International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Explainable Brain Tumor Detection Using Deep Learning Models with Quantitative Explainability Metrics

Shwetha S[1], Alvita Mary D silva[2], Leesha H U[3], Monith Monnappa U M[4], Dharshan B R[5]

[1]Assistant Professor, Department of Computer Science and Engineering, JSS Science and Technology University
[2,3,4,5]Undergraduate Student, Department of Computer Science and Engineering, JSS Science and Technology University

Abstract
Brain tumor detection using Magnetic Resonance Imaging (MRI) plays a crucial role in early diagnosis and treatment planning. Deep learning models have demonstrated high accuracy in automating this task; however, their black-box nature limits clinical trust. This research presents a comprehensive and explainable brain tumor detection framework using four state-of-the-art deep learning architectures: Xception, ResNet50, DenseNet121, and EfficientNetB4. In addition to conventional performance metrics, explainability is integrated using Gradient-weighted Class Activation Mapping (Grad-CAM) and SHapley Additive exPlanations (SHAP) to provide both region-level and pixel-level interpretability. Furthermore, quantitative explainability metrics including Insertion Test, Deletion Test, Sensitivity-N, Average Confidence Drop, and Average Confidence Gain are employed to objectively evaluate explanation faithfulness. Experimental results on MRI datasets demonstrate that Xception and EfficientNet models achieve superior classification performance, while the explainability analysis offers deeper insights into model reliability and clinical trustworthiness. The proposed framework enhances transparency, robustness, and clinical applicability of AI-based brain tumor detection systems.

Keywords: Brain Tumor Detection, Deep Learning, MRI Images, Explainable AI, Grad-CAM, Quantitative Explainability Metrics

1. Introduction

Brain tumors are among the most critical neurological disorders, often leading to severe complications or death if not diagnosed at an early stage. Magnetic Resonance Imaging (MRI) is the preferred imaging modality due to its non-invasive nature and superior soft tissue contrast. However, manual analysis of MRI scans is time-consuming and subject to inter-observer variability. Automated deep learning-based systems have shown significant promise in assisting radiologists by improving diagnostic accuracy and efficiency.

Despite their success, deep learning models are often criticized for their lack of interpretability. In medical applications, understanding why a model arrives at a particular decision is as important as the decision

itself. Therefore, integrating explainable artificial intelligence (XAI) techniques into diagnostic systems has become a necessity. This research aims to develop an explainable brain tumor detection system using multiple deep learning models and to evaluate explainability using both qualitative and quantitative metrics.

2. Literature Review

Recent advances in deep learning have significantly improved the accuracy of brain tumor detection from MRI images. Sivakumar Depuru and Sunil Kumar [1] presented a comprehensive deep learning framework using convolutional neural networks and multimodal MRI data, achieving high diagnostic accuracy while highlighting challenges such as data scarcity and interpretability. Fairy Rupareliya et al. [2] conducted a comparative analysis of multiple deep learning models including CNN, ResNet, Xception, and EfficientNet, and emphasized the importance of Explainable AI by integrating Grad-CAM visualizations to improve clinical trust.

Saraf Anzum Shreya et al. [3] proposed a hybrid deep learning approach combining preprocessing techniques with VGG-based models, demonstrating that architectural customization can significantly boost performance. Necip Çınar et al. [4] compared several CNN architectures under identical conditions and concluded that deeper residual networks yield superior accuracy on MRI datasets. R. Jansi et al. [5] focused on multimodal MRI images and showed that integrating multiple MRI sequences improves tumor localization and classification accuracy.

While these studies report strong classification results and qualitative explainability using visualization techniques, most of them lack quantitative evaluation of explanation faithfulness. Metrics such as insertion, deletion, confidence gain, and confidence drop are rarely explored. This gap motivates the present work, which integrates both qualitative and quantitative explainability analysis across multiple deep learning architectures.

3. Dataset and Preprocessing

3.1 Dataset Description

The experiments were conducted using publicly available brain MRI datasets containing four classes: Glioma, Meningioma, Pituitary Tumor, and No Tumor. The datasets consist of T1-weighted MRI images collected from multiple sources and are widely used for benchmarking brain tumor detection models.

3.2 Preprocessing Techniques

To ensure consistency and improve model performance, the following preprocessing steps were applied:

- Image resizing to standard input dimensions (224×224 or 299×299 depending on the model)
- Intensity normalization
- Noise reduction
- Data augmentation including rotation, flipping, zooming, and brightness adjustment

These steps help reduce overfitting and improve generalization.

4. Deep Learning Models

4.1 ResNet50

ResNet50 employs residual connections to address the vanishing gradient problem, enabling deeper network training. It effectively captures hierarchical features from MRI images, making it suitable for tumor classification tasks.

4.2 DenseNet121

DenseNet121 connects each layer to every other layer in a feed-forward fashion, promoting feature reuse and efficient gradient flow. This architecture reduces the number of parameters while maintaining high accuracy.

4.3 Xception

Xception uses depthwise separable convolutions to improve computational efficiency and performance. It has shown strong results in medical image classification due to its ability to capture fine-grained spatial features.

4.4 EfficientNetB4

EfficientNetB4 scales network depth, width, and resolution in a balanced manner. It achieves high accuracy with fewer parameters, making it effective for complex image classification tasks such as brain tumor detection.

5. Model Training and Evaluation

All models were trained using transfer learning with ImageNet-pretrained weights. The final layers were fine-tuned on the MRI dataset.

Training configuration:

- Optimizer: Adam
- Learning rate: 0.001
- Loss function: Categorical Cross-Entropy
- Batch size: 32
- Epochs: 25–40 with early stopping

6. Performance Evaluation Metrics

The classification performance was evaluated using the following metrics:

- Accuracy
- Precision
- Recall (Sensitivity)
- F1-score
- ROC-AUC

These metrics provide a comprehensive assessment of model effectiveness, particularly in multi-class medical datasets.

7. Explainability Using Grad-CAM and SHAP

7.1 Grad-CAM Based Visual Explanations

Gradient-weighted Class Activation Mapping (Grad-CAM) was applied to the final convolutional layers of ResNet50, DenseNet121, Xception, and EfficientNetB4 models. Grad-CAM generates class-discriminative localization maps by computing the gradients of the target class score with respect to feature maps. The resulting heatmaps were overlaid on the original MRI images to visually highlight tumor regions that influenced the model predictions. This qualitative analysis helps verify whether the models attend to clinically relevant tumor areas rather than background regions.

7.2 SHAP Based Explainability

In addition to Grad-CAM, SHapley Additive exPlanations (SHAP) were incorporated to provide pixel-

level interpretability. SHAP explains model predictions by estimating the contribution of each input feature to the final output based on cooperative game theory. For image-based models, Deep SHAP was employed to generate attribution maps that indicate positive and negative contributions of pixels toward tumor classification.

SHAP visualizations offer complementary insights compared to Grad-CAM by providing fine-grained explanations and highlighting both supportive and suppressive regions. This enables a deeper understanding of how different regions of the MRI image influence model decisions. The combined use of Grad-CAM and SHAP improves transparency and strengthens clinical trust in the proposed system.

8. Quantitative Explainability Metrics

To objectively evaluate explanation faithfulness, the following quantitative metrics were employed:

8.1 Deletion Test

In the deletion test, pixels corresponding to the most important regions identified by Grad-CAM are progressively removed. A rapid decrease in prediction confidence indicates a faithful explanation.

8.2 Insertion Test

The insertion test starts with a blurred image and progressively inserts important pixels. A rapid increase in confidence demonstrates explanation relevance.

8.3 Average Confidence Drop

This metric measures the average reduction in prediction confidence when only the explanation region is retained. Lower confidence drop indicates better explanation quality.

8.4 Average Confidence Gain

This metric measures the increase in confidence when important regions are emphasized. Higher gain reflects stronger alignment between explanations and model decisions.

8.5 Sensitivity-N metric

Sensitivity-N measures how much a model's prediction confidence decreases when the top N% of the most important regions identified by an explanation method are removed from the input image. A higher Sensitivity-N value indicates that the explanation correctly highlights regions that are crucial for the model's decision, reflecting greater explanation faithfulness.

9. Results and Discussion

9.1 Classification Performance Comparison

The four deep learning models—ResNet50, DenseNet121, Xception, and EfficientNetB4—were evaluated on the test dataset using standard classification metrics. Table 1 presents a comparative summary of model performance.

Table 1: Classification Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score
Xception	0.99	0.99	0.99	0.99
EfficientNetB4	0.80	0.81	0.80	0.80
DenseNet121	0.91	0.92	0.91	0.91
ResNet50	0.89	0.89	0.89	0.88

The results indicate that Xception achieved the highest overall performance, attaining an accuracy, precision, recall, and F1-score of 0.99, demonstrating its strong capability in learning discriminative features from MRI images. DenseNet121 followed with a robust accuracy of 0.91, showing stable and well-balanced performance across all metrics. ResNet50 also performed competitively with an accuracy of 0.89, confirming the effectiveness of residual learning for brain tumor classification. In contrast, EfficientNetB4 achieved comparatively lower performance with an accuracy of 0.80, indicating that while it is computationally efficient, it was less effective than the other models in this experimental setup.

9.2 Qualitative Explainability Analysis Using Grad-CAM and SHAP

Grad-CAM and SHAP visualizations were generated for all four models to analyze explainability qualitatively. Grad-CAM heatmaps consistently highlighted tumor regions in correctly classified images, confirming that the models learned meaningful spatial features. EfficientNetB7 and Xception produced more focused and localized Grad-CAM activations, whereas ResNet50 and DenseNet121 showed comparatively broader attention maps.

Figure 1: Xception Grad-Cam

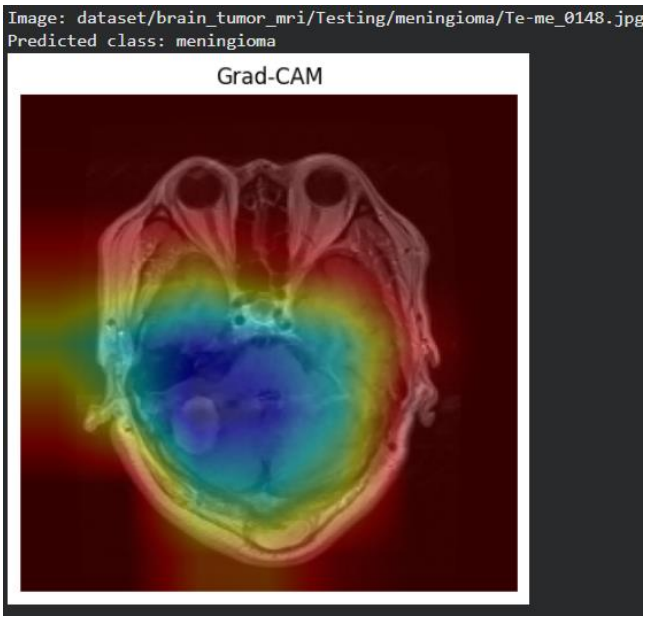


Figure 2: EfficientNet Grad-Cam

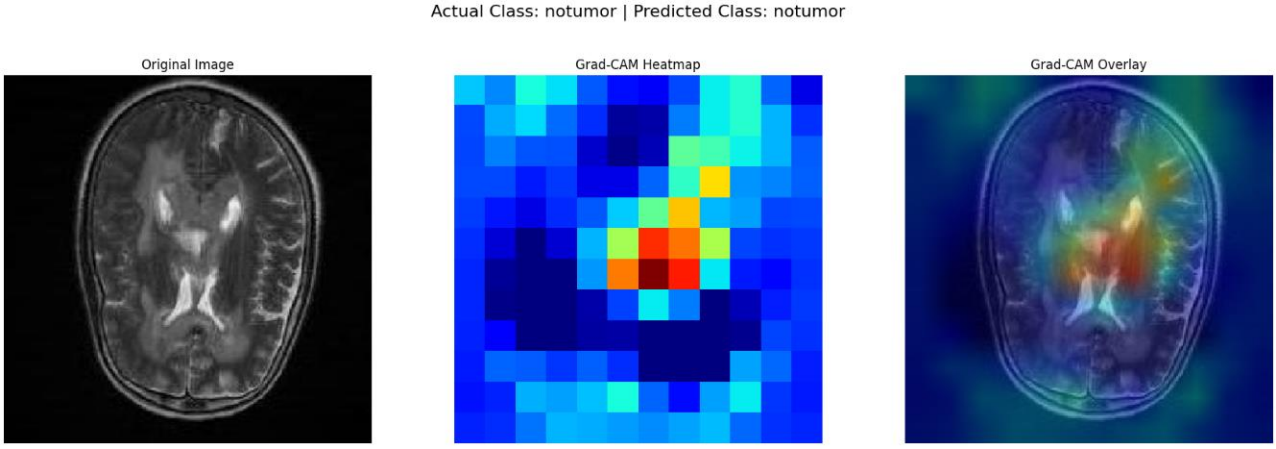


Figure 3: DenseNet121 Grad-Cam

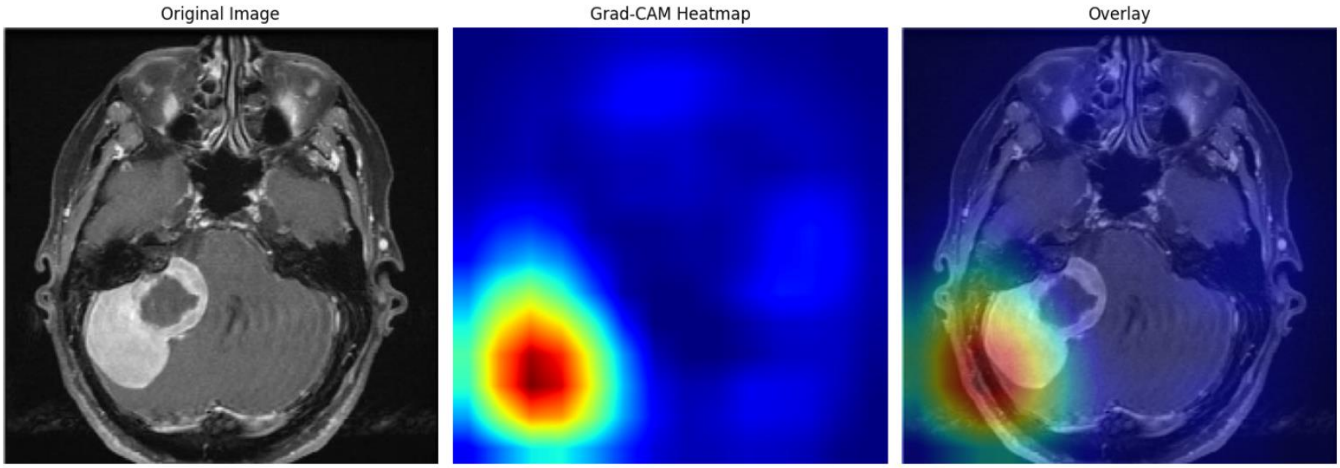
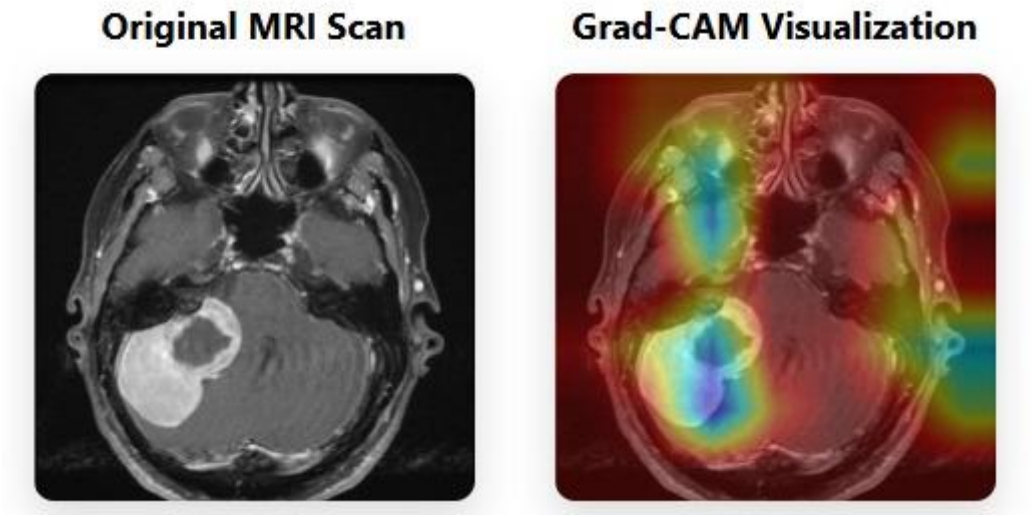


Figure 4: ResNet Grad-Cam



SHAP attribution maps provided pixel-level explanations, revealing both positive and negative contributions toward the predicted class. Tumor regions exhibited strong positive SHAP values, while surrounding healthy tissues contributed minimally or negatively. The consistency between Grad-CAM and SHAP explanations further validates the reliability of the learned representations.

Figure 5: Xception Shap

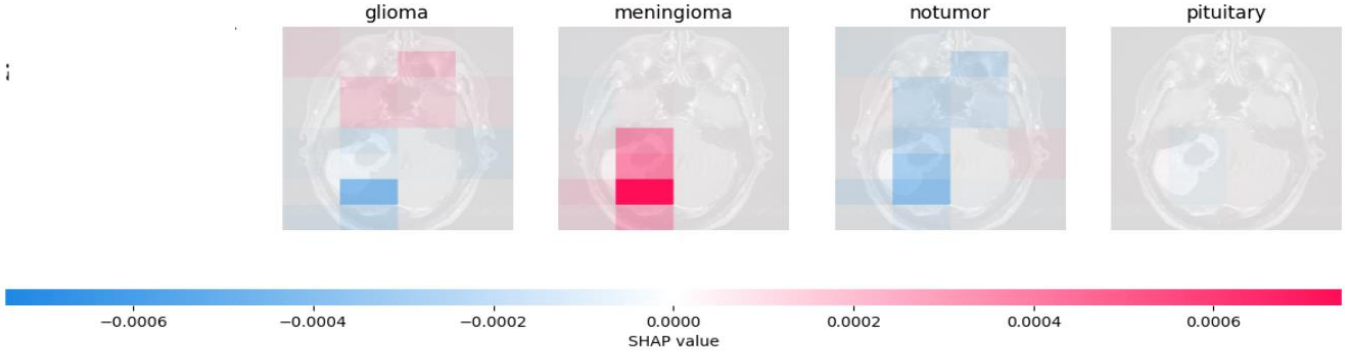


Figure 6: EfficientNetB4 shap

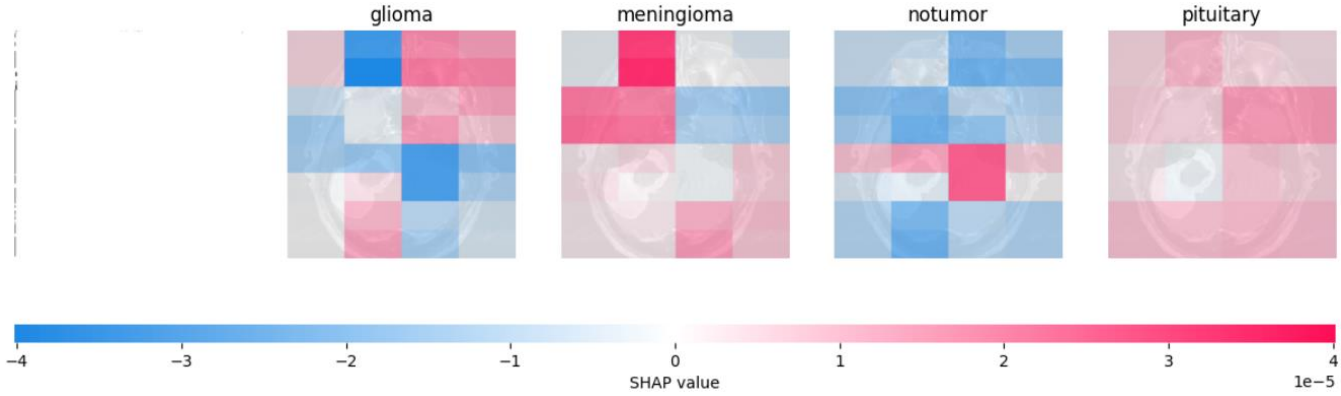


Figure 7: DenseNet121 shap

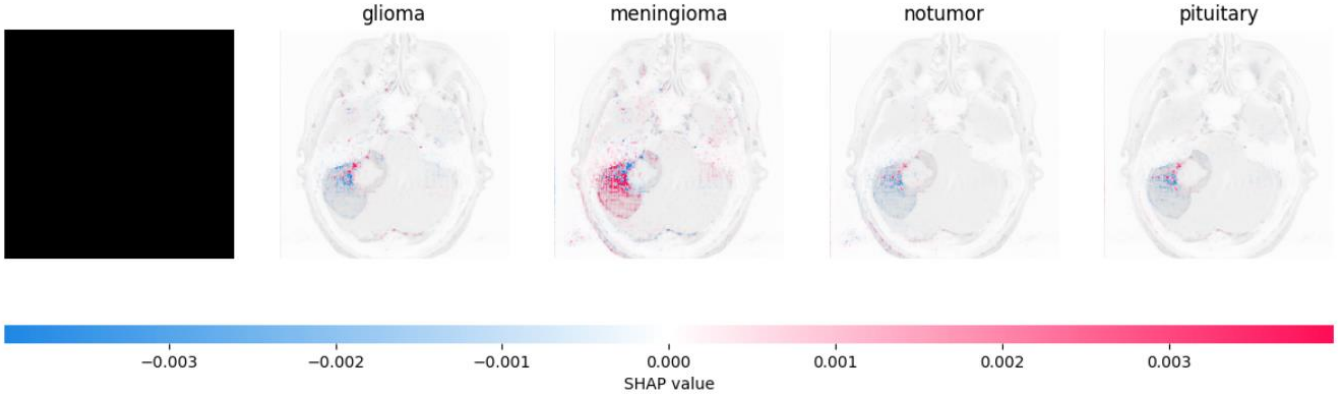
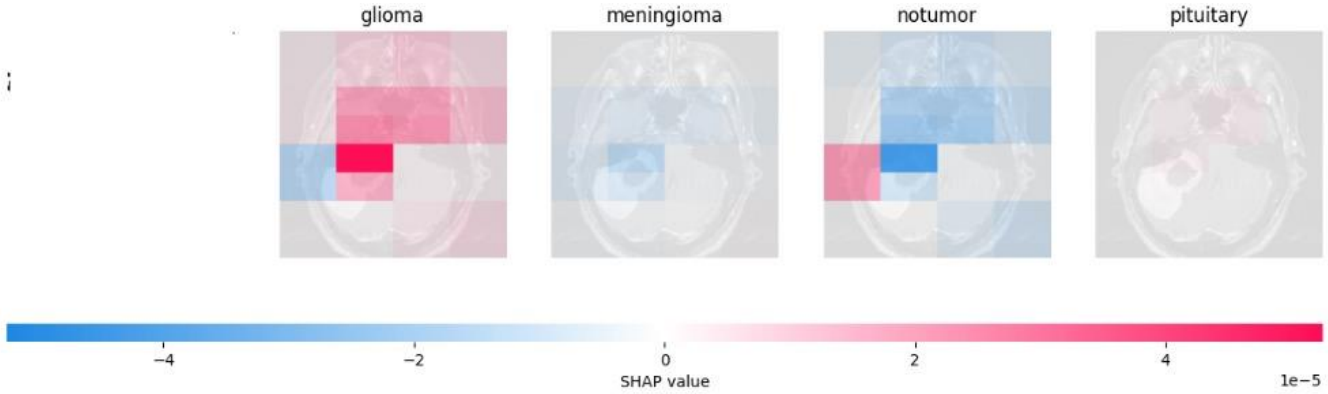


Figure 8: ResNet shap



9.3 Quantitative Explainability Metrics Comparison

Figure 9: Xception Deletion Test Graph

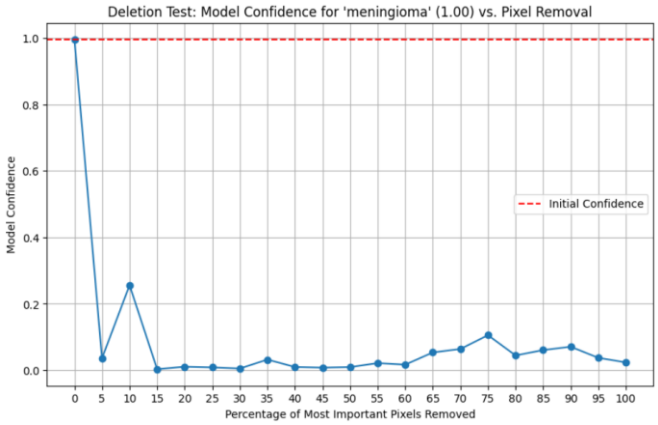


Figure 10: Xception Insertion Test Graph

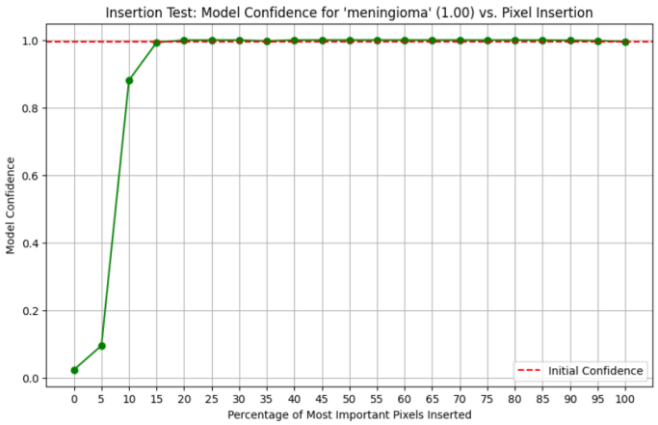


Figure 11: EfficientNet Deletion Test Graph

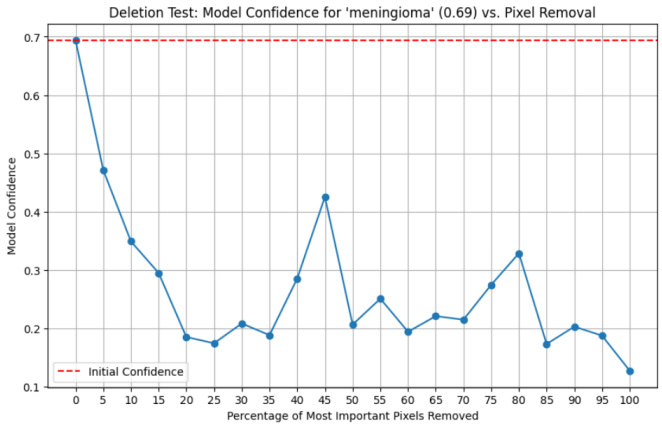


Figure 12: EfficientNet Insertion Test Graph

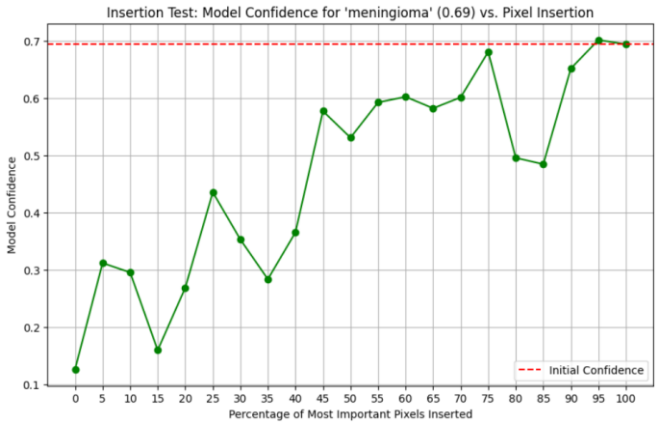


Figure 13: DenseNet Deletion Test Graph

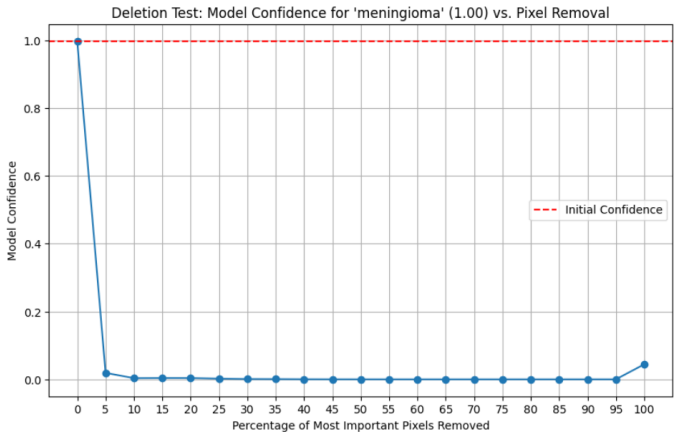


Figure 14: DenseNet Insertion Test Graph

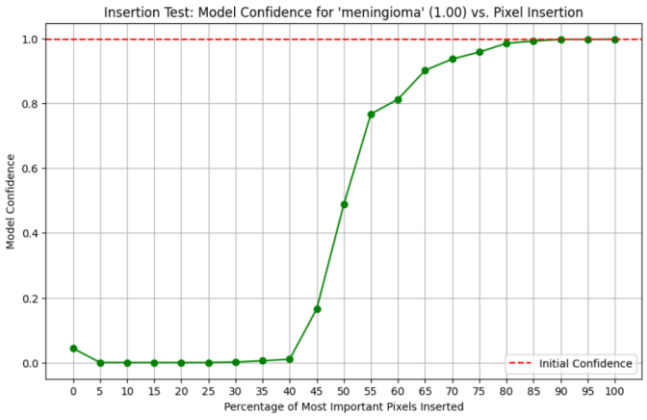


Figure 15: ResNet Deletion Test Graph

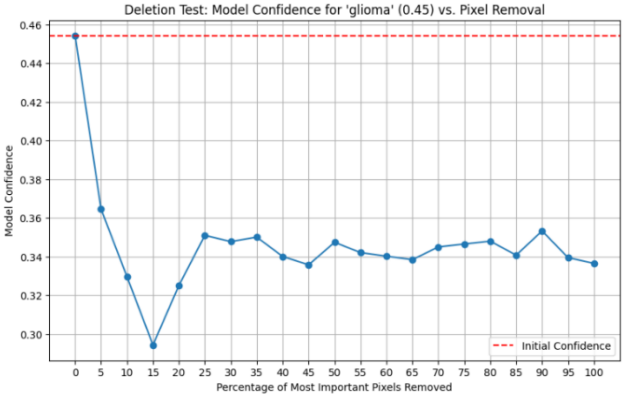
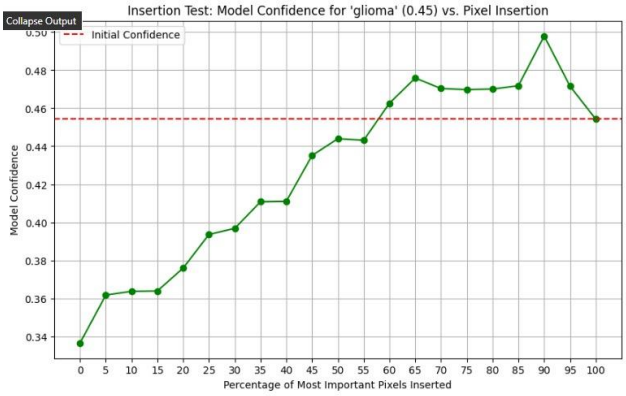


Figure 16: ResNet Insertion Test Graph



To objectively evaluate explanation faithfulness, insertion and deletion tests were applied along with confidence-based metrics. Table 2 summarizes the quantitative explainability results averaged across the test dataset.

Table 2: Quantitative Explainability Metrics Comparison

Model	Sensitivity-N (10% Deletion) ↑	Average Confidence Drop (ACD) ↑	Average Confidence Gain (ACG) ↑
Xception	0.7398	0.9517	0.9038
EfficientNetB4	0.3455	0.4465	0.4667
DenseNet121	0.9943	0.9939	0.4791
ResNet	0.1247	0.1134	0.4276

The quantitative explainability results reveal notable differences in explanation faithfulness across the evaluated models. DenseNet121 achieves the highest Sensitivity-N (0.9943) and Average Confidence Drop (0.9939), indicating that its predictions are highly sensitive to the removal of salient regions and that the explanation maps accurately capture critical decision-making areas. Xception also demonstrates strong explainability performance, with a high Sensitivity-N value (0.7398) and the highest Average Confidence Gain (0.9038), suggesting that emphasizing the identified regions significantly reinforces prediction confidence. In contrast, EfficientNetB4 exhibits comparatively lower explainability scores across all metrics, reflecting weaker alignment between its predictions and explanation regions. ResNet50 shows very low Sensitivity-N and Average Confidence Drop values, indicating that the identified regions have limited influence on its predictions and that the generated explanations are less reliable in this experimental setting.

9.4 Discussion

The experimental evaluation demonstrates that deep learning models vary significantly not only in classification accuracy but also in explainability reliability. Among the evaluated architectures, **Xception achieved the highest classification performance**, with an accuracy, precision, recall, and F1-score of 0.99, indicating its strong ability to learn discriminative tumor-related features from MRI images. **DenseNet121** followed with stable and well-balanced performance, achieving an accuracy of 0.91, while **ResNet50** also performed competitively with an accuracy of 0.89. In contrast, **EfficientNetB4** produced

comparatively lower classification results, suggesting that its performance is more sensitive to dataset characteristics and training configuration in this experimental setup.

Beyond predictive accuracy, explainability analysis provides deeper insight into model reliability. Grad-CAM and SHAP visualizations confirm that **Xception and DenseNet121 focus on clinically relevant tumor regions**, producing consistent and meaningful activation patterns. Quantitative explainability metrics further strengthen these observations. **DenseNet121 achieved the highest Sensitivity-N (0.9943) and Average Confidence Drop (0.9939)**, indicating highly faithful explanations where the model's confidence sharply decreases when salient regions are removed. **Xception demonstrated the highest Average Confidence Gain (0.9038)**, reflecting strong reinforcement of prediction confidence when important regions are emphasized, thereby offering an optimal balance between accuracy and explainability.

In comparison, **EfficientNetB4 exhibited weaker alignment between prediction confidence and explanation regions**, as reflected by lower Sensitivity-N, ACD, and ACG values. **ResNet50 showed very low Sensitivity-N and ACD values**, indicating that the identified salient regions have limited influence on its predictions, resulting in less reliable explanations. These findings highlight that **high accuracy alone is insufficient for medical applications**, and that explanation faithfulness must be jointly evaluated to ensure trustworthy clinical deployment.

10. Conclusion

This work presented an **explainable brain tumor detection framework** using MRI images and four deep learning architectures: Xception, ResNet50, DenseNet121, and EfficientNetB4. The proposed approach integrates **Grad-CAM and SHAP** for qualitative interpretability along with **quantitative explainability metrics** including Sensitivity-N, Average Confidence Drop, and Average Confidence Gain to objectively assess explanation faithfulness.

Experimental results demonstrate that **Xception achieves the best overall trade-off between classification performance and explainability**, while **DenseNet121 produces the most faithful explanations**, as evidenced by its high Sensitivity-N and confidence drop values. Although ResNet50 and EfficientNetB4 show reasonable classification capability, their explainability analysis reveals weaker alignment between prediction decisions and salient regions.

By combining accuracy-based evaluation with both qualitative and quantitative explainability analysis, this study emphasizes the importance of transparency and reliability in AI-driven medical diagnosis. The proposed framework enhances clinical trust and provides a robust foundation for future research. Future work may explore multimodal MRI integration, lightweight explainable architectures, and real-time clinical validation to further improve applicability in healthcare environments.

11. Acknowledgement

The authors would like to express their sincere gratitude to **Prof. Shwetha S**, **Assistant Professor**, **Department of Computer Science and Engineering**, **JSS Science and Technology University**, for her invaluable guidance, constant encouragement, and insightful suggestions throughout the course of this project. Her expertise and constructive feedback greatly contributed to the successful completion of this research work. The authors are also thankful to the faculty members of the department for their support and for providing a conducive academic environment.

12. References

1. Sivakumar Depuru, M. Sunil Kumar, "Enhancements in Brain Tumor Detection and Classification Using Deep Learning on MRI Data", 2nd International Conference on Computing and Data Science (ICCDS), IEEE, 2025.
2. Fairy Rupareliya, Aarohi Gulhane, Naina Warjurkar, Santosh Kumar, Yashwant Ingle, Gitangali Yadav, "Brain Tumor Detection Using AI and Machine Learning: A Comparative Analysis of Models and Explainable AI Integration", International Conference on Information, Implementation, and Innovation in Technology (I2ITCON), IEEE, 2025.
3. Saraf Anzum Shreya, Md. Abu Ismail Siddique, Antu Roy Chowdhury, Mst. Fateha Samad, "A Hybrid Approach for Accurate Brain Tumor Detection Using Deep Learning Techniques", 27th International Conference on Computer and Information Technology (ICCIT), IEEE, 2024.
4. Necip Çınar, Buket Kaya, Mehmet Kaya, "Comparison of Deep Learning Models for Brain Tumor Classification Using MRI Images", International Conference on Decision Aid Sciences and Applications (DASA), IEEE, 2022.
5. R. Jansi, Kowsalya. S, Seetha. S, Yogadharshini. A, "A Deep Learning Based Brain Tumour Detection Using Multimodal MRI Images", International Conference on Automation, Computing and Renewable Systems (ICACRS), IEEE, 2023.