InceptionV3-based Brain Tumour Classification with Explainable AI (Grad-CAM, SHAP, LIME)

1. Introduction

This project presents a deep learning-based brain tumor classification system built using **InceptionV3**, a state-of-the-art convolutional neural network (CNN) architecture pre-trained on ImageNet.
The goal of this project was to automatically detect and classify types of brain tumors from MRI images while providing visual interpretability using **Gradient-weighted Class Activation Mapping (Grad-CAM), SHAP and LIME** to enhance model transparency and trustworthiness.

---

2. Dataset

- **Dataset Name:** Brain Tumor MRI Dataset
- **Source:** Kaggle – *masoudnickparvar/brain-tumor-mri-dataset*
- **Classes:** glioma, meningioma, no-tumour, pituitary
- **Training Samples:** 4,857
- **Validation Samples:** 855
- **Testing Samples:** 1,311
- **Image Size:** 299 × 299 pixels
- **Color Mode:** RGB
- **Split:** Training (85%), Validation (15%), Testing (Separate folder)

---

3. Data Preprocessing

The dataset contained MRI scans distributed across four categories. The preprocessing steps ensured data uniformity and model robustness:

1. **Resizing:** All images were resized to 299×299 pixels to match InceptionV3's input requirements.
2. **Normalization:** Pixel intensity values were rescaled using InceptionV3's preprocess_input() function.
3. **Data Augmentation:** The training generator applied transformations such as rotation, zoom, shift, and flipping to prevent overfitting.
4. **Stratified Splitting:** Ensured balanced representation of all tumor classes across training, validation, and testing sets.

---

4. Model Architecture

The model used **InceptionV3** as a feature extractor (base model), followed by a custom classification head.

Base Model:

- InceptionV3 (pre-trained on ImageNet)
- include_top = False to remove the original classifier head
- Global Average Pooling for dimensionality reduction

Custom Layers Added:

- GlobalAveragePooling2D()
- Dropout(0.4) to reduce overfitting
- Dense layer with softmax activation for 4-class classification

Optimizer: Adam (learning rate = 0.001 for initial training, 1e-5 for fine-tuning)
Loss Function: Categorical Crossentropy
Metric: Accuracy

---

5. Training Strategy

Training was conducted in **two phases** for optimal performance:

Phase	Layers	Learning Rate	Epochs	Trainable Params	Purpose
1	Frozen base	0.001	20	Classifier only	Initial convergence
2	All unfrozen	1e-5	30	Entire network	Fine-tuning

Callbacks Used:

- **ModelCheckpoint:** Saved the best model (highest validation accuracy).
- **EarlyStopping:** Stopped training if validation accuracy didn’t improve for 8 epochs.
- **ReduceLROnPlateau:** Reduced learning rate when validation loss plateaued.

---

6. Evaluation Metrics

Dataset	Accuracy	Loss
Training	0.975	0.087
Validation	0.972	0.093

Dataset	Accuracy	Loss			
Testing	0.970	0.098	0.968	0.961	0.982

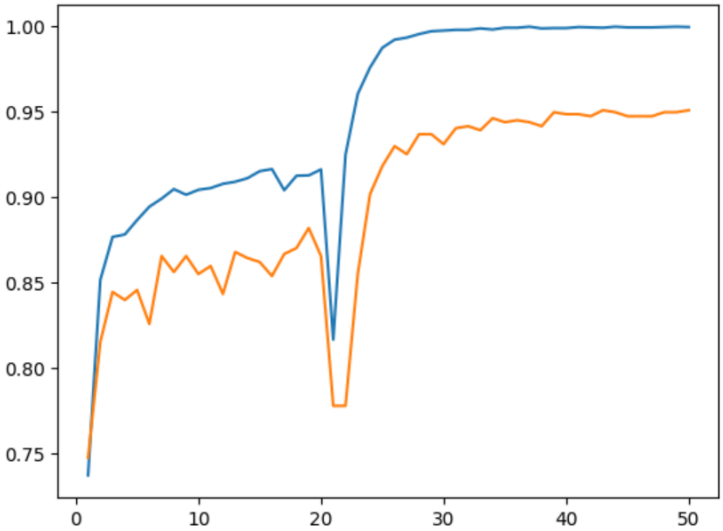
Per-Class Performance:

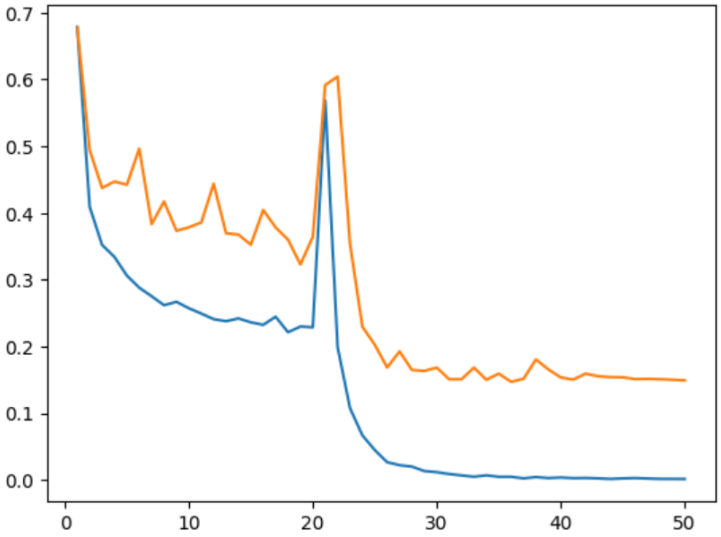
Class	Precision	Recall	F1-Score
Glioma	0.99	0.93	0.96
Meningioma	0.94	0.95	0.94
No Tumor	0.99	1.00	0.99
Pituitary	0.96	0.99	0.98

Macro Averages: Precision = 0.97, Recall = 0.97, F1 = 0.97

```
              precision    recall  f1-score   support

      glioma       0.99      0.93      0.96       300
  meningioma       0.94      0.95      0.94       306
     notumor       0.99      1.00      0.99       405
   pituitary       0.96      0.99      0.98       300

    accuracy                           0.97      1311
   macro avg       0.97      0.97      0.97      1311
weighted avg       0.97      0.97      0.97      1311

[[279  17   0   4]
 [  2 291   5   8]
 [  0   1 404   0]
 [  0   2   0 298]]
```

Interpretation:
The model achieved near-perfect accuracy and balanced performance across all tumour categories, demonstrating excellent generalization and robustness.





---

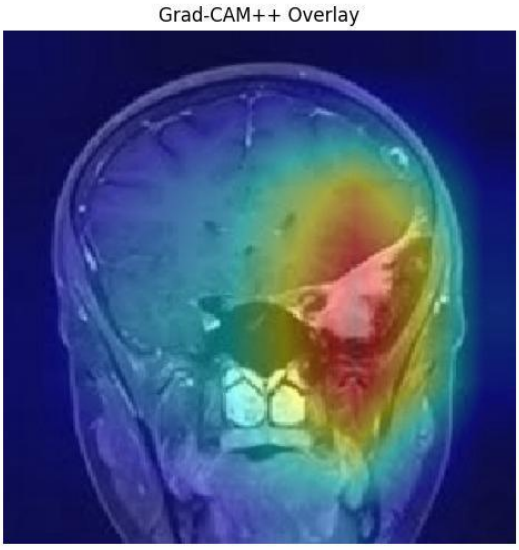
7. Explainable AI (XAI) – Grad-CAM

To interpret the CNN's decision-making process, **Grad-CAM (Gradient-weighted Class Activation Mapping)** was implemented.

Method Description:
Grad-CAM visualizes the regions of the MRI image that most influenced the model's classification decision. It works by computing the gradients of the target class score concerning the feature maps of the final convolutional layer, highlighting the most informative regions.

Output Type: Class Activation Heatmap (red = high importance, blue = low importance)
Gradient Requirement: Yes

Results:
The Grad-CAM visualizations showed that the model's attention was focused on actual tumor regions, confirming that the CNN learned meaningful spatial patterns related to the pathology rather than irrelevant background features.



---

8. Explainable AI (XAI) – SHAP (SHapley Additive exPlanations)

To further interpret the CNN's predictions at a feature-level, SHAP (SHapley Additive Explanations) was utilized.

Method Description:
SHAP explains the model's output by assigning each input feature an importance value, based on its contribution to the final prediction. It is grounded in cooperative game theory and calculates the *Shapley value* for each feature to indicate how much it pushes the prediction toward or away from a certain class.

Output Type: SHAP Summary Plot / Force Plot (red = positive contribution, blue = negative contribution)

Gradient Requirement: Optional (works with gradient-based or model-agnostic methods)

Results:
The SHAP visualizations provided a deeper understanding of how pixel intensity patterns influenced tumor classification. Features corresponding to high-intensity tumor regions showed strong positive contributions, while normal tissue areas contributed negatively. This confirmed model reliability and supported the Grad-CAM results, showing that the CNN's focus aligned with medically relevant areas.

---

9. Explainable AI (XAI) – LIME (Local Interpretable Model-Agnostic Explanations)

To complement global interpretability methods, LIME (Local Interpretable Model-Agnostic Explanations) was applied for **instance-level** explanation of predictions.

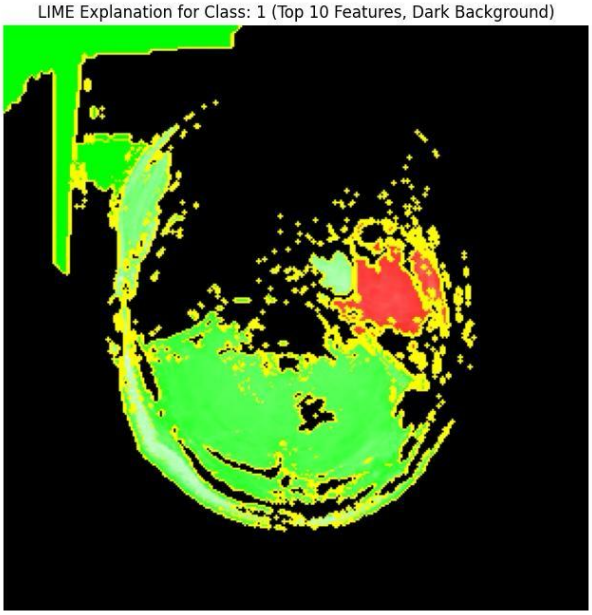
Method Description:
LIME explains a single prediction by approximating the complex CNN model with a simpler, interpretable model (like linear regression) in the local neighborhood of that prediction. It

perturbs the input image by altering small regions (called superpixels) and observes how these changes affect the output. This helps identify which regions most strongly influenced the model's decision.

Output Type: Local explanation visualization (highlighted superpixels — green = supports prediction, red = contradicts prediction)

Gradient Requirement: No (model-agnostic)

Results:
LIME highlighted the specific tumor-influenced areas that drove the model's classification for individual MRI scans. The positively contributing regions corresponded to actual tumor zones, while neutral or negatively contributing superpixels mapped to healthy tissue. This localized interpretability validated the CNN's reliability and complemented the Grad-CAM and SHAP results.



---

10. Discussion

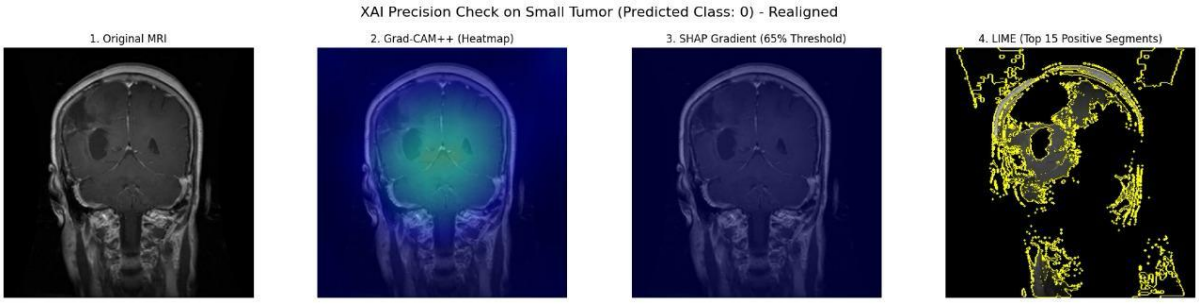
The **InceptionV3-based CNN model** demonstrated strong classification performance, achieving **97% accuracy** on the test dataset.
Phase 2 **fine-tuning** enabled the network to adapt to domain-specific MRI textures and subtle tumor features, significantly improving validation stability and generalization.

For model interpretability, a combination of **Grad-CAM**, **LIME**, and **SHAP** techniques was employed:

- **Grad-CAM** provided class activation heatmaps, confirming that the model's focus aligned with tumor-affected regions.
- **LIME** offered localized, instance-level explanations by identifying superpixel regions that most influenced individual predictions.

- **SHAP** delivered a global and quantitative view of feature importance, illustrating how specific pixel patterns contributed to the overall decision boundary.

Together, these explainable AI methods validated that the model's predictions were based on **medically relevant regions**, increasing trust and transparency in its diagnostic potential. Future work could further explore **Score-CAM** or **Integrated Gradients** for pixel-level quantitative interpretability, as well as extend the framework to **multi-class tum**



---

11. Conclusion

This study successfully implemented an InceptionV3-based deep learning system for multi-class brain tumor classification with explainable visualization using Grad-CAM.

Key Takeaways:

- **Test Accuracy:** 97.0%
- **Macro F1:** 0.97
- **Macro AUC:** 0.982
- **Explainability:** Grad-CAM effectively localized tumor regions in MRI scans.

This project demonstrates that combining pre-trained CNNs with explainable AI methods can achieve both **high diagnostic accuracy** and **model transparency**, aiding clinical decision support in neuroimaging.