CAUCHY – Jurnal Matematika Murni dan Aplikasi
Volume 10 (2) (2025), Pages 890-903
p-ISSN: 2086-0382; e-ISSN: 2477-3344

# An Explainable Deep Learning Approach for Brain Tumor Detection Using MobileNet and Grad-CAM Visualization

Amalan Fadil Gaib[1,2], Safrizal Ardana Ardiyansa[1,3*], Anggito Karta Wijaya[1,4], Eric Julianto[1], I Gusti Ngurah Bagus Ferry Mahayudha[1], and Ando Zamhariro Royan[1]

[1] *Braincore, Jakarta, Indonesia*
[2] *Department of Industrial Engineering, Faculty of Engineering, Gorontalo State University*
[3] *Department of Mathematics, Faculty of Mathematics and Natural Sciences, Brawijaya University*
[4] *Department of Computer Science, Faculty of Information System, Jember University*

**Abstract**

Brain tumor detection remains a significant challenge due to the complex variations in tumor appearance. Although deep learning models have demonstrated high accuracy, their limited interpretability hinders clinical adoption. To address this issue, this study integrates Gradient-weighted Class Activation Mapping (Grad-CAM) into Convolutional Neural Networks (CNNs) to enhance the visual interpretability of predictions. Grad-CAM extends Class Activation Mapping (CAM) and is applicable to a wide range of deep learning architectures. The primary contribution of this work is the demonstration that combining Grad-CAM with MobileNet architectures yields an interpretable and efficient framework for diagnosis of brain tumor, effectively balancing accuracy, computational efficiency, and clinical transparency. Using a Brain Tumor MRI dataset, MobileNetV4 achieved an accuracy of 98.29% with the shortest training time (1738.82 seconds) and an ROC accuracy of 99.96%. MobileNetV3 achieved 99.62% accuracy with an ROC accuracy of 99.92%. Grad-CAM effectively highlighted tumor regions while showing uniform attention in non-tumor cases, thereby reducing false positives. These results demonstrate that lightweight models can achieve a strong balance between predictive performance, training efficiency, and interpretability. The proposed framework thus supports the development of explainable and efficient diagnostic tools for clinical practice.

**Keywords:** Brain tumor detection; Grad-CAM; MobileNet; Visual interpretability



## 1 Introduction

A brain tumor refers to an abnormal proliferation of cells within the brain or the Central Nervous System (CNS) [1]. Globally, the incidence of brain tumors is on the rise, with the highest rates observed in high-income regions such as Europe and North America, while the lowest rates are reported in Africa [2]. These tumors can be classified as malignant (cancerous) or non-malignant (benign), and their clinical impact depends on multiple factors, including tumor type, size, and anatomical location. Among all malignant tumors, glioblastoma becomes the most prevalent, whereas benign tumors are reported more frequently in females [3]. In the United States, the

*Corresponding author. E-mail: safrizal@student.ub.ac.id

Submitted: August 2, 2025 Reviewed: August 20, 2025 Accepted: September 1, 2025
DOI: https://doi.org/10.18860/cauchy.v10i2.35901

annual mortality rate for malignant brain and other CNS tumors stands at 4.43 per 100,000 population, translating to approximately 16,606 deaths annually between 2014 and 2018 [3].

Brain and central nervous system (CNS) tumors rank among the most aggressive and life-threatening cancers. Malignant types have a five-year relative survival rate of only 35.7% [4], with improvements from 23% in 1975–1977 to 36% in 2009–2015 largely limited to younger patients. In children, brain tumors often result in early neuropsychological impairments [5]. The annual incidence in individuals aged 0–19 years is about 6.02 per 100,000, with a five-year survival rate near 75%; however, up to 40% of survivors experience persistent cognitive deficits, particularly in processing speed and working memory [5], [6]. Involvement of eloquent brain regions may further impair language, motor, and visual functions [7]. Long-term survivors also face an elevated risk of ischemic stroke, with an incidence of 267 per 100,000 person-years [8].

Early detection of brain tumors is critical to enhancing survival outcomes and improving the efficacy of therapeutic interventions [9]. Nevertheless, several challenges hinder early diagnosis, including insufficient understanding of disease onset, the lack of risk factors, and the need for more sensitive and accurate detection technologies [10]. Therefore, prompt and accurate diagnosis is essential to prevent irreversible neural damage and to preserve essential cognitive and motor functions. Magnetic Resonance Imaging (MRI) remains the gold standard for visualizing brain tumors due to its superior contrast resolution and non-invasive nature.

MRI employs strong magnetic fields and radiofrequency waves to generate detailed three-dimensional images of internal anatomical structures, particularly soft tissues such as the brain, muscles, and organs [11]. It plays an important role not only in tumor detection but also in therapy planning and monitoring treatment response [12]. However, manual tumor segmentation on MRI scans is widely recognized as a time-consuming and labor-intensive process for radiologists, often requiring significant expertise and attention to detail [13]. The complexity of tumor morphology and variability in imaging appearance further compound the challenges of manual delineation.

To address these limitations, Machine Learning (ML) models have increasingly been adopted in medical image analysis to support disease detection and classification, showing considerable promise in reducing mortality rates associated with cancer and tumors [14]. Previous studies have successfully demonstrated the effectiveness of ML in critical healthcare applications, such as heart attack diagnosis [15], diabetic retinopathy [16], and breast cancer classification [17]. Moreover, ML has been utilized beyond the medical domain, including in insurance analytics [18] and optimization-related tasks [19], highlighting its broad applicability and versatility. Although effective in certain contexts, the performance of traditional ML methods is often constrained by their reliance on handcrafted features and their limited capacity to model the high-dimensional and nonlinear characteristics inherent in complex medical imaging data.

To overcome these constraints, Deep Learning (DL) models, particularly Convolutional Neural Networks (CNNs), have emerged as the most effective tools for brain tumor segmentation and classification on digital MRI scans [20]. Unlike traditional DL models, CNNs offer a significant advantage by automatically learning and extracting complex hierarchical features directly from raw imaging data, thereby reducing the need for manual intervention [21], [22]. These networks have demonstrated exceptional performance in detecting and delineating tumor boundaries, significantly improving diagnostic precision [23]. Recent studies report that state-of-the-art CNN architectures can achieve classification accuracies as high as 98.7%, underscoring their robustness and reliability in neuro-oncological imaging tasks [24]. As a result, CNN-based models are becoming increasingly integrated into clinical decision support systems, paving the way for more efficient and accurate brain tumor diagnosis.

Lightweight CNN architectures such as MobileNet have gained attention in medical imaging for offering high diagnostic accuracy with low computational complexity. Recent enhancements to MobileNetV4 reduce model parameters by up to 64.8% without sacrificing performance [25]. While its application in neuro-oncological imaging remains limited, related studies using MobileNetV3 report diagnostic accuracies up to 99.75%, rivaling more complex models [26]. However, most

existing work focuses on classification, with limited exploration of visual interpretability and precise tumor segmentation, highlighting the need for interpretable models in MRI analysis.

To address this limitation, Gradient-weighted Class Activation Mapping (Grad-CAM) presents a promising strategy. Grad-CAM is an adopted visualization technique that generates heatmaps highlighting the most discriminative regions in an input image that influence a CNN model's decision [27]. By providing visual explanations for the model's predictions, Grad-CAM enhances interpretability and facilitates clinical trust. While Grad-CAM has been extensively utilized in classification contexts, its application within segmentation frameworks for brain tumor analysis remains underexplored. Integrating Grad-CAM with lightweight CNN architectures such as MobileNet has the potential to not only enable accurate tumor classification but also to provide clinically meaningful visual explanations of tumor boundaries.

In response to these challenges, this study proposes a novel framework entitled "MobileNet Assisted Brain Tumor Detection with Visual Interpretability Using Grad-CAM". The primary contributions of this work are as follows.

- Proposing a novel integration of Grad-CAM with MobileNet to enable interpretable brain tumor detection and localization from MRI images.
- This paper contributes as an application-driven and comparative study, demonstrating how established methods (Grad-CAM with MobileNet, ResNet50, and VGG19) can be systematically evaluated for the detection of brain tumors.

The remainder sections of this article is organized as follows. Section 2 describes the research methodology, including the dataset and research flow. Section 3 presents the experimental results, highlighting the performance of the Grad-CNN approach and its comparative evaluation against other models. Finally, Section 4 provides the conclusions and recommendations for future work.

## 2 Methods

This section outlines the methodological framework adopted in this study. Subsection 2.1 presents a detailed description of the research dataset, including its sources and structure. Subsection 2.2 explains the overall research flow, encompassing the sequence of preprocessing, model development, training configuration, and evaluation procedures.

### 2.1 Research MRI Dataset

This research utilizes the publicly available Brain Tumor MRI dataset obtained on Kaggle platform[1]. The dataset is a curated combination of three primary sources, including Figshare dataset, SARTAJ dataset, and Br35H dataset. In total, it contains 7,023 brain MRI images, categorized into four classes, including glioma, meningioma, pituitary tumor, and no tumor.

Specifically, no tumor category dataset are sourced from the Br35H dataset, while glioma, meningioma, and pituitary tumor images are primarily derived from the Figshare and SARTAJ datasets. All images are provided in JPEG format and exhibit variability in both resolution and background margins, reflecting their origin from multiple clinical sources. This diversity contributes to the dataset's robustness and realism, making it a widely adopted benchmark for brain tumor classification studies in medical imaging research.

### 2.2 Research Flow

This section presents and explains the research flow designed for this study. The process begins with a literature review aimed at understanding recent developments in DL techniques for medical image analysis. This review specifically focuses on the performance and architectural differences

[1] https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset

of widely used CNNs. Additionally, the literature review covers the working principles of Grad-CAM, which is crucial for model interpretability in medical diagnostics. Techniques for image augmentation are also studied to understand their impact on improving model generalization and robustness, particularly when dealing with limited and imbalanced medical imaging datasets.

Following the literature review, the dataset was obtained from Kaggle platform and uploaded to the Google Colaboratory for further processing and CNN model development. To improve the generalization capability of the CNN, image preprocessing and augmentation techniques were applied. For the training dataset, each image was first resized to a fixed resolution of $224 \times 224$ pixels using bicubic interpolation to ensure compatibility with standard CNN input dimensions. Data augmentation strategies were then employed, including random horizontal flipping and random rotation within a $\pm 15$ degree range. The augmented images were subsequently converted into tensor format and normalized across all three RGB channels using a mean and standard deviation of 0.5, resulting in a pixel value distribution scaled to the range $[-1, 1]$.

In contrast, the testing MRI dataset underwent only deterministic preprocessing to ensure a consistent and unbiased evaluation of model performance. This included resizing using bicubic interpolation, tensor conversion, and the same normalization process as applied to the training data. Following these preprocessing steps, the dataset was partitioned into 70% training, 15% testing and 15% validation subsets, allowing for systematic model training and independent performance evaluation on unseen data samples. To address potential class imbalance, stratified sampling was employed during the partitioning process to preserve the original class distribution across training, testing, and validation sets, thereby ensuring that each subset contained a representative proportion of all tumor categories.

A CNN model, including MobileNetV3, MobileNetV4, VGG19, and ResNet50, is constructed and iteratively trained on the training dataset. The training process is conducted for 50 epochs, utilizing the Adam optimizer with a learning rate of `(lr = 1e-4)` and the cross-entropy loss function as the objective criterion. In each epoch, the model's weights and biases are updated using the backpropagation algorithm, and performance is evaluated on the testing set. Key metrics such as training and validation loss, classification accuracy, and computation time are recorded and analyzed throughout the training process to monitor convergence and performance trends. This iterative process continues until the maximum epochs is reached.

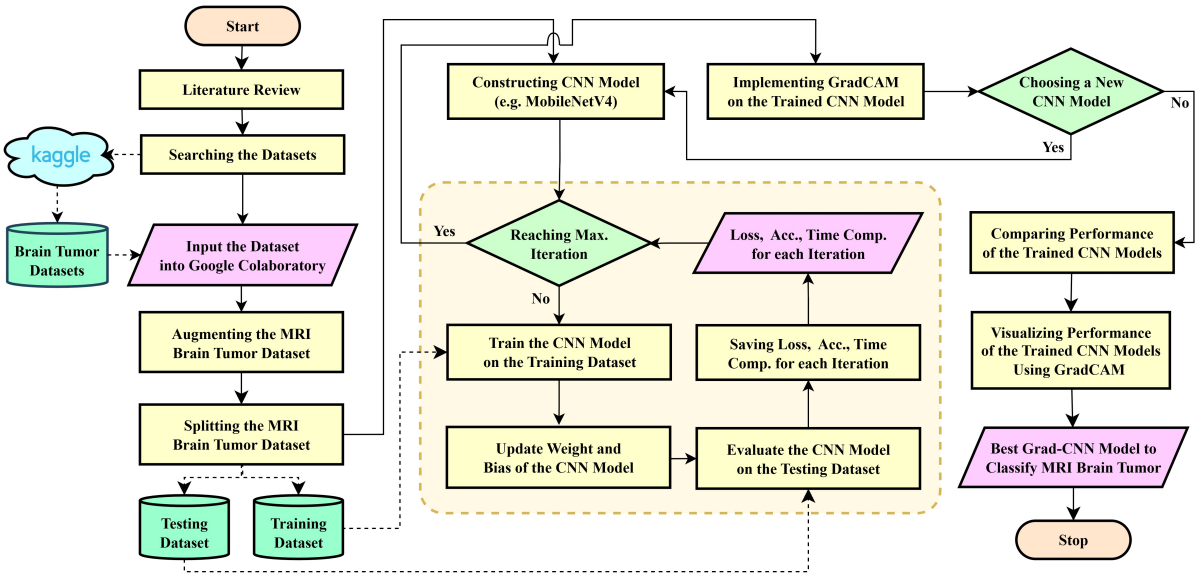



**Figure 1:** Research flow

Upon completion of model training, Grad-CAM is applied to the trained CNN to produce class-discriminative visual explanations, allowing for the interpretation of spatial attention and feature localization during classification. If necessary, alternative CNN architectures can be

selected and re-evaluated under the same experimental protocol to ensure optimal performance and explainability. Finally, the trained CNN models are compared based on their performance and interpretability, and the best-performing Grad-CAM enhanced CNN model is selected for classifying MRI-based brain tumor images. This CNN model is intended to support clinical decision-making by providing accurate classifications along with visual explanations. The overall process is illustrated in Figure 1, which outlines a comprehensive pipeline for the classification of brain tumors from MRI images using a CNN architecture integrated with Grad-CAM to improve the interpretability of the model.

### 2.3 Software Configuration

All experiments in this study were conducted using the Google Colaboratory environment to ensure reproducibility and scalability of deep learning workflows. The computational setup included a CPU-only configuration with a 4-core processor and 30 GB of RAM per editing session, which was primarily utilized for general preprocessing tasks and non-accelerated model operations. In addition, a GPU-accelerated configuration with dual NVIDIA Tesla T4 GPUs supported by 4 CPU cores and 29 GB of RAM, was employed to significantly reduce training time and enhance parallel computation efficiency for CNN-based models. This configuration enabled seamless integration of preprocessing, model training, and Grad-CAM visualization while ensuring computational efficiency and effective resource utilization for large-scale MRI data.

## 3 Results and Discussion

This section presents the experimental results and analysis of CNN architectures applied to brain tumor classification. The discussion is organized into four key subsections. Subsection 3.1 describes the integration of Grad-CAM into different CNN models to enhance explainability. Subsection 3.2 provides a comparative evaluation of model performance based on quantitative metrics such as accuracy, precision, and recall. Subsection 3.3 explores the visual interpretability of model predictions through Grad-CAM heatmaps. Finally, Subsection 3.4 analyzes Grad-CAM performance across all model architectures.

### 3.1 Integration of Grad-CAM into CNN Architectures

Gradient-weighted Class Activation Mapping (Grad-CAM) is a widely adopted post-hoc interpretability technique designed to produce class-discriminative visual explanations from CNNs. In this study, Grad-CAM was applied to four CNN architectures, including MobileNetV3, MobileNetV4, VGG19, and ResNet50 to enhance clinical interpretability in brain tumor classification.

For a given input image and a target class $c$, let $y^c$ denote the pre-softmax score for class $c$. Let $A^k \in \mathbb{R}^{u \times v}$ denote the $k^{\text{th}}$ feature map in the last convolutional layer of CNN, where $u$ and $v$ represent spatial dimensions and $k = 1, 2, ..., K$. The importance weight $\alpha_k^c$ for each feature map $A^k$ is computed as the global average of the gradient of $y^c$ with respect to $A^k$ as follows.

$$\alpha_k^c = \frac{1}{uv} \sum_{i=1}^{u} \sum_{j=1}^{v} \frac{\partial y^c}{\partial A_{ij}^k} \tag{1}$$

These weights reflect the contribution of each feature map to the target class prediction. The Grad-CAM heatmap $L^c_{\text{Grad-CAM}}$ is then generated as a weighted linear combination of the feature maps, followed by a ReLU activation to retain only the features with a positive influence:

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_{k=1}^{K} \alpha_k^c A^k\right) \tag{2}$$

This heatmap localizes the relevant regions in the input image that contribute to the class score $y^c$. By upsampling $L^c_{\text{Grad-CAM}}$ to the original input image size, clinicians can visually inspect which areas of the brain MRI image influenced the CNN's decision.

**Lemma 1.** *Grad-CAM is a generalization of Class Activation Mapping (CAM). Specifically, when the CNN model ends with a global average pooling (GAP) layer followed by a fully connected (FC) layer for classification, and the gradients $\partial y^c / \partial A^k_{ij}$ are constant across spatial locations, then Grad-CAM reduces to the original CAM formulation.*

*Proof.* In the CAM, the class score $y^c$ is computed as a weighted sum of GAP features as follows.

$$y^c = \sum_{k=1}^{K} w_k^c \cdot \text{GAP}(A^k), \quad \text{where} \quad \text{GAP}(A^k) = \frac{1}{uv} \sum_{i=1}^{u} \sum_{j=1}^{v} A^k_{ij} \tag{3}$$

The weights $w_k^c$ are learned directly in the final FC layer. However, $\alpha_k^c$ is computed dynamically using the gradient of $y^c$ with respect to $A^k$ in Grad-CAM. If the model ends with a GAP layer, the gradient $\partial y^c / \partial A^k_{ij}$ is constant for all $(i, j)$, and equal to $w_k^c$. Thus the Equation (4) is obtained.

$$\alpha_k^c = \frac{1}{uv} \sum_{i=1}^{u} \sum_{j=1}^{v} \frac{\partial y^c}{\partial A^k_{ij}} = \frac{1}{uv} \cdot uv \cdot w_k^c = w_k^c \tag{4}$$

Intuitively, $\alpha_k^c$ indicates the contribution of the $k^{\text{th}}$ feature map to the prediction of class $c$. Substituting Equation (4) into Equation (2) yields the following expression.

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left( \sum_{k=1}^{K} \alpha_k^c A^k \right) = \text{ReLU}\left( \sum_{k=1}^{K} w_k^c A^k \right) \tag{5}$$

□

Equation (5) shows that Grad-CAM is equivalent to CAM up to the application of the ReLU function. This confirms that Grad-CAM generalizes CAM when the underlying CNN architecture satisfies the CAM specific assumptions. The Grad-CAM method is architecture-agnostic and does not require any modifications or retraining of the original CNN models. In this study, feature maps were extracted from specific layers for each architecture: the final convolutional block (`block5_conv4`) in VGG19, the last residual block in ResNet50, and the final convolutional layers before classification in MobileNetV3 and MobileNetV4.

Figure 2 illustrates the overall Grad-CAM mechanism as applied in this work. The process begins with a forward pass through the CNN to compute class scores, followed by gradient computation with respect to the feature maps of a selected convolutional layer. These gradients are globally averaged to obtain importance weights, which are then used to compute a weighted sum of the feature maps. After applying the ReLU activation, the resulting Grad-CAM heatmap highlights spatial regions in the input image that most strongly contribute to the model's decision.

### 3.2 Comparative Performance of CNN Architectures

Comparative evaluation across five widely was adopted convolutional and transformer-based neural network architectures, including MobileNetV3, MobileNetV4, VGG19, ResNet50, and Vision Transformer (ViT). All models were initialized using transfer learning with pre-trained weights, ensuring that the networks benefited from robust feature representations learned from large-scale natural image datasets. This strategy enhances generalization, accelerates convergence, and provides a fair and reproducible basis for comparison across architectures when applied to the relatively smaller MRI dataset. The evaluation includes both qualitative and quantitative perspectives, integrating epoch-wise training loss and accuracy curves, as illustrated in Figure 3.

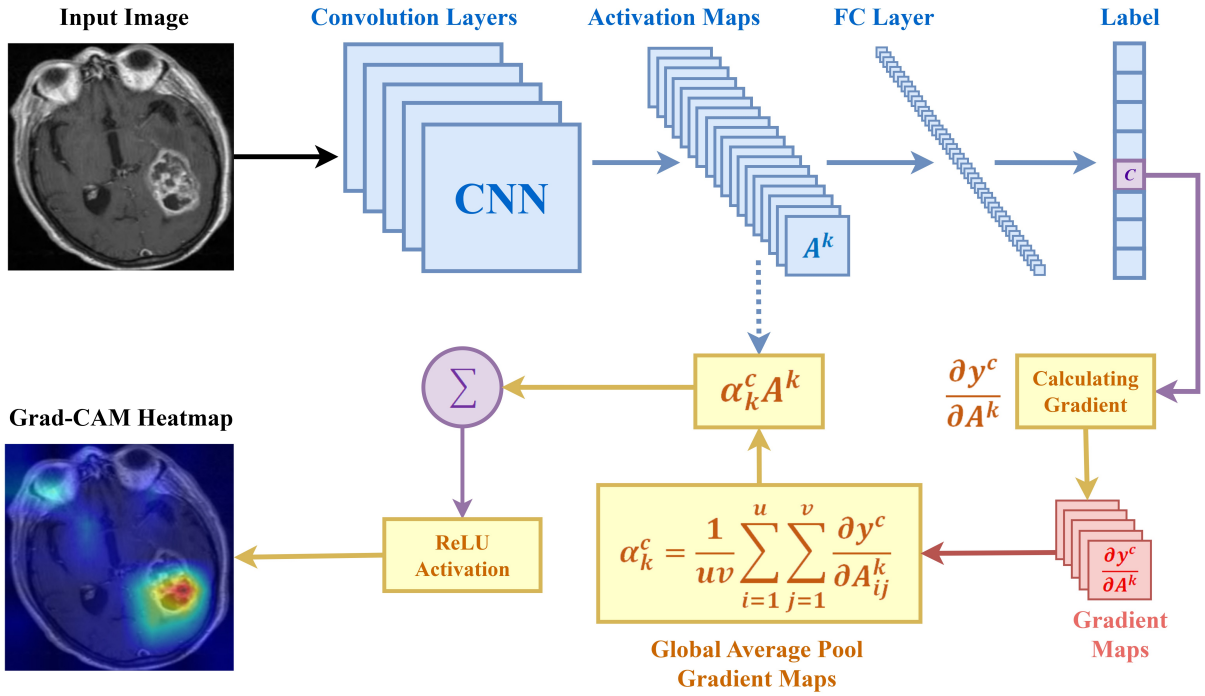



**Figure 2:** Schematic diagram of Grad-CAM integration into CNN models

As shown in Figure 3a, MobileNetV3-Small demonstrated a consistently low training loss throughout the learning process, converging below 1% after approximately 30 epochs. This trend aligns with its near-perfect training accuracy (exceeding 99.5% after epoch 20), as depicted in Figure 3b. Likewise, ResNet50 and ViT achieved competitive convergence with final training losses below 2% and stable accuracy levels ranging from 99.4% to 99.7%.

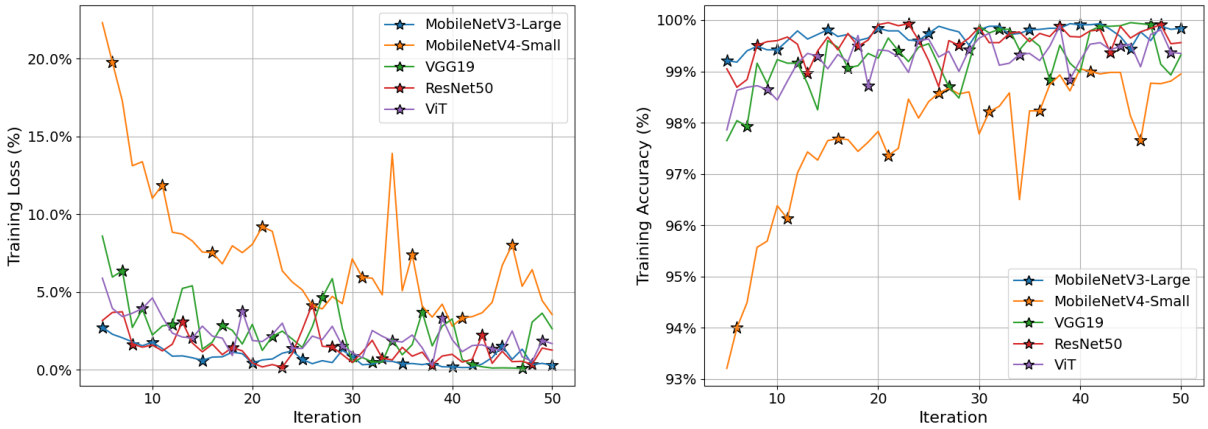
MobileNetV4, while exhibiting a more fluctuating training pattern, showed an initially sharp drop in loss and a steady increase in accuracy, ultimately reaching approximately 99.0% accuracy. Notably, it converged faster than most other models and maintained a moderate final loss range (3–5%), indicating adequate learning without signs of excessive overfitting. In contrast, VGG19 presented relatively slower convergence and more oscillatory behavior, stabilizing at a final accuracy of just above 99% with loss under 2.5%.

In addition to model accuracy and convergence trends, computational efficiency was also evaluated to assess suitability for real-world deployment. The total training time for each model was recorded as follows: ViT (7176.40 s), VGG19 (4037.52 s), MobileNetV3 (2419.43 s), MobileNetV3-Small (2377.95 s), and ResNet50 (1984.62 s). Among these, MobileNetV4 stood out with the shortest training time of just 1738.82 s, demonstrating its computational efficiency while retaining competitive performance. Its faster convergence and lower complexity make it advantageous for scenarios with limited hardware resources or real-time constraints, despite being marginally outperformed in final accuracy by heavier models like MobileNetV3 or ViT.

Furthermore, classification performance on the held-out test dataset was compared between the best-performing MobileNetV3 and the efficient MobileNetV4 models, as presented in Table 1. MobileNetV3 model achieved perfect accuracy across all tumor classes 99.62%, reflecting both convergence and generalization. Meanwhile, MobileNetV4 achieved a respectable 98.29% accuracy, though it displayed a lower recall of 0.89 for glioma cases, suggesting potential underdetection possibly due to intra-class heterogeneity or feature overlap with other tumor types.

In addition to the learning curves, the classification performance on the test dataset across two key models, MobileNetV4 and MobileNetV3, was also evaluated, as presented in Table 1. Both models demonstrated strong and consistent performance on the test dataset. MobileNetV3 achieved an overall accuracy of 99.62%, reflecting robust generalization across tumor classes.

In comparison, MobileNetV4 obtained an accuracy of 98.29%, still indicating competitive performance and confirming its effectiveness as a lightweight architecture.




**(a)** **(b)**

**Figure 3:** (a) Training loss and (b) training accuracy from epoch 5 to 50 across CNN architectures

These results indicate that while MobileNetV4 provides an efficient balance of accuracy and computational cost, MobileNetV3 demonstrates superior stability in class-wise generalization. Nevertheless, the close performance between the two models highlights that lightweight architectures can achieve nearly the same level of diagnostic reliability as their larger counterparts.

The high accuracy achieved by MobileNetV3 (99.62%) does not indicate overfitting, as there was no significant difference between the training and testing accuracies. This consistency suggests that the model was able to generalize well rather than simply memorizing the training data. Although additional validation, such as k-fold cross-validation or testing on external datasets, would further strengthen the evidence, the results confirm that the model maintains reliable performance without signs of overfitting.

**Table 1:** MobileNetV4 and MobileNetV3 classification performance per tumor class.

| Class | MobileNetV4 | | | MobileNetV3 | | | Support |
|---|---|---|---|---|---|---|---|
| | Precision | Recall | F1-score | Precision | Recall | F1-score | |
| Glioma | 0.98 | 0.99 | 0.99 | 0.98 | 0.99 | 0.99 | 249 / 266 |
| Meningioma | 0.98 | 0.98 | 0.99 | 0.99 | 0.99 | 0.99 | 240 / 239 |
| No Tumor | 0.99 | 0.99 | 0.98 | 0.99 | 0.99 | 0.99 | 296 / 291 |
| Pituitary | 0.99 | 0.99 | 0.99 | 0.99 | 0.98 | 0.98 | 268 / 257 |
| **Accuracy** | | 0.99 | | | 0.99 | | 1053 |
| **Macro Avg** | 0.99 | 0.99 | 0.99 | 0.98 | 0.99 | 0.99 | 1053 |
| **Weighted Avg** | 0.99 | 0.99 | 0.99 | 0.98 | 0.99 | 0.99 | 1053 |

### 3.3 Visual Interpretability through Grad-CAM Integration

Figures 4 and 5 illustrate the interpretability of Grad-CAM for two tumor classes. In Figure 4, the coronal MRI slice of a glioma case (left) is overlaid with a Grad-CAM heatmap (right), which highlights strong activation in the upper central mass, consistent with the known tumor location. The color intensities, predominantly in red and yellow, indicate that the CNN model concentrated its attention on the actual lesion area during classification.

Similarly, in Figure 5, Grad-CAM highlights tumor region in transverse MRI of meningioma. The heatmap reveals a strong correlation between the CNN's focus and the tumor boundaries, suggesting that the model successfully identifies class-discriminative anatomical structures.

Notably, minor activations in peripheral zones may reflect secondary spatial features learned during training. These results demonstrate the model's clinically relevant attention patterns, reinforcing Grad-CAM's role in providing transparent and trustworthy diagnostic support.

Figures 6 and 7 provide further insight into the model's interpretability performance for pituitary tumor and no tumor cases. In Figure 6, the sagittal MRI image displays moderate Grad-CAM activation near the sella turcica, which is anatomically appropriate for pituitary tumors. Some activation also appears in the upper frontal lobe, potentially due to attention spillover or artifacts, but the dominant focus remains consistent with clinical expectations.

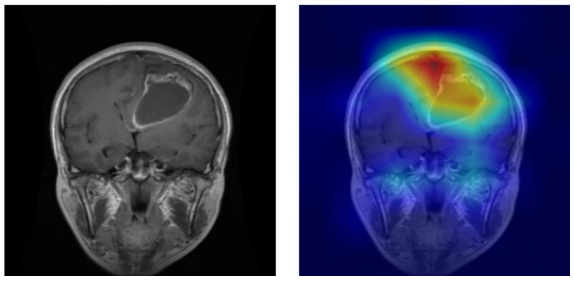


**Figure 4:** Grad-CAM for glioma tumor

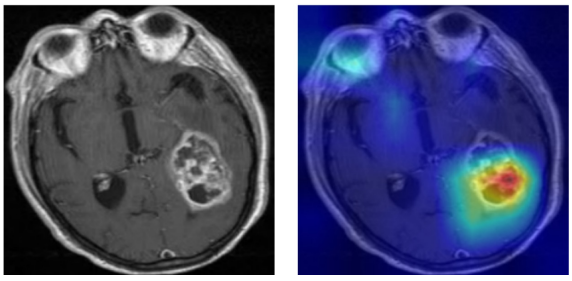


**Figure 5:** Grad-CAM for meningioma tumor

In contrast, Figure 7 presents the Grad-CAM result for a no tumor case, where the heatmap displays uniformly distributed across the entire brain. This suggests that the model considers all brain regions as relevant for verifying the absence of abnormalities. The model exhibits appropriate sensitivity to normal anatomical structures. Such behavior indicates high specificity and supports the model's reliability in distinguishing healthy scans from pathological ones.

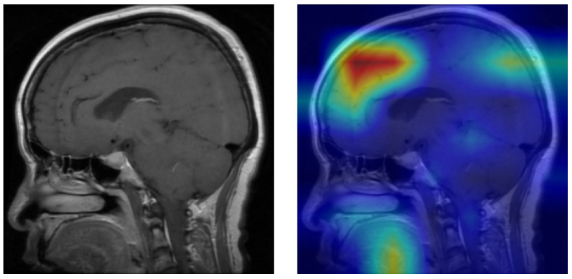


**Figure 6:** Grad-CAM for pituitary tumor

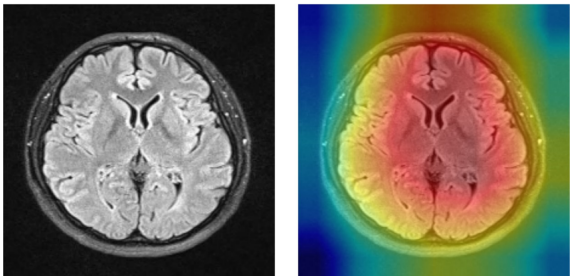
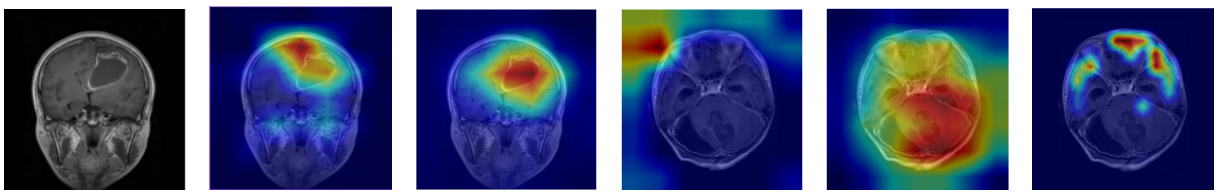


**Figure 7:** Grad-CAM for no tumor

### 3.4 Grad-CAM Performance on CNN Architectures

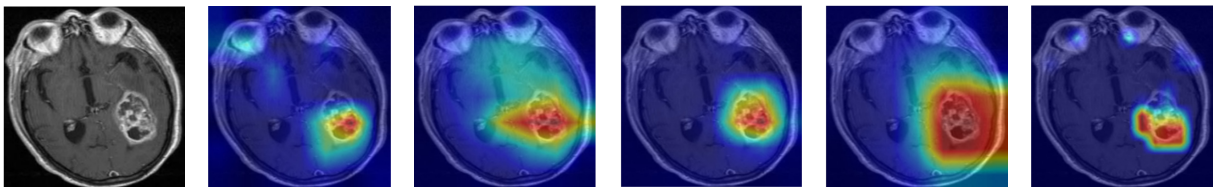
The Grad-CAM visualizations presented in Figure 8 offer a comparative insight into the interpretability of various architectures when applied to glioma tumor localization. Among the six models evaluated, MobileNetV3 and MobileNetV4 (Figures 8c and 8b) exhibit highly localized and concentrated attention around the tumor region, aligning closely with the ground truth observed in the original MRI (Figure 8a). In contrast, ResNet50 and VGG19 (Figures 8e and 8d) demonstrate broader activation regions, potentially indicating a lower specificity in feature attribution. Notably, the ViT model (Figure 8f) shows sparse and dispersed activations, which may reflect challenges in spatial localization due to its patch-based tokenization.



**(a)** Original **(b)** MobileNetV4 **(c)** MobileNetV3 **(d)** VGG19 **(e)** ResNet50 **(f)** ViT

**Figure 8:** Grad-CAM heatmaps for a glioma tumor from different models

These findings indicate that lightweight CNNs like MobileNetV3 and MobileNetV4 not only offer computational efficiency but also produce clinically interpretable heatmaps that sharply highlight the tumor mass. Therefore, in the context of explainable AI for glioma diagnosis, MobileNetV3 integrated with Grad-CAM offers a compelling balance between diagnostic accuracy and interpretability, making it a promising candidate for deployment in real-world clinical decision-support systems.

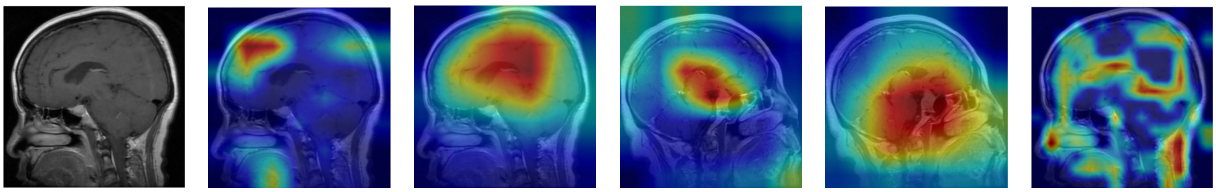


**(a)** Original **(b)** MobileNetV4 **(c)** MobileNetV3 **(d)** VGG19 **(e)** ResNet50 **(f)** ViT

**Figure 9:** Grad-CAM heatmaps for a meningioma tumor from different models

The Grad-CAM visualizations in Figure 9 illustrate the attention maps of various CNN architectures in detecting a meningioma tumor. All models show the capability to localize the tumor region, albeit with varying degrees of precision and interpretability. Notably, MobileNetV4, VGG19, and ViT (Figures 9b, 9d, and 9f) exhibit strong performance in highlighting the tumor area with high specificity. Their attention maps are well-centered and focused, effectively aligning with the tumor boundary observed in the original MRI (Figure 9a). In particular, ViT benefits from its global self-attention mechanism, enabling compact and well-defined localization, while VGG19 maintains consistent gradient transitions conducive to clinical interpretability.

In contrast, MobileNetV3 and ResNet50 (Figures 9c and 9e) also capture the tumor but with comparatively broader and more diffused activation zones, which may reduce spatial specificity despite adequate coverage. These findings suggest that architectures like MobileNetV4 and ViT offer a favorable balance between diagnostic accuracy and explainability in meningioma cases, particularly in scenarios that demand high precision for clinical decision-making. Consequently, these models stand out as viable candidates for deployment in real-world diagnostic support systems where both localization fidelity and interpretability are crucial.



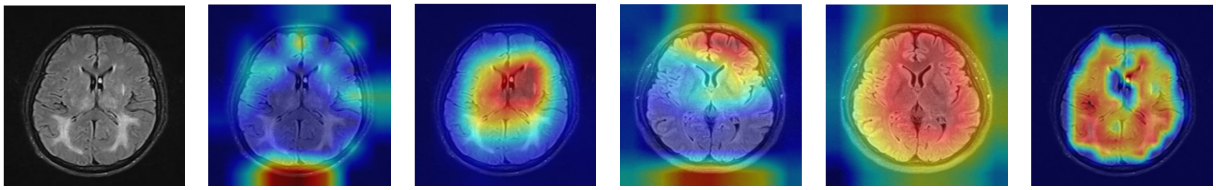
**(a)** Original **(b)** MobileNetV4 **(c)** MobileNetV3 **(d)** VGG19 **(e)** ResNet50 **(f)** ViT

**Figure 10:** Grad-CAM heatmaps for a pituitary tumor from different models

The Grad-CAM visualizations in Figure 10 illustrate how different CNN architectures localize pituitary tumors, which are typically small and centrally located. MobileNetV3 and VGG19 show the most focused and accurate attention around the tumor region, with clear and smooth heatmaps. MobileNetV4 localizes near the target but with slightly shifted focus, while ResNet50 covers the region broadly, indicating less spatial precision. ViT produces dispersed activations, suggesting difficulty in capturing small tumor structures. These results suggest that MobileNetV3 and VGG19 offer a good balance between localization accuracy and interpretability.

The Grad-CAM results in Figure 11 illustrate the response of different CNN architectures without the presence of a tumor. In such cases, a desirable model behavior is to produce a diffuse or uniformly distributed heatmap, indicating no specific focus due to the absence of abnormalities. MobileNetV3, ResNet50, and ViT exhibit such characteristics, with broad and even attention across the brain region. In contrast, MobileNetV4 and VGG19 display more

concentrated activations in localized areas, which may reflect false positive tendencies. These findings underscore the importance of verifying model interpretability not only on tumor cases but also on healthy controls to ensure robust and trustworthy deployment in clinical environments.



**(a)** Original **(b)** MobileNetV4 **(c)** MobileNetV3 **(d)** VGG19 **(e)** ResNet50 **(f)** ViT

**Figure 11:** Grad-CAM heatmaps for a no tumor from different models

The integration of Grad-CAM into MobileNet architectures provides substantial clinical value by combining model transparency with computational efficiency in automated brain tumor detection. Compared to conventional CNNs, which often function as "black boxes," MobileNet enhanced with Grad-CAM produces highly focused visual heatmaps that emphasize tumor-specific regions, thereby enabling clinicians to validate predictions against medically relevant features. The lightweight nature of MobileNet ensures faster training and inference without sacrificing diagnostic accuracy, making it particularly suitable for deployment in healthcare facilities with limited computational resources. This dual advantage of interpretability and efficiency not only strengthens clinical trust and acceptance of AI-assisted diagnostics but also facilitates error analysis and reduces the likelihood of false positives or negatives. Furthermore, MobileNet with Grad-CAM can support radiologists in delineating tumor boundaries, prioritizing critical regions for assessment, and improving diagnostic throughput in high-volume clinical settings. Ultimately, this synergy establishes MobileNet as a superior candidate for reliable, interpretable, and resource-efficient AI-based brain tumor detection in real-world medical practice.

### 3.5 Limitations of Grad-CAM Evaluation

Although the Grad-CAM visualizations demonstrated qualitatively meaningful attention patterns that aligned with tumor regions, no quantitative evaluation of the heatmaps was performed in this study. In particular, metrics such as Dice Similarity Coefficient (DSC), Intersection-over-Union (IoU), or localization accuracy against ground truth tumor masks were not employed. The primary reason for this limitation is that the publicly available Brain Tumor MRI Dataset used in our experiments does not provide pixel-level segmentation masks or region-level annotations required for such quantitative comparisons. As a result, we were unable to perform statistical testing or compute overlap-based interpretability metrics to validate the consistency of Grad-CAM outputs.

This absence of quantitative interpretability evaluation restricts the ability to rigorously compare Grad-CAM with alternative explanation methods on a statistical basis. Future research should therefore employ datasets with available tumor segmentation labels, enabling robust statistical testing and the application of quantitative metrics to strengthen the reliability and clinical trust of Grad-CAM–based visual explanations.

## 4 Conclusion

Brain tumor detection remains a highly challenging task due to the complex and heterogeneous morphology of tumors and the lack of interpretability in most deep learning models. While CNNs have demonstrated strong diagnostic accuracy, their "black-box" nature and computational demands limit clinical adoption. This study addressed this gap by integrating Gradient-weighted Class Activation Mapping (Grad-CAM) with lightweight CNN architectures, particularly MobileNetV3 and MobileNetV4, to provide interpretable, accurate, and computationally efficient

tumor classification from MRI scans.

The proposed framework successfully met its objectives by demonstrating that Grad-CAM can highlight clinically relevant tumor regions while maintaining diagnostic performance comparable to or exceeding heavier models such as ResNet50, VGG19, and ViT. Specifically, MobileNetV3 achieved the highest accuracy of 99.62% with an ROC accuracy of 99.96%, while MobileNetV4 provided the fastest training time (1738.82 seconds) with a competitive accuracy of 98.29%. These results confirm that lightweight CNNs can achieve a favorable balance between predictive performance, interpretability, and efficiency, supporting their potential for clinical deployment.

Despite these promising findings, several limitations should be acknowledged. First, the evaluation relied on a single publicly available MRI dataset, which may not fully capture variability across diverse clinical populations. Second, only Grad-CAM was explored as the interpretability method, without systematic comparison to alternative techniques such as SHAP or LIME. Third, quantitative evaluation of visual explanations (e.g., using Dice or IoU metrics) was not performed, and the models were limited to 2D slice-based MRI analysis. These limitations highlight the need for further validation before deployment in real-world clinical workflows.

Future research should therefore expand to multi-institutional and multi-modal datasets (e.g., T1, T2, and FLAIR sequences), integrate alternative or hybrid interpretability techniques, and explore volumetric (3D CNN) approaches with rigorous quantitative evaluation of heatmaps. Such directions will enhance both the robustness and clinical reliability of explainable AI systems for brain tumor detection.

## CRediT Authorship Contribution Statement

**Amalan Fadil Gaib, Anggito Karta Wijaya:** Conceptualization, Data Curation, Resources, Software, Writing-Original Draft. **Safrizal Ardana Ardiyansa:** Supervision, Validation, Formal Analysis, Investigation, Writing–Review Editing. **Eric Julianto:** Project Administration, Funding Acquisition, Resources. **I Gusti Ngurah Bagus Ferry Mahayudha, Ando Zamhariro Royan:** Visualization, Software, Writing–Review & Editing. **Ando Zamhariro Royan:** Writing–Review & Editing.

## Declaration of Generative AI and AI-assisted technologies

AI-assisted technologies were utilized during the preparation of this manuscript. Specifically, ChatGPT was employed to assist in writing refinement, code structuring, and enhancing the clarity of technical descriptions. Grammarly was used to perform grammar and language quality checks to ensure linguistic consistency and correctness throughout the manuscript.

## Declaration of Competing Interest

Authors confirm that there are no conflicts of interest associated with this publication, including employment, consultancies, stock ownership, honoraria, paid expert testimony, patent applications, or any other financial or non-financial interests relevant to the subject matter of the research.

## Funding and Acknowledgments

The authors express their sincere appreciation to Braincore Indonesia who provided support throughout the completion of this research. We acknowledge the contributors of the publicly available Brain Tumor MRI dataset on the Kaggle platform, which served as the source for our experimental analysis. Additionally, we thank to our colleagues and mentors for their valuable feedback during the design, implementation, and refinement phases of this research.

## Data and Code Availability

The Brain Tumor MRI dataset utilized in this study is publicly accessible via the Kaggle repository[2]. The source code supporting the findings of this study is available from the corresponding author upon reasonable request.

## References

[1] S. Rasheed, K. Rehman, and M. Akash, "An insight into the risk factors of brain tumors and their therapeutic interventions," *Biomed Pharmacother*, vol. 143, p. 112 119, 2021. DOI: 10.1016/j.biopha.2021.112119.

[2] I. Ilic and M. Ilic, "International patterns and trends in the brain cancer incidence and mortality: An observational study based on the global burden of disease," *Heliyon*, vol. 9, no. 7, e18222, 2023. DOI: 10.1016/j.heliyon.2023.e18222.

[3] Q. Ostrom, M. Price, C. Neff, *et al.*, "Cbtrus statistical report: Primary brain and other central nervous system tumors diagnosed in the united states in 2015–2019," *Neuro-Oncology*, vol. 24, no. Supplement$_5$, pp. v1–v95, 2022. DOI: 10.1093/neuonc/noac202.

[4] K. Miller, Q. Ostrom, C. Kruchko, *et al.*, "Brain and other central nervous system tumor statistics, 2021," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 5, pp. 381–406, 2021. DOI: 10.3322/caac.21693.

[5] I. Liguoro, C. Pilotto, F. Tuniz, M. Toniutti, P. Cogo, and T. Zilli, "Prospective analysis on possible changes of cognitive functions in children on follow-up for brain tumor," *Child's Nervous System*, vol. 41, no. 1, p. 97, 2025. DOI: 10.1007/s00381-025-06751-2.

[6] N. Sahrizan, H. Manan, H. Hamid, J. Abdullah, and N. Yahya, "Functional alteration in the brain due to tumour invasion in paediatric patients: A systematic review," *Cancers*, vol. 15, no. 7, p. 2168, 2023. DOI: 10.3390/cancers15072168.

[7] S. Mekler, S. Virtue-Griffiths, and K. Pike, "Self- and informant-reported cognitive concerns associated with primary brain tumour: Systematic review," *Supportive Care in Cancer*, vol. 33, no. 310, 2025. DOI: 10.1007/s00520-025-09345-5.

[8] K. Figuracion, W. Jung, and S. Martha, "Ischemic stroke risk among adult brain tumor survivors: Evidence to guide practice. the journal of neuroscience nursing," *Journal of the American Association of Neuroscience Nurses*, vol. 53, no. 5, pp. 202–207, 2021. DOI: 10.1097/JNN.0000000000000606.

[9] M. Ijaz, I. Hasan, B. Aslam, *et al.*, "Diagnostics of brain tumor in the early stage: Current status and future perspectives," *Biomaterials Science*, vol. 13, pp. 2580–2605, 2025. DOI: 10.1039/d4bm01503g.

[10] D. Crosby, S. Bhatia, K. M. Brindle, *et al.*, "Early detection of cancer," *Science*, vol. 375, no. 6586, eaay9040, 2022. DOI: 10.1126/science.aay9040.

[11] Y. Zhao, Y. Ding, V. Lau, *et al.*, "Whole-body magnetic resonance imaging at 0.05 tesla," *Science*, vol. 384, no. 6696, eadm7168, 2024. DOI: 10.1126/science.adm7168.

[12] M. Martucci, R. Russo, F. Schimperna, *et al.*, "Magnetic resonance imaging of primary adult brain tumors: State of the art and future perspectives," *Biomedicines*, vol. 11, no. 2, p. 364, 2025. DOI: 10.3390/biomedicines11020364.

[13] D. Veiga-Canuto, L. Cerdá-Alberich, C. Nebot, *et al.*, "Comparative multicentric evaluation of inter-observer variability in manual and automatic segmentation of neuroblastic tumors in magnetic resonance images," *Cancers*, vol. 14, no. 15, p. 3648, 2022. DOI: 10.3390/cancers14153648.

[2] https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset

[14] M. Rana and M. Bhushan, "Machine learning and deep learning approach for medical image analysis: Diagnosis to detection," *Multimedia Tools and Applications*, pp. 1–39, 2022. DOI: 10.1007/s11042-022-14305-w.

[15] S. Ardiyansa, N. Maharani, S. Anam, and E. Julianto, "Optimizing heart attack diagnosis using random forest with bat algorithm and greedy crossover technique," *BAREKENG: Jurnal Ilmu Matematika Dan Terapan*, vol. 18, no. 2, pp. 1053–1066, 2022. DOI: 10.30598/barekengvol18iss2pp1053-1066.

[16] P. Yang and B. Yang, "Development and validation of predictive models for diabetic retinopathy using machine learning," *PLOS One*, vol. 20, no. 2, e0318226, 2025. DOI: 10.1371/journal.pone.0318226.

[17] H. Chen, N. Wang, X. Du, K. Mei, Y. Zhou, and G. Cai, "Classification prediction of breast cancer based on machine learning," *Computational Intelligence and Neuroscience*, p. 6 530 719, 2023. DOI: 10.1155/2023/6530719.

[18] L. P. D. Jayanti, S. Anam, S. A. Ardiyansa, and N. C. Maharani, "Health insurance claim classification using support vector machine with velocity pausing particle swarm optimization," *CAUCHY: Jurnal Matematika Murni dan Aplikasi*, vol. 10, no. 2, pp. 698–710, 2025. DOI: 10.18860/cauchy.v10i2.31914.

[19] S. A. Ardiyansa, M. Muslikh, and A. R. Alghofari, "Binary hippopotamus algorithm with random forest for optimizing feature selection problem," *Numerical Algebra, Control and Optimization*, vol. 15, no. 4, pp. 1176–1191, 2025. DOI: 10.3934/naco.2025009.

[20] P. Jyothi and A. Singh, "Deep learning models and traditional automated techniques for brain tumor segmentation in mri: A review," *Artificial Intelligence Review*, vol. 56, pp. 2923–2969, 2022. DOI: 10.1007/s10462-022-10245-x.

[21] J. Herr, R. Stoyanova, and E. A. Mellon, "Convolutional neural networks for glioma segmentation and prognosis: A systematic review," *Critical Reviews in Oncogenesis*, vol. 29, no. 3, pp. 33–65, 2024. DOI: 10.1615/critrevoncog.2023050852.

[22] C. M. L. Zegers, J. Posch, A. Traverso, *et al.*, "Current applications of deep-learning in neuro-oncological mri," *Physica Medica*, vol. 83, pp. 161–173, 2021. DOI: 10.1016/j.ejmp.2021.03.003.

[23] T. Soomro, L. Zheng, A. Afifi, *et al.*, "Image segmentation for mr brain tumor detection using machine learning: A review," *IEEE Reviews in Biomedical Engineering*, vol. 16, no. 2, pp. 70–90, 2022. DOI: 10.1109/RBME.2022.3185292.

[24] D. Reyes and J. Sanchez, "Performance of convolutional neural networks for the classification of brain tumors using magnetic resonance imaging," *Heliyon*, vol. 10, no. 3, e25468, 2024. DOI: 10.1016/j.heliyon.2024.e25468.

[25] Y. Xiao, Y. Guo, Q. Pang, X. Yang, Z. Zhao, and X. Yin, "Star-detr: A lightweight real-time detection transformer for space targets in optical sensor systems," *Sensors*, vol. 25, no. 4, p. 1146, 2025. DOI: 10.3390/s25041146.

[26] S. K. Mathivanan, S. Sonaimuthu, S. Murugesan, H. Rajadurai, B. D. Shivahare, and M. A. Shah, "Employing deep learning and transfer learning for accurate brain tumor detection," *Scientific Reports*, vol. 14, no. 1, p. 7232, 2024. DOI: 10.1038/s41598-024-57970-7.

[27] R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *IEEE International Conference on Computer Vision*, pp. 618–626, 2016. DOI: 10.48550/arXiv.1610.02391.