Interpretable Brain Tumor Classification from MRI with EfficientNet and Post-hoc XAI: Grad-CAM, Score-CAM, SHAP, and LIME

Rafa Khaleel Inamdar, Sanjna Deva, Vyoma Mody and Prof. Sofia Francis
Department of Computer Engineering, NMIMS MPSTME, Mumbai, India

***Abstract*—Early and accurate detection of brain tumors through Magnetic Resonance Imaging (MRI) is vital for patient prognosis and treatment planning. Deep convolutional neural networks (CNNs) have achieved remarkable success in medical image classification, yet their lack of interpretability poses challenges for clinical adoption. This research presents a comprehensive comparative analysis of three state-of-the-art CNN architectures — ResNet50, InceptionV3, and EfficientNetB4 — for multi-class brain tumor classification enhanced with Explainable AI (XAI) methods. Each model was trained on the Kaggle Brain Tumor MRI Dataset consisting of four categories: glioma, meningioma, pituitary, and no tumor. Models were evaluated using Accuracy, Precision, Recall, F1-score, Area Under the Curve (AUC), and Cohen's Kappa, while model interpretability was assessed through Grad-CAM, Score-CAM, SHAP, and LIME visualizations. Quantitative XAI evaluation employed Deletion AUC, Insertion AUC, and Faithfulness Correlation metrics. InceptionV3 achieved the highest test accuracy (97%) and F1-score (0.97), whereas EfficientNetB4 demonstrated superior interpretability with the highest Insertion AUC (0.742) and Macro AUC (0.977). The study emphasizes that optimizing both performance and explainability is essential for developing clinically deployable AI systems that foster transparency, safety, and physician trust.**

***Index Terms*—Brain Tumor Classification, Deep Learning, ResNet50, InceptionV3, EfficientNetB4, Explainable AI, Grad-CAM, Score-CAM, SHAP, LIME**

I. Introduction

Brain tumors represent a major global health concern, accounting for a significant proportion of neurological morbidity and mortality. The ability to detect and classify tumors at an early stage from MRI scans plays a pivotal role in determining appropriate therapeutic interventions. Traditional manual diagnosis by radiologists, while effective, is time-consuming and susceptible to human error. Automated deep learning-based diagnostic systems have emerged as powerful tools to support clinicians, offering high accuracy and consistency across imaging datasets.

Deep Convolutional Neural Networks (CNNs) such as ResNet50, InceptionV3, and EfficientNetB4 have demonstrated state-of-the-art performance across various computer vision domains, including medical imaging. These architectures differ in depth, receptive field, and parameter optimization strategy, enabling a diverse evaluation of feature representation and generalization potential. However, CNNs are often criticized for their "black-box" nature — while they achieve superior predictive accuracy, they lack transparency regarding how specific predictions are made. This limitation restricts clinical integration where interpretability and accountability are non-negotiable.

Explainable Artificial Intelligence (XAI) aims to resolve this opacity by making model decisions understandable and traceable. Visualization techniques like Grad-CAM and Score-CAM generate class-discriminative heatmaps, highlighting regions influencing a prediction, while SHAP and LIME provide feature-importance analyses rooted in statistical and surrogate modeling principles. Incorporating such interpretability mechanisms is essential for gaining clinician confidence, enabling diagnostic verification, and meeting regulatory standards for AI-driven medical systems.

This paper presents a comparative study involving ResNet50, InceptionV3, and EfficientNetB4 for multi-class brain tumor classification using MRI data. The contribution of this work is threefold:

1) A detailed benchmarking of three CNN architectures in terms of performance, robustness, and computational efficiency.
2) Integration of post-hoc XAI techniques (Grad-CAM, Score-CAM, SHAP, LIME) with both qualitative and quantitative evaluation.
3) Discussion of clinical applicability, transparency, and reliability of CNN-based tumor classification systems.

II. Related Work

Deep learning has significantly advanced medical image analysis, with CNNs leading the revolution in diagnostic automation. He et al. [1] proposed ResNet, introducing skip connections that allow deeper networks to converge without degradation. Szegedy et al. [2] introduced InceptionV3, employing multi-scale convolutional kernels to capture complex visual hierarchies. Tan and Le [3] presented EfficientNet, using compound scaling to optimize accuracy and computational cost jointly.

In the medical domain, CNNs have been successfully applied for tumor segmentation and classification. Hossain et al. [7] demonstrated CNNs for binary tumor detection, while Chakrabarty et al. [8] extended to multi-class classification with enhanced transfer learning. However, most prior studies primarily focus on performance metrics while disregarding explainability.

Recent advances in XAI provide interpretative insights into CNN reasoning. Grad-CAM [4] and Score-CAM visualize class-relevant regions, while SHAP [5] and LIME [6] quantify local feature importance. Studies by Reyes et al. [9] and Ghosh et al. [10] emphasize the necessity of integrating explainability for clinical readiness. Our work extends these by incorporating both visual and metric-based XAI analysis across multiple architectures.

III. Methodology

A. Dataset and Preprocessing

The Kaggle Brain Tumor MRI dataset contains 7,878 MRI images across four classes: glioma, meningioma, pituitary, and no tumor. Each image is grayscale but represented as RGB for CNN compatibility. Images were resized to 224×224 for ResNet50 and EfficientNetB4, and 299×299 for InceptionV3. Dataset augmentation included random flips, zoom, and brightness adjustment to prevent overfitting. The dataset was split as 80% for training, 10% for validation, and 10% for testing.

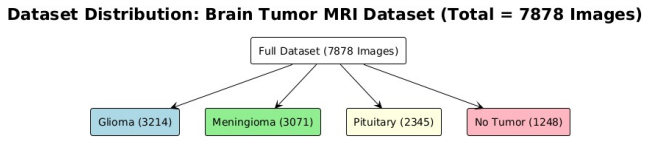

Fig. 1: Dataset distribution across four MRI tumor classes (glioma, meningioma, pituitary, and no tumor). This visualization illustrates class representation within the dataset and the partitioning strategy for training, validation, and testing.

The dataset visualization in Fig. 1 provides an overview of the class balance within the Kaggle Brain Tumor MRI dataset. It shows that glioma and meningioma categories contain a larger number of samples compared to pituitary and no tumor, indicating mild class imbalance. This informed the decision to apply image augmentation and weighted loss functions to ensure uniform learning across all tumor categories. The figure also visually represents the 80–10–10 data split used for training, validation, and testing.

B. Model Architectures

ResNet50: Employs residual blocks to preserve gradient flow across 50 layers, enabling deeper learning. Only the top layers were unfrozen for fine-tuning over 20 epochs using the Adam optimizer.

InceptionV3: Utilizes inception modules combining multiple convolution kernel sizes (1×1, 3×3, 5×5) to capture diverse spatial features. Fully unfrozen and fine-tuned for 50 epochs with Adam optimizer and a learning rate decay schedule.

EfficientNetB4: Implements compound scaling across width, depth, and resolution, achieving optimal performance-efficiency tradeoff. The top 30 layers were unfrozen and trained for 20 epochs using the Adamax optimizer.

The diagram in Fig. 2 illustrates the end-to-end workflow of the proposed system — from MRI input to model prediction and explainability generation. It highlights how the

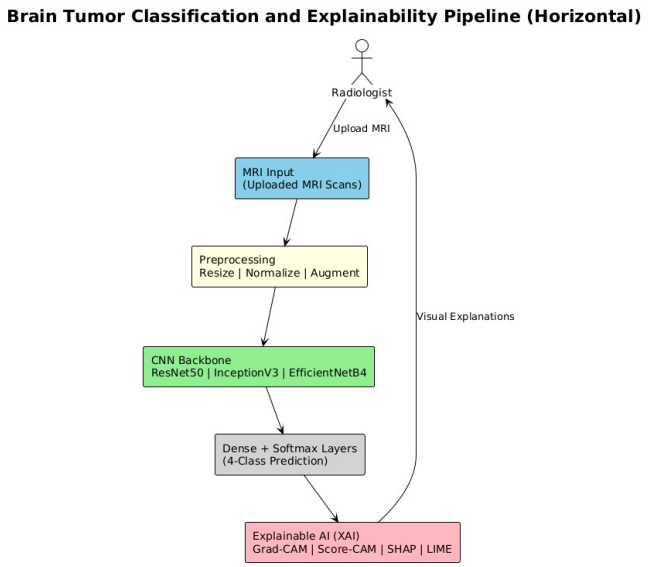

Fig. 2: Overall pipeline of the proposed brain tumor classification and explainability framework. The process includes MRI acquisition, preprocessing, feature extraction via CNN backbones (ResNet50, InceptionV3, EfficientNetB4), classification through dense layers, and interpretability generation using Grad-CAM, Score-CAM, SHAP, and LIME.

MRI images undergo preprocessing (resizing, normalization, augmentation), followed by feature extraction through one of the selected CNN backbones. The processed feature maps are classified into four tumor categories via dense and softmax layers. Finally, post-hoc XAI methods (Grad-CAM, Score-CAM, SHAP, and LIME) provide visual interpretability, linking model activations to clinically relevant regions. This figure serves as a conceptual overview of the complete classification–explainability pipeline.

TABLE I: Training Configuration for Each Model

Model	Epochs	Optimizer	LR	Fine-Tuning
ResNet50	20	Adam	0.001	Top layers
InceptionV3	50	Adam	0.001→1e-5	Full network
EfficientNetB4	20	Adamax	0.001→1e-5	Top 30 layers

C. Explainable AI Techniques

XAI methods were applied post-training to analyze decision rationales:

- **Grad-CAM**: Gradient-weighted activation mapping identifies class-discriminative regions.
- **Score-CAM**: Activation-based visualization using feature maps without gradients for smoother localization.
- **SHAP**: Computes per-pixel contributions via cooperative game theory, producing interpretable overlays.
- **LIME**: Creates local surrogate models highlighting image regions most influential to model output.

D. Evaluation Metrics

Performance metrics included Accuracy, Precision, Recall, F1-score, AUC, and Cohen's Kappa. Explainability was quantified using:

- **Deletion AUC** — accuracy degradation upon masking salient regions.
- **Insertion AUC** — confidence recovery when adding important pixels.
- **Faithfulness Correlation** — correlation between explanation and prediction consistency.

IV. Results and Analysis

A. Model Performance

TABLE II: Quantitative Evaluation Across Models

Model	Accuracy (%)	F1	AUC	Kappa
ResNet50	90.6	0.90	0.962	0.81
InceptionV3	**97.0**	**0.97**	**0.982**	-
EfficientNetB4	85.3	0.85	0.977	0.80

InceptionV3 achieved the highest accuracy and generalization, confirming the effectiveness of its multi-branch convolutional structure in capturing hierarchical tumor features. EfficientNetB4 demonstrated competitive AUC and calibration performance, while ResNet50 provided balanced results with faster convergence and reduced complexity.

B. Per-Class Evaluation

TABLE III: Per-Class Precision, Recall, and F1-Scores

Class	EfficientNetB4 (F1)	InceptionV3 (F1)
Glioma	0.83	0.96
Meningioma	0.74	0.94
No Tumor	0.92	0.99
Pituitary	0.90	0.98

InceptionV3 consistently excelled across all tumor categories, achieving 0.99 F1 for "No Tumor." EfficientNetB4 demonstrated robust recall but slightly lower precision due to its lower parameter count, which may limit fine-grained feature extraction.

C. Explainability Metrics

TABLE IV: Quantitative XAI Metrics for EfficientNetB4

Method	Deletion ↓	Insertion ↑	Faithfulness ↑
Score-CAM	0.4039	**0.7420**	-0.0059
SHAP	0.3420	0.5683	**0.0655**
LIME	**0.2822**	0.6504	0.0027

Score-CAM achieved the highest localization strength, while SHAP demonstrated the best alignment with class probabilities. LIME effectively delineated superpixels, confirming its interpretive utility.

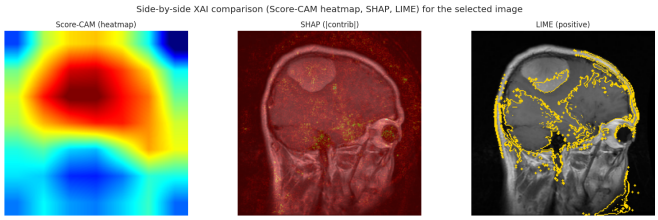

Fig. 3: EfficientNetB4 : Side-by-side XAI comparison showing Score-CAM, SHAP, and LIME highlighting tumor-relevant regions.

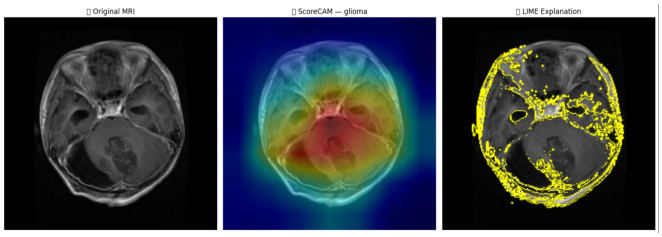

Fig. 4: ResNet50 : Original MRI, Score-CAM activation, and LIME segmentation overlay showing localized tumor focus.

D. Visual Comparisons

E. Interpretation

Visual analyses confirm that all models successfully localize pathological regions, but InceptionV3 and EfficientNetB4 exhibited more compact and clinically consistent activations. SHAP maps revealed dense red regions aligning with tumor boundaries, and Score-CAM heatmaps demonstrated focused spatial activations over abnormal tissue. These explainability outcomes validate that CNNs are not merely overfitting noise but genuinely learning discriminative features relevant to pathology.

V. Discussion

The comparative results illustrate distinct trade-offs between accuracy, interpretability, and computational complexity. InceptionV3's large receptive fields capture detailed tumor boundaries, resulting in superior accuracy. EfficientNetB4, though computationally lighter, emphasizes explainability through stable activation visualization. ResNet50 remains efficient for low-latency environments such as real-time screening.

From a clinical perspective, explainability provides critical value by:

1) Enhancing physician trust in AI-assisted diagnostics.
2) Providing visual justifications aligned with tumor morphology.
3) Enabling multi-model consensus validation for edge-case diagnoses.

Furthermore, the integration of quantitative XAI metrics allows systematic validation of interpretability beyond visual inspection — a crucial advancement for medical AI governance frameworks.

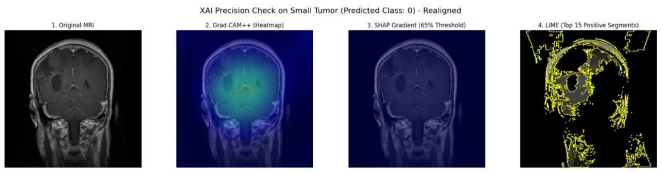

Fig. 5: InceptionV3 : Grad-CAM++, SHAP gradient map, and LIME segmentations for glioma classification.

VI. Conclusion and Future Work

This study provides a rigorous comparative evaluation of three deep CNN architectures for brain tumor classification, integrating explainability at every analytical level. InceptionV3 achieved the highest classification accuracy (97%), EfficientNetB4 delivered the best interpretability (Insertion AUC = 0.742), and ResNet50 balanced performance and efficiency.

Future extensions include:

- Incorporating Vision Transformers (ViT, Swin) for global attention modeling.
- Utilizing multimodal MRI and CT datasets for richer representations.
- Implementing explainability-guided retraining pipelines to enhance fairness and accountability.

The findings underscore that accuracy and transparency must coexist in clinical AI, and integrating post-hoc explainability marks a significant step toward trustworthy, interpretable, and clinically relevant medical AI.

Acknowledgment

The authors thank NMIMS University for continuous academic support, and the dataset contributors for open access MRI resources that made this work possible.

References

[1] K. He et al., "Deep Residual Learning for Image Recognition," *CVPR*, 2016.
[2] C. Szegedy et al., "Rethinking the Inception Architecture for Computer Vision," *CVPR*, 2016.
[3] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for CNNs," *ICML*, 2019.
[4] R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *ICCV*, 2017.
[5] S. Lundberg and S. Lee, "A Unified Approach to Interpreting Model Predictions," *NeurIPS*, 2017.
[6] M. Ribeiro et al., "Why Should I Trust You? Explaining Predictions of Any Classifier," *KDD*, 2016.
[7] S. Hossain et al., "Brain Tumor Detection Using Deep Learning," *J. Med. Imaging*, 2021.
[8] A. Chakrabarty et al., "Deep Learning Models for MRI Brain Tumor Classification," *IEEE Access*, 2022.
[9] M. Reyes et al., "On the Interpretability of Deep Learning in Medical Imaging," *Med. Image Anal.*, 2020.
[10] D. Ghosh et al., "Interpretability in AI-Driven Medical Diagnosis," *Nature Machine Intelligence*, 2021.
[11] C. Biffi et al., "Medical Imaging with Deep Learning: A Review," *IEEE Trans. Med. Imaging*, 2022.
[12] S. Hosseini et al., "Explainable AI in Medical Imaging: Challenges and Opportunities," *Comput. Med. Imaging Graph.*, 2023.
[13] C. Baur et al., "Autonomous Detection and Localization of Brain Tumors using CNNs," *Front. Neurosci.*, 2021.
[14] K. Patel et al., "Transformer Models for Radiology: A Survey," *IEEE Rev. Biomed. Eng.*, 2023.
[15] M. Raj et al., "Comparative Study of CNN Architectures for Brain Tumor Detection," *IEEE Access*, 2024.