Layer-Wise Relevance Propagation for Classifying Brain MRI Images

Ganesh Naik, Shivyogi Bendegerimath, Vijeth Kawari,
Gautam Narajji and Prashant Narayankar
School of Computer Science and Engineering, KLE Technological University, Hubli, Karnataka, India

Keywords: Brain Tumor Classification, Explainable AI (XAI), Medical Imaging, Layer-Wise Relevance Propagation (LRP), Automated Medical Report Generation.

Abstract: Accurate diagnosis and explainable predictions are important in effective planning and monitoring treatment in brain tumor analysis using medical imaging. To enhance the capabilities of tumor detection and interpretation in Brain MRI scans, the proposed work presents a comprehensive framework that combines Explainable AI (XAI) with brain tumor classification. The framework, based on ResNet18, a deep learning model, classifies MRI images into four categories: glioma tumor, meningioma tumor, pituitary tumor, and no tumor. The system incorporates Layer-wise Relevance Propagation (LRP) to highlight regions influencing predictions, providing richer interpretability and visual explanations of the decision-making process. The proposed work has demonstration of 3 approaches for explainable decision making process 1) LRP with heatmaps 2) LRP using overlayed heatmaps 3) Pixel-wise Relevance of presence of tumor. Additionally, the proposed approach includes Automated Medical Report Generation, summarizing categorization results and presenting visual explanations to assist physicians effectively. The proposed model has reached 85% accuracy with strong prediction capabilities and superior explainability in performance to adequately fulfill the fundamental demand of AI-based health solutions to provide more transparent and reliable performance.

1 INTRODUCTION

Brain tumors are among the most serious and dangerous types of cancer; they disturb the delicate balance of the brain and impair vital physical and cognitive functions. The diagnosis is especially challenging because of the complex anatomy of the brain and because the tumors have a tendency to simulate healthy tissue in imaging studies (Litjens et al., 2017a). In Magnetic Resonance Imaging (MRI), despite advancements, diagnoses remain time-consuming and heavily reliant on human judgment. This reliance delays critical treatments, particularly for aggressive cancers. Chemotherapy, radiation, and surgery also carry significant risks, making surgical precision crucial to avoid damaging surrounding tissues. These challenges prolong the emotional and mental stress experienced by patients and their families, emphasizing the need for timely and precise diagnostic solutions.

Conventional MRI analysis requires interpretation by radiologists, which is subject to human error and delays the early diagnosis and treatment of brain tumors (Ronneberger et al., 2015). Furthermore, many existing AI algorithms function as "black boxes," providing predictions without explaining their rationale. This lack of transparency hampers medical professionals, who require clear, interpretable information to make informed decisions. Addressing these limitations is critical to advancing diagnostic accuracy and building trust in AI-powered tools.

The proposed work aims to address these challenges using advanced machine learning techniques, such as Deep Learning and Layer-wise Relevance Propagation (LRP) (Bach et al., 2015a). By leveraging the ResNet architecture, the model effectively processes high-dimensional MRI datasets and handles complex analysis tasks, such as identifying irregular shapes and varying intensities in tumors (He et al., 2016). The integration of LRP provides interpretable heatmaps that highlight critical MRI regions influencing classification decisions, fostering transparency and trust among medical professionals. The approach further reduces dependence on traditional manual feature engineering, making it suitable for low-resource settings where expert radiologists may be scarce. Additionally, the system supports accurate tumor subtype characterization and visual expla-

Naik, G., Bendegerimath, S., Kawari, V., Narajji, G. and Narayankar, P.
Layer-Wise Relevance Propagation for Classifying Brain MRI Images.
DOI: 10.5220/0013607800004664

In *Proceedings of the 3rd International Conference on Futuristic Technology (INCOFT 2025) - Volume 3*, pages 5-11
ISBN: 978-989-758-763-4

nations of tumor-specific properties, such as heterogeneity and border irregularity, enabling personalized therapeutic interventions (Isensee et al., 2021).

Brain tumors affect an estimated 40,000 to 50,000 adults annually in India, with children constituting 20% of these cases. This prevalence, combined with the unique challenges posed by brain tumors, emphasizes the urgent need for improved diagnostic tools. Many existing algorithms remain opaque, operating as “black boxes” and providing predictions without explaining the rationale. The lack of transparency hampers medical professionals, who rely on clear and precise information for making informed decisions. Addressing this gap is crucial to advancing patient care and fostering confidence in AI-powered solutions.

The rest of the paper is organized as follows: Section II reviews the relevant literature and highlights existing gaps in the domain. Section III details the methodology, including the ResNet architecture, LRP integration along with Report Generation. Section IV presents experimental results and analysis, showcasing the model’s effectiveness in addressing diagnostic challenges. Section V discusses the clinical applicability of the proposed approach and its potential impact along with future research directions.

2 BACKGROUND STUDY

2.1 Related Work and Prior Studies

Recent advancements in medical imaging have focused on multi-class classification of brain MRI images, with deep learning models achieving significant breakthroughs in accuracy and efficiency. Traditional machine learning techniques, such as Support Vector Machines (SVMs) and K-Nearest Neighbors (KNNs), relied heavily on handcrafted features like Gray-Level Co-occurrence Matrices (GLCM) and Principal Component Analysis (PCA) (Bach et al., 2015b). These methods often struggled with the complexity and variability inherent in medical imaging datasets. In contrast, Convolutional Neural Networks (CNNs) and transfer learning frameworks, including ResNet, AlexNet, and GoogLeNet, have demonstrated superior robustness and scalability in classifying MRI images into multiple classes (Vankdothu and Hameed, 2022). The incorporation of preprocessing techniques such as data augmentation, skull stripping, and morphological operations further enhances the effectiveness of these models, showcasing their potential for clinical applications (Kulkarni and Sundari, 2020).

Despite their success in achieving high classification accuracy, deep learning models often suffer from a black-box nature, which hinders their interpretability and transparency in medical imaging. Explainability is critical in multi-class classification, as understanding the reasoning behind predictions fosters reliability and trust among clinicians. Visualization methods like Grad-CAM have been employed to highlight tumor regions in MRI images, adding a layer of interpretability to these models (Pang et al., 2023). Additionally, techniques such as Layer-wise Relevance Propagation (LRP) have emerged as powerful tools for explaining classifier decisions by providing pixel-wise decomposition of predictions. LRP generates heatmaps that highlight regions most relevant to a given class prediction, enhancing transparency and interpretability (Bach et al., 2015b). Studies have validated LRP’s utility in multi-class medical imaging tasks by confirming predictions and identifying biologically meaningful features, reinforcing its value in AI-driven diagnostic systems (Babu Vimala et al.,).

2.2 Gaps in Current Research and How proposed work Addresses Them

While existing deep learning models have achieved remarkable performance in classifying brain MRI images, they often lack adequate interpretability. Methods like Grad-CAM, though widely used, focus primarily on high-level feature activations and lack the precision required for fine-grained analysis. Additionally, they may fail to distinguish subtle differences among multiple classes, a critical need in medical imaging (Pang et al., 2023). Furthermore, techniques like SHAP (SHapley Additive exPlanations), which emphasize feature importance, are computationally expensive and do not provide the spatial visualizations necessary for medical diagnostics.

These limitations highlight the need for more advanced explainability methods, such as Layer-wise Relevance Propagation (LRP), which combines computational efficiency with detailed interpretability. LRP addresses the black-box challenge by generating pixel-wise heatmaps that pinpoint the regions contributing most to model predictions, providing a fine-grained understanding of decision-making processes (Bach et al., 2015b). Unlike Grad-CAM, LRP ensures granularity in analyzing multi-class predictions, making it suitable for distinguishing subtle differences in tumor characteristics. Additionally, the proposed integration of LRP with advanced architectures like ResNet leverages the strengths of deep learning for robust multi-class classification while enhancing transparency. By addressing these gaps, the current work bridges the divide between high-performing deep learning models and the clinical need for explainable AI solutions, promoting trust and adoption among medical professionals.

3 PROPOSED WORK

The proposed model for brain tumor classification uses ResNet-18 while for explaining the decision-making process, it uses Layer-wise Relevance Propagation, a technique used in Explainable Artificial Intelligence (XAI) to make deep learning models, especially neural networks, more interpretable. For making the classification and explanation process more readable, the model uses a report generation module that summarizes the intricate details of tumor detection and analysis. The workflow for the proposed model is shown in Figure 1.

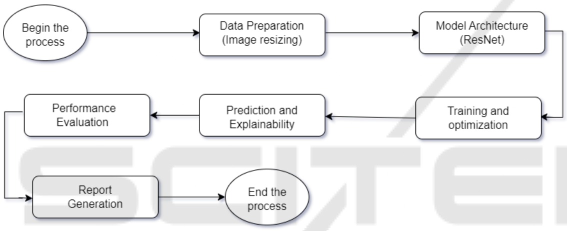

Begin the process
Data Preparation (Image resizing)
Model Architecture (ResNet)
Training and optimization
Prediction and Explainability
Performance Evaluation
Report Generation
End the process

Figure 1: Proposed Model Workflow.

The following steps make up the methodology:

3.1 Dataset Preparation

Dataset consists of training and testing folders of Brain MRI images of namely 4 classes: glioma tumor, pituitary tumor, meningioma tumor and no tumor.

During the Dataset Preparation process, photos are categorized by their class names and arranged into distinct training and testing folders. The image is shrunk to 224 × 224 and normalized using standard ImageNet mean and standard deviation data. PyTorch's ImageFolder tool is used to load the dataset, and data loaders are made with a batch size of 32 for processing efficiency.

The formula for normalization is shown in Equation 1:

$$x_{\text{normalized}} = \frac{x-\mu}{\sigma}, \tag{1}$$

3.2 Resnet-18

Resnet-18 architecture as shown in Figure 2 is a deep learning architecture that is very powerful in handling complex image classification tasks. The residual connections make ResNet-18 particularly effective, as they allow the network to learn deeper and more intricate features without running into problems like vanishing gradients. These connections enable the model to process and understand detailed patterns in images, which is crucial when working with medical images like MRIs. ResNet-18 has been used in a number of image recognition applications, and its ability to capture simple features such as edges and more complex patterns that differentiate tumor types makes it a good choice for this task. Besides, ResNet-18 is computationally efficient, and it can be fine-tuned for specific tasks like brain tumor classification without requiring much more resources. This makes ResNet-18 a good model for this classification task due to its accuracy, efficiency, and ability to handle complex data.

The fundamental architecture we use is the ResNet-18 model pretrained on the ImageNet dataset. The fully connected layer is swapped out for a new linear layer that generates probabilities for four classes in order to modify the model for brain tumor classification. The brain tumor dataset is then used to refine the model.

The formula for the output of the linear layer is shown in Equation 2:

$$\mathbf{y} = \text{softmax}(\mathbf{Wh}+\mathbf{b}) \tag{2}$$

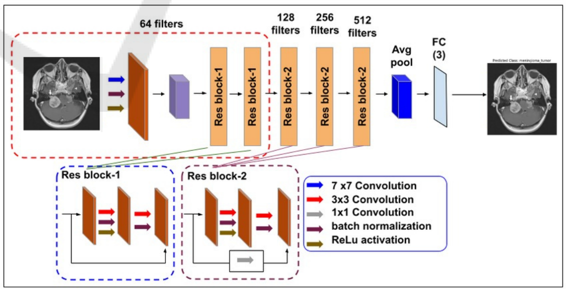


Figure 2: Resnet18 Architecture for Brain Tumor Classification.

3.3 Training and Optimization

In Training and Optimization, the Adam optimizer with a learning rate of 0.001 is used to optimize the model parameters during the training phase. The error between the true and anticipated class labels is calculated using the cross-entropy loss function. Over the

course of 50 epochs, the training loop iteratively modifies the model's parameters while tracking accuracy and loss, among other performance metrics, throughout training and validation.
The formula for Cross-Entropy Loss is shown in Equation 3:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log(\hat{y}_{ij}) \tag{3}$$

The formula for Adam Optimizer is shown in Equation 4:

$$\theta_t = \theta_{t-1} - \eta\frac{m_t}{\sqrt{v_t}+\varepsilon} \tag{4}$$

3.4 Predictability and Explainability

The model uses the training weights to predict the class of a particular MRI picture in this step of predictability and explainability. Layer-wise Relevance Propagation (LRP) is used to make the data interpretable. By breaking down the model's predictions into pixel-by-pixel relevance scores, LRP highlights the areas of the MRI that have the most influence on the outcome.
The formula for relevance propagation is shown in Equation 5:

$$R_i = \sum_j \frac{a_i w_{ij}}{\sum_k a_k w_{kj}} R_j \tag{5}$$

3.5 Performance Evaluation

Lastly, accuracy is the primary metric used to measure the performance of the model during performance evaluation. Both training and validation accuracy are monitored throughout the training process to ensure robustness and generalization.
The formula for accuracy is given by Equation 6:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{6}$$

3.6 Report Generation

A comprehensive PDF report is generated at the end of the model evaluation process.The report provides a summary of the visualizations, explanations, and key findings in the experiment. It includes patient-specific information such as whether or not a tumor was detected, the type of tumor, and the region highlighted by the heatmap corresponding to the tumor. Moreover, the report includes medical advice for the patient according to the type of tumor; it might include suggestions for periodic monitoring or potential surgical removal. It also serves as a useful tool for clinicians and researchers, providing not only visual outputs but also contextual information to be used in the decision-making and further analysis.

4 RESULTS AND ANALYSIS

In this section, we present and analyze the results of the proposed model, which demonstrated high accuracy in classifying brain tumors into glioma tumor, meningioma tumor, pituitary tumor, and non-tumorous cases. The performance metrics, such as precision and F1-score, validate the model's ability to perform the task effectively. Additionally, Layer-wise Relevance Propagation (LRP) was utilized to enhance interpretability, providing heatmaps that highlight the location of the tumor.

4.1 Presentation of Results through Visualizations

To provide clear and effective insights into the model's performance, we use three different types of visualizations based on LRP. These visualizations not only demonstrate the model's focus but also help assess its interpretability.
The following results are for the type of Meningioma tumor:

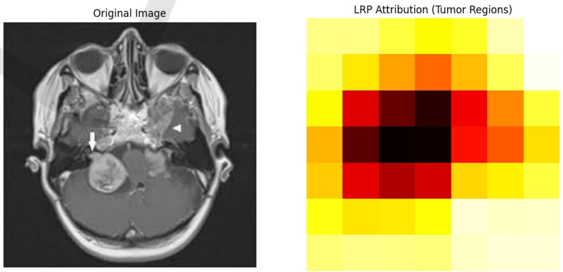


Figure 3: Layer-wise Relevance Propagation with heatmaps.

The heatmap shown in Figure 3 highlights areas of high relevance using warm colors, where the darkest red indicates the strongest attention by the model. The areas that correspond to the tumor region are clearly visible, making it evident that the model is focusing on clinically significant structures.

Figure 4 shows the overlayed heatmap, which combines the heatmap with the original MRI image. This allows us to visualize the anatomical structure along with the relevance information. This technique

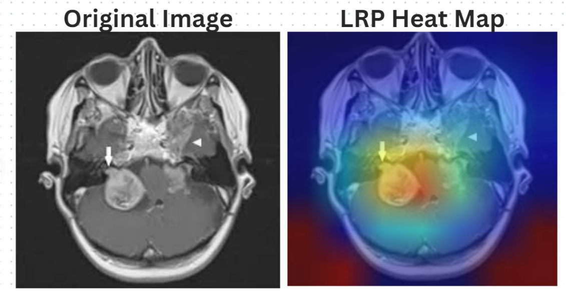


Figure 4: Layer-wise Relevance Propagation Using Overlayed Heatmaps.

provides a spatial context for the model's focus, ensuring that the relevance aligns with clinical expectations and helping users understand the relationship between the tumor and surrounding tissues.

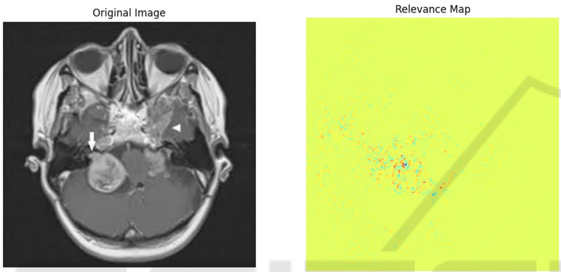


Figure 5: Pixel-wise Relevance of Presence of Tumor.

Figure 5 presents LRP without the heatmap overlay, offering a simpler depiction of the relevant areas. This visualization provides a straightforward representation of relevance, using binary masks to identify tumor regions, as well as showing non-contributing areas and tumor contours with distinct color coding: blue for tumor regions, red for tumor borders, and green for non-relevant areas.

Similarly the relevance maps for pituitary (Figure 6, Figure 7) and glioma (Figure 8, Figure 9) type of tumors are also depicted. The model does not generate any type of map if tumor is not detected.

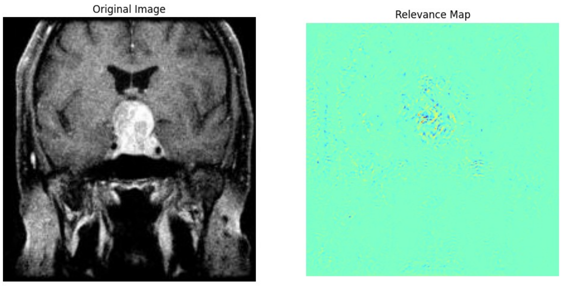


Figure 6: Pixel-wise Relevance of Pituitary Tumor

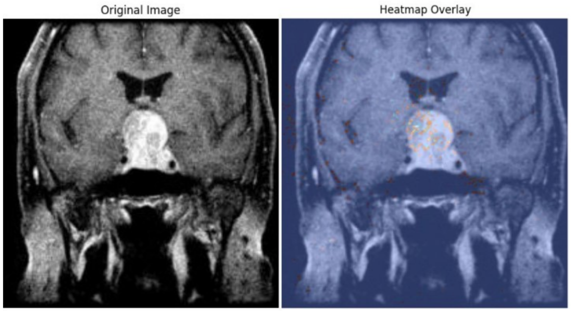


Figure 7: Overlayed heatmap of Pituitary Tumor

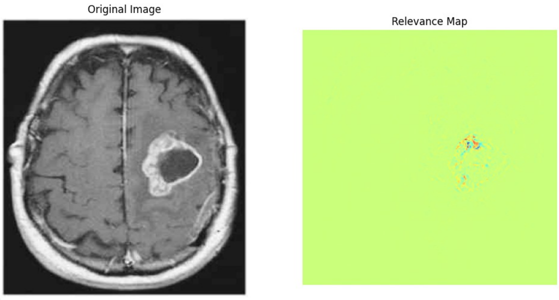


Figure 8: Pixel-wise Relevance of Glioma Tumor

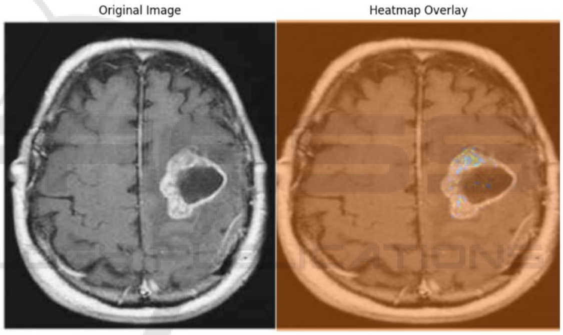


Figure 9: Overlayed heatmap of Glioma Tumor

4.2 Detailed Analysis of Model Performance

The model demonstrated high accuracy in identifying and classifying brain tumors. Performance metrics, including precision and F1-score, validate its effectiveness. The use of LRP heatmaps further enhances the confidence in the model's decision-making process by visually confirming that it focuses on the correct areas, such as the tumor regions.

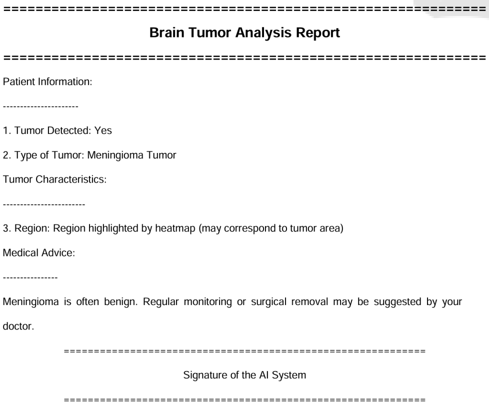
The heatmaps provide an intuitive understanding of the model's attention to specific areas, while the overlayed heatmaps give a more contextual view, allowing clinicians to assess the spatial relationship between the tumor and its surrounding tissue. These visualizations not only validate the model's predictions but also serve as a reliable interpretability tool that can support clinical decision-making.

4.3 Comparison with Previous Approaches

When compared with previous methods in brain tumor classification, the proposed model provides both high performance and excellent interpretability through LRP visualizations. Traditional methods may lack interpretability or offer limited visual insights into model predictions. The proposed approach, by contrast, offers a comprehensive understanding of how the model arrives at its predictions, which is crucial in medical applications where trust in the model's decision-making is essential.

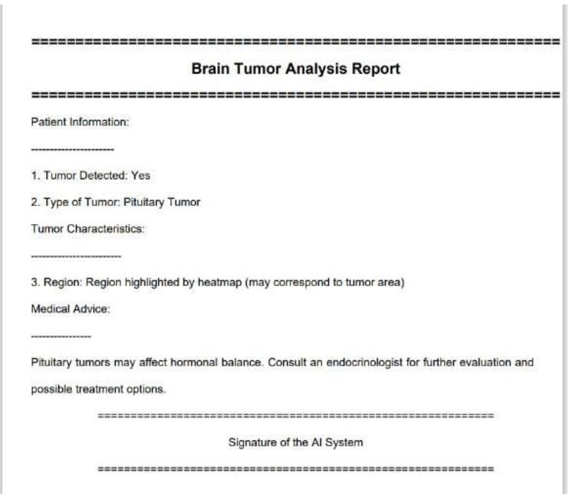
4.4 Interpretation of Results and Implications

The results underscore the ability of the proposed model to focus on clinically relevant regions, offering interpretability through heatmaps and overlayed heatmaps. This helps clinicians to trust the model's predictions and provides assurance that the areas identified by the model correspond to the tumor regions in the MRI images. Additionally, the simplified visualization without heatmaps ensures that clinicians can rely on accurate, quantitative data when making decisions. The report generated as shown in Figure 6 at the end of the process also facilitates patient care by summarizing important tumor-related information and offering treatment recommendations based on the tumor type.



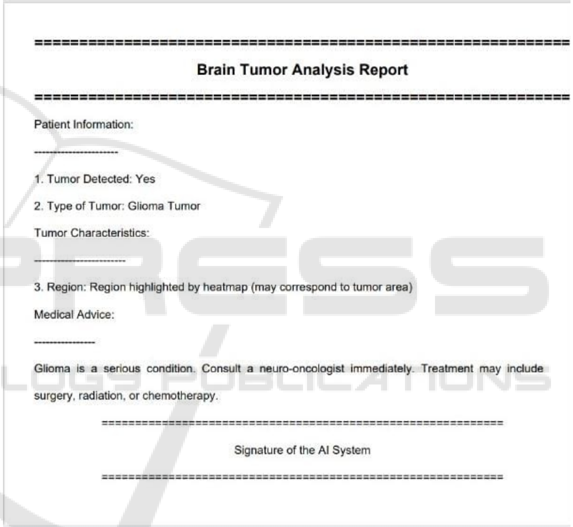
```
============================================================
                 Brain Tumor Analysis Report
============================================================
Patient Information:
--------------------
1. Tumor Detected: Yes
2. Type of Tumor: Meningioma Tumor
Tumor Characteristics:
----------------------
3. Region: Region highlighted by heatmap (may correspond to tumor area)
Medical Advice:
--------------
Meningioma is often benign. Regular monitoring or surgical removal may be suggested by your
doctor.
        ============================================================
                        Signature of the AI System
        ============================================================
```

Figure 10: PDF Report of Meningioma Tumor



```
============================================================
                 Brain Tumor Analysis Report
============================================================
Patient Information:
--------------------
1. Tumor Detected: Yes
2. Type of Tumor: Pituitary Tumor
Tumor Characteristics:
----------------------
3. Region: Region highlighted by heatmap (may correspond to tumor area)
Medical Advice:
--------------
Pituitary tumors may affect hormonal balance. Consult an endocrinologist for further evaluation and
possible treatment options.
        ============================================================
                        Signature of the AI System
        ============================================================
```

Figure 11: PDF Report of Pituitary Tumor

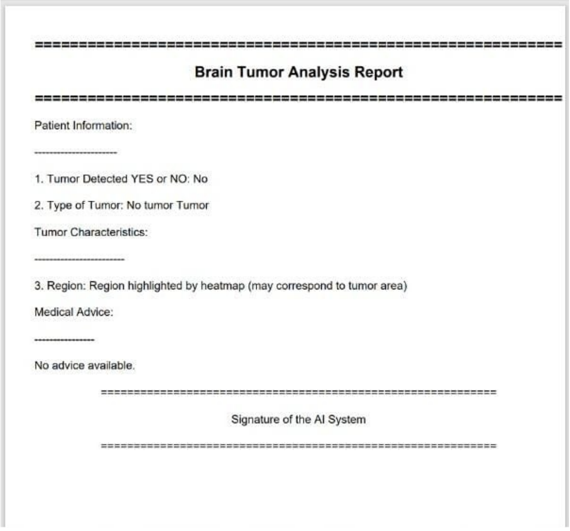


```
============================================================
                 Brain Tumor Analysis Report
============================================================
Patient Information:
--------------------
1. Tumor Detected: Yes
2. Type of Tumor: Glioma Tumor
Tumor Characteristics:
----------------------
3. Region: Region highlighted by heatmap (may correspond to tumor area)
Medical Advice:
--------------
Glioma is a serious condition. Consult a neuro-oncologist immediately. Treatment may include
surgery, radiation, or chemotherapy.
        ============================================================
                        Signature of the AI System
        ============================================================
```

Figure 12: PDF Report of Glioma Tumor

4.5 Limitations of the Current Approach

While the proposed model performs well in classifying brain tumors and providing interpretability, there are some limitations. One challenge arises when the relevance is distributed across multiple areas, which can reduce the clarity of heatmap visualizations. In such cases, the model's focus may become ambiguous, making it harder to interpret the results clearly. Additionally, although the heatmaps provide valuable insights, they may not always offer the level of precision required in all medical applications, particularly when multiple tumors or complex cases are involved.



Brain Tumor Analysis Report

Patient Information:

1. Tumor Detected YES or NO: No

2. Type of Tumor: No tumor Tumor

Tumor Characteristics:

3. Region: Region highlighted by heatmap (may correspond to tumor area)

Medical Advice:

No advice available.

Signature of the AI System

Figure 13: PDF Report of No Tumor

5 CONCLUSION AND FUTURE WORK

The proposed approach bridges the gap by applying Layer-wise Relevance Propagation (LRP) to generate heatmaps that explain model predictions, enabling medical practitioners to verify the rationale behind AI outputs. This is crucial for understanding tumor properties, which vary in size, location, and type. LRP enhances the clarity and reliability of the system, helping clinicians make better decisions, develop personalized treatment plans, and increase trust in AI-driven diagnostic tools. Using ResNet-18 and LRP, the system classifies brain tumors in MRI scans into four categories: glioma, meningioma, pituitary, and no tumor. The LRP results are intuitive and easy to understand, making the system suitable for medical use. Future work could improve performance by expanding the dataset and exploring more complex architectures, such as EfficientNet (Litjens et al., 2017b). Incorporating multi-modal data and refining LRP for clearer visual explanations could further enhance model accuracy. An intuitive interface with real-time predictions could help healthcare providers quickly and accurately identify patients in clinical settings.

REFERENCES

Babu Vimala, B., Srinivasan, S., Mathivanan, S., et al. Detection and classification of brain tumor using hybrid deep learning models. sci rep 13, 23029 (2023).

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015a). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015b). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., and Maier-Hein, K. H. (2021). nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211.

Kulkarni, S. M. and Sundari, G. (2020). A framework for brain tumor segmentation and classification using deep learning algorithm. *International Journal of Advanced Computer Science and Applications*, 11(8).

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I. (2017a). A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I. (2017b). A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88.

Pang, T., Li, P., and Zhao, L. (2023). A survey on automatic generation of medical imaging reports based on deep learning. *BioMedical Engineering OnLine*, 22(1):48.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer.

Vankdothu, R. and Hameed, M. A. (2022). Brain tumor mri images identification and classification based on the recurrent convolutional neural network. *Measurement: Sensors*, 24:100412.