pp. 24–28.

[10] A. K. Anaraki, M. Ayati, and F. Kazemi, "Magnetic resonance imaging-based brain tumor grades classification and grading via convolutional neural networks and genetic algorithms," *Biocybern. Biomedi. Engg.*, vol. 39, no. 1, pp. 63–74, 2019.

[11] S. Chitnis, R. Hosseini, and P. Xie, "Brain tumor classification based on neural architecture search," *Scint. Reports*, vol. 12, no. 1, p. 19206, 2022.

[12] A. Pashaei, H. Sajedi, and N. Jazayeri, "Brain tumor classification via convolutional neural network and extreme learning machines," in *Proce. 8th Int. Conf. Computa. Know. Engg. (ICCKE).* IEEE, 2018, pp. 314–319.

[13] E. Simon and A. Briassouli, "Vision transformers for brain tumor classification." in *BIOIMAG.*, 2022, pp. 123–130.

[14] S. Sagheer, M. KH, P. Ameer *et al.*, "Transformers for multi-modal image analysis in healthcare." *Compu. Mate. Continu.*, vol. 84, no. 3, 2025.

[15] S. Azizi, B. Mustafa, F. Ryan *et al.*, "Big self-supervised models advance medical image classification," in *Proc. IEEE/CVF Int. Confe. Compute. Visi.*, 2021, pp. 3478–3488.

[16] Y. Sun, B. Xue, M. Zhang *et al.*, "Completely automated cnn architecture design based on blocks," *IEEE Transa. Neural Net. Learn. Sys.*, vol. 31, no. 4, pp. 1242–1254, 2019.

[17] M. Sandler, A. Howard, M. Zhu *et al.*, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proce. IEEE Conf. Compute. Visi. Patt. Recog.*, 2018, pp. 4510–4520.

[18] A. Mustapha, L. Mohamed, H. Hamid *et al.*, "Diabetic retinopathy classification using resnet50 and vgg-16 pretrained networks," *Int. Jour. Compute. Engg. Data Scie.*, vol. 1, no. 1, pp. 1–7, 2021.