Explainable Deep Learning for Brain Tumor Classification: An EfficientNet-CBAM with XAI Integration

MD.Amanour Rahman*, Md.Noman Ehsan†, Rowzatul Zannath Prerona‡
*Department of Computer Science, Ahsanullah University of Science and Technology
Dhaka, Bangladesh
ID: 20210204010, 202210204019, 20210204018

***Abstract*—**
***Index Terms*—IEEEtran, conference, template, text-only, LaTeX**

I. Introduction

The human brain, an organ of unparalleled complexity, serves as the epicenter of cognition, consciousness, and motor control. However, its intricate neural architecture is vulnerable to the anomalous proliferation of cells, leading to the formation of brain tumors. This condition represents a significant global health challenge, with brain and other central nervous system cancers contributing substantially to cancer-related mortality worldwide. According to the World Health Organization (WHO), the prognosis for patients with brain tumors is critically dependent on early and accurate diagnosis, which dictates the subsequent therapeutic pathway and ultimately, patient outcomes. Magnetic Resonance Imaging (MRI) has emerged as the non-invasive gold standard for visualizing brain structures, offering exceptional soft-tissue contrast that is indispensable for identifying neoplastic growths.

Despite the diagnostic power of MRI, the manual interpretation of scans by radiologists is a demanding task fraught with challenges. These include the subtle morphological similarities between different tumor types (e.g., glioma, meningioma, and pituitary tumors), variations in tumor size, shape, and location, and the sheer volume of images requiring analysis. This intricate process is not only time-consuming but also susceptible to inter-observer variability, which can impact diagnostic consistency. Consequently, there is a pressing need for automated, accurate, and reliable Computer-Aided Diagnosis (CAD) systems to augment clinical workflows and support radiologists in making timely and precise decisions.

In recent years, the field of medical image analysis has been revolutionized by the advent of deep learning, particularly Convolutional Neural Networks (CNNs). Transfer learning, leveraging pre-trained models such as ResNet50, MobileNetV2, and EfficientNet, has proven exceptionally effective, enabling the development of high-performance classifiers even with limited medical data. While these models have set new benchmarks in classification accuracy, a significant challenge remains: their "black box" nature. For a model to be adopted in high-stakes clinical environments, exceptional accuracy alone is insufficient; it must also be interpretable. Clinicians need to trust and understand the reasoning behind a model's prediction. This has catalyzed research into Explainable AI (XAI), a field dedicated to making AI systems more transparent. Furthermore, a new frontier in model optimization involves the integration of attention mechanisms, which empower models to mimic human expert focus by dynamically highlighting the most salient features within an image, thereby enhancing diagnostic precision.

This paper addresses these critical gaps by proposing a novel, attention-guided deep learning framework that is not only highly accurate but also transparent and trustworthy. We present a comprehensive study that compares multiple state-of-the-art architectures and introduces a new model that significantly advances the task of multi-class brain tumor classification. The primary contributions of our research are summarized as follows:

1) **Proposed a Novel Attention-Guided Architecture:** We introduce the EfficientNet-CBAM model, which strategically integrates a Convolutional Block Attention Module (CBAM) into a powerful EfficientNetB3 backbone. This hybrid architecture enhances feature discrimination by focusing on the most pathologically relevant regions within the MRI scans.
2) **Achieved State-of-the-Art Performance:** Through rigorous experimentation, our model achieved a classification accuracy of **99.29%**. We present a comparative analysis against baseline models including ResNet50, MobileNetV2, and EfficientNetB3.
3) **Ensured Clinical Interpretability with Explainable AI (XAI):** We implement the gradient-based visualization technique HiResCAM, generating heatmaps that visually explain the model's predictions by highlighting tumor regions that most influenced its decision.
4) **Validated Model Reliability:** We analyze prediction confidence scores, visualizing correctly and incorrectly classified images with their confidence levels, providing

deeper insight into model certainty.

II. Related Work

Brain tumor classification from MRI images has been extensively studied using both traditional machine learning and deep learning approaches. Classical image segmentation techniques, such as watershed, fuzzy logic, and optimization-based methods, rely on handcrafted features and often require manual intervention, which limits scalability and generalizability. Deep learning models, particularly convolutional neural networks (CNNs) such as VGG, ResNet, DenseNet, MobileNet, and EfficientNet, automatically learn hierarchical features, achieving superior performance but often at the cost of high computational complexity and limited interpretability.

Recent studies have focused on enhancing both performance and explainability:

- M, M.M. et al. [1] proposed a framework using ResNet50 combined with Grad-CAM, achieving 98.52% testing accuracy and precision/recall above 97%. Grad-CAM visualizations provided explainable insights aligning with radiological assessments, demonstrating robustness while highlighting challenges related to dataset size and diversity.
- Nahiduzzaman et al. [2] introduced a hybrid explainable model combining a lightweight Parallel Depthwise Separable CNN (PDSCNN) with a Hybrid Ridge Regression Extreme Learning Machine (RRELM), enhanced with CLAHE preprocessing and SHAP-based interpretability. Their model reduced parameters to 0.53M while achieving 99.22% accuracy, outperforming pseudoinverse-based ELM and several state-of-the-art CNN models, demonstrating efficiency, robustness, and transparency.
- An ensemble approach [3] combined MobileNetV2 and DenseNet121 via soft voting and integrated Grad-CAM++ with a Clinical Decision Rule Overlay (CDRO) based on tumor size, location, and enhancement patterns. Evaluated on the Figshare brain tumor dataset (3,064 images), the ensemble achieved 91.7% accuracy and F1-score of 91.6%, with explainability maps strongly aligned with expert annotations (Dice: 0.88, IoU: 0.78), and radiologists rated the explanations as clinically useful (4.4/5).
- A. Rahman et al. [4] proposed a customized CNN enhanced with multiple XAI techniques—SHAP, LIME, and Grad-CAM—for brain tumor detection, trained on the BR35H dataset (3,060 images). The model achieved 100% training accuracy and 98.67% validation accuracy, with strong precision, recall, and F1-scores. On an external dataset, it demonstrated 92% accuracy, providing robust and clinically interpretable visual explanations.

These studies highlight a clear trend toward hybrid, ensemble, and explainable architectures that combine high accuracy with interpretability, paving the way for clinically trustworthy AI-based brain tumor detection systems.

A. Dataset Description

To develop a reliable and generalizable deep learning model for brain tumor classification, this study employed a composite dataset created by merging two distinct and publicly available Magnetic Resonance Imaging (MRI) collections. The integration of multiple sources was deliberately chosen to ensure heterogeneity in terms of imaging protocols, patient demographics, and clinical variations, thereby improving the robustness of the model.

1) Dataset A: Brain Tumor Classification MRI (Kaggle): The first dataset was obtained from Kaggle, consisting of 7,023 human brain MRI scans sourced from three repositories: Figshare, the SARTAJ dataset, and Br35H. These scans were categorized into four classes: glioma, meningioma, pituitary tumor, and no tumor.

During curation, it was identified that the glioma subset in the SARTAJ repository contained potentially mislabeled cases. To maintain data integrity, these questionable samples were excluded and replaced with verified glioma images from the Figshare collection. The no tumor class was sourced exclusively from the Br35H dataset.

A key characteristic of Dataset A is its variability in image dimensions and background margins, necessitating further preprocessing steps to standardize the inputs prior to model training.

2) Dataset B: PMRAM Bangladeshi Brain Cancer MRI Dataset: To further enhance clinical and demographic diversity, the PMRAM Bangladeshi Brain Cancer MRI Dataset was incorporated. This dataset was collected across multiple hospitals in Bangladesh with the direct involvement of four senior medical professionals, ensuring high-quality clinical validation.

It originally contained 1,600 raw MRI scans, which were subsequently augmented to achieve a balanced distribution of 6,000 images (1,500 per class across glioma, meningioma, pituitary tumor, and no tumor). Unlike Dataset A, all images in this collection were uniformly pre-processed to 512×512 pixels, making them directly suitable for deep learning applications.

3) Final Merged Corpus: The curated versions of Dataset A and Dataset B were merged to construct the final experimental corpus consisting of 11,716 MRI scans equally distributed across the four classes. The fusion of a globally aggregated dataset with a geographically localized dataset was designed to enhance both accuracy and clinical applicability, enabling the model to generalize well across diverse populations and imaging conditions.

B. Data Preprocessing and Augmentation

A rigorous preprocessing and augmentation pipeline was designed to prepare the heterogeneous dataset for effective model training. These steps ensure standardized inputs, reduce bias, and improve the generalization capability of the model.

1) Image Standardization: All images, irrespective of their source, were subjected to the following preprocessing steps:

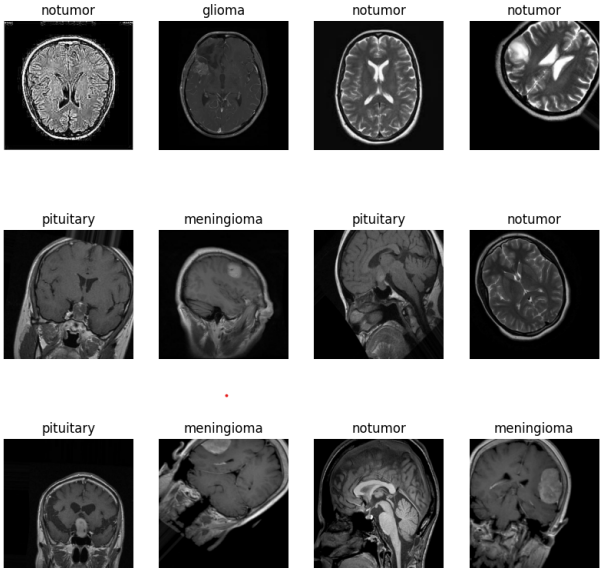

Fig. 1. Representative MRI sample images from the merged dataset across different tumor classes.

- Resizing: Every image was resized to 224×224 pixels to match the input requirements of Convolutional Neural Networks (CNNs).
- Normalization: Pixel intensities were rescaled from the original range [0, 255] to [0, 1] by applying a normalization factor of 1/255. This step stabilizes gradient updates and accelerates convergence during training.

2) Data Augmentation: To combat overfitting and increase robustness to real-world variations, on-the-fly augmentation was applied to the training set. The augmentation pipeline included:

- Random horizontal flipping (mirroring along the vertical axis).
- Random rotation (slight angle variations up to ± 0.02).
- Random contrast adjustment (± 0.1 intensity variation).
- Random zoom and translation (minor rescaling and shifting to mimic positional variations).

This augmentation strategy ensured that the network learned discriminative tumor features instead of memorizing orientation, illumination, or scanner-specific artifacts.

3) Dataset Splitting: The final merged corpus (11,716 images) was partitioned into three disjoint sets to enable training, validation, and unbiased performance evaluation:

- Training Set: 9,373 images (80%)
- Validation Set: 2,343 images (20%)
- Test Set: 1,311 images (independent hold-out set)

All images were fed to the model in mini-batches of 32 during training and evaluation. This systematic division ensures robust performance estimation while minimizing data leakage across the experimental pipeline.

4) Dataset Composition Summary: The overall distribution of images across Dataset A, Dataset B, and the final merged corpus is summarized in Table I.

TABLE I
CLASS-WISE DISTRIBUTION OF MRI IMAGES IN DATASET A, DATASET B, AND THE MERGED DATASET.

Class	Dataset A	Dataset B	Final Merged
Glioma	1,626	1,500	3,126
Meningioma	1,644	1,500	3,144
Pituitary Tumor	1,910	1,500	3,410
No Tumor	1,843	1,500	3,343
Total	**7,023**	**6,000**	**11,716**

C. Hyperparameter Settings

The training of the proposed EfficientNet-CBAM model was carried out using the hyperparameters listed in Table II. These hyperparameters were selected through empirical tuning to balance performance and computational efficiency.

Hyperparameter	Value
Input Image Size	$224 \times 224 \times 3$
Base Architecture	EfficientNetB3 (ImageNet pre-trained)
Attention Module	CBAM (Channel + Spatial)
Number of Classes	4
Optimizer	Adam
Learning Rate (initial)	1×10^{-3} (with ReduceLROnPlateau)
Loss Function	Sparse Categorical Crossentropy
Batch Size	32 (default of `tf.data`)
Epochs	40
Dropout Rate	0.5 (after Global Average Pooling)
Early Stopping Patience	5 epochs
Learning Rate Reduction Factor	0.2
Minimum Learning Rate	1×10^{-7}
Validation Monitor Metric	Accuracy (maximization)
Checkpoint Saving	Best model (`val_accuracy`)
Carbon Emission Logging	Enabled (custom callback)

TABLE II
SUMMARY OF HYPERPARAMETER SETTINGS USED FOR TRAINING THE PROPOSED EFFICIENTNET-CBAM MODEL.

III. PROPOSED METHODOLOGY

A. *Base Network: EfficientNetB3*

In this study, EfficientNetB3 was chosen as the backbone due to its superior scaling efficiency and favorable balance between accuracy and computational cost. Unlike conventional convolutional neural networks (CNNs), EfficientNet employs a *compound scaling method* that uniformly scales network depth, width, and input resolution in a balanced manner. This design allows EfficientNet to achieve high accuracy with significantly fewer parameters compared to traditional architectures. EfficientNetB3, in particular, is well-suited for medical image analysis tasks, as it captures both fine-grained tumor features and global contextual information.

B. *Convolutional Block Attention Module (CBAM)*

To enhance feature representation, we integrated the Convolutional Block Attention Module (CBAM) with EfficientNet. CBAM improves model performance by learning to emphasize "*what*" features are important (channel attention) and "*where*" in the image to focus (spatial attention).

1) Channel Attention (What to focus on): Channel attention refines the feature maps by computing global average pooling and global max pooling across the spatial dimensions. These pooled features are processed through shared multilayer perceptrons and then merged. The result is a channel-wise attention map that highlights discriminative features (e.g., tumor textures, lesion boundaries) while suppressing irrelevant ones.

2) Spatial Attention (Where to focus on): Spatial attention identifies the most relevant regions of the feature maps. It applies global average pooling and max pooling across channels, concatenates the results, and passes them through a convolutional layer to generate a spatial attention map. This map directs the model to tumor regions and reduces background noise.

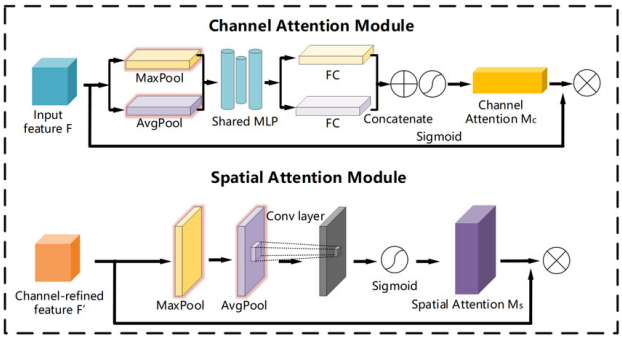

Fig. 2. The CBAM module, consisting of sequential Channel Attention and Spatial Attention.

C. Explainable AI with HiResCAM

To enhance the interpretability of the proposed EfficientNet-CBAM model, we implemented an Explainable AI (XAI) approach using HiResCAM, a gradient-based visualization technique. HiResCAM highlights the regions in an MRI image that contribute most significantly to the model's prediction, providing insights into model decision-making and aiding clinical validation.

The process involves the following steps:

1) **Identify the last convolutional layer:** The last Conv2D layer of the trained model is located, as it contains the richest spatial features.
2) **Compute gradients:** Using a gradient tape, the gradient of the predicted class score with respect to the output feature map of the last convolutional layer is computed.
3) **Generate the HiResCAM heatmap:** Element-wise multiplication of the gradients and feature maps is performed, followed by a summation across channels. The resulting heatmap is normalized to highlight salient regions.
4) **Superimpose on the original image:** The heatmap is resized to match the input image dimensions and overlaid on the original MRI scan, providing a visual explanation of the model's focus.
5) **Visualization:** For multiple images, the original and heatmap-superimposed images are displayed side-by-side, along with predicted class labels and confidence scores.

D. Proposed EfficientNet-CBAM Architecture with Explainable AI

The proposed EfficientNet-CBAM architecture integrates CBAM directly after the EfficientNet feature extraction stage to enhance feature representation. Additionally, we incorporate an *Explainable AI (XAI)* module using HiResCAM to visualize model attention on relevant tumor regions. The processing pipeline is as follows:

1) EfficientNetB3 extracts rich hierarchical features from MRI scans.
2) The extracted feature maps are refined using the CBAM block, which sequentially applies channel and spatial attention.
3) Global Average Pooling is applied to aggregate the refined features.
4) A dropout layer is included to reduce overfitting.
5) A dense softmax classifier predicts one of the four target classes: glioma, meningioma, pituitary tumor, or no tumor.
6) HiResCAM visualization: The last convolutional layer of the model is used to compute a heatmap of class-specific gradients, which is then superimposed on the original MRI scan. This highlights the regions contributing most to the predicted class, providing interpretability and clinical validation.

By combining the efficient feature scaling of EfficientNet, the refinement ability of CBAM, and the interpretability provided by HiResCAM, the proposed model achieves enhanced accuracy, robustness, and transparency in brain tumor classification.

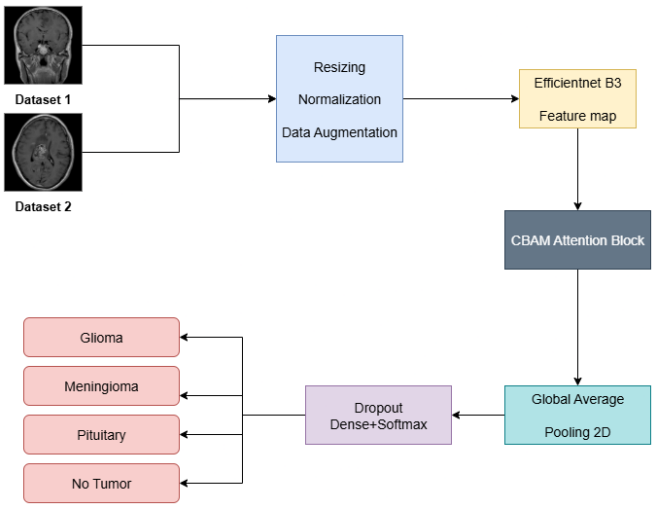

Fig. 3. Proposed EfficientNet-CBAM architecture with HiResCAM explainability.

IV. Results

A. Evaluation Metrics

To assess the performance of the proposed model and baseline models, several evaluation metrics were computed, including Accuracy, Precision, Recall, and F1-score.

Accuracy is the ratio of correctly predicted observations to the total observations:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision (Macro-Average) across all classes:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

Recall (Macro-Average) across all classes:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

F1-Score (Macro-Average):

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Where: TP = True Positive, FP = False Positive, TN = True Negative, FN = False Negative

B. Training and Validation Performance

The learning curves of the proposed model are shown in Figure 4, displaying both loss and accuracy trends over epochs.

C. Confusion Matrix

The confusion matrix in Figure 5 provides a detailed view of model performance for each class, showing correct predictions along the diagonal and misclassifications off-diagonal.

D. Experimental Results

Table III summarizes the performance of baseline models and the proposed model. The proposed EfficientNet-CBAM with HiResCAM achieves the highest precision, recall, and F1-score, demonstrating superior classification performance.

TABLE III
PERFORMANCE COMPARISON OF BASELINE MODELS AND PROPOSED MODEL.

Model Name	Precision	Recall	F1-Score	Accuracy
ResNet50	0.88	0.88	0.87	0.88
EfficientNetV2B3	0.91	0.90	0.90	0.90
MobileNetV2	0.76	0.77	0.75	0.78
EfficientNetB3+CBAM	0.99	0.99	0.99	0.99

TABLE IV
CARBON EMISSION (KG CO2 EQ) AND RUNTIME FOR DIFFERENT MODELS.

Model Name	Carbon Emission	Runtime (m)
ResNet50	0.0274	18.51
EfficientNetV2B3	0.0458	31.69
MobileNetV2	0.0240	16.92
EfficientNetB3+CBAM	0.6426	53.55

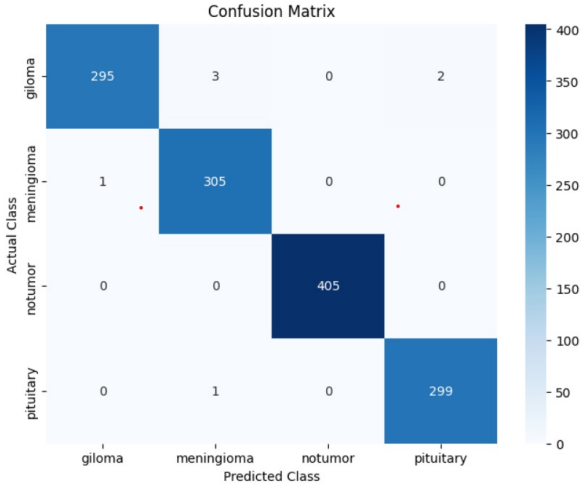

Fig. 5. Confusion matrix of the proposed model on the test dataset.

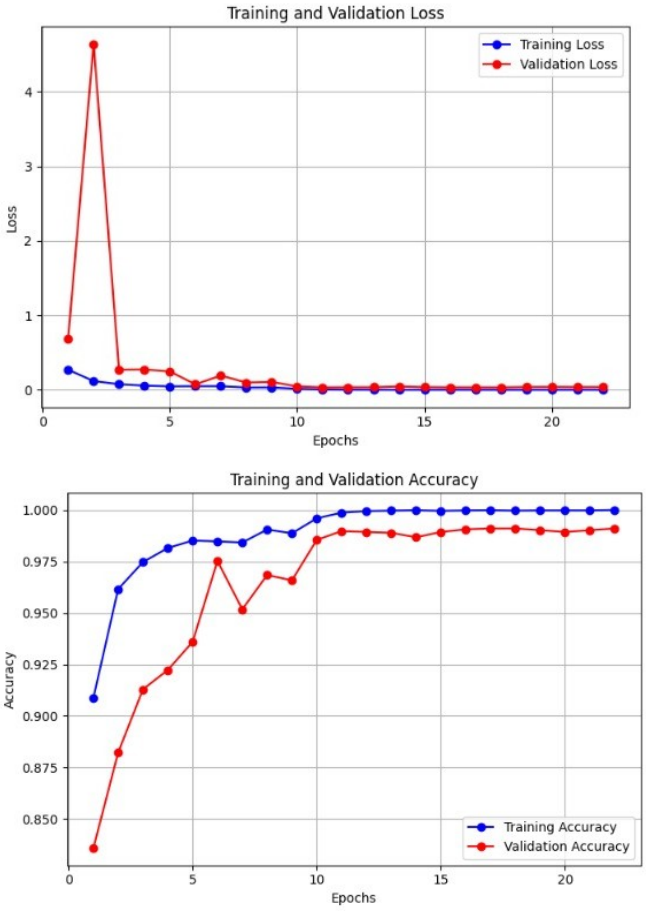

Fig. 4. Training and validation performance: Left – loss; Right – accuracy.

E. Explainable AI Visualization

To interpret the decision-making process of the proposed EfficientNet-CBAM model, we generated visual explanations using HiResCAM. Figure 6 shows the model's focus on MRI scans for each brain tumor class. The heatmaps highlight

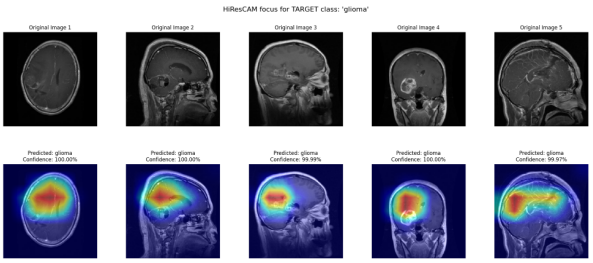

Fig. 6. HiResCAM-based Explainable AI visualization.

the regions that contributed most significantly to the predicted class, enabling clinicians to understand and validate the model's predictions.

V. Conclusion

In this study, we proposed a novel attention-guided deep learning framework, EfficientNet-CBAM, for multi-class brain tumor classification from MRI scans. By integrating the Convolutional Block Attention Module (CBAM) with the EfficientNetB3 backbone, the model effectively focuses on the most pathologically relevant regions, enhancing discriminative feature learning. Additionally, we incorporated Explainable AI (XAI) through HiResCAM to provide visual interpretability, allowing clinicians to validate the model's predictions and gain insights into its decision-making process.

Extensive experiments on a merged dataset of 11,716 MRI scans demonstrated the superiority of the proposed model over several state-of-the-art baselines, including ResNet50, MobileNetV2, and EfficientNetV2B3. The EfficientNet-CBAM achieved a remarkable classification accuracy of **99.29%**, along with high precision, recall, and F1-score across all classes, establishing it as a highly reliable and clinically interpretable solution.

The combination of high performance, attention-guided feature refinement, and explainability underscores the potential of the proposed framework for deployment in real-world clinical settings, facilitating timely and accurate brain tumor diagnosis while maintaining transparency and trust in AI-assisted medical decision-making.

Future work may explore further optimization using lightweight attention modules, multi-modal MRI sequences, and integration with automated segmentation to enhance generalizability and clinical utility across diverse patient populations.

References

[1] M, M.M., T. R, M., V, V.K. et al. Enhancing brain tumor detection in MRI images through explainable AI using Grad-CAM with ResNet50. *BMC Medical Imaging*, 24, 107 (2024). https://doi.org/10.1186/s12880-024-01292-7

[2] Nahiduzzaman, M., Abdulrazak, L.F., Kibria, H.B. et al. A hybrid explainable model based on advanced machine learning and deep learning models for classifying brain tumors using MRI images. *Scientific Reports*, 15, 1649 (2025). https://doi.org/10.1038/s41598-025-85874-7

[3] arXiv:2508.06891 Ensemble-based deep learning framework for brain tumor classification using MobileNetV2 and DenseNet121 with Grad-CAM++ and Clinical Decision Rule Overlay. *arXiv preprint* (2025).

[4] A. Rahman, S. S. Sohail, M. S. Alam, A. Sharma and W. Mansoor, Detecting Brain Cancer Using Explainable AI, in *2024 7th International Conference on Signal Processing and Information Security (ICSPIS)*, Dubai, United Arab Emirates, 2024, pp. 1-6, doi: 10.1109/ICSPIS63676.2024.10812596.