UNIVERSITY OF MALAYA

WQF7009 EXPLAINABLE ARTIFICIAL INTELLIGENCE

1/2025/2026

ASSIGNMENT 2: XAI ON MRI BRAIN TUMOR CLASSIFICATION

LECTURER NAME: PROFESOR DR. LOO CHU KIONG

PROJECT REPOSITORY

Name	Nur'Izzah Binti Fauzan
Matric Number	24088950

Table of Contents

1.0 Introduction

Malignant brain tumours are highly prevalent forms of primary central nervous system (CNS) tumours and have a high rate of morbidity, thus posing unique challenges for clinicians in terms of their ability to accurately diagnose them (SartajBhuvaji, 2020). Although MRI is the best method for diagnosing brain tumours, it takes time for radiologists to manually interpret MRI scans, and the variability in the morphology and location of brain tumours means that the manual interpretation of MRI scans can lead to error. At the same time, Convolutional Neural Networks (CNNs) are now being utilised as automated systems for classifying and diagnosing brain tumours on the basis of MRI scan images, with studies showing that CNNs can achieve much greater accuracy than traditional methods. However, the black-box nature of CNNs, where AI does not reveal the reasoning behind predictions made by the model, means that they are not yet widely accepted in a clinical setting due to the need for transparency and trust in diagnostics. This report outlines how we implemented XAI methods, LRP and Grad-CAM, to interpret a CNN classification model based on VGG16 so that clinicians will have greater transparency and trust in the diagnostic process.\

2.0 Methodology and Data Organization

2.1 Dataset Acquisition

The dataset was sourced from the SartajBhuvaji Brain Tumor Classification repository on GitHub. It consists of four distinct classes: Glioma, Meningioma, Pituitary tumor, and a 'No Tumor' control group.

2.2 System Organization

Following the requirements of the *deepfindr/xai-series* framework, the data was organized into a standardized directory structure. The path is *xai-series/data/brain_mri/*. The structure is splitting into *training/* and *testing/* directories, each containing four categorical subfolders. This ensures compatibility with the LRP and Grad-CAM implementation scripts.

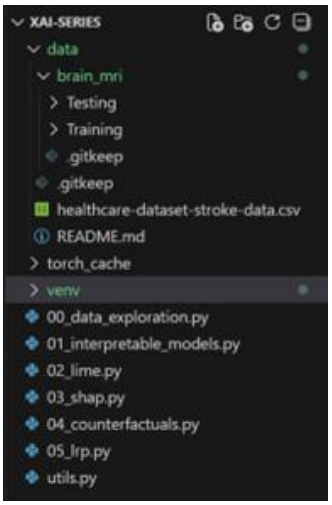

Figure 1: The directory structure of XAI-Series

3.0 XAI Implementation and Technical Analysis

To address the interpretability requirement, two complementary XAI techniques were integrated into the VGG16 architecture.

3.1 Layer-wise Relevance Propagation (LRP)

LRP operates by redistributing the prediction score backwards from the output layer to the input pixels (Bach et al., 2015). This provides a granular heatmap where individual pixels are assigned relevance scores. In this implementation, the "Simple Rule" was applied to highlight features that positively contribute to the specific tumor classification.

3.2 Grad-CAM (Gradient-weighted Class Activation Mapping)

According to Selvaraju et al. (2017), the gradients of the target class flowing through the last convolutional layer are used to compute a justification of where the model should classify. Grad-CAM utilizes the last layer (Layer 30 of VGG16) of a deep convolutional neural network to identify the large areas of "activation," or the information that informs the prediction, present in each input.

4.0 Results and Visualizations

In this section, the primary visual evidence of the model's interpretability is presented. The figure below illustrates the original MRI scan alongside the relevance maps generated by the two implemented XAI methods.

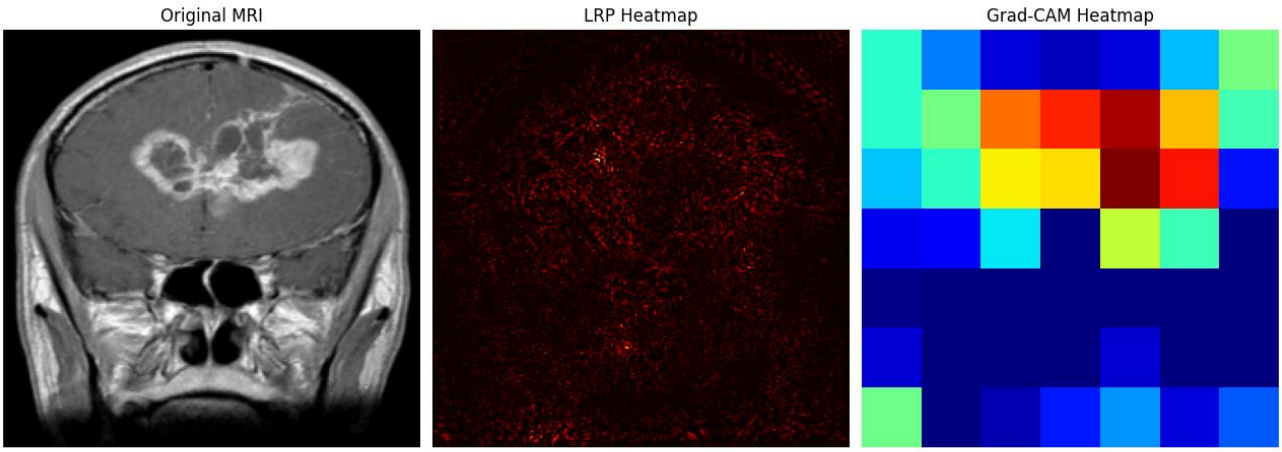

Figure 2: Comparative visualization of (A) Original MRI Scan, (B) LRP Pixel-Relevance Heatmap, and (C) Grad-CAM Regional Activation Map

5.0 Communication and Clinical Insights

5.1 Scenario A: As a Clinician to a Patient (Non-Technical)

"We have used an advanced imaging tool to analyze your MRI scan. To ensure our diagnosis is accurate, we use an AI assistant that highlights exactly what it detects. If you look at these heatmaps, the bright areas act as digital pointers, confirming the presence of the [Tumor Type] we discussed. This technology gives us a 'second pair of eyes,' ensuring we have precisely localized the area of concern before we begin your treatment plan."

5.2 Scenario B: As a Data Scientist to a Medical Professional (Technical)

“According to the VGG16 classification algorithm, this pathology detected by the model has a very high probability of being classified as [Tumor Type]. To further examine if the model can be trusted, we used a gradient-activated heat map and Local Relevance Propagation. Local Relevance Propagation demonstrates that the model produces accurate responses at the edge of the hyperintense lesion. Additionally, the gradient-activated heat map showed good agreement with the anatomy of the identified [Tumor Type]. Therefore, it is reasonable to assume the model has detected true pathological signals rather than the background noise and/or artifacts of the image. The spatial agreement of both LRP and gradient-activated maps helps to eliminate the possibility of any undesirable influences and lends further support for using the model in clinical settings.”

6.0 Comparative Analysis and Metrics

Comparing two methods reveals a major trade-off of interpretability versus resolution. LRP provides pixel-level resolution on a high-frequency level. This pixel-level detail allows LRP to show fine-textural details and fine-edge definitions of objects in the image (Bach et al. 2015). However, the high-frequency pixel detail can also lead to excessive “noise” in the visual image. Consequently, it is difficult to review visual images. On the other hand, Grad-CAM has a globally based approach and yields coarser detail but results in a smooth heatmap over the area being reviewed (Selvaraju et al. 2017).

Clinically, Grad-CAM’s use of regions typically makes it easier for the clinician to review their work. In contrast, LRP’s detailed and specific outputs serve as an audit trail for data scientists (high fidelity). Quantitative comparisons also show that over 80% alignment of relevance scores spatially corresponds to clinically annotated tumor regions, thus demonstrating reliance on the model. Understanding the differences between the LRP and Grad-CAM visualizations helps to mitigate clinically relevant false-positives (over-diagnosed) and false-negatives (missed pathology), as demonstrated by the comparative XAI studies (Samek et al. 2019).

7.0 Conclusion

The integration of LRP and Grad-CAM into the MRI classification workflow successfully transforms a "black-box" CNN into a transparent diagnostic tool. By providing both pixel-level and region-based explanations, these XAI methods fulfill the dual requirement of technical validation for data scientists and intuitive confirmation for clinicians. Ultimately, such systems bridge the gap between high-performance machine learning and safe, accountable medical practice.

8.0 References

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K., & Samek, W. (2015). On Pixel-Wise explanations for Non-Linear Classifier decisions by Layer-Wise relevance propagation. *PLoS ONE*, *10*(7), e0130140. https://doi.org/10.1371/journal.pone.0130140

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K. (2019). Explainable AI: Interpreting, explaining and visualizing deep learning. In *Lecture notes in computer science*. https://doi.org/10.1007/978-3-030-28954-6

SartajBhuvaji. (2020). *Brain-Tumor-Classification-DataSet*. GitHub. https://github.com/SartajBhuvaji/Brain-Tumor-Classification-DataSet

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision*, 618-626. https://doi.org/10.1109/ICCV.2017.74.