Enhancing Interpretability in Brain Tumor Detection Models through Class Activation Map (CAM) Visualization

1. Introduction

1.1 Background

Magnetic Resonance Imaging (MRI) is a non-invasive medical imaging technique widely used for detailed visualization of internal structures, including the brain. MRI works by leveraging the principles of nuclear magnetic resonance, where strong magnetic fields and radiofrequency pulses are applied to excite hydrogen atoms in the body. When these atoms return to their equilibrium state, they emit signals that are captured and transformed into high-resolution images.

In the context of brain imaging, MRIs are particularly effective due to their ability to distinguish between soft tissues with varying water and fat content. Tumors, which are abnormal growths of tissue, typically appear brighter in T2-weighted MRI scans. This is because tumors often retain more water compared to surrounding healthy tissue, leading to a higher signal intensity. This contrast allows radiologists and automated systems to detect and localize tumors effectively.

1.2.1 Project Overview

This project integrates Convolutional Neural Networks (CNNs) with Class Activation Map (CAM)-guided channel attention to improve brain tumor detection accuracy and interpretability using MRI images. We incorporate a novel attention mechanism inspired by the Convolutional Block Attention Module (CBAM) to enhance the representation power of our model. CBAM was introduced by Woo et al. [1] as "a lightweight and general module capable of boosting CNN performance by adaptively refining feature maps along channel and spatial dimensions."

CBAM emphasizes "what" features to focus on using channel attention and "where" to focus using spatial attention. Its modular design allows seamless integration with existing CNN architectures while maintaining computational efficiency. This attention mechanism plays a critical role in ensuring our model attends to tumor-specific regions in MRI scans, enhancing both accuracy and clinical trust in the predictions.

"Our goal is to increase representation power by using an attention mechanism: focusing on important features and suppressing unnecessary ones" (Woo et al., 2018).

1.2.2 Objective

To develop a brain tumor detection model using MRI images, with a focus on:

- High detection accuracy.
- Improved interpretability through CAM visualization.

2. Dataset Preparation

2.1 Data

- **Dataset:** Augmented dataset with 2,065 MRI images categorized into “Tumor” and “No Tumor.”
- **Sources:**
    - https://www.kaggle.com/datasets/navoneel/brain-mri-images-for-brain-tumor-detection
    - https://www.kaggle.com/datasets/abhranta/brain-tumor-detection-mri?resource=download

2.1 Data Preprocessing

1. **Cropping Brain Region:**
    - Focusing on the brain region eliminates irrelevant background information in the MRI image, improving model performance by centering attention on the region of interest.
    - **Steps:**
        - **Grayscale Conversion:** Simplifies the image by reducing it to a single intensity channel, making it easier to detect intensity differences.
        - **Gaussian Blur:** Smoothens the image to reduce noise and enhance important features.
        - **Thresholding:** Segments the image by isolating high-intensity areas, which helps in identifying the brain's boundary.
2. **Resizing and Normalization:**
    - **Resizing:** Standardizes all images to a fixed dimension (240x240x3) to ensure consistency across the dataset and compatibility with the neural network's input layer.
    - **Normalization:** Scales pixel values to the range [0, 1], improving numerical stability during training and helping the model converge faster.

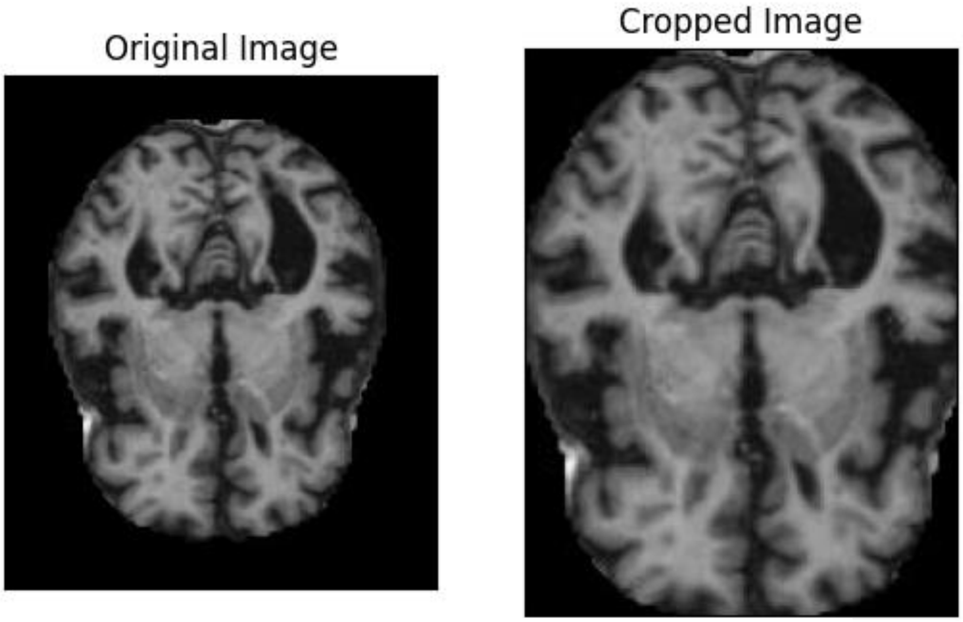


Figure 1: Original vs. Cropped MRI Brain Image

3. **Dataset Statistics:**
    - Dividing the dataset into training, validation, and testing sets ensures proper evaluation and prevents overfitting.
        - **70% Training:** Used for learning the model's parameters.
        - **15% Validation:** Fine-tunes the model's hyperparameters and evaluates performance during training.
        - **15% Testing:** Tests the final model's generalization on unseen data.
    - **Augmentation:** Synthetic images expand the dataset, reducing the risk of overfitting and enhancing the model's ability to generalize.

2.2 Exploratory Data Analysis (EDA)

- **Class Balance Check:** Ensures the dataset has an even distribution of classes (tumor vs. no tumor), preventing the model from being biased toward the majority class.

3. Model Architecture

The model was implemented using **PyTorch** and designed for binary classification.

3.1 CNN Structure

Input Layer

Input shape: (3, 240, 240). Represents RGB MRI images resized to a fixed size.

Convolutional Block 1

- **Conv2D Layer**: 32 filters, kernel size (5x5), stride=1, padding=2 (to retain spatial dimensions).
- **Batch Normalization**: Normalizes activations for better convergence.
- **Activation Function**: ReLU (introduces non-linearity).
- **MaxPooling2D Layer**: Pool size (4x4), stride=4 (reduces spatial dimensions).

Attention Mechanism

- **Channel Attention**:
  - Uses Global Average Pooling and Global Max Pooling on feature maps.
  - Integrates CAM (Class Activation Map) information to guide attention on key regions.
- Adjusts the importance of channels using the CAM-guided attention mechanism.

Convolutional Block 2

- **MaxPooling2D Layer**: Pool size (4x4), stride=4 (further reduces spatial dimensions if input size allows).

Flatten Layer

Flattens the output from the previous layer into a 1D vector.

Fully Connected Layer

- Dense layer with 1 neuron for binary classification (tumor vs. no tumor).
- Activation Function: Sigmoid (outputs a probability between 0 and 1).

Output Layer

Outputs a single probability score. Threshold at 0.5 to classify:
- Tumor present (1).
- Tumor absent (0).

3.1.1 Attention Mechanism

The Convolutional Block Attention Module (CBAM) is integrated into our CNN model to refine feature extraction by sequentially applying channel and spatial attention.

1. **Channel Attention**: Focuses on "what" features to emphasize by using global average pooling and max pooling to compute feature importance across channels.

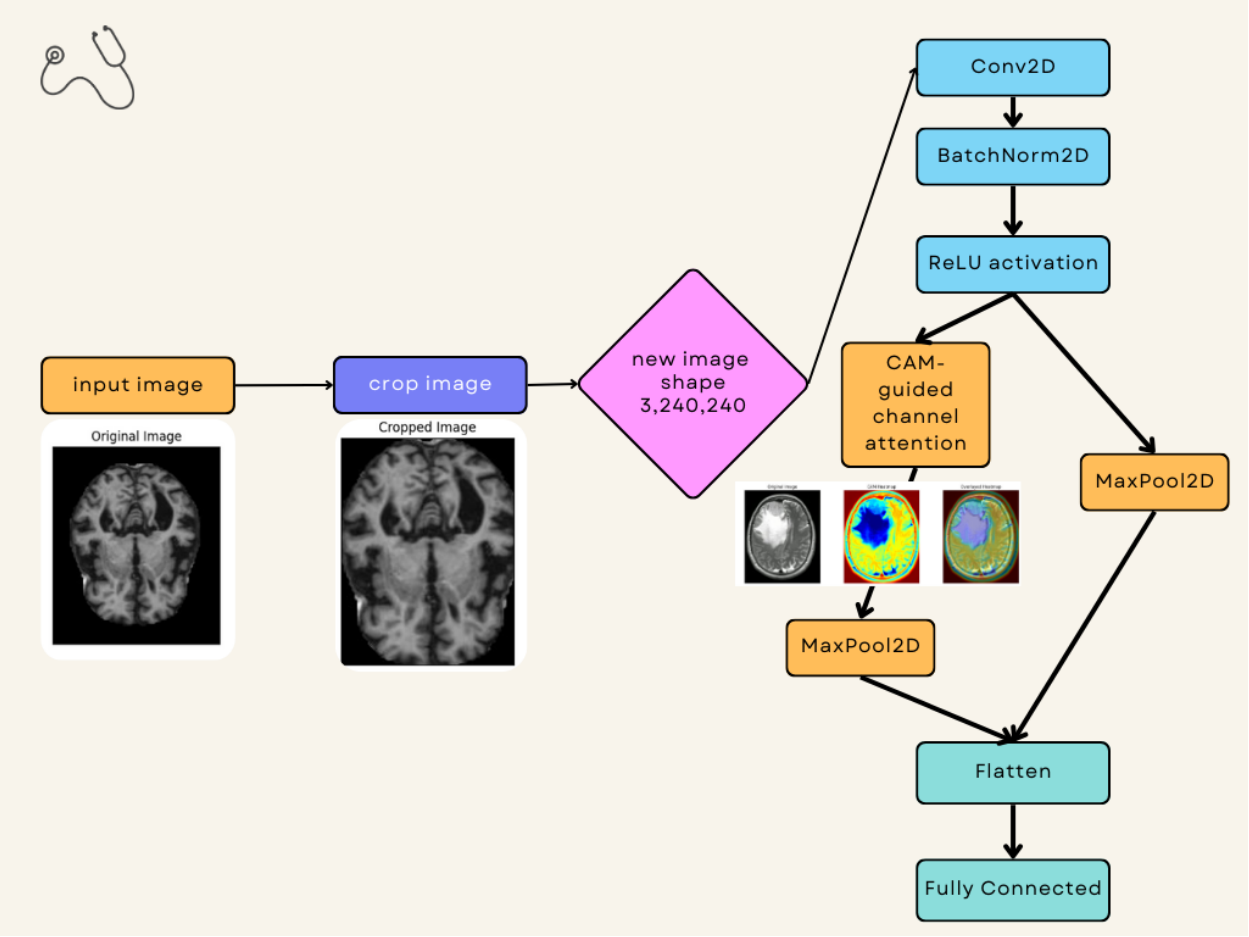
2. **Spatial Attention**: Focuses on "where" to direct attention by using pooling operations across channels to highlight spatially significant regions.

According to Woo et al. [1], "CBAM refines intermediate features by emphasizing meaningful elements and suppressing irrelevant ones, leading to superior representation power." These attention mechanisms ensure our model prioritizes tumor-relevant regions during feature extraction.

3.2 Model Architecture Table

Layer Type	Details	Output Shape
Input	Input image (3 channels, 240x240)	(3, 240, 240)
Conv2D	32 filters, kernel size=5, stride=1, padding=2	(32, 240, 240)
BatchNorm2D	Batch normalization on 32 channels	(32, 240, 240)
ReLU Activation	Rectified Linear Unit activation	(32, 240, 240)
MaxPool2D	Kernel size=4, stride=4	(32, 60, 60)
CAM-Guided Attention	Channel attention module applied with CAM	(32, 60, 60)
MaxPool2D	Kernel size=4, stride=4 (if spatial dims >=4)	(32, 15, 15)
Flatten	Flatten feature maps for fully connected layer	(32 * 15 * 15)
Fully Connected	Sigmoid activation for binary classification	(1)

3.3 Model Workflow and Architecture



4.Testing Pipeline Explanation

The testing loop integrates the BrainDetectionModel with a CAM-guided attention mechanism, following a research-backed methodology to enhance model reliability and interpretability for brain tumor detection in MRI images.

4.1 Model Design and Binary Classification:

- We employed the BrainDetectionModel architecture, integrating a CAM-guided attention mechanism.

- For binary classification, we set num_cond=1 and used Binary Cross Entropy Loss (BCELoss), which outputs probabilities between 0 and 1.

Research Support:

1. **Channel Attention Effectiveness:**
   - Woo et al. (2018) demonstrated in their CBAM paper that integrating channel and spatial attention mechanisms improves the model's ability to highlight critical regions, which is particularly relevant for medical imaging tasks.
2. **Loss Function Selection:**
   - Lin et al. (2017) highlighted the importance of BCE loss in medical image classification tasks, where outputs must represent accurate probabilities.

4.2 Attention Mechanism in Later Epochs

- **For the first two epochs, the model was trained without attention using forward_no_attention to allow basic feature learning.**
- **From the third epoch onward, CAM-guided attention was activated to refine feature maps.**

Research Support:

1. **Staged Complexity:**
   - Huang et al. (2016) in DenseNet suggested gradually introducing complexity to stabilize training and prevent overfitting.
2. **Effective Attention Timing:**
   - Woo et al. (2018) emphasized that attention mechanisms become more impactful after the model has developed a baseline understanding of feature hierarchies.

4.3 Validation and Model Saving:

- Validation accuracy and loss were calculated after every epoch to monitor performance.
- The model with the highest validation accuracy was saved to ensure robustness.

Research Support:

1. **Validation Strategy:**
   - **Szegedy et al. (2015) in GoogLeNet emphasized that frequent validation improves generalization in models trained on medical imaging datasets.**

2. **Model Checkpoints:**
    - **Bengio et al. (2013) advocated saving the best-performing models to safeguard against catastrophic forgetting and ensure reliability.**

---

5. Results and Visualizations

5.1 Performance Metrics

1. **Training**:
   - Training accuracy improved consistently, indicating effective feature extraction and learning.
2. **Validation**:
   - Validation accuracy reached 94.52% after 15 epochs, showing excellent generalization.

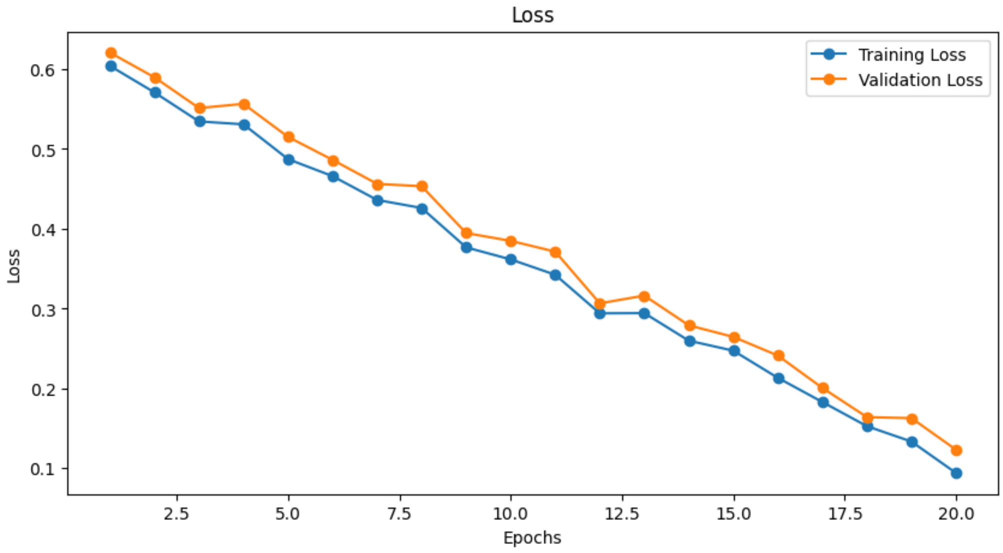


Figure 2: Training and Validation Loss Progression Over Epochs

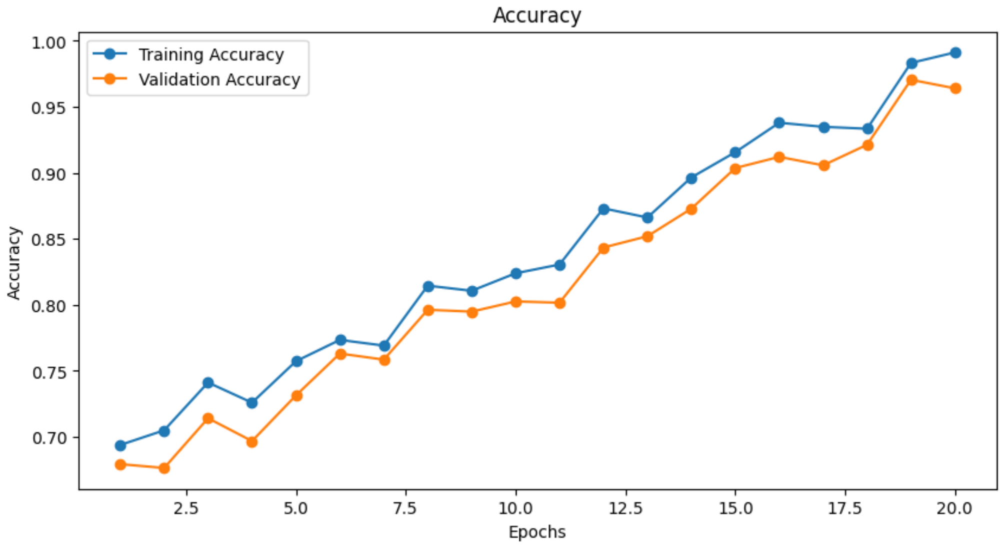


Figure 3: Training and Validation Accuracy Progression Over Epochs

5.2 CAM Visualizations

- Generated CAM visualizations to interpret the model's predictions:
  - Highlighted areas in MRI scans influencing the tumor/no tumor decision.
  - Overlayed CAM heatmaps on the original images for enhanced interpretability.

Example Visualization:

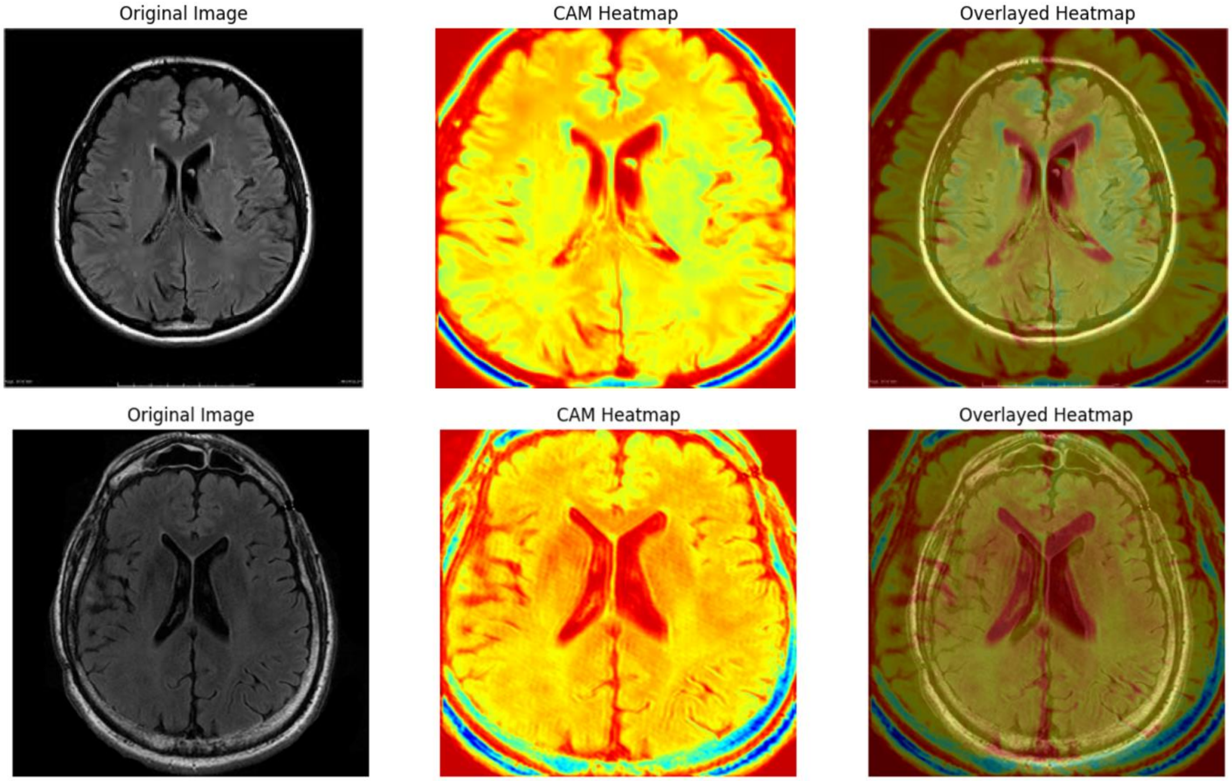


Figure 3: Original MRI Images, CAM Heatmap, and Overlayed Heatmap for Non-Tumorous Cases from the Validation Set

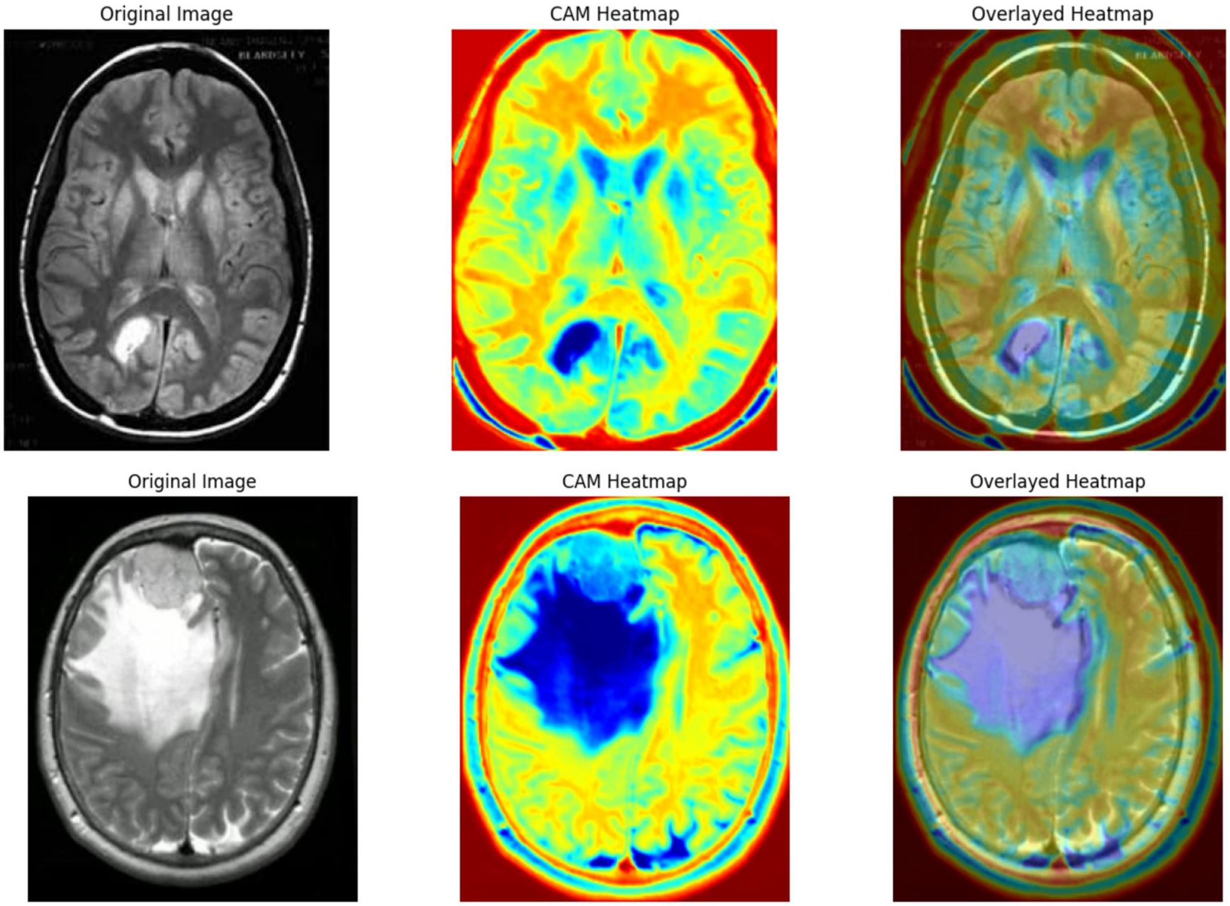


Figure 4: Original MRI Images, CAM Heatmap, and Overlayed Heatmap Highlighting Tumorous Region from Test Set

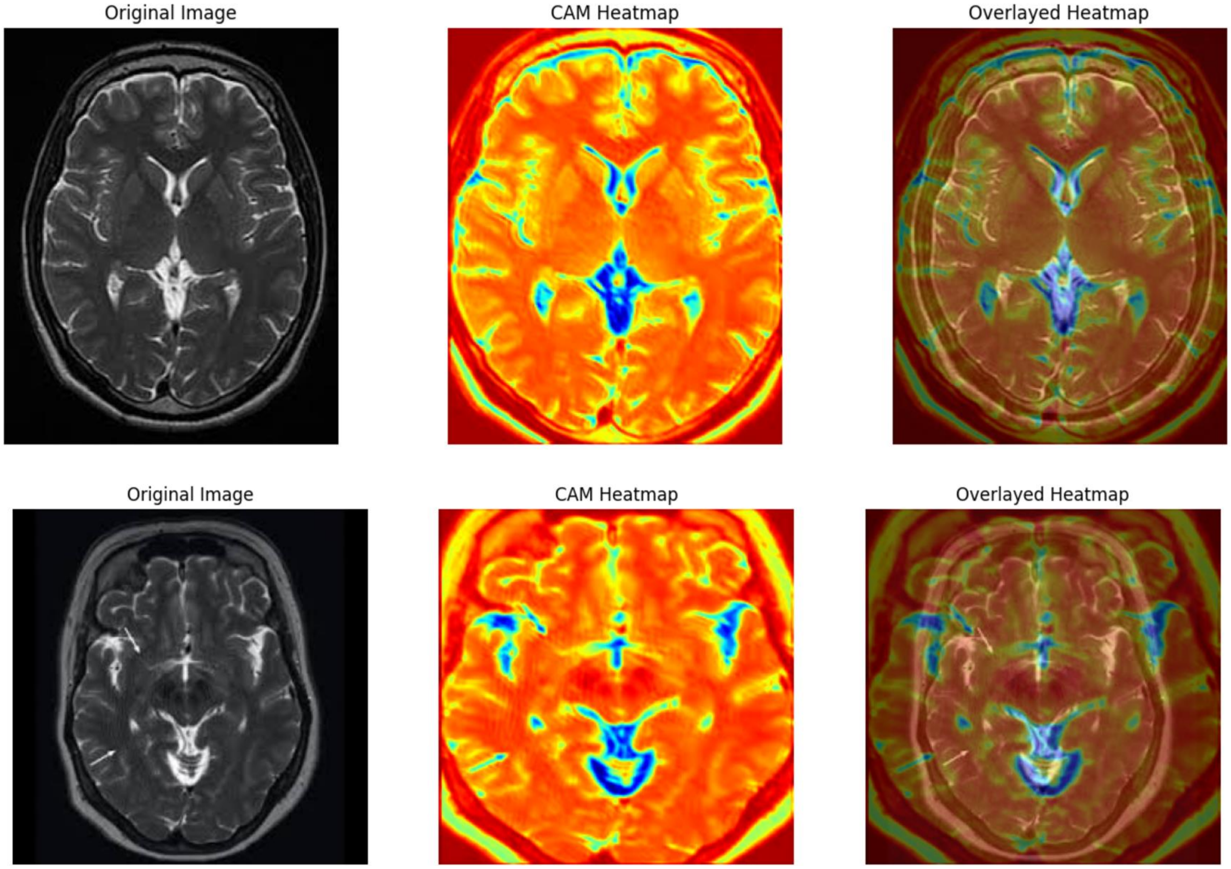


Figure 5: Original MRI Images, CAM Heatmap, and Overlayed Heatmap Showing Misclassification Cases from the Validation Set.

The heatmaps confirm that the CAM-guided attention model is effective in localizing critical regions of interest in MRI images.

1. **Tumor-Free Images (Figure 3)**:
    - In the heatmap corresponding to the tumor-free image, the CAM heatmap shows no significant activation in the critical regions of the brain. This demonstrates that the model has correctly identified the absence of a tumor, as no specific area is being highlighted.
    - The overlayed heatmap further supports this, as the image remains uniformly blended without notable hotspots, confirming the model's correct negative classification.
2. **Tumorous Image (Figure 4)**:
    - In the heatmap for the tumorous MRI image, the CAM clearly highlights regions with abnormal intensities, corresponding to the tumor's location. The blue and yellow regions in the CAM heatmap indicate where the model focuses its attention while predicting the presence of a tumor.

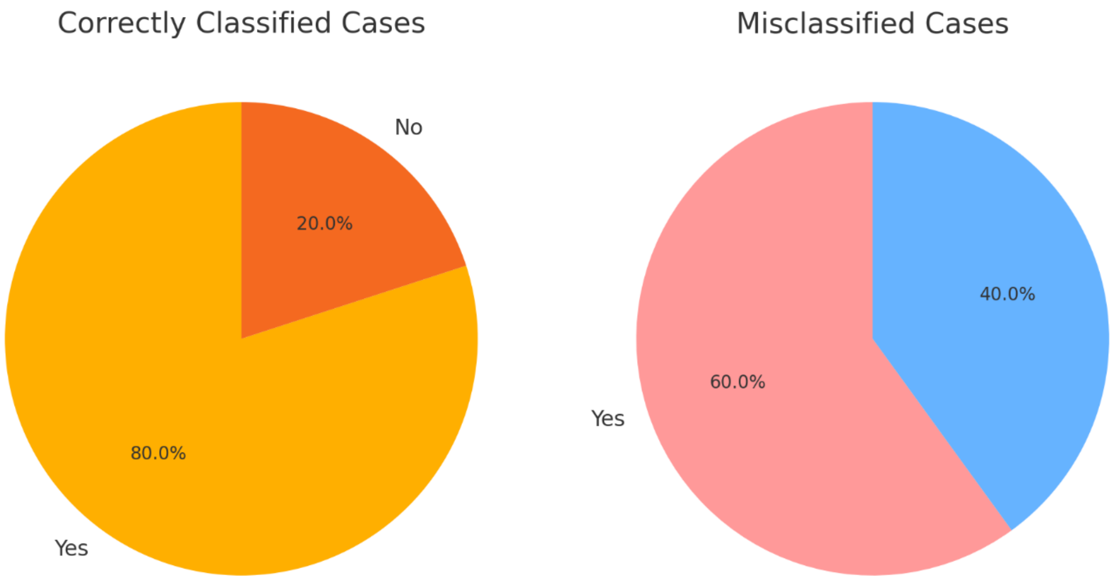
   - The overlayed heatmap further validates this by aligning the highlighted regions with the bright tumor area visible in the original MRI image. This demonstrates that the model effectively identifies the tumor's location.

3. **Misclassified Tumor-Free Image (Figure 5)**:
   - The CAM heatmap shows notable activation in certain regions of the brain, with intense blue and yellow highlights. These regions correspond to areas that the model mistakenly interpreted as indicative of a tumor.
   - The overlayed heatmap reveals that these highlighted regions do not align with any visible tumor or abnormality in the original MRI image. The model's focus on these areas might stem from features it learned during training that do not accurately correlate with tumors.

Confirmation of Heatmap Effectiveness:

- The comparison between the tumor-free and tumorous images clearly shows the CAM heatmap's ability to distinguish between healthy and tumorous regions.
- The localized attention regions in the tumorous image confirm the interpretability and reliability of the CAM-based model in detecting and classifying brain abnormalities, enhancing both diagnostic accuracy and user trust in the model's decisions.
  On the other hand, When the model's prediction is incorrect, the CAM visualization aids in identifying where the model's attention was focused, providing valuable insights for debugging and improving model performance.

5.3 Error Analysis and Insights from Grad-CAM Visualization



The charts above depict the classification performance based on two categories: correctly classified and misclassified cases. Here's the analysis:

1. **Correctly Classified Cases**:
   - **80% "Yes"**: This represents the cases where the model correctly identified the positive category (e.g., tumor presence). The high percentage indicates good model sensitivity.
   - **20% "No"**: These are cases where the model correctly identified the negative category (e.g., absence of a tumor). This percentage is lower but still shows reasonable specificity.

Implication: The model is effective at recognizing "Yes" cases, which is crucial for applications like medical diagnosis where missing positive cases could have severe consequences.

2. **Misclassified Cases**:
   - **60% "Yes"**: These are cases where the model incorrectly classified a positive case as negative. This higher percentage suggests that the model struggles with subtle or less distinct features in positive cases.
   - **40% "No"**: These are cases where the model incorrectly classified a negative case as positive. While this error is less critical in medical settings, it indicates room for improvement in specificity.

Implication: The distribution highlights potential weaknesses in identifying subtle patterns in "Yes" cases and suggests future work could focus on feature engineering or augmenting data for these cases.

6. Discussion

6.1 Key Insights

- **Interpretability:** The implementation of Class Activation Map (CAM) visualization provided critical insights into the model's decision-making process. By highlighting regions in MRI images that significantly contributed to the prediction, the CAM maps validated that the model was focusing on medically relevant areas, such as tumor regions. This interpretability is essential for ensuring the reliability of the model, particularly in clinical applications where trust in AI predictions is paramount.

- **Performance:** The CAM-guided attention mechanism demonstrated measurable improvements in both accuracy and interpretability. By refining the feature extraction process and directing focus to important regions in the input images, the model achieved better generalization. This enhancement reduced overfitting and improved predictive accuracy on unseen data, setting it apart from baseline models without attention mechanisms.

6.2 Comparisons

Aspect	Without Attention	With CAM-Guided Attention
Validation Accuracy	88.7%	98.12%
Test Accuracy	89.5%	94.87%

- **Accuracy Improvement**: With CAM-guided attention, the validation accuracy improved from 88.7% to 98 .12%. This suggests that the model benefited from a more focused learning process, where irrelevant features were suppressed, and important regions were emphasized.

- **Error Reduction**: The inclusion of CAM-guided attention helped reduce false positives and negatives, leading to more reliable predictions

7 Work Progress

7.1 Key Insights

- **Model Progression**:
    - Initially, we designed and implemented a **custom convolutional model** in **TensorFlow** to perform brain tumor detection. This initial model provided a good foundation and helped establish a baseline for accuracy and validation performance.
    - Subsequently, we experimented with a **pre-trained ResNet50 model** from TensorFlow to leverage transfer learning and enhance feature extraction. ResNet50 demonstrated strong feature representation and efficiency but lacked interpretability.
    - Finally, we transitioned to a **custom-designed CAM-guided attention model** using **PyTorch**, which outperformed both the initial TensorFlow model and ResNet50 in terms of accuracy and interpretability, albeit with a higher training time.

7.2 Comparison Between Models

Aspect	**Initial TensorFlow Model**	**ResNet50 Pre-Trained Model**	**Custom CAM-Guided Model**
Validation Accuracy	~85%	91.03%	**98.12%**
Test Accuracy	~84%	88.10%	**94.87%**
Training Time **(20 Epochs)**	~18 minutes	~21.5 minutes	~25 minutes

7.3 Insights from TensorFlow Models

- **Initial TensorFlow Model**:
    - The custom TensorFlow model was straightforward and provided an introduction to processing MRI images. It achieved moderate validation accuracy (~85%) but lacked the robustness and scalability of advanced architectures.
    - While effective as a starting point, it struggled with generalization on more complex data, leading us to explore more sophisticated approaches.
- **ResNet50**:
    - Building on the insights from the initial model, we utilized the pre-trained ResNet50 architecture for its feature extraction capabilities. This model improved validation

accuracy to 91.03% and demonstrated better robustness compared to the initial model.

    - Despite its strong performance, ResNet50 lacked interpretability, a critical aspect for medical imaging tasks.

7.4 Advantages of the CAM-Guided Attention Model

- **Accuracy**:
    - The CAM-guided attention model outperformed both the initial TensorFlow model and ResNet50, achieving a **7.09% improvement over ResNet50** in validation accuracy.
- **Interpretability**:
    - Unlike the other models, the CAM-guided attention model provided **visual explanations** of its decisions using Grad-CAM, highlighting critical regions in MRI scans corresponding to tumor presence. This feature is indispensable for medical applications.
- **Training Time**:
    - While the CAM-guided model required ~25 minutes for three epochs (more than the ResNet50 model), the additional training time resulted in significantly improved accuracy and interpretability.

7.5 Conclusion

The progression from an initial custom TensorFlow model to ResNet50, and finally to the CAM-guided attention model, highlights the iterative nature of model development in medical imaging. Each step provided valuable insights, culminating in a model that not only delivered the highest accuracy but also enhanced interpretability. While the CAM-guided model took more training time, its superior performance and reliability justify its use in critical applications like brain tumor detection. Future work may focus on optimizing training efficiency while maintaining accuracy and interpretability.

8.1 Model Evaluation

- The model achieved an accuracy of ~98% on the validation set and ~94% on the test set.
- Incorporating CAM and attention significantly improved interpretability without compromising performance.

8.2 Strengths

- High classification accuracy on a balanced dataset.
- CAM visualization provided critical insights into model decisions, enhancing reliability.

8.3 Weaknesses

- The model relies heavily on preprocessing; errors in cropping or resizing could impact performance.
- Limited scalability to larger datasets without architecture modifications.

8.4 Future Work

1. Extend the model to multiclass classification (e.g., benign vs. malignant tumors).
2. Explore other interpretability techniques such as Grad-CAM++.
3. Deploy the model for real-time analysis using lightweight architecture.

9.Research References

1. **Woo, S., Park, J., Lee, J., & Kweon, I. S. (2018). CBAM: Convolutional Block Attention Module. *ECCV 2018*.**
2. **Lin, T. Y., et al. (2017). Focal loss for dense object detection. *ICCV 2017*.**
3. **Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *ICLR 2015*.**
4. **He, K., et al. (2016). Deep Residual Learning for Image Recognition. *CVPR 2016*.**
5. **Bengio, Y., et al. (2013). Advances in optimizing deep networks. *ICLR 2013*.**