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

1st Abdullah Al Mahmud Joy, 2nd Md Faizer Islam, 3rd Md Mamun Hossain

Department of Computer Science and Engineering

Bangladesh University of Business and Technology

Dhaka, Bangladesh

Email: {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 owing to the variability in tumor morphology. This study presents a comparative evaluation of four deep learning architectures—Custom CNN, MobileNetV2, VGG16, and EfficientNetB0—alongside a soft-voting ensemble method for brain tumor classification using MRI scans. The dataset included balanced images across four classes: glioma, meningioma, pituitary tumors, and no tumors. Preprocessing involved intensity normalization, spatial resizing, and label encoding, with training optimized through class weighting, adaptive learning rate scheduling, and early termination. Model interpretability was supported using Grad-CAM visualizations to highlight the discriminative 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

1. 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; 4]. 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 [5; 6]. 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 [7]. CNN can automatically extract features from images, so the classification process doesn't require human feature engineering or prior domain knowledge. Deep learning-based architectures such as VGG16, MobileNet, and EfficientNet have demonstrated impressive results in medical image classification tasks in recent years. These models offer a variety of trade-offs between accuracy and computational cost, and they differ in complexity and efficiency. Given these developments, we have put forth a reliable CNN-based method that makes use of these architectures to reliably and accurately classify brain tumors from MRI scans. The contributions of this study are as follows:

- 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 interclass confusion.
- Incorporated Grad-CAM visualizations to enhance interpretability and support clinical applicability.

2. 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. [8] 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 1. COMPARATIVE ANALYSIS OF EXISTING STUDIES

Study	Technique	Results
Veeranki et al. [8]	VGG-16, Transfer Learning	Accuracy: 89%, Precision: 81%, Recall: 89%, F1 Score: 84%
Remzan et al. [9]	Custom CNN Model	Accuracy: 89%
Lakshmi et al. [10]	Inception-V3, Transfer Learning	Accuracy: 89%
Hashan et al. [11]	Custom CNN Model	Accuracy: 90%, F1 Score: 89%
Kabir Anakari et al. [12]	Genetic Algorithm, CNN	Accuracy: 90.9%
Chitnis et al. [13]	LeaSE	Accuracy: 88.87%, Precision: 90.62%, Recall: 88.63%, F1 Score: 89.61%
Pashaei et al. [14]	Custom CNN Model, KELM	Accuracy: 93.68%, Precision: 94.6%, Recall: 91.43%, F1 Score: 93%

A novel CNN model was proposed by Remzan et al. [9] 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. [10]. The authors assessed the VGG-16, ResNet50, and Inception-v3 pretrained CNN models. With Inception-V3, the highest accuracy of 89% on validation data was attained. Using a dataset of 400 images, Hashan et al. [11] created a novel CNN model to distinguish between 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. [12] 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. [13] proposed the Learning-by-Self-Explanation (LeaSE) method, which automatically searches for effective neural architectures and achieved an F1 score of 89.61% and accuracy of 88.87% on brain tumor data. Similarly, Pashaei et al. [14] 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%. An overview of previous research is given in Table 1.

3. Methodology

3.1. Dataset Analysis

The original Brain Tumor Dataset contained 10,287 MRI images divided into four classes: 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 images evenly distributed across 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

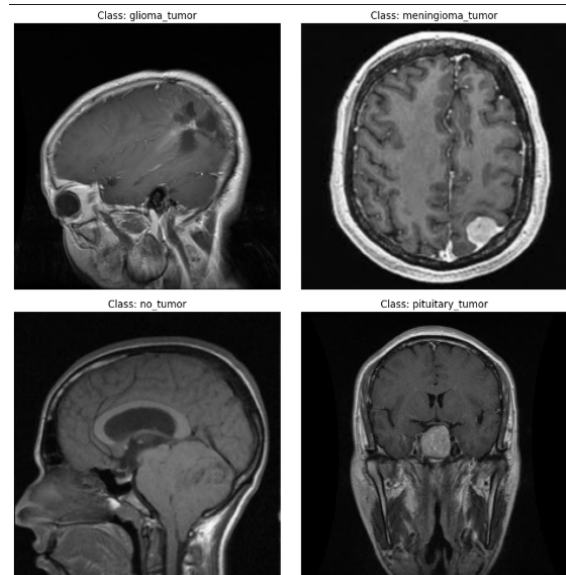Figure 1. Class Distribution of Brain Tumour Dataset

Figure 2. Sample MRI Images of Class Distribution

3.2. Data Preprocessing

To make the MRI images ready for training, a series of preprocessing tasks was executed to establish

consistency and enhance the model performance. To begin with, all images were resized to a standard size of 224×224 pixels, catering to the input shape expectations of CNN-based models such as VGG16, MobileNetV2, and EfficientNetB0. Normalization was achieved by rescaling pixel values to the [0, 1] range, which speeds up convergence during training and prevents gradient-related problems. The class labels were converted to one-hot encoded vectors to facilitate multi-class classification using the softmax activation function at the output layer. Data augmentation strategies were utilized to artificially boosted dataset variation and limited overfitting. These involved random horizontal and vertical flipping, slight rotations, zoom levels, and brightness levels, enabling models to be more generalizable to new unseen data. While the class distribution was quite balanced, class weights were calculated and used during training to tweak the learning process and maintain sensitivity across all classes. The dataset was split into training and testing datasets according to their original folder hierarchies, and precautions were taken to avoid data leakage between datasets. This extensive preprocessing pipeline helped ensure that all models received a uniform and varied input space for effective learning.

3.3. Model Architecture

To solve the issue of classifying brain tumours in MRI images into four classes, we have chosen four different convolutional neural network (CNN) architectures for this research: Custom CNN, MobileNetV2, VGG16, and EfficientNetB0. To take advantage of their unique strengths for feature extraction and managing processing capacity, we chose these to balance a homemade baseline with top-performing pre-trained models. To handle our four classes (glioma, meningioma, pituitary tumour, and no tumour), each was modified with a softmax layer. To avoid overfitting, we added dropout at 0.5, following dense layers. Where possible, we adapted the pre-trained weights of ImageNet to suit our specific requirements for medical imaging. The details of each model and its implementation are discussed in the following sections.

3.3.1. Custom CNN. Our Custom CNN, which we implemented from scratch, offers a versatile method for understanding and classifying tumors. It includes three convolutional blocks, each of which has a convolutional layer, batch normalization for stability, ReLU activation for non-linearity, and a 2×2 max-pooling layer to downsample spatial dimensions while preserving key features. The number of filters increased throughout the network, from 32 to 64 to 128, allowing the model to progressively learn from simple to complex patterns. The output of the convolutional layers was flattened and fed through two fully connected layers of 256 and 128 neurons, respectively, with a dropout layer rate of 0.5.

The last layer employs softmax activation to classify the four types of tumors.
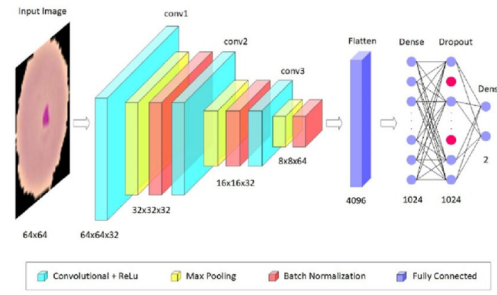

Figure 3. CNN architecture [15]

3.3.2. MobileNetV2. MobileNetV2, introduced by Sandler et al. in 2018, was selected because of its efficiency in resource-constrained applications, such as medical imaging on embedded hardware. It employs depth-wise separable convolutions and inverted residual blocks with linear bottlenecks to reduce computational complexity while maintaining accuracy. The model begins with a 3×3 convolutional block with 32 filters, followed by 17 bottleneck blocks with expansion factors of 1 or 6, and variable channel dimensions. We modified the pre-trained model by freezing the top 20 layers and adding a global average pooling layer, a dense layer with 128 neurons, a 0.5 dropout, and a softmax classifier for our task.
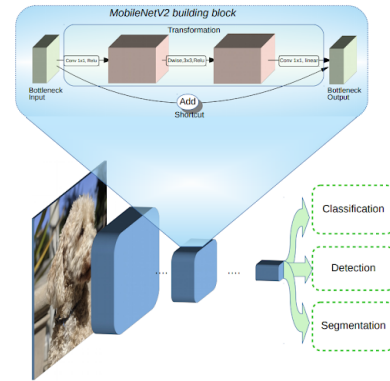

Figure 4. MobileNetV2 architecture [16]

3.3.3. VGG16. VGG16, introduced by Simonyan and Zisserman in 2014, is known for its simple and uniform architecture. It consists of 13 convolutional layers organized into five blocks, each followed by a 2×2 max pooling layer and three fully connected layers. The filter size progressively increased from 64 to 512, enabling the model to learn hierarchical spatial features. For our task, we customized the final classifier by replacing the initial dense layers with a 256-neuron fully connected layer, followed by global average pooling

and a softmax output. Dropout and batch normalization were incorporated to enhance generalization and the stability of training. In addition, we unfroze the last five convolutional blocks to facilitate domain-specific fine-tuning.
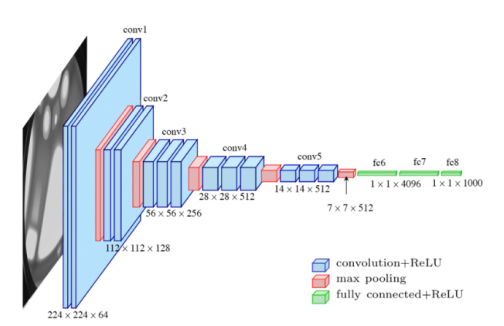

Figure 5. VGG16 architecture [17]

3.3.4. EfficientNetB0.
EfficientNetB0, proposed by Tan and Le in 2019, utilizes compound scaling to balance network width, depth, and resolution. It starts with a 3×3 convolutional layer with 32 filters, followed by a series of MBConv blocks incorporating depthwise convolutions, Squeeze-and-Excitation optimization, and residual connections. These blocks progressively scale from 16 to 1280 channels. For our task, we modified the pre-trained model by adding a global average pooling layer, a 128-unit dense layer, a 0.5 dropout layer, and a softmax classifier. We fine-tuned the top 15% of the network with a reduced learning rate to adapt to MRI-specific features.

3.3.5. 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.

3.4. Grad-CAM Visualization

To interpret the model decisions in brain tumor classification, we used Grad-CAM to visualize the important regions in the MRI scans. Heatmaps showed that EfficientNetB0 and VGG16 consistently focused on tumor-affected areas, particularly gliomas and pituitary tumors. MobileNetV2 and the Custom CNN showed a broader focus owing to their lighter architectures. The ensemble model combined their strengths to produce sharp, localized heatmaps. These results enhance trust in the model predictions and support their clinical applicability by providing visual insights for radiologists.

4. Results & Discussion

This study evaluated five deep learning models for brain tumor classification using MRI scans. The Custom CNN achieved 96.79% accuracy, demonstrating that even simpler architectures can perform well with proper regularization. MobileNetV2 reached 97.38% accuracy by leveraging transfer learning, whereas VGG16 attained 96.99% despite its deeper architecture. EfficientNetB0 outperformed all individual models with an accuracy of 97.62%, benefiting from its optimized design. However, the ensemble model, which combined predictions from all four models via soft voting, achieved the highest performance, with an accuracy of 98.93%, along with superior precision, recall, and F1-scores.

TABLE 2. PERFORMANCE METRICS OF DIFFERENT 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 Custom-CNN and VGG16 exhibited slight fluctuations in the validation performance compared to MobileNetV2.

The test performance Figure 6 comparison of the five models was based on accuracy, precision, recall, and F1-score. The Ensemble Model achieved the best results across all metrics, followed closely by EfficientNetB0 and VGG16, indicating a strong and balanced classification performance.
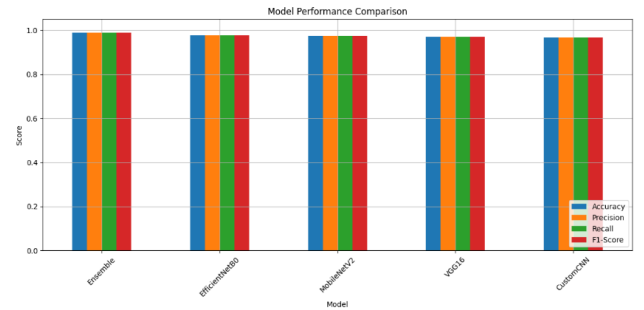

Figure 6. Model Performance Comparison

The Custom CNN model performed well, particularly in detecting pituitary tumors, but struggled to distinguish gliomas from meningiomas. MobileNetV2 also exhibited strong classification capability with a balanced performance across classes, although it exhibited some confusion between gliomas and meningiomas. VGG16 in Figure 9 reliably classified menin-
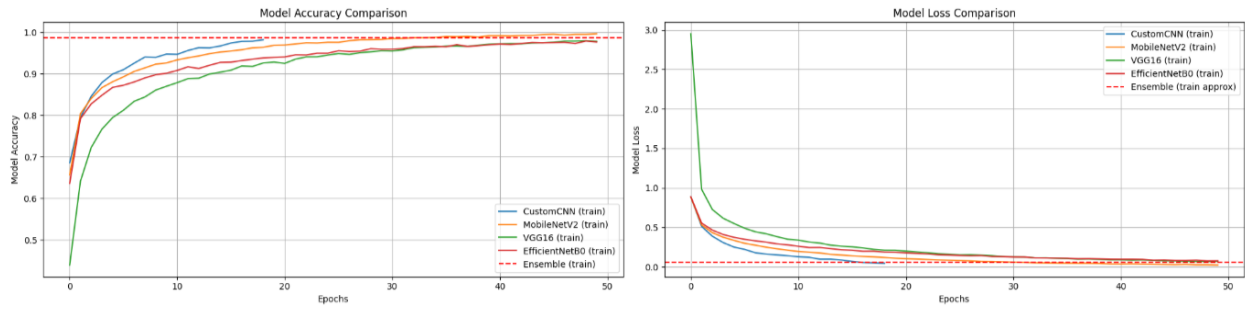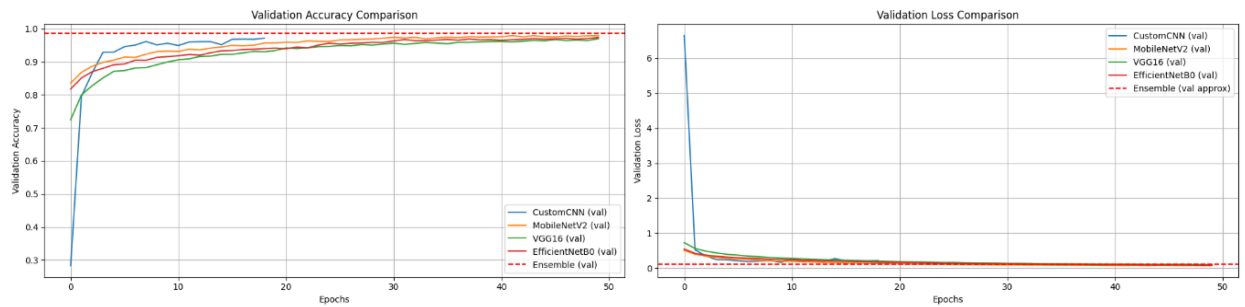
Figure 7. Training Accuracy and Loss Curves

Figure 8. Validation Accuracy and Loss Curves

giomas and pituitary tumors but had slight difficulty with glioma classification, indicating a feature overlap. EfficientNetB0 delivered a higher accuracy across all tumor types, notably reducing misclassifications between gliomas and meningiomas. Overall, the Ensemble Model in Figure 10 achieved the best results, offering the highest accuracy and most balanced predictions with minimal misclassifications across all categories.
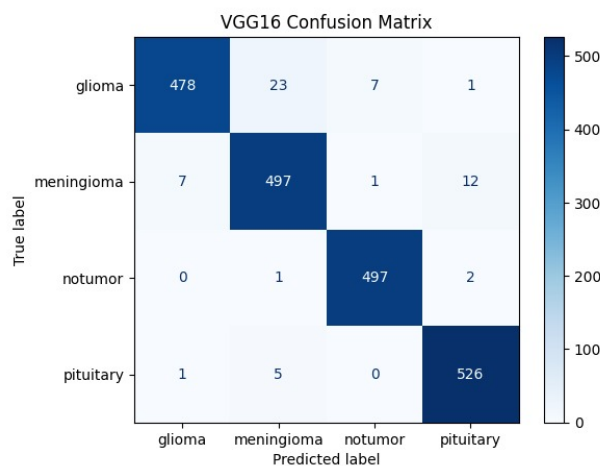

Figure 10. Confusion Matrix in Ensemble Model
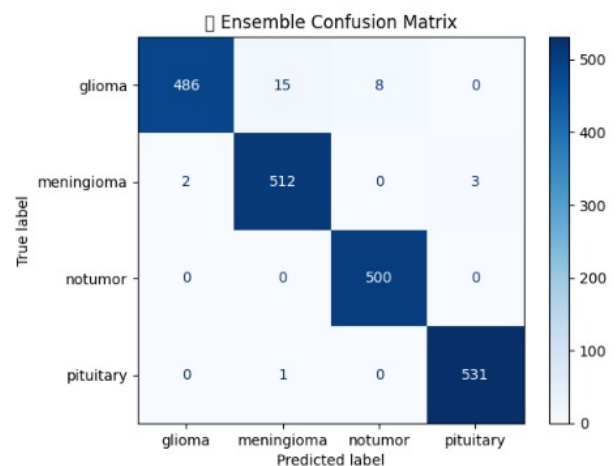

Figure 9. Confusion Matrix in VGG16

Grad-CAM visualizations Figure 11 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.
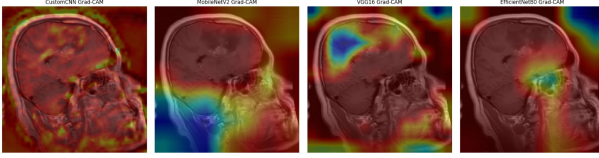
Figure 11. GradCAM Visualization

5. Conclusion & Future Directions

This study presents a comparative analysis of four CNN models—Custom CNN, MobileNetV2, VGG16, and EfficientNetB0—for classifying brain tumors from MRI scans. Through fine-tuning, data augmentation, and class balancing, all models achieved robust classification performance. EfficientNetB0 emerged as the top single model, while the soft-voting ensemble model attained the highest accuracy of 98.93%. Grad-CAM visualizations confirmed model interpretability by highlighting tumor-related regions in MRIs. These results validate the efficacy of transfer learning and ensemble techniques for medical image classification, establishing a strong foundation for clinical decision support systems.

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 68, no. 6, pp. 394–424, 2018.

[2] Y. Liu, A. B. Carpenter, C. J. Pirozzi, H. Yuan, M. S. Waitkus, Z. Zhou, L. Hansen, M. Seywald, R. Odion, and P. K. Greer, "Non-invasive sensitive brain tumor detection using dual modality bioimaging nanoprobe," *Nanotechnology*, vol. 30, no. 27, p. 275101, 2019.

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

[4] S. M. Boby and S. Sharmin, "Medical Image Denoising Techniques against Hazardous Noises: An IQA Metrics Based Comparative Analysis," *International Journal of Image, Graphics and Signal Processing*, vol. 13, no. 2, pp. 25–43, 2021.

[5] N. Arunkumar, M. AbedMohammed, S. A. Mostafa, D. Ahmed Ibrahim, J. J. Rodrigues, and V. H. C. de Albuquerque, "Fully automatic model-based segmentation and classification approach for MRI brain tumor using artificial neural networks," *Concurrency and Computation: Practice and Experience*, vol. 32, no. 1, pp. 1–9, 2018.

[6] P. Afshar, K. N. Plataniotis, and A. Mohammadi, "Capsule networks for brain tumor classification based on MRI images and coarse tumor boundaries," in *Proc. IEEE Int. Conf. on Image Processing*, pp. 1368–1372, 2019.

[7] Y. Xie, F. Zaccagna, L. Rundo, C. Testa, R. Agati, R. Lodi, D. N. Manners, and C. Tonon, "Convolutional neural network techniques for brain tumor classification (from 2015 to 2022): Review, challenges, and future perspectives," *Diagnostics*, vol. 12, no. 8, p. 1850, 2022.

[8] P. S. L. Veeranki, G. L. Banavath, and P. R. Devi, "Detection and Classification of Brain Tumors using Convolutional Neural Network," in *Proc. IEEE Int. Conf.*, pp. 780–786, 2023.

[9] N. Remzan, K. Tahiry, and A. Farchi, "Deep Learning Approach for Brain Tumor Classification Implemented in Raspberry Pi," in *Proc. Springer Conf.*, pp. 136–147, 2022.

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

[11] A. M. Hashan, E. Agbozo, A. A. K. Al-Saeedi, S. Saha, A. Haidari, and M. N. F. Rabi, "Brain Tumor Detection in MRI Images Using Image Processing Techniques," in *Proc. IEEE Int. Conf.*, pp. 24–28, 2021.

[12] 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," *Biocybernetics and Biomedical Engineering*, vol. 39, no. 1, pp. 63–74, 2019.

[13] S. Chitnis, R. Hosseini, and P. Xie, "Brain tumor classification based on neural architecture search," *Scientific Reports*, vol. 12, no. 1, p. 1920, 2022.

[14] A. Pashaei, H. Sajedi, and N. Jazayeri, "Brain tumor classification via convolutional neural network and extreme learning machines," in *Proc. IEEE Conf.*, pp. 314–319, 2018.

[15] Sun, Y., Xue, B., Zhang, M., & Yen, G. G. (2019). Completely automated CNN architecture design based on blocks. IEEE transactions on neural networks and learning systems, 31(4), 1242-1254.

[16] ——, "Mobilenetv2: Inverted residuals and linear bottlenecks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4510–4520.

[17] A. Mustapha, L. Mohamed, H. Hamid, and K. Ali, "Diabetic retinopathy classification using resnet50 and vgg-16 pretrained networks," Interna tional Journal of Computer Engineering and Data Science, vol. 1, no. 1, pp. 1–7, 2021.