# CNN-LCA: A Lightweight Attention-Based Network with Confidence-Guided Grad-CAM for Explainable Brain Tumor Classification

**Capstone Project**

**Course Code: ICE - 3210**

**B.Sc. (Engineering) Examination - 2024**

*A project paper submitted to the Department of Information and Communication Engineering, Pabna University of Science and Technology in partial fulfillment of the requirements for the degree of Bachelor of Science in Engineering in Information and Communication Engineering*


Prepared by:

**MD. RIFAT HOSSEN**

Roll No: 220614

Reg. No: 1065461

Session: 2021-2022

---

**Department of Information and Communication Engineering**

**Pabna University of Science and Technology**

**Pabna-6600, Bangladesh**


**July 2026**

# Certificate of the Supervisor

It is certified that the research work incorporated in this project entitled **"Confidence with Confusion: A Lightweight CNN-LCA Network with Confidence-Guided Grad-CAM for Brain"** is the original work carried out by **Md. Rifat Hossen** under my supervision, and it fulfills the conditions laid out in the Pabna University of Science and Technology Ordinances. The research work included in this project forms a distinct contribution to knowledge. The project contains work worthy of consideration for the award of the degree of Bachelor of Science in Engineering in Information and Communication Engineering (ICE).

..........................................

**Dr. Md. Sarwar Hosain**
Professor
Department of Information and Communication Engineering
Pabna University of Science and Technology, Pabna-6600, Bangladesh.

# Declaration

This is to certify that the project entitled **"Confidence with Confusion: A Lightweight CNN-LCA Network with Confidence-Guided Grad-CAM for Brain"**, has been carried out by me under the course entitled **"Capstone Project (ICE-3210)"**. I also declare that this work has not been submitted elsewhere in part or full for the requirement of any degree or for any other purpose, except for academic publication.

............................................

**MD. RIFAT HOSSEN**
Roll No: 220614
Reg. No: 1065461
Department of Information and Communication Engineering
Pabna University of Science and Technology, Pabna-6600, Bangladesh.

# Dedicated to...

***My parents***

***and***

***My Teachers, whose guidance and support shaped my journey.***

# Acknowledgements

Firstly, I am sincerely thankful to **Pabna University of Science and Technology** for providing a supportive and intellectually stimulating environment that fostered my academic and personal growth throughout the Bachelor's program.

It is a great honor and privilege to have conducted this project as a part of the **Bachelor of Science in Information and Communication Engineering**, under the **Faculty of Engineering and Technology**, at **Pabna University of Science and Technology**.

I want to extend my deepest gratitude to my respected supervisor for his constant guidance, encouragement, and insightful advice throughout the course of this work. His unwavering support and belief in my potential have been instrumental in completing this research successfully.

Finally, I am genuinely thankful to everyone who supported me throughout this journey: my teachers, friends, and family. Your motivation, assistance, and kind words have played a significant role, and I am truly grateful for all the encouragement I have received.

The Author

# Abstract

Brain tumors remain among the most lethal diseases, profoundly affecting patients' physical, mental, and emotional well-being. Early and accurate detection is critical for effective treatment and improved survival outcomes. Deep learning methods combined with Grad-CAM have shown promise in MRI-based tumor classification, yet challenges persist in low- and middle-income countries due to high healthcare costs and limited radiological expertise. Traditional diagnostic approaches suffer from subjectivity and low interpretability, while many deep models operate as opaque "black boxes" that undermine clinical trust, and existing attention mechanisms like CBAM and ECA introduce computational overhead that hinders deployment in resource-limited settings.

To address these issues, we propose a novel, lightweight CNN architecture incorporating a Lite Convolutional Attention (LCA) mechanism and adaptive explainable AI techniques. The LCA module performs channel-wise feature recalibration via grouped 1-D convolutions with adaptively determined kernel sizes, substantially reducing parameter overhead while preserving cross-channel dependency modeling. Additionally, the model leverages softmax-derived prediction confidence to generate confidence-aware Grad-CAM visualizations, stratified into High, Medium, and Low reliability zones, enabling adaptive interpretation and preventing misleading explanations for uncertain predictions.

The model was trained on a Kaggle-hosted dataset of 7,200 brain MRI images across four classes, achieving 98.40% accuracy and a macro F1-score of 98.35%. Ablation studies confirmed that LCA improves recall and F1 over the non-attention baseline. External validation on a melanoma skin cancer dataset yielded 93.70% accuracy, demonstrating generalizability beyond neuro-oncology. Qualitative assessment employed confidence-stratified Grad-CAM visualizations, while quantitative evaluation included precision, recall, F1-score, sensitivity, specificity, ROC curves, and confusion matrix analysis. Despite its lightweight design, the model maintains strong performance and computational efficiency, making it suitable for real clinical deployment. This confidence-aware framework enables real-time inference in edge-based medical imaging and federated learning environments, supporting accurate, interpretable brain tumor detection and enhanced clinical decision-making.

# Contents

# List of Figures

# List of Tables

Chapter 1

# Introduction

## 1.1 Reason of Tumor

A critical disease among the millions of diseases is Brain Tumors (BTs), which develop inside the brain due to uneven growth of the skull or the brain tissue surrounding it [1]. These tumors apply pressure on the brain as they grow erratically throughout the brain. Under pressure, it induces various brain disorders that affect the body, such as dizziness, headaches, fainting attacks, paralysis, etc which are among the most serious neurological disorders. They are approximately 70% tumor cells, designed to metastasize only within the brain. Gliomas are malignant tumors that originate from glial cells, which support the brain and spinal cord. In contrast, pituitary tumors develop in the pituitary gland, impacting hormonal balance and causing symptoms that vary widely depending on the hormone affected [2].

## 1.2 Background study of Tumor

Over 250,000 people worldwide have died from brain and central nervous system tumours, which account for around 308,000 new cases annually [3]. Death among children by central nervous system tumors (27%), one third of which are benign or borderline malignant brain tumors. Cancer types and their distribution differ in adolescents, among whom the most common cancer is central nervous system tumors (22%), two thirds of which are benign or borderline malignant brain tumors [4].

## 1.3 Significance of MRI

Various imaging techniques are used to localize different types of brain tumors, each with a specific objective. The widely used imaging techniques consist of positron emission tomography (PET), radiography, ultrasonographic (US) etc. MRI (magnetic resonance imaging) appears to be among the techniques that have shown high efficiency in both tumor detection and further treatment of cancer patients. MRI is the main diagnostic tool for brain cancer classification because of its ability to provide high-resolution images and detailed information about the tumor [2]. MRI has emerged as a primary modality for brain tumor characterization owing to its superior soft-tissue contrast. The precise analysis of MRI data now serves as a cornerstone for brain tumor classification [5].

## 1.4 Significant Challenge in Diagnosis

Brain tumor detection poses a significant challenge in the field of medical diagnosis and treatment, necessitating accurate and efficient detection methods for timely intervention. During clinical practice, medical images are traditionally analyzed qualitatively by skilled medical professionals [1]. Accurate and automated classification of brain tumors from Magnetic Resonance Imaging (MRI) images presents a significant challenge in medical image analysis. The complex anatomical structure of the brain and the considerable inter-subject variability render manual diagnosis time-consuming and prone to subjectivity [5].

## 1.5 Problem Statement

Accurate and timely intervention is crucial for the diagnosis of tumours and improving clinical outcomes [6]. In real-world clinical settings, healthy radiographic images significantly outnumber those with fractures. Standard Deep Learning models trained on such datasets tend to be biased toward the majority class (Normal), leading to a dangerously low Recall (Sensitivity). In a medical context, missing a fracture (False Negative) is far more costly in a medical context than a false alarm [7]. Although mortality data cover the complete Native American population, racial misclassification is mitigated using factors developed by the NCHS for all cancers combined and may overestimate or underestimate rates for individual [4]. While deep learning algorithms have substantial limitations in data requirements, generalization, computational costs, noise sensitivity, and boundary accuracy, overcoming these limitations often requires additional approaches such as transfer learning, data augmentation, loss function modification, or hybrid architectures. Deep learning algorithms, often called ”black boxes,” make it difficult to understand the reasoning behind a particular prediction. This interpretability problem may jeopardize the model's legitimacy in medical settings and affect regulatory approvals. The primary limitations of this study include a small, imbalanced dataset, reliance on extensive augmentation, lack of external validation, and potential sensitivity to training configurations. These factors limit the generalizability of the findings and preclude claims of immediate clinical deployment [8]. As new architectures and components are introduced, research seems to focus on improving the average accuracy and speed of networks. Yet, far less attention has been given to determining when predictions can be trusted and when they cannot. In other words, neural networks are good at providing answers that are correct most (but not all) of the time, but not good at telling when such answers can be trusted, or when, instead, they amount to little more than an educated guess.

However, in some cases, viewing conditions are much poorer. A good example is pedestrian detection at night (Figure 1), in which poor, unusual, and rapidly changing illumination, shadows, reduced color sensitivity, and the tendency of colors to fuse in the background make reliable detection exponentially harder mismatch between confidence and correctness is what we call calibration [9]. A calibrated model that claims to be 90% certain should be wrong about one out of ten times, not half the time. In multiclass problems, they are often applied per class or in a one-vs-rest fashion, and can yield excellent calibration when the label space is small and well covered by the data [10]. The increasing integration of AI into healthcare highlights the critical need for interpretability, ensuring not only the “what” but also the “why” behind algorithmic predictions is clear. Current solutions often involve post hoc interpretability methods such as Gradient- Weighted Class Activation Mapping (Grad-CAM). While these techniques have advanced our understanding of AI predictions, they exhibit critical limitations. Grad-CAM provides broad, generalized visual explanations but lacks [11]. Cancer has been a monumental global health challenge, with a significant and disproportionate impact in low-and middle-income countries (LMICs). According to recent estimates, there were approximately 20 million new cancer cases and 9.7 million deaths in 2022 alone (2). Central Nervous System (CNS) tumours are among the most lethal malignancies, causing substantial morbidity and mortality [12].

Recent advances in deep learning have introduced cutting-edge technologies such as transfer learning models (e.g., ResNet, DenseNet, VGG) and hybrid architectures that combine CNNs with SVMs or attention mechanisms.While these methods achieve competitive accuracy, they often require high computational resources, making them less suitable for real-time clinical use. To address this limitation, we propose an optimized lightweight CNN model that leverages modern training strategies (Adam optimizer, cross-entropy loss) to balance high accuracy with computational efficiency. This positions our work within the scope of cutting-edge research, while addressing gaps in clinical applicability.

## 1.6 Limitations in Other Methods

High computational depth, diffused and less localized distribution of 'attention ' on GradCAM [7]. Although the CNN achieved high accuracy, interpretability of deep learning models remains a challenge, which may hinder clinical acceptance. Lastly, the computational requirements for training the model could limit its deployment in resource-constrained healthcare settings, which rely heavily on high-quality labeled MRI data and may be sensitive to class imbalance [13]. Conventional biopsy techniques are painful, time-consuming, and prone to sampling inaccuracies [14]. Similarly, using X-rays on the skull can lead to an increase in the risk of cancer due to the radiation [15]. The utilization of Gradient-weighted Class Activation Mapping adds an interpretive layer to model predictions; however, it generates heatmaps for all predictions regardless of confidence, which may lead to misleading explanations when the model's prediction is uncertain.

## 1.7 The novelty of the proposed work

In this paper, our model adopts an adaptive explainability approach. a method for training deep neural networks with a novel loss function that explicitly regularizes class probabilities to produce more reliable confidence estimates based on ordinal ranking according to prediction confidence [10]. After classification, the model evaluates the confidence score obtained from the Softmax Function. If the confidence score exceeds a predefined threshold, the model then applies Gradient-weighted Class Activation Mapping to generate a heatmap highlighting the most influential regions in the MRI image. Unlike conventional Grad-CAM approaches that generate explanations for every prediction, our model introduces a confidence-based mechanism in which visualization is triggered only when the prediction confidence surpasses the defined threshold. If the confidence score does not exceed the predefined threshold, the model does not generate Grad-CAM visualizations, indicating that the prediction is uncertain and preventing potentially misleading explanations.

The proposed method is easy to implement and can be applied to the existing architectures without any modification. These techniques can recover probability estimates more advanced and better reflect what a model has learned. For structured tasks with defined answer sets, these approaches allow users to extract calibrated probabilities they can trust. A clinician deciding whether to act on a model's suggestion, or a researcher weighing its outputs against other evidence, benefits greatly from knowing when the system is uncertain [10]. Also, it has nearly the same computational cost as training conventional deep classifiers and produces reliable predictions in a single inference [16]. This architectural simplicity with fewer parameters leads to an even greater reduction in memory and training time, especially in edge-based medical imaging. Lightweight, the architecture will be well-suited for real-time medical use. The LCA attention mechanism is much lighter than other newly developed mechanisms, such as CBAM and ECA [17, 18]. Grad-CAM highlights the regions most critical for the model's decision-making process. By comparing these interpretability techniques, the analysis highlights their effectiveness in identifying critical diagnostic features and their potential impact on clinical decision-making.

## 1.8 Classification of Brain Tumors

The World Health Organization classifies brain tumors into primary and secondary types, graded from I (benign) to IV (malignant). This study focuses on four clinically significant MRI classes:

**Glioma:** Arising from glial cells, gliomas are among the most common and aggressive primary brain tumors. On T1-weighted contrast-enhanced MRI, they appear as irregular, infiltrative lesions with heterogeneous enhancement, making automated segmentation challenging.

**Meningioma:** Originating from the meninges, meningiomas are typically benign, well-circumscribed extra-axial masses with homogeneous enhancement. Their proximity to critical structures necessitates accurate localization.

**Pituitary Tumor:** Developing in the pituitary gland, these tumors can disrupt hormonal regulation and cause visual impairment. They usually appear as small, well-defined masses in the sellar region.

**No Tumor:** Healthy scans without pathological abnormalities serve as the negative class, where correct identification prevents unnecessary patient anxiety and delayed treatment.

Table 1.1 summarizes the radiological characteristics of these classes, which are implicitly learned by the proposed CNN-LCA model through its attention mechanism.

Table 1.1: Radiological characteristics of the four MRI classes used in this study

| **Class** | **Origin** | **Typical MRI Appearance** | **Malignancy Potential** |
|---|---|---|---|
| Glioma | Glial cells | Irregular, infiltrative, heterogeneous enhancement | Low–High (Grade I–IV) |
| Meningioma | Meninges | Well-circumscribed, homogeneous enhancement | Mostly benign |
| Pituitary Tumor | Pituitary gland | Small, well-defined sellar/suprasellar mass | Mostly benign |
| No Tumor | – | Normal brain parenchyma | – |

## 1.9 Research Objectives

The primary objectives of this thesis are as follows:

1. To design a lightweight CNN-LCA architecture with efficient channel-wise attention for accurate brain tumor classification on edge devices;

2. To develop a Lite Convolutional Attention (LCA) module that performs efficient channel-wise feature recalibration with significantly lower computational complexity and fewer trainable parameters than existing attention mechanisms such as CBAM and ECA.

3. To introduce a confidence-aware Grad-CAM framework that generates visual explanations only when the softmax prediction confidence exceeds a predefined clinical threshold, categorizing predictions into *High*, *Medium*, and *Low* reliability levels.

4. develop a confidence-aware Grad-CAM framework with stratified reliability levels for trustworthy interpretability; and to comprehensively evaluate the model against state-of-the-art methods while validating its generalization on an external melanoma dataset

## 1.10 Scope and Contributions

The scope of this research is limited to the classification of two-dimensional (2-D) axial and coronal brain MRI slices into four tumor categories. More advanced tasks, such as three-dimensional (3-D) tumor segmentation and multi-modal MRI fusion (e.g., integrating T1-, T2-, and FLAIR-weighted sequences), are beyond the scope of this study and are identified as potential directions for future research in Chapter 5.

The major contributions of this thesis are summarized as follows:

1. A novel Lite Convolutional Attention (LCA) module combining Global Average Pooling with grouped 1-D convolutions and adaptive kernel size for efficient channel-wise feature recalibration, significantly reducing parameter overhead while preserving cross-channel dependencies.

2. A confidence-guided explainability framework integrating softmax scores with Grad-CAM, generating visual explanations only for sufficiently confident predictions to prevent misleading saliency maps.

3. Comprehensive experimental evaluation through ablation studies, cross-domain validation on melanoma dataset, and comparative analysis against six state-of-the-art architectures.

4. Demonstration that the proposed CNN-LCA achieves 98.40% accuracy with substantially lower computational complexity than conventional approaches, enabling deployment in resource-constrained clinical environments.

## 1.11 Organization of the Thesis

The remainder of this thesis is organized as follows. Chapter 2 reviews related work on CNN-based brain tumor classification, lightweight attention mechanisms, and XAI techniques, highlighting limitations and positioning our framework. Chapter 3 details the methodology, covering dataset, preprocessing, CNN-LCA architecture, LCA formulation, confidence-aware classification, Grad-CAM explainability, and experimental setup. Chapter 4 presents results and evaluation using standard metrics, ablation studies, cross-domain validation on melanoma dataset, and comparative analysis with six state-of-the-art models. Chapter 5 concludes with major findings, contributions, limitations, and future directions including multi-modal MRI fusion, 3-D volumetric analysis, and edge device deployment.

Chapter 2

# Literature Review

Recent studies show that convolutional neural networks (CNNs), when integrated with attention mechanisms and explainable artificial intelligence (XAI), can significantly improve image classification accuracy and interpretability [7]. MobileNetV2 model significantly outperformed the baseline, achieving a remarkable 99% accuracy, a 1.00 Recall for fracture detection, and a Macro F1-score of 0.95 with efficiency-optimized.There is an ongoing debate regarding whether ”deeper” networks (like ResNet50) are superior for medical tasks, or if more ”efficient” architectures (like MobileNetV2) can generalize better under data constraints [1]. ResNet50 augmented with Grad-CAM, for brain tumor detection in MRI images. Achieving a testing accuracy of 98.52% [7]. Deep feature extraction: ResNet50 was selected for its ability to train very deep networks through ”skip connections” or ”identity shortcuts.” These connections allow the gradients to flow through the network more effectively, mitigating the vanishing gradient problem. Despite these promising results, several limitations [7] exist in the activation regions for the ResNet50 model using Grad-CAM. Although the model nominally identifies certain features, the heat map reveals a diffused and less localized distribution of ’attention’. The activation often spreads to non-clinical background areas or irrelevant anatomical structures. This lack of spatial precision, combined with the erratic fluctuations observed in its accuracy and loss curves, explains its tendency to overfit the majority class.

Attention mechanisms have been extensively studied to boost classification accuracy. Similarly, An et al. (2021) developed a multiscale CNN with a visual attention mechanism that attained 99.89% accuracy, demonstrating that attention-guided feature learning can significantly improve classification results [19]. Zhou et al. proposed LP-CANet, which combines positional and channel attention mechanisms, reaching 83.15%

accuracy on high-resolution histopathological datasets and surpassing several standard CNN models [15]. Additionally, another study presented HCCANet, incorporating an attention mechanism for colorectal cancer classification, with an overall accuracy of 87.3%, further illustrating improvements in feature learning and classification performance [20]. Channel attention mechanisms have been widely used to enhance feature representation in convolutional neural networks. The channel-wise architecture generalises extremely effectively across different datasets. Mir et al. [5] present an advanced methodology for brain tumour classification using magnetic resonance imaging (MRI), combining the EfficientNet-B0 architecture with Convolutional Block Attention Modules (CBAM). It attains an outstanding accuracy of 99%, with weighted F1 and macro scores exceeding 98%. Although the proposed framework demonstrates strong performance, it has several limitations. The reliance on transfer learning from natural image datasets may limit domain adaptation for medical images. The relatively small dataset may affect model generalization.

Moreover, CBAM uses fully connected layers and a 7×7 spatial convolution, resulting in slightly higher complexity and more parameters. Ajay et al. [6] introduced BoneVisionNet, a hybrid deep learning model that combines ECA attention mechanism-based feature extraction with Dense-169-customized components, achieving an accuracy of 84.35% and providing interpretable results via XAI techniques. Although ECA is considered a lightweight mechanism, using Dense-169 made the architecture heavier. Squeeze-and-Excitation" (SE) block adjusts channel features to improve CNN performance with little extra cost [21]. ECA is nothing but a simplified architecture of SE that replaces fully connected layers with 1D convolution. While a unique approach presented by Prem et al. demonstrates that efficient channel reweighting can improve CNN performance while maintaining low computational complexity [18]. Lite Channel Attention (LCA) achieves competitive accuracy with minimal parameters; however, its latency on MobileNetV2 exceeds ECA's. This highlights the importance of hardware-aware design: group convolutions may incur kernel launch overhead on certain GPU architectures. Experimental results show that LCA achieves competitive accuracy (94.68% on ResNet-18, 93.10% on MobileNetV2) while matching ECA in parameter efficiency and maintaining favorable inference latency. LCA builds upon ECA's design philosophy while incorporating group convolutions for further parameter reduction [21]. As shown, LCA is much lighter than other newly adapted mechanisms such as CBAM and ECA.

Additionally, Skliarov et al. (2025) compared various XAI techniques and showed that gradient-based methods offered greater fidelity and stability in CNN-driven image classification across benchmark datasets [8], comparing heavy and lightweight architectures, implementing a sensitivity- focused strategy on imbalanced data and utilizing Grad-CAM visualizations to ensure clinical trust [7].The proposed multi-model deep learning framework, based on a modified CBAMVGG19 with selective fine-tuning, provides significant improvements in leukemia classification, outperforming complex models such as DenseNet121, InceptionV3, and MobileNetV2. Syed et al. explored the VGG19 architecture and selected it for its homogeneous convolutional design, which enables controlled analysis of attention mechanisms without introducing confounding architectural variations on the blood cancer dataset. In contrast to modern residual or transformer-based architectures, VGG19 provides a stable baseline that is less prone to overfitting in small-scale medical imaging scenarios [8].

A custom lightweight CNN backbone extracts local features from frequency-domain representations [17]. SSPANet-VGG16 achieves a 4.5% improvement in classification accuracy over the standard VGG16 [22]. Similarly, demonstrates a 7.78% accuracy gain when SSPANet is combined with ResNet50. These results confirm that SSPANet enhances the representational capacity of CNNs by introducing more precise and adaptive attention mechanisms. This enables the network to focus more on tumor-relevant regions, particularly in challenging cases such as gliomas, where accurate, localized feature discrimination is essential. It motivates exploring more customized VGG-style models [17]. The primary contribution of the proposed model focuses on the overall performance with a classification accuracy of 95.88%,which is higher than the benchmark models of MobileNetV2 (93.1%) and MobileNetV3Small(94.80%), while preserving fewer trainable parameters and a smaller memory footprint, and Grad-CAM visualizations focused on explaining the model's predictions and attending to tumor-relevant regions. Moreover,Deploying the model on a Raspberry Pi highlights the potential for real-time, point-of-care, edge-based medical imaging [5].

Abu et al. [13] introduced a multi-class CNN architecture with several modifications over conventional deep CNNs.The three transfer learning models, namely ResNet50, DenseNet121, and VGG19, obtained accuracies of 93.2%, 89.1%, and 75.0%, respectively. Category-wise accuracy of the proposed CNN model as (a) No Tumor: 98% (b) Pituitary Tumor: 96% (c)Meningioma Tumor: 89% (d) Glioma Tumor: 86%.Unlike standard models such as VGG-16 or ResNet, our framework employs a reduced number of convolutional layers optimized for lightweight computation, an adaptive kernel configuration that balances feature extraction and model complexity, and a hybrid activation and normalization pipeline, along with four Conv Blocks, designed to prevent overfitting on limited MRI datasets. Similarly, Sieradzki et al. (2024) assessed several pretrained CNN models for industrial image classification, finding that VGG19 achieved 67.03%

accuracy and outperformed other architectures tested [11]. Emphasizing interpretability, Ennab et al. (2025) introduced the Pixel-Level Interpretability (PLI) model based on VGG19, which scored 0.82 on interpretability—far exceeding baseline models—while still delivering robust classification performance [23]. Our model architecture is inspired by Yang et al. [24], which integrates deeper convolutional layers, batch normalization, and dropout to improve hierarchical feature extraction and mitigate overfitting. To build a more cost-friendly lightweight model on a real-time dataset.

Neumann et al. [9] introduced a modification of the standard softmax layer in which a probabilistic confidence score is explicitly premultiplied by the incoming activations to modulate confidence [16]. The quality of confidence estimates associated with a model's predictions can be assessed from two perspectives: confidence calibration and ordinal ranking based on confidence values. Confidence calibration is the problem of predicting probability estimates that reflect the true correctness likelihood. Thus, a well-calibrated classifier outputs predictive probabilities that can be directly interpreted as predictions' confidence level. It is known that modern neural networks generate miscalibrated outputs despite their high accuracy.Prior studies are summarized in Table 2.1, highlighting methods, performance Metrics, and Interpretability Limitations.

Table 2.1: Comparative analysis of state-of-the-art CNN architectures in recent literature

| Ref. Paper | Year | Method | Accuracy | Limitation |
|---|---|---|---|---|
| Guluwadi et al. [1] | 2024 | ResNet50, Grad-CAM | 98.52% | High computational depth; diffused and less localized distribution of attention in Grad-CAM. |
| Ajay et al. [6] | 2025 | DenseNet-169, ABMLFE-Net, Efficient Channel Attention, Grad-CAM, LIME | 84.35% | High computational depth. |
| Mir et al. [5] | 2025 | EfficientNetB0 + CBAM | 99% | Small datasets may affect model generalization. |
| Abu Alkebash et al. [13] | 2026 | ResNet50, DenseNet121, VGG19 (Transfer Learning) | 98.5% | Interpretability of deep learning models remains challenging, and computational requirements may limit deployment in resource-constrained environments. |
| Salman et al. [7] | 2026 | MobileNetV2, ResNet50 | 99% | Overfitting and model bias issues. |
| Kanaparthi et al. [18] | 2026 | ResNet-18, MobileNetV2 | 94.68%, 93.10% | Applied only on the limited CIFAR-10 dataset. |
| Binish et al. [14] | 2026 | CBAM–SMK | 97.0% | The model relies heavily on high-quality labeled MRI data and may be sensitive to class imbalance. |
| Neena et al. [17] | 2026 | PWT + CNN + Self Attention | 95.88% | Loss of fine tumor boundaries due to resizing MRI slices to $150 \times 150$, which may affect detection of small lesions. |
| Hasan et al. [22] | 2026 | ResNet50 + SS-PANet + Grad-CAM++ Hybrid Attention | 97% | Rare and ambiguous tumor cases with low contrast or irregular boundaries may challenge model reliability. |
| Syed Ijaz et al. [8] | 2026 | VGG19 + CBAM | 98.73% | Depends solely on training data and risks overfitting. |

Chapter 3

# Materials and Methods

## 3.1 Dataset

This dataset contains 7,200 human brain MRI images kaggle.which is curated from multiple publicly available sources, including the figshare brain MRI dataset, the SARTAJ dataset, and the Br35H dataset.The dataset was removed unreliable or duplicate to ensure label accuracy for research in medical image analysis, deep learning, multi-class classification.Originally, the dataset was split into 80% training and 20% testing data with balanced class distributions. The training set contains 1,400 images per class (5,600 images total), and the testing set contains 400 images per class (1,600 images total).Which was precisely divided into four classes: Glioma, Meningioma, Pituitary tumor, and No tumor.The Fig. 3.1 and Table 3.1 illustrates the distribution of the dataset classification and training-testing sets.

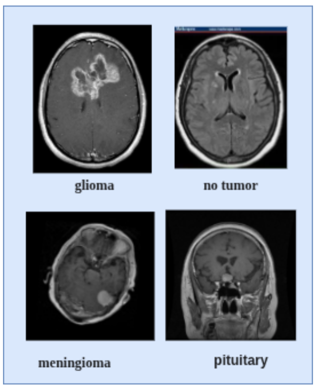


Figure 3.1: Brain Tumor MRI Dataset classification.

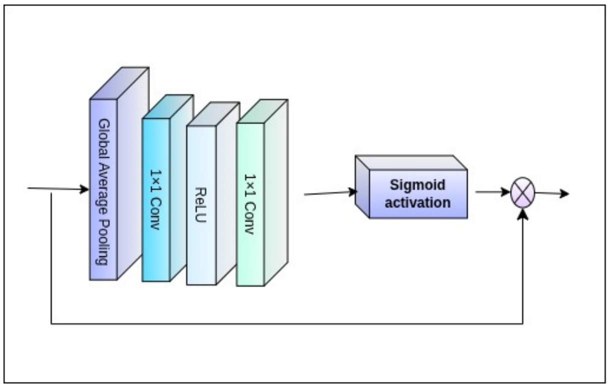


Figure 3.2: Architecture integrating the Lite Convolutional Attention (LCA) mechanism.

Table 3.1: Distribution of brain tumor MRI dataset across training and test sets

| **Tumor Category** | **Training Set** | **Test Set** |
|---|---|---|
| Glioma | 1321 | 300 |
| Meningioma | 1339 | 306 |
| Pituitary | 1457 | 300 |
| No Tumor | 1595 | 405 |

The constituent datasets contribute as follows. The *Figshare* brain tumor dataset provides T1-weighted contrast-enhanced MRI images of *glioma*, *meningioma*, and *pituitary tumor* classes. The *SARTAJ* dataset contributes additional *glioma* and *meningioma* samples together with healthy (*No Tumor*) MRI images. The *Br35H* dataset primarily supplements the *No Tumor* class. After merging the three datasets, duplicate images were identified using file hashing and removed. Images containing corrupted headers, annotations, or poor visual quality were also excluded, resulting in the final balanced dataset distribution presented in Table 3.1.

All MRI images were anonymized by the original dataset providers and contained no patient-identifiable information. Prior to inclusion in this study, the datasets were further verified to ensure compliance with standard ethical practices for publicly available medical imaging datasets.The the complete data curation and class-merging pipeline, beginning with the three publicly available source datasets and ending with the final balanced corpus of 7,200 MRI images used for training, validation, and testing.

## 3.2 Pre-processing

Proper preprocessing can significantly improve classification performance. During training, augmentation methods such as random shuffle,rotations (max 10○), horizontal flipping were applied to maximize generalization.Resizing, scaling were applied for the validation and test datasets.Each image was resized to 224×224 pixels and and normalized to [0,1] per channel.

Intensity normalization is applied during the preprocessing, which adjusts the pixel values of an image to a standardized range [14]. Additionally, normalization helps balance the influence of each color channel, preventing one from dominating the others due to differences in magnitude [24]. This can improve the consistency of input data and potentially enhance the convergence qualities of the learning algorithm, and Images were converted to PyTorch tensors using torchvision.transforms.ToTensor(), which scales pixel intensities to the interval [0, 1] as follows:

$$I_{\text{norm}}(x, y) = \frac{I(x, y) - I_{\min}}{I_{\max} - I_{\min}} \tag{3.1}$$

Table 3.2: Data augmentation operations applied during training

| **Operation** | **Parameter Range** | **Purpose** |
|---|---|---|
| Random Rotation | $\pm 10°$ | Simulate patient head-tilt variation. |
| Horizontal Flip | $p = 0.5$ | Account for left-right anatomical symmetry. |
| Random Shuffle | Batch-level | Prevent class-order learning bias during training. |
| Resize | $224 \times 224$ pixels | Match the input dimensions required by the proposed CNN model. |
| Normalization | $[0, 1]$ per channel | Stabilize gradient magnitudes and improve training convergence. |

Data augmentation was applied exclusively to the training dataset, whereas the validation and test datasets underwent only deterministic resizing and normalization (Table 3.2) to ensure that the reported performance metrics accurately reflect the model's generalization ability on previously unseen and unaltered images. Aggressive geometric transformations, including large-angle rotations, shear transformations, and elastic deformations, were intentionally avoided because excessive distortion of tumor morphology may introduce label-inconsistent artifacts, particularly for small pituitary tumors and irregular glioma boundaries.

## 3.3 Proposed Method

This study proposes an approach for evaluating custom lightweight convolutional neural network (CNN) models applied to brain tumor classification, with an emphasis on interpretability. The core method involves integrating the Lite Convolutional Attention Module (LCA) into selected lightweight CNN backbones and employing Grad-CAM and techniques to visualize and assess confidence scores for localization. The aim is to explore how LCA affects both classification performance and model interpretability workflow of proposed models as shown in Fig. 3.3.

## 3.4 Proposed CNN-LCA Architecture Overview

This paper proposes the CNN-LCA model,to integrate a convolutional neural network with a lightweight channel attention mechanism for enhancing the extraction of discriminative features from brain MRI images. The architecture is designed to capture multi-level spatial features through a series of three convolutional blocks composed of convolution layers, batch normalization [1]

$$y = \gamma \frac{x - E[x]}{\sqrt{Var[x] + \epsilon}} + \beta \tag{3.2}$$

and non-linear activation functions. With the overall framework shown in Fig. 3.4.

## 3.5 Feature Extraction

In order to overcome the limitations of traditional CNN feature extraction capabilities,these convolution layers enable the network to learn hierarchical representations, starting from low-level patterns such as edges and textures to more complex structures relevant to tumor classification. Each convolutional block in the proposed CNN-LCA architecture consists of a sequence of layers including a Convolutional layer (Conv), Batch Normalization layer (BN), Rectified Linear Unit activation (ReLU), and a MaxPooling layer (MaxPool). Two consecutive $3 \times 3$ convolution kernels are used in each block to effectively capture spatial features while maintaining computational efficiency. Although deep architecture has been widely used in too many network layers,that may lead to some problems such as overfitting and oversize. Therefore, in this paper, we stack only 3 block layers of grouped convolution and attention submodules to build the network branch.Here, the batch normalization stabilizes the training process and accelerates convergence, while the ReLU activation introduces non-linearity to enhance feature learning. The $2 \times 2$ max-pooling operation reduces the spatial dimensions of the feature maps and helps the network focus on the most prominent spatial patterns and are incorporated after each convolutional block to reduce spatial dimensions.Dropout layer 0.2 has been employed to prevent early overfitting each block .Since the convolution kernel size is relatively small compared to the spatial dimensions of the input images, single-scale convolution is sufficient to capture meaningful local patterns such as edges,

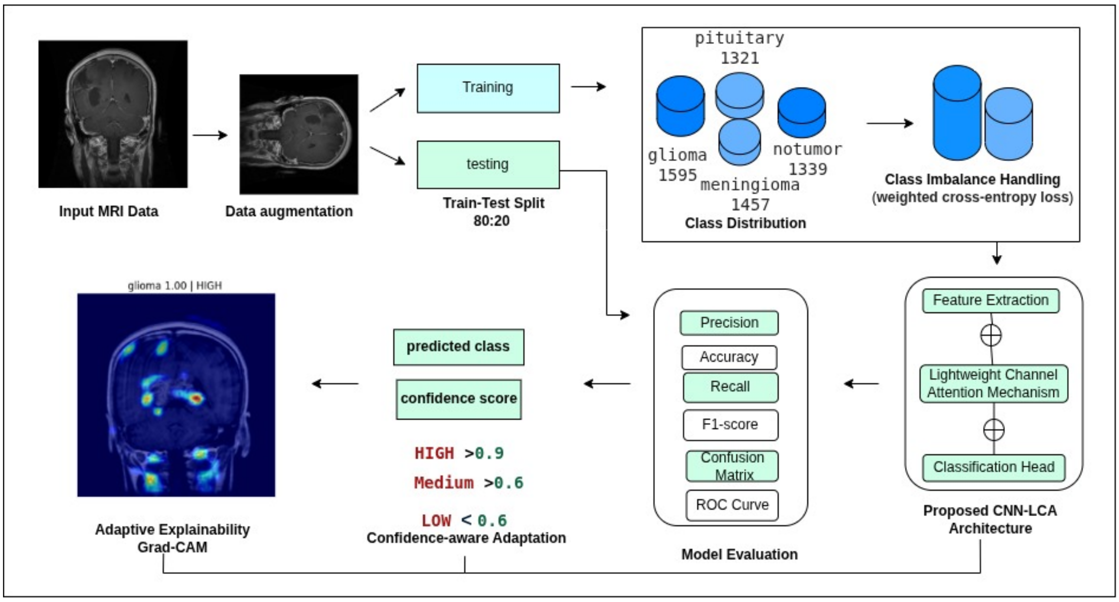


Figure 3.3: Workflow of the proposed model.

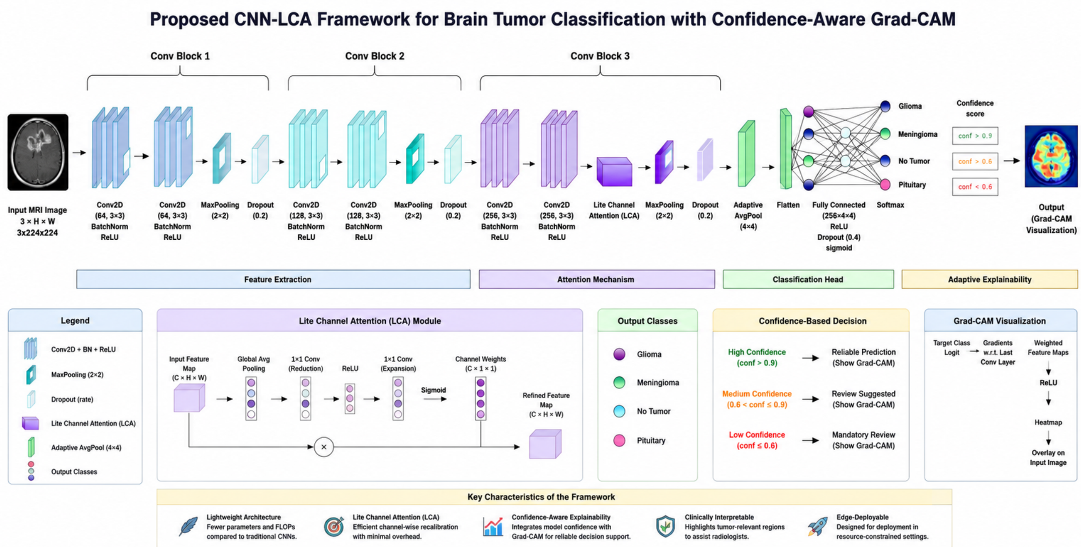


Figure 3.4: Framework of the proposed CNN-LCA model.

textures, and structural information. For feature extraction, the proposed model employs a sequence of single-scale convolutional blocks inspired by the VGG-style architecture, which has demonstrated strong performance in visual feature learning.

## 3.6 Lightweight Convolution Attention Machinism(LCA)

The proposed model incorporates a Lite Channel Attention (LCA) mechanism Fig. 3.2 inspired by the Squeeze-and-Excitation (SE) block and builds upon the design philosophy of Efficient Channel Attention (ECA) while incorporating group convolutions for further parameter reduction. The LCA module focuses on channel-wise feature recalibration through a squeeze-and-excitation mechanism.

Given an input feature map $X \in \mathbb{R}^{C\times H\times W}$, the LCA module first applies global average pooling to obtain a channel descriptor $z \in \mathbb{R}^{C}$. Specifically, the network extracts the feature map $X$ through convolutional operations and compresses the spatial information using global average pooling to capture global contextual information. The $c$-th element of $z$, denoted as $z_c$, is computed as:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_c(i,j) \tag{3.3}$$

where $u_c$ denotes the $c$-th channel of the feature map $X$, while $H$ and $W$ represent its height and width, respectively. Following the Excitation step, the interdependencies between channels are modeled to produce the channel-wise weight coefficients, computed as follows. The descriptor is then processed with a 1D convolution with reduction ratio $r = 8$ [18]:

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2\delta(W_1 z)) \tag{3.4}$$

To further enhance features into the deeper layers of the network, LCA uses global average pooling followed by 1D convolution 32:

$$s = \sigma(\text{GroupConv1D}_{k,g}(z)) \tag{3.5}$$

the where g = 4 groups are employed. The adaptive kernel size follows the same formulation as The kernel size k is determined adaptively based on channel dimension C:

$$k = \left| \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right|_{odd} \tag{3.6}$$

where $\gamma = 2$ and $b = 1$ are hyperparameters, and $|\cdot|_{\text{odd}}$ denotes rounding to the nearest odd number. This design ensures only $k$ parameters per block, dramatically reducing overhead. By using grouped convolutions, LCA reduces parameters to $k \times C/g$ while maintaining local cross-channel interaction within each group. LCA explicitly constrains attention to localized channel groups. This design choice enables an empirical examination of whether full channel connectivity is necessary for effective channel recalibration. LCA is motivated by three observations: full cross-channel interaction may be unnecessary for attention, grouped processing can capture local channel dependencies efficiently, and adaptive kernel sizing preserves receptive fields across different layer depths. This design achieves a favorable balance between expressiveness and efficiency. where $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$, $W_2 \in \mathbb{R}^{C \times \frac{C}{r}}$, $\delta$ is ReLU, and $\sigma$ is sigmoid. The output is obtained by:

$$\tilde{X} = s \odot X \tag{3.7}$$

In our propsoed model, the LCA module is applied after the final convolutional layer (256× 256) .The global average pooling followed by two lightweight 1×1 convolution layers with ReLU and sigmoid (nonlinear) activation functions to generate channel-wise attention weights. These weights emphasize on constantly allowing the model to emphasize on which channels are important for informative features and suppressing less informative channels.This attention-guided feature refinement improves the model's ability By adding minimal computational overhead.The model uses a reduction ratio of 8 within the LCA block for dimensionality reduction hidden = max(channels // reduction, 4))which keeps the attention mechanism computationally lite.To focus on tumor-relevant regions and improves the discriminative capability of the extracted features within MRI image are crucial for accurate tumor classification.

### 3.6.1 Computational Complexity of the LCA Module

To quantitatively evaluate the efficiency of the proposed Lite Convolutional Attention (LCA) module, its computational complexity is analytically compared with several widely used channel attention mechanisms, including Squeeze-and-Excitation (SE), CBAM, and Efficient Channel Attention (ECA). Consider an input feature map with $C$ channels.

- **Squeeze-and-Excitation (SE):** An SE block employs two fully connected layers with a reduction ratio $r$, resulting in approximately

$$P_{\mathrm{SE}} = \frac{2C^2}{r}, \tag{3.8}$$

trainable parameters.

- **CBAM:** CBAM incorporates an SE-style channel attention branch together with an additional spatial attention branch using a $7 \times 7$ convolution. Consequently, its parameter count is higher than that of the SE block because of the added spatial convolution.

- **Efficient Channel Attention (ECA):** ECA eliminates fully connected layers and replaces them with a single one-dimensional convolution having an adaptive kernel size $k$. The resulting parameter count is

$$P_{\mathrm{ECA}} = k, \tag{3.9}$$

which is independent of the channel dimension $C$.

- **Proposed Lite Convolutional Attention (LCA):** The proposed LCA module extends the ECA mechanism by introducing grouped one-dimensional convolutions, thereby maintaining lightweight computation while improving cross-channel feature interaction. The corresponding parameter count is given by

$$P_{\mathrm{LCA}} = \frac{kC}{g}, \tag{3.10}$$

where $g$ denotes the number of channel groups. In this work, $g = 4$.

For the final convolutional block of the proposed CNN-LCA architecture, where $C = 256$, $r = 8$, $g = 4$, and the adaptive kernel size obtained from Eq. **??** is $k = 5$, the corresponding parameter counts are

$$P_{\mathrm{SE}} = \frac{2 \times 256^2}{8} = 16\,384, \tag{3.11}$$

$$P_{\mathrm{ECA}} = 5, \tag{3.12}$$

$$P_{\mathrm{LCA}} = \frac{5 \times 256}{4} = 320. \tag{3.13}$$

Table 3.3 summarizes the parameter requirements of different attention mechanisms together with the total parameter counts of several commonly used CNN backbones.

This comparison highlights the lightweight nature of the proposed CNN-LCA framework relative to conventional transfer learning models.

Table 3.3: Parameter comparison of attention mechanisms and backbone models

| **Component / Model** | **Approx. Parameters** | **Relative Size** |
|---|---|---|
| SE Block ($C = 256$, $r = 8$) | 16,384 | 1$\times$ |
| CBAM Block ($C = 256$) | $\sim$18,500 | 1.13$\times$ |
| ECA Block ($C = 256$) | $\sim$5 | 0.0003$\times$ |
| LCA Block ($C = 256$, $g = 4$) | $\sim$320 | 0.02$\times$ |
| Proposed CNN-LCA | **500,000** | – |
| VGG19 | $\sim$143,000,000 | – |
| ResNet50 | $\sim$25,600,000 | – |
| EfficientNet-B0 | $\sim$5,300,000 | – |

The total parameter count of the proposed CNN-LCA model should be obtained using a network analysis tool such as `torchinfo.summary()` or `thop.profile()`. If available, floating-point operations (FLOPs) may also be reported to provide a more comprehensive evaluation of computational efficiency. The results presented in Table 3.3 provide quantitative evidence supporting the lightweight design of the proposed architecture and will be further discussed in Chapter 4 during comparative performance analysis.

## 3.7 Fully Connected Classification Head

Following the attention mechanism, the refined feature maps are processed through an adaptive average pooling layer 4×4, Flatten size 4×4×256 into Relu activation and a fully connected classification head.The fully connected layers transform the high-level features into class probabilities using the softmax function [1], enabling the model to perform multi-class tumor classification according to the score hierarchy: high, low, and medium. Dropout regularization 0.4 is applied to mitigate overfitting and enhance the generalization capability of the network

$$\text{Confidence score} = \max P(c \mid x) = \text{Softmax}(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}} \tag{3.14}$$

In our model, the softmax function also provides logits xi, or raw scores [10], known as confidence scores. The core component that generates the confidence score is the Softmax activation function, which transforms the raw logits from the final fully connected layer into the class probability distribution. The maximum class probability represents the model's confidence, which is then heuristically categorized into High (confidence score $> 0.9$), Medium (confidence score 0.6-0.9), or Low (confidence score $<$ 0.6) zones to provide an adaptive approach for reliability [16]. Rescaling of logits is done before the softmax to better match predicted probabilities to observations. Because it is easy to implement and does not change accuracy, only the sharpness of the probabilities [10]. By combining hierarchical feature extraction, lightweight channel attention, and a robust classification head, the proposed CNN-LCA architecture provides an effective framework for accurate and interpretable brain tumor classification. Table 3.4 Architectural Overview of the Proposed CNN-LCA Model

Table 3.4: Architectural Overview of the Proposed CNN-LCA Model

| **Layer** | **Output Size (H×W×C)** | **Kernel** | **Filters** | **Activation** |
|---|---|---|---|---|
| Input Image | $224 \times 224 \times 3$ | – | – | – |
| Conv1 + BN | $224 \times 224 \times 64$ | $3 \times 3$ | 64 | ReLU |
| Conv2 + BN | $224 \times 224 \times 64$ | $3 \times 3$ | 64 | ReLU |
| MaxPooling | $112 \times 112 \times 64$ | $2 \times 2$ | – | – |
| Conv3 + BN | $112 \times 112 \times 128$ | $3 \times 3$ | 128 | ReLU |
| Conv4 + BN | $112 \times 112 \times 128$ | $3 \times 3$ | 128 | ReLU |
| MaxPooling | $56 \times 56 \times 128$ | $2 \times 2$ | – | – |
| Conv5 + BN | $56 \times 56 \times 256$ | $3 \times 3$ | 256 | ReLU |
| Conv6 + BN | $56 \times 56 \times 256$ | $3 \times 3$ | 256 | ReLU |
| LiteChannelAttention | $56 \times 56 \times 256$ | $1 \times 1$ | 256 | Sigmoid |
| MaxPooling | $28 \times 28 \times 256$ | $2 \times 2$ | – | – |
| AdaptiveAvgPool | $4 \times 4 \times 256$ | – | – | – |
| Flatten | $4 \times 4 \times 256$ | – | – | – |
| Fully Connected | 512 | – | – | ReLU |
| Fully Connected | 4 | – | – | Softmax |

## 3.8 Class Imbalance Handling

As presented in Table 3.1, the training dataset exhibits a mild class imbalance, where the *No Tumor* class (1,595 images) contains approximately 21% more samples than the *Glioma* class (1,321 images). If left unaddressed, such an imbalance may bias the classifier toward the majority class, resulting in inflated overall accuracy while reducing recall for minority classes. This issue is particularly critical in medical diagnosis because false-negative predictions may delay treatment and have more serious clinical consequences than false-positive predictions.

To alleviate this problem, class weights were computed inversely proportional to the frequency of each class according to

$$w_c = \frac{N}{K \times n_c}, \tag{3.15}$$

where $N$ denotes the total number of training samples, $K$ represents the total number of classes, and $n_c$ is the number of samples belonging to class $c$.

The computed class weights were incorporated into the categorical cross-entropy loss function to assign larger penalties to misclassified minority-class samples during backpropagation. Furthermore, label smoothing with $\varepsilon = 0.05$ was applied to improve model calibration and reduce overconfidence. The combination of weighted cross-entropy loss and label smoothing constitutes the final optimization objective used throughout the training process, as described in Section 3.8.

## 3.9 Grad-CAM Visualization

Grad-CAM is a widely used method of visualizing feature maps. It is displayed as weights. The weights represent the degree of importance. The greater the weight is, the greater the correlation with this category. The smaller the weight is, the bluer the location, indicating that these places are not important and are unrelated to do the current class. The CNN model has been regarded as a "black box" because it is difficult to understand how it makes predictions [15]. Grad-CAM provides intuitive, visually interpretable explanations that identify the areas of the image that contributed most to a specific class prediction. First, for each feature map $A^k$ in the last convolutional layer, a gradient with respect to a predicted class $c$, $y^c$ is calculated, which can formally be expressed as:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{3.16}$$

where weight $\alpha_k^c$ is the importance of feature map $A^k$ for class $c$. Then, a weighted combination of the resulting tensors is performed, followed by a ReLU (Rectified Linear Unit activation function, which produces the final sensitivity map:

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha^c_k A^k\right) \tag{3.17}$$

By utilizing a multi-staged approach to interpret model decisions by combining probabilistic certainty with spatial localization. Simultaneously, the Grad-CAM algorithm targets the last layer, computing the gradients of the predicted class score with respect to the feature maps to produce a weighted heat map. This heat map is normalized and overlaid on the original input image, thereby visually grounding the model's high- and low-confidence predictions in specific spatial regions of the data [23].

## 3.10 Experimental Setup

The experiments were implemented using the PyTorch deep learning framework with Python 3.7. All models were trained and evaluated in Google Colab, using an NVIDIA T4 GPU (16 GB of VRAM) with CUDA acceleration to enable efficient deep learning computation.

A robust accuracy can be achieved by optimizing precise hyperparameter tuning.The AdamW optimizer parameters were set to the initial learning rate lr=$3e^{-4}$ and weight decay $1e^{-4}$. The learning rate is dynamically adjusted using the Cosine AnnealingLR schedule optimized at a maximum of 50 epochs, which smoothly decays the rate to zero, allowing the model to converge effectively to optimal local minima during the final stages of training. So, the total number of epochs was 50.A class-weighted cross-entropy loss was employed in the target distributions to improve generalization and mitigate model overconfidence. Additionally, label smoothing $(\varepsilon = 0.05)$ was incorporated to enhance generalization and reduce overconfidence. Class-Weighted Cross-Entropy Loss [1] function is implemented, where weights are dynamically calculated as weights = $total\_samples/class\_counts$ to penalize errors in minority classes to address class imbalance. Table 3.5 provides an overview of the hyperparameter configuration.

$$L = -\sum_i y_i \log(p_i) \tag{3.18}$$

Table 3.5: Optimal Hyperparameter Configurations for the Proposed CNN Model

| **Hyperparameter** | **Value** |
|---|---|
| Optimizer | AdamW |
| Learning rate | $3e^{-4}$ |
| Weight decay | $1e^{-4}$ |
| Total epoch | 50 |
| Loss Function | Weighted Cross-Entropy |
| Label smoothing rate | 0.05 |
| Scheduler | Cosine AnnealingLR |
| Batch_size | 16 |

Chapter 4

# Results and Discussion

As illustrated in Table 4.1, various optimizers including Adam, RMSProp, and AdamW were evaluated to determine the most effective optimization strategy. While RMSProp has shown robust performance in prior literature, our experimental results demonstrate that AdamW achieved the superior performance with an accuracy of 98.40% and an F1-score of 98.35%. This improvement can be attributed to AdamW's decoupled weight decay, which provides more effective regularization compared to the standard Adam or RMSProp, particularly in deep CNN architectures.

Table 4.1: Evaluation of Convergence and Accuracy Metrics across Different Optimization Algorithms

| **Optimizer** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
|---|---|---|---|---|
| Adam | 0.9809 | 0.9808 | 0.9794 | 0.9800 |
| RMSProp | 0.9825 | 0.9836 | 0.9815 | 0.9824 |
| AdamW | 0.9840 | 0.9844 | 0.9827 | 0.9835 |

To evaluate the scalability of the proposed framework, the model was further validated using an external Melanoma skin cancer dataset containing 9,600 images. As shown in Table 4.2, the model achieved a notable accuracy of 93.70%. Despite this dataset being significantly larger—containing 2,400 more images than our primary experimental dataset—the model maintained consistent performance. This demonstrates that the proposed architecture is not only effective for specific tasks but also possesses the capacity to generalize and scale efficiently as the data volume increases.

Table 4.2: Analysis of the proposed CNN model on external datasets

| **Source of Dataset** | **No. of Total Images** | **Accuracy** |
| --- | --- | --- |
| Melanoma Skin Cancer Dataset | 9600 | 93.70% |
| Brain Tumor Dataset | 7200 | 98.40% |

## 4.1 Experiment Evaluation and Analysis

The analysis aims to quantify the models' performance and interpretability, highlighting their diagnostic reliability and relevance for clinical applications.The key metrics such as accuracy, precision, recall, F1-score, sensitivity, specificity Table 4.3.

Table 4.3: Evaluation metrics used for performance assessment

| **Metric** | **Equation** | **Description (Ref)** |
| --- | --- | --- |
| Accuracy | $\frac{TP+TN}{TP+TN+FP+FN}$ | Ratio of total correct predictions to all predictions [13]. |
| Precision | $\frac{TP}{TP+FP}$ | Fraction of positive predictions that are actually correct 12. |
| Recall | $\frac{TP}{TP+FN}$ | Percentage of correct positive predictions among all possible positive cases . |
| F1-Score | $\frac{2\times Precision\times Recall}{Precision+Recall}$ | Harmonic mean of precision and recall, providing a balance between the two [1]. |

*Where:* TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives.

Fig. 4.1 a and b shows the visual outputs of Grad-CAM heatmap on a pituitary MRI, focusing on the clarity, granularity, and diagnostic relevance of their heatmaps.The color scale on the right indicates activation intensity, ranging from low (blue) to high (red), providing the interpretability of the model's focus areas. The model focuses most strongly on the tumor region (red/yellow areas), indicating high confidence in its classification. Peripheral regions exhibit lower attention (green/blue), reflecting areas with less influence on the decision. This demonstrates that the Adaptive XAI effectively identifies the most relevant features based on the prediction confidence score. It highlights the regions most critical for the model's decision-making process. By comparing these interpretability techniques, the analysis highlights their effectiveness in identifying critical diagnostic features and their potential impact on clinical decision-making [25].

Table 4.4, the proposed model demonstrates exceptional classification performance

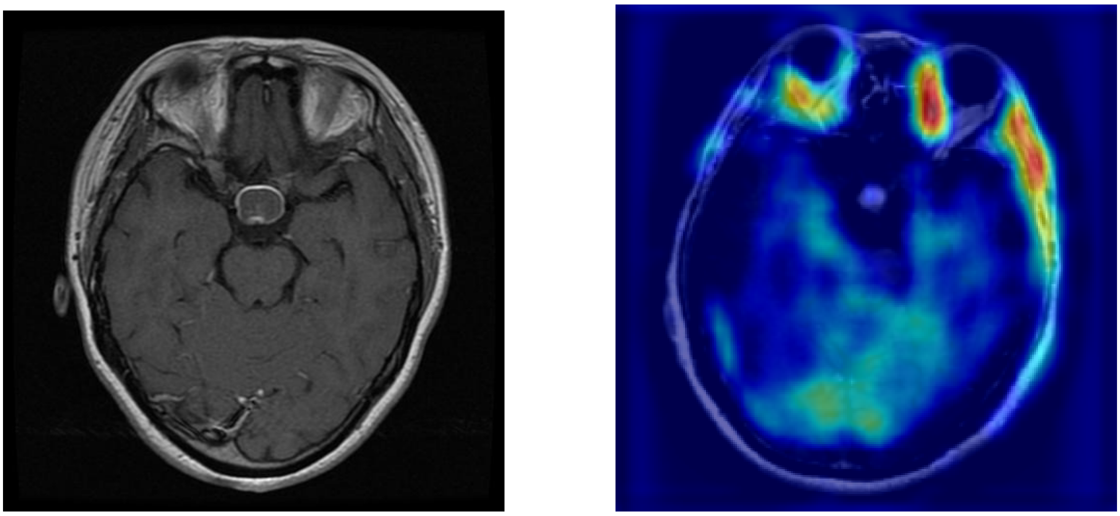

Figure 4.1: (a) Original MRI (T1-weighted) image. (b) Grad-CAM heatmap highlighting the pituitary tumor region.

across all four categories. Notably, the Glioma class achieved a precision of 0.9966, while the Pituitary and No Tumor classes exhibited recall rates exceeding 99.6%. The high Specificity scores across all classes (above 0.99) further underscore the model's reliability in medical diagnostics, effectively minimizing false positive results. Furthermore, the high Average Confidence Scores indicate that the model's decision-making process is statistically robust.

Table 4.4: Classification report of accuracy, precision, recall, F1-score, sensitivity, and specificity

| **Class** | **Precision** | **Recall** | **F1-score** | **Sensitivity** | **Specificity** | **Avg Confidence Score** |
|---|---|---|---|---|---|---|
| Glioma | 0.9966 | 0.9800 | 0.9882 | 0.9800 | 0.9990 | 0.983 |
| Meningioma | 0.9767 | 0.9575 | 0.9670 | 0.9575 | 0.9930 | 0.978 |
| No Tumor | 0.9830 | 0.9975 | 0.9902 | 0.9975 | 0.9923 | 0.998 |
| Pituitary | 0.9803 | 0.9967 | 0.9884 | 0.9967 | 0.9941 | 0.998 |
| Accuracy | — | — | — | 0.9840 | — | — |
| Macro Avg | 0.9841 | 0.9829 | 0.9835 | — | — | — |
| Weighted Avg | 0.9840 | 0.9840 | 0.9839 | — | — | — |

The confusion matrix in Fig. 4.2 shows that glioma was 294, meningioma 293, no tumor 404, and pituitary was 290. Here, glioma was consecutively misclassified in 6 cases, meningioma in 13 cases, no tumor in 1 case, and pituitary in 1 case. The error analysis reveals that Meningioma had the highest misclassification rate (13 instances), which is typical in neuroimaging due to the subtle structural similarities between Meningioma and surrounding tissues. In contrast, the No Tumor and Pituitary classes showed exceptional reliability, with only 1 misclassification each. The ablation study presented demonstrates that without LCA, the model achieved a baseline accuracy of approximately 96% in accuracy, with only higher precision, but recall and F1 score are much lower than with LCA. That shows with LCA promotes better performance overall. Here, the proposed attention mechanism justifies that the average confidence score is calculated as the mean of the confidence values provided by the SoftMax function for each class. It reflects the model's certainty in its predictions, providing assurance of the validity of the predicted class. As shown in Fig. 4.2, the proposed model achieves strong classification performance. The results of the ablation study are presented in Fig. 4.3, demonstrating the effectiveness of the compression and attention modules.

While the precision remained relatively high, there was a noticeable degradation in Recall and F1-score. The integration of LCA directly addressed these gaps by enabling the model to focus on the most discriminative pathological features. This results in a more balanced performance, proving that LCA is essential for achieving the high sensitivity required in medical diagnostics.

As shown in Fig. 4.4, the proposed CNN model exhibits a steady improvement in both training and validation accuracy throughout the training process. The training accuracy increases consistently and reaches 98.63% at Epoch 50, while the validation accuracy converges to 97.50%. The close agreement between the two curves indicates that the model learns effectively without significant overfitting, demonstrating strong generalization capability on unseen data.

As illustrated in Fig. 4.5, the proposed CNN model achieves stable convergence after approximately the 30th epoch. The best performance is observed at Epoch 50, where the training loss reaches a minimum of 0.0403, and the validation loss decreases to 0.0689. Moreover, the close alignment between the training and validation loss curves indicates minimal overfitting, confirming that the proposed architecture, together with the AdamW optimizer, effectively stabilizes the learning process while maintaining strong generalization performance.

Notably, the close alignment between the training and validation loss values indicates a lack of overfitting, confirming that the architecture and optimization strategy—particularly the use of AdamW—successfully stabilized the learning process.

As illustrated by the ROC curves Fig. 4.6, the model achieves an Area Under the Curve (AUC) of 1.000 for the Glioma, Pituitary, and 'No Tumor' classes, with a marginal drop to 0.998 for Meningioma. This indicates that the learned representations are statistically robust and effectively capture complex tumor morphologies with minimal class overlap, providing a reliable foundation for automated clinical diagnostic support

## 4.2 Model Performance Analysis

The results indicate that CNN–LCA significantly improves classification accuracy. The table presents a comparative performance analysis of the CNN–LCA model against well-established deep learning models such as EfficientNetB0, EfficientNetV2B0, ResNet50, VGG19, etc on the brain tumor dataset.The models were evaluated using statistical significance analysis, including accuracy, precision, recall, F1-score, sensitivity, and specificity. Moreover, the confusion matrix,loss curve, accuracy curve, and ROC curve are used for qualitative analysis, and Grad-CAM is used for these metrics. These metrics provide a comprehensive evaluation of model performance in terms of both the ability to correctly classify.

The classification results in Table 8 demonstrate the model's high reliability, achieving an Overall Accuracy of 0.9840. The Glioma and Pituitary classes show exceptional performance, with Glioma reaching a precision of 0.9966 and Pituitary achieving a recall of 0.9967, indicating near-perfect identification and minimal false negatives for these categories. The Notumor class also performs at an elite level, yielding the highest F1-score (0.9902) and recall (0.9975), which highlights the model's effectiveness in correctly identifying healthy scans. While the Meningioma class shows the relatively lowest metrics, its F1-score of 0.9670 and precision of 0.9767 still represent strong classification capability. Furthermore, the close alignment between the Macro Average (0.9835) and Weighted Average (0.9839) confirms that the model is well-balanced across all tumor types, maintaining consistent predictive power regardless of the variation in support. The average confidence score is calculated as the mean of the confidence values provided by the SoftMax function for each class. It reflects the model's certainty in its predictions, providing assurance of the validity of the predicted class. In addition to quantitative performance, the proposed Adaptive XAI framework provides confidence-aware visual explanations. High-confidence predictions (e.g., confidence $> 0.98$) produce more concentrated heatmap regions, highlighting the tumor area precisely. In contrast, lower confidence predictions generate broader heatmap distributions, indicating greater model uncertainty.

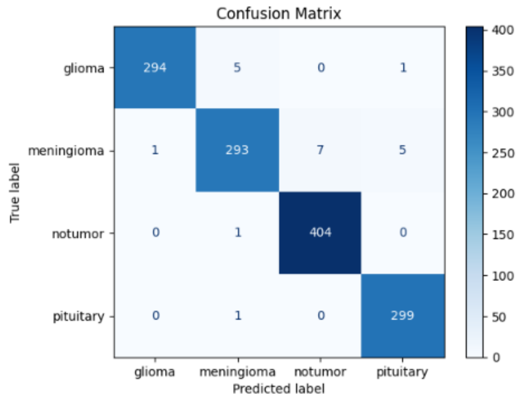


Figure 4.2: Confusion matrix of the proposed model.

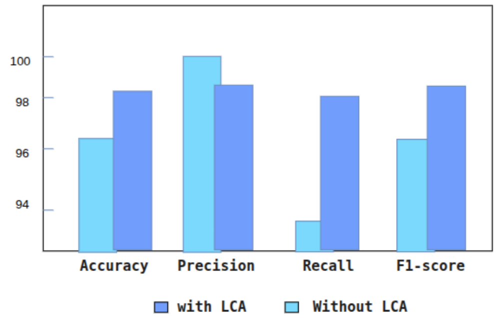


Figure 4.3: Ablation study of the proposed model variants.

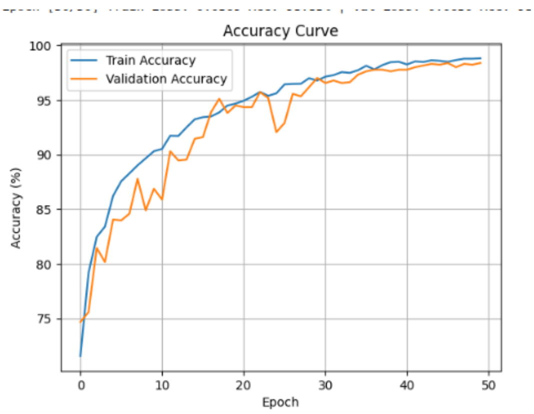


Figure 4.4: Accuracy curve of the proposed CNN model.

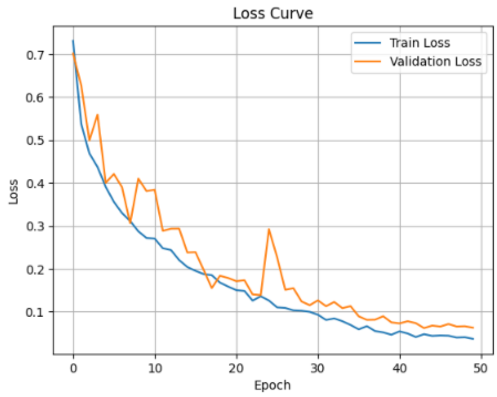


Figure 4.5: Loss curve of the proposed CNN model.

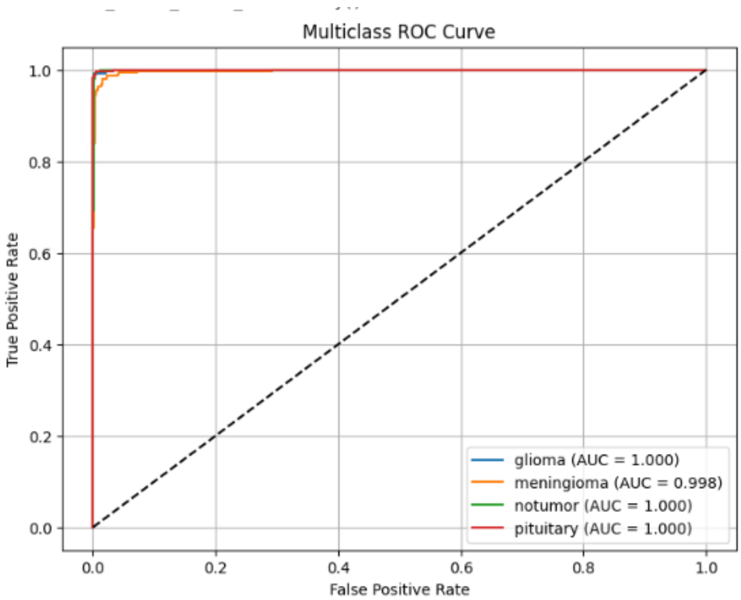


Figure 4.6: ROC curve of the proposed CNN-LCA model.

Table 4.5: Comparative summary of recent deep learning models for brain tumor classification using MRI datasets

| **Author** | **Model** | **Attention Mechanism** | **Accuracy** |
|---|---|---|---|
| Mir et al. [5] | EfficientNetB0 + CBAM | CBAM | 99% |
| Anas et al. [26] | EfficientNetV2B0 + CBAM | CBAM | 97% |
| Binish et al. [14] | CBAM–SMK | CBAM | 97.0% |
| Neena et al. [17] | PWT + CNN + Self Attention | Self Attention | 95.88% |
| Jahid et al. [22] | ResNet50 + SSPANet + GradCAM++ | Hybrid Attention Mechanism | 97% |
| Syed Ijaz et al. [8] | CBAM + VGG19 | CBAM | 98.73% |
| **Proposed Model** | Customized VGG19-Style + LCA + GradCAM | LCA | **98.40%** |

### 4.2.1 Error Case Analysis

A closer examination of the confusion matrix presented in Fig. 4.2 reveals that the majority of misclassifications occurred in the *Meningioma* class. Specifically, seven meningioma samples were incorrectly classified as *No Tumor*, while five were misclassified as *Pituitary Tumor*. This observation can be explained from a clinical perspective. Small, low-contrast meningiomas located near the skull base may closely resemble normal dural thickening, leading to confusion with the *No Tumor* class. Similarly, meningiomas located near the sellar region may share anatomical characteristics with pituitary tumors, increasing the likelihood of confusion between these two classes.

Furthermore, these challenging cases generally exhibited lower prediction confidence than correctly classified samples, indicating a higher degree of model uncertainty. As discussed in Section 3.7, the proposed confidence-aware Grad-CAM framework is designed to associate prediction confidence with visual explanations, thereby reducing the risk of presenting misleading interpretations for uncertain predictions. This behavior supports the clinical safety objective outlined in Section 1.5 by providing more reliable and trustworthy model explanations during the diagnostic decision-making process.

# Chapter 5

# Conclusions and Future Research

Brain tumors represent a significant clinical challenge, with meningiomas and gliomas being the most prevalent among adults, while gliomas predominate in the paediatric population. In this study, we propose a cost-efficient, lightweight, and low-parameterized Convolutional Neural Network (CNN) architecture that combines Lite Convolution Attention Mechanism(LCA) and adaptive explainable AI (XAI) techniques, achieving remarkable classification accuracy.This study has employed three convolutional block. Each convolutional block in the proposed CNN-LCA architecture consists of a sequence of layers including a Convolutional layer (Conv), Batch Normalization layer (BN), Rectified Linear Unit activation (ReLU), and a MaxPooling layer (MaxPool). Two consecutive 3 × 3 convolution kernels are used in each block to effectively capture spatial features while maintaining computational efficiency. Although deep architectures have been widely used with many network layers, that may lead to some problems such as overfitting and over-parameterization; the model uses a reduction ratio of 8 within the LCA block for dimensionality reduction hidden = max(channels // reduction, 4)), which keeps the attention mechanism computationally light. To focus on tumor-relevant regions and improves the discriminative capability of the extracted features within MRI image are crucial for accurate tumor classification. To further strengthen generalization, label smoothing $(\varepsilon = 0.05)$ and Class-Weighted Cross-Entropy Loss were incorporated to reduce overconfidence and address class imbalance. The model's versatility was validated on a skin cancer dataset, where it achieved an accuracy of 93.70%, demonstrating its transferability across medical imaging domains. For brain tumor classification, the model attains an Area Under the Curve (AUC) of 1.000 for Glioma, Pituitary, and No Tumor classes, with a marginal reduction to 0.998 for Meningioma.

A key contribution of this work is the integration of confidence-aware Grad-CAM visualizations, which leverage prediction confidence derived from softmax probabilities to generate adaptive visual explanations. Unlike conventional approaches that produce static saliency maps regardless of model certainty, the proposed framework dynamically adjusts its explanations according to confidence zones — classified as HIGH, MEDIUM, and LOW — providing clinically meaningful and context-sensitive reasoning. This stands in contrast to most existing models, which compute confidence scores but rarely incorporate them into the explanation pipeline. The proposed confidence-aware XAI framework lies in a simple but powerful distinction: confidence for what to look at, and confidence for where to look. By highlighting the regions most critical to the model's decision-making process, the adaptive XAI module effectively bridges the gap between model performance and clinical interpretability, supporting pathologists' decision-making and enhancing the framework's transparency, reliability, and clinical relevance.

# Bibliography


[1] L. Neumann, A. Zisserman, and A. Vedaldi, "Relaxed softmax: Efficient confidence auto-calibration for safe pedestrian detection," 2018.

[2] V. Anand, A. Khajuria, R. K. Pachauri, and V. Gupta, "Multi-class classification of brain tumors using optimized cnn and transfer learning techniques," *Scientific Reports*, 2026.

[3] S. Ferdousi, U. Das, and T. M. Zihan, "An explainable multi-class brain tumor classifier using efficientnetv2 with hybrid attention mechanisms," in *2025 IEEE 2nd International Conference on Computing, Applications and Systems (COMPAS)*, pp. 1–6, IEEE, 2025.

[4] R. L. Siegel, T. B. Kratzer, N. S. Wagle, H. Sung, and A. Jemal, "Cancer statistics, 2026," *Ca*, vol. 76, no. 1, p. e70043, 2026.

[5] A. N. Mir and D. R. Rizvi, "Exploring explainable ai in brain tumor detection: a hybrid approach using efficientnet and cbam," in *2025 3rd International Conference on Device Intelligence, Computing and Communication Technologies (DICCT)*, pp. 39–44, IEEE, 2025.

[6] A. Ajay, R. Karthik, A. S. Bisht, and P. Uppuluri, "Bonevisionnet: A deep learning approach for the classification of bone tumours from radiographs using a triple fusion attention network of transformer and cnns with xai visualizations," *Computerized Medical Imaging and Graphics*, p. 102668, 2025.

[7] F. M. Salman and S. S. Abu-Naser, "Comparative analysis of deep learning architectures for bone fracture detection: Mobilenetv2 vs. resnet50," *International Journal of Academic Information Systems Research (IJAISR)*, vol. 10, no. 1, pp. 39–51, 2026.

[8] S. I. U. Rahman, N. Abbas, S. Ali, S. Jan, I. E. Lee, M. Salman, A. Salhi, and A. Alkhyyat, "Improving acute lymphoblastic leukemia diagnosis through cbam-enhanced vgg19 deep learning," *Scientific Reports*, 2026.

[9] L. Neumann, A. Zisserman, and A. Vedaldi, "Relaxed softmax: Efficient confidence auto-calibration for safe pedestrian detection," 2018.

[10] J. S. Berkowitz, J. R. Patock, A. Nawaz, G. Gonzalez-Hernandez, and N. P. Tatonetti, "A crisis of overconfidence: Why confidence, not accuracy, is the real risk in clinical ai," *BioData Mining*, vol. 19, no. 1, p. 10, 2026.

[11] M. Ennab and H. Mcheick, "Advancing ai interpretability in medical imaging: a comparative analysis of pixel-level interpretability and grad-cam models," *Machine Learning and Knowledge Extraction*, vol. 7, no. 1, p. 12, 2025.

[12] B. Ojakapeli, "A narrative review of brain tumour epidemiology, surgical management, and system-level challenges in kenya (2000-2024)," *East African Journal of Neurological Sciences*, vol. 5, no. 1, pp. 62–66, 2026.

[13] H. AbuAlkebash, R. A. Saleh, and H. M. Ertunç, "Explainable depth-wise and channel-wise fusion models for multi-class skin lesion classification," *PLoS One*, vol. 21, no. 1, p. e0340901, 2026.

[14] M. Binish, S. R. RS, and V. Thomas, "Cbam–smk: Integrating convolution block attention module with separable multi-resolution kernels in deep neural networks for brain tumor classification," *Biomedical Signal Processing and Control*, vol. 112, p. 108483, 2026.

[15] X. Zhou, C. Tang, P. Huang, F. Mercaldo, A. Santone, and Y. Shao, "Lpcanet: Classification of laryngeal cancer histopathological images using a cnn with position attention and channel attention mechanisms," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 13, no. 4, pp. 666–682, 2021.

[16] J. Moon, J. Kim, Y. Shin, and S. Hwang, "Confidence-aware learning for deep neural networks," in *international conference on machine learning*, pp. 7034–7044, PMLR, 2020.

[17] N. KA and A. K. MN, "Haar-initialized parametric wavelet compression with attention-driven lightweight cnn for brain tumor classification on edge devices," *Biomedical Physics & Engineering Express*, vol. 12, no. 1, p. 015051, 2026.

[18] P. B. Kanaparthi and T. V. S. V. Padamata, "Lightweight channel attention for efficient cnns," *arXiv preprint arXiv:2601.01002*, 2026.

[19] F. An, X. Li, and X. Ma, "Medical image classification algorithm based on visual attention mechanism-mcnn," *Oxidative Medicine and Cellular Longevity*, vol. 2021, no. 1, p. 6280690, 2021.

[20] P. Zhou, Y. Cao, M. Li, Y. Ma, C. Chen, X. Gan, J. Wu, X. Lv, and C. Chen, "Hccanet: histopathological image grading of colorectal cancer using cnn based on multichannel fusion attention mechanism," *Scientific reports*, vol. 12, no. 1, p. 15103, 2022.

[21] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

[22] M. J. Hasan, M. Hasan, S. Akter, A. B. S. Mahi, and M. P. Uddin, "Enhancing brain tumor classification with a novel attention based explainable deep learning framework," *Biomedical Signal Processing and Control*, vol. 112, p. 108636, 2026.

[23] M. Skliarov, R. E. Shawi, C. Dhaoui, and N. Ahmed, "A comparative evaluation of explainability techniques for image data," *Scientific Reports*, vol. 15, no. 1, p. 41898, 2025.

[24] X. Yang, S. Yu, and W. Xu, "Enhanced convolutional neural networks for improved image classification," *arXiv preprint arXiv:2502.00663*, 2025.

[25] M. Ennab and H. Mcheick, "Advancing ai interpretability in medical imaging: a comparative analysis of pixel-level interpretability and grad-cam models," *Machine Learning and Knowledge Extraction*, vol. 7, no. 1, p. 12, 2025.

[26] N. Sengodan, "Cbam-efficientnetv2 for histopathology image classification using transfer learning and dual attention mechanisms," *arXiv preprint arXiv*, vol. 2410, 2024.