An Explainable Deep Learning Approach for Brain Tumor Classification and Localization Using MRI Images:

by Ankit Dutta and Anusha Saha

INTRODUCTION

Brain tumors are one of the most critical neurological disorders and can be life-threatening if not detected at an early stage. Magnetic Resonance Imaging (MRI) is widely used for brain tumor diagnosis due to its high contrast and ability to capture detailed structural information of brain tissues. However, manual analysis of MRI scans is time-consuming, prone to human error, and requires expert radiologists.

With the rapid advancement of artificial intelligence, deep learning techniques have shown remarkable performance in medical image analysis. Convolutional Neural Networks (CNNs) are highly effective in extracting spatial features from medical images, while Long Short-Term Memory (LSTM) networks are capable of learning sequential and contextual patterns from extracted features. Additionally, explainable AI techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM) help visualize the regions of an image that contribute most to the model's decision, increasing trust and transparency.

This project proposes a CNN–LSTM hybrid deep learning model for automatic brain tumor classification from MRI images, along with Grad-CAM-based tumor localization for interpretability.

PROBLEM STATEMENT

Manual detection and classification of brain tumors from MRI images pose several challenges, including dependency on expert radiologists, subjective interpretation, and high diagnosis time. Traditional machine learning methods rely on handcrafted features, which may fail to capture complex tumor patterns and variations in MRI scans.

Moreover, many existing automated systems focus only on classification accuracy without providing insights into *why* a particular prediction was made.

The lack of explainability limits the clinical acceptance of AI-based diagnostic systems.

Therefore, there is a need for an automated, accurate, and interpretable brain tumor detection system that can classify different tumor types and localize tumor regions from MRI images.

OBJECTIVE

The main objectives of this project are:

1. To develop an automated deep learning model for brain tumor classification using MRI images.
2. To classify MRI scans into four categories: **Glioma, Meningioma, Pituitary Tumor, and No Tumor**.
3. To utilize a **CNN–LSTM hybrid architecture** for effective spatial feature extraction and sequential pattern learning.
4. To evaluate the model using performance metrics such as **accuracy, confusion matrix, classification report, ROC–AUC curves, and loss graphs**.
5. To implement **Grad-CAM** for tumor localization and visual explanation of model predictions.
6. To enhance model transparency and reliability for potential clinical decision support.

SCOPE OF THE PROJECT

The scope of this project includes:

- Automated classification of brain MRI images into multiple tumor categories.
- Integration of CNN for spatial feature extraction and LSTM for sequential learning.
- Visualization of tumor-affected regions using Grad-CAM without affecting model accuracy.

- Performance evaluation using standard medical imaging metrics.
- Development of an interpretable AI system suitable for academic and clinical research.

METHODOLOGY

The proposed brain tumor detection system follows a structured deep learning pipeline that integrates image preprocessing, feature extraction, sequence learning, classification, and explainability. The complete methodology is divided into the following stages:

1. Data Acquisition and Preprocessing

Brain MRI images are collected from a labeled dataset consisting of four classes: Glioma, Meningioma, Pituitary Tumor, and No Tumor. Each image is converted to grayscale to reduce computational complexity and resized to a fixed dimension of 64 × 64 pixels. Pixel values are normalized to the range [0,1] to improve model convergence and training stability.

2. Dataset Splitting

The preprocessed dataset is divided into training and testing sets using an 80:20 split. Stratified sampling is applied to ensure equal class distribution across both sets, thereby preventing class imbalance during training and evaluation.

3. Feature Extraction using CNN

Convolutional Neural Networks (CNNs) are used to extract spatial features from MRI images. Multiple convolution and max-pooling layers are employed to capture low-level and high-level tumor characteristics such as edges, textures, and structural patterns. These layers generate rich feature maps representing tumor-relevant information.

4. Sequence Learning using LSTM

The extracted CNN feature maps are reshaped into a sequential format and passed to a Long Short-Term Memory (LSTM) network. The LSTM learns dependencies and relationships among spatial features, improving the classification performance by modeling contextual information within the image.

5. Classification

The output of the LSTM layer is fed into fully connected dense layers followed by a Softmax activation function. This produces probability scores for each tumor class, and the class with the highest probability is selected as the final prediction.

6. Model Evaluation

The trained model is evaluated using multiple performance metrics, including accuracy, confusion matrix, classification report (precision, recall, F1-score), ROC–AUC curves, and training-validation loss graphs. These metrics provide a comprehensive assessment of the model's effectiveness.

7. Tumor Localization using Grad-CAM

Grad-CAM (Gradient-weighted Class Activation Mapping) is applied to the final convolutional layer of the CNN to generate heatmaps highlighting the regions of the MRI image that contribute most to the model's prediction. These heatmaps are overlaid on the original MRI images to visually localize tumor-affected regions, enhancing interpretability.

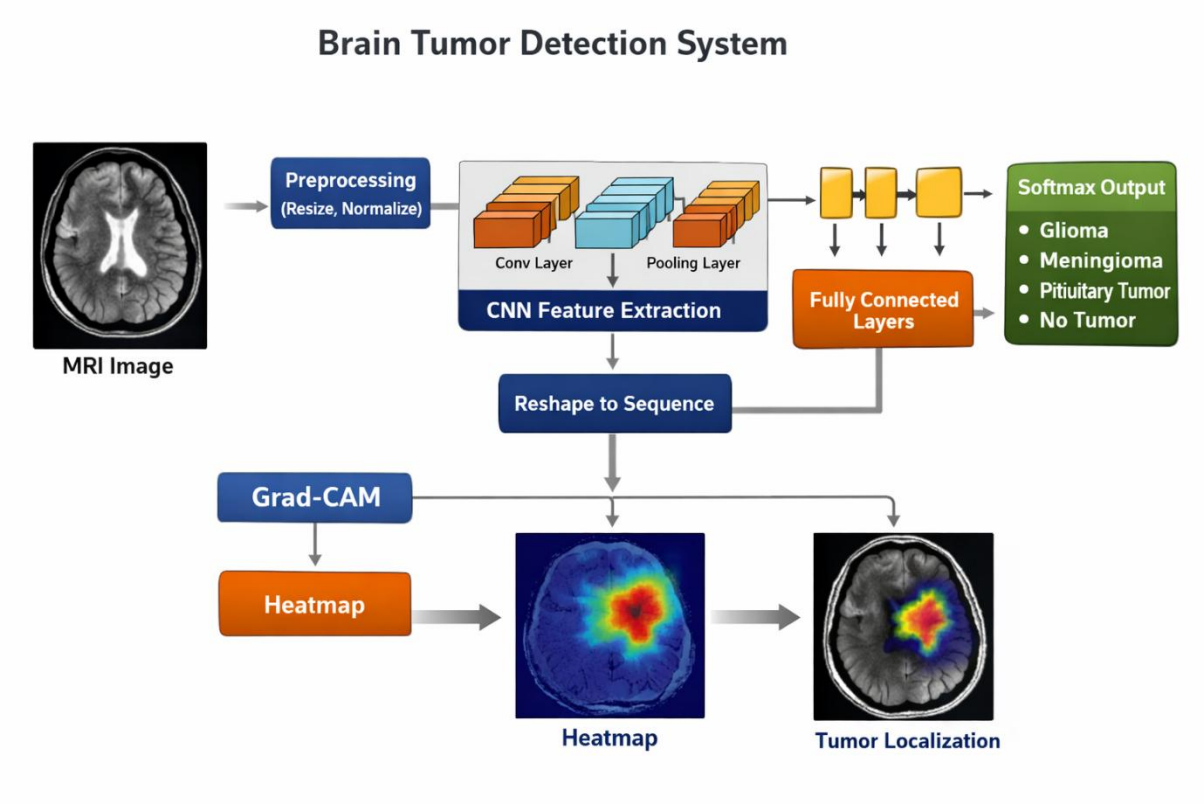
SYSTEM ARCHITECTURE

The system follows a modular and layered architecture to ensure clarity, scalability, and interpretability.

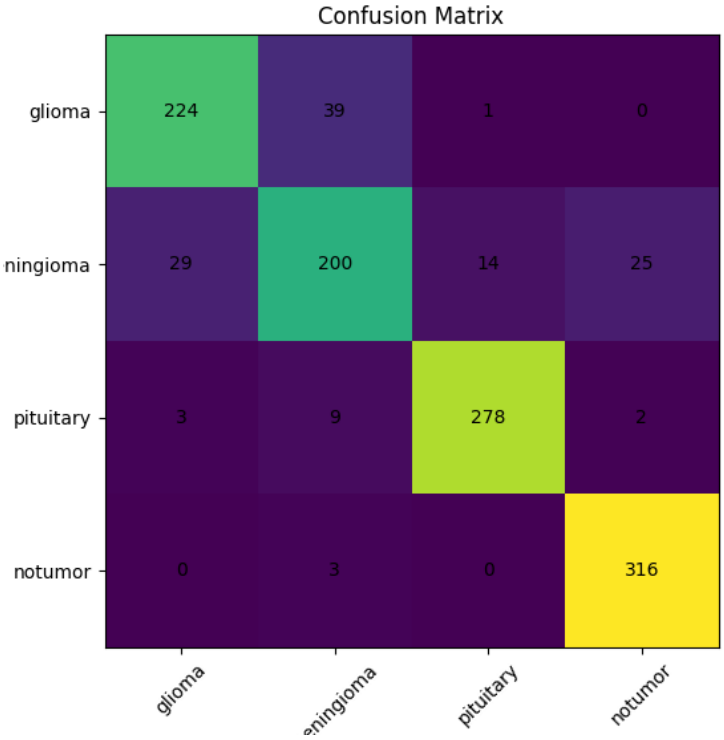
Architecture Description:

1. **Input Layer**
   Receives grayscale MRI images resized to 64 × 64 pixels.
2. **CNN Feature Extraction Layer**
   Multiple convolution and pooling layers extract spatial features such as tumor shape, texture, and intensity variations.
3. **Feature Reshaping Layer**
   The CNN feature maps are reshaped into sequences suitable for LSTM input.
4. **LSTM Layer**
   Learns sequential dependencies from the reshaped feature vectors to enhance classification accuracy.
5. **Fully Connected Layer**
   Dense layers perform high-level reasoning based on extracted features.

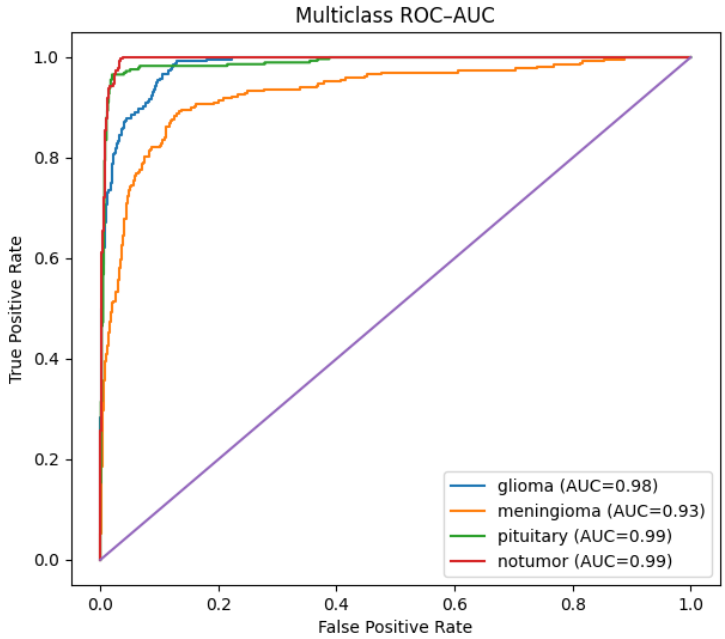
6. **Output Layer**
   Softmax activation produces probabilities for four tumor classes.
7. **Grad-CAM Module**
   Generates visual explanations by highlighting tumor-relevant regions on MRI images.



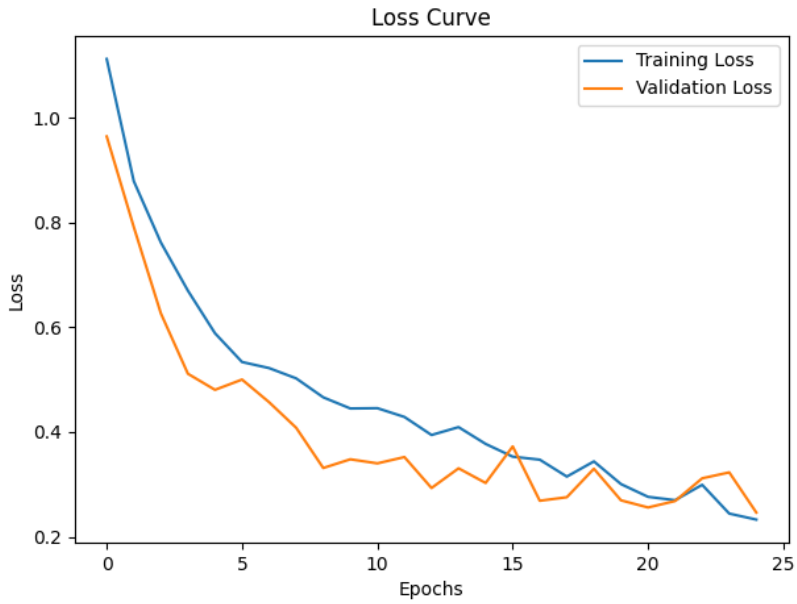
RESULT:



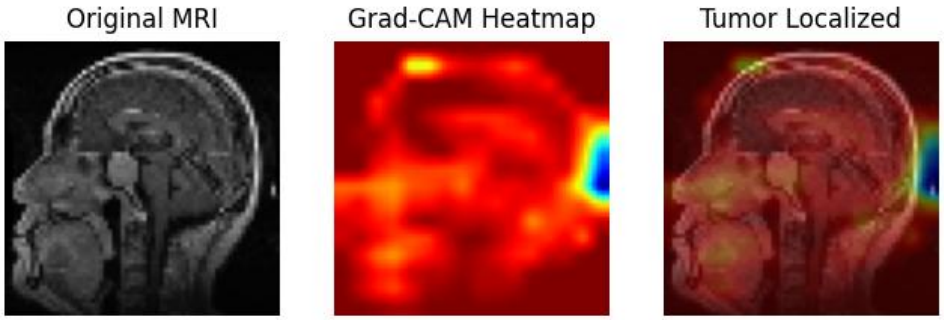
CONFUSION MATRIX FOR MULTI-CLASS CLASSIFICATION



ROC-AUC CURVE FOR MULTI-CLASS CLASSIFICATION



LOSS CURVE FOR MULTI-CLASS CLASSIFICATION



LOCALISATION OF TUMOR USING Grad-CAM

Future Scope:

- Extension to multi-modal medical imaging such as CT and histopathology images.
- Integration of attention mechanisms for improved localization accuracy.
- Deployment as a web-based or hospital-integrated diagnostic tool.
- Clinical validation using real-world datasets and regulatory approval.

CONCLUSION

This project presents an efficient and interpretable deep learning approach for brain tumor detection and classification using MRI images. By combining Convolutional Neural Networks (CNNs) for spatial feature extraction with Long Short-Term Memory (LSTM) networks for sequential learning, the proposed system achieves accurate multi-class tumor classification.

The integration of Grad-CAM enhances the transparency of the model by visually localizing tumor regions, making the predictions more trustworthy and clinically meaningful. Comprehensive evaluation using accuracy, confusion matrix, classification report, ROC–AUC curves, and loss graphs demonstrates the robustness and reliability of the system.

Overall, the proposed CNN–LSTM with Grad-CAM framework highlights the potential of explainable artificial intelligence in medical image analysis and serves as a strong foundation for future clinical decision-support systems.