

Recent Advancements in Grad-CAM and variants: Enhancing Brain Tumor Detection, Segmentation, and Classification

Krishan Kumar

krishan21060051@gndec.ac.in

Guru Nanak Dev Engineering College

Dr.Kiran Jyoti



Guru Nanak Dev Engineering College

Systematic Review

Keywords: Blackbox, Deep Learning, Explainable AI, Brain tumor Detection, Segmentation, Classification, Diagnostic imaging

Posted Date: December 3rd, 2024

DOI: <https://doi.org/10.21203/rs.3.rs-5485128/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Brain tumors, caused by abnormal tissue growth in the brain, can significantly impair brain functions and pose significant health risks. As the tumor progresses to higher stages, the patient's prognosis and survival decline, resulting in a high mortality rate. With advances in medical imaging, particularly the use of MRI, AI approaches have emerged as powerful tools for detecting, segmenting, and classifying brain tumors. CNN and hybrid models such as Vision Transformers (ViTs) have provided promising insights in this area. Despite their high accuracy, but the reasoning behind the prediction often remained unanswered. AI models lack transparency and interpretability, paving the way for the development of explainable AI methods in diagnosing brain diseases. In recent years, gradient weighted class activation mapping (Grad-CAM) and its variants have emerged as powerful techniques for visualizing and interpreting deep learning models in medical imaging tasks, including brain tumor detection, segmentation, and classification. This paper provides a comprehensive overview of Grad-CAM and its improvements, with particular emphasis on their applications in brain tumor analysis. In this article, we reviewed the 31 research papers based on data from various research databases. This review highlights the importance of interpretability in deep learning for clinical applications and explores how Grad-CAM can complement traditional metrics to provide deeper insights into model decisions. The results provide valuable guidance for researchers developing explainable AI frameworks for accelerating brain tumor analysis, with the objective of developing trust, efficiency and clinical utility in resource-limited settings.

1. Introduction

A brain tumor is an abnormal mass of tissue that develops within brain cells and disrupts normal brain functions. It is classified as benign or malignant. Benign tumors are usually classified as grades 1 or 2, while malignant tumors are more serious and fall into grades 3 and 4. Malignant tumors are also categorized by their level of aggressiveness, with some being less aggressive and others being very aggressive. Important factors that help determine the grade of a tumor are its vascularity, invasiveness and growth rate. When a tumor reaches a higher stage, the patient's prognosis and chances of survival decrease significantly. Therefore, diagnosis of brain tumor and early treatment will certainly improve the patient's chances of survival [1-3].

In 2023, an estimated 24,810 adults (14,280 men and 10,530 women) in the United States were diagnosed with primary brain or spinal cord cancer. These brain tumors account for approximately 85-90% of central nervous system (CNS) tumors. It is estimated that 5,230 children under 20 years of age will be diagnosed with CNS cancer in the United States by 2023 [4].

MRI is widely used to diagnose primary tumors because it can provide extremely detailed images, which are crucial for the clinical diagnosis and detection of brain tumors [5]. This imaging technique is very accurate and provides clear images, which is why it is an important tool in clinical practice for diagnosing brain diseases [6]. Properly diagnosing diseases requires a lot of medical training and can take a long time. Many areas in the medical field suffer from a shortage of medical professionals, which leads to delays in life-saving measures [7] [8].

The field of artificial intelligence (AI) has made significant progress in recent years, particularly in medical imaging and diagnostics. AI models based on deep learning have shown remarkable performance in detecting, segmenting and classifying brain tumors from medical images such as MRI and CT scans [1]. The most advanced models for image segmentation and classification are Convolutional Neural Networks (CNN) [9], although encoder-decoder-transformer architectures have also been proposed [10-12]. These AI-driven tools promise to expand the capabilities of physicians by providing rapid, accurate, and reproducible diagnoses that are important for early intervention and treatment planning in brain tumor cases. In addition, AI algorithms such as deep learning have many disadvantages.

They are also a black box for predictive interpretation, making decisions that are not easy for human experts to interpret. This lack of transparency represents a major challenge in healthcare, where understanding the reasons behind a diagnosis is critical to gaining the trust of doctors and patients alike. Clinicians need to understand why a particular diagnosis was suggested in order to trust AI recommendations and effectively integrate them into their treatment strategies. Furthermore, imaging of brain tumors has intrinsic complications due to the wide range of tumor appearance, location and morphology [13]. This variability makes it difficult for AI systems to effectively generalize to unknown scenarios in the absence of large and diverse training data. Errors in tumor classification can also lead to inappropriate treatment plans, which can negatively impact patient outcomes.

To solve this problem, the science of explainable AI (XAI) has made rapid progress with its successful application in brain tumor diagnosis. XAI aims to make AI results more understandable to humans and provide insights into algorithmic decision-making. Explainable AI (XAI) helps build trust among healthcare professionals and strengthens collaboration between AI systems and healthcare teams by making AI decisions more transparent [14]. Gradient-Weighted Class Activation Mapping (Grad-CAM) techniques are useful in this regard because they highlight exactly the parts of an image that influenced the model's predictions. This allows physicians to visually review and validate AI-powered diagnostic recommendations, ensuring the technology enhances their expertise rather than obscuring it. This technique not only improves diagnostic accuracy, but also bridges the gap between AI capabilities and actual clinical needs, turning AI into a useful diagnostic assistant rather than a "black box." In neuro-oncology, where accurate and transparent diagnoses are critical, the inclusion of XAI can lead to better decision making and better patient outcomes.

Therefore, in this article, we will discuss the Grad Cam based techniques used in the detection, segmentation and classification of brain tumors, which will help practitioners and researchers gain knowledge and understanding in this field. Therefore, this systematic review is conducted.

The main contributions of this review are:

1. Detailed taxonomy of Grad-CAM variants

2. Evaluation of Grad-CAM's Role in Brain Tumor Detection, Segmentation, and Classification.

3.Impact of Grad-CAM on model interpretability and clinical decision-making.

4.Future Directions for Grad-CAM research in neuro-oncology.

This paper presents a thorough analysis of the most recent techniques for brain tumor detection, segmentation, classification and Grad-Cam methods that are intended to assist medical professionals. Rest paper is organised as follows: Section 2 presents methodology, Section 3 Literature review, Section 4 Results and discussion, section 5 presents conclusions of paper.

2. Methodology

This section presents the paper's methodology for conducting the review study on brain tumor detection, segmentation, classification, and Explainable AI (XAI) techniques that help make AI more transparent in medical diagnoses. This study follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) [15] guidelines for conducting systematic reviews. PRISMA 2020 was adopted because of the precise parameters it provides for conducting rigorous systematic reviews. This review paper therefore follows the recommendations of the guidelines.

2.1 Papers selection

We carefully selected the articles for review by conducting keyword searches and closely analyzing the titles, abstracts, and conclusions. We prioritized recent studies that offered something new or significantly different, all based on specific inclusion and exclusion criteria. Only the articles that met these criteria are discussed in this article. A detailed breakdown of these criteria can be found in Table II. Peer-reviewed manuscripts and conference proceedings from the published online library databases PubMed, IEEE Explore, ScienceDirect, Google Scholar, Springer, and Wilay were searched.

2.2 Search strategy

Six databases (PubMed, IEEE Explore, ScienceDirect, Google Scholar, Springer and Wilay online library) were searched systemically to filter papers.

2.3 Inclusion and exclusion criteria

The criteria for inclusion and exclusion is mentioned in Table I.

Table I Inclusions and Exclusions criteria

Criteria	No.	Description
Inclusions	IC1	Studies related to brain tumor detection, segmentation, classification employing ML, DL and XAI techniques.
	IC2	Published within 2020 to 2024.
	IC3	Published in English language.
	IC4	Paper published in SCI/ ESCI, Scopus indexed journals were selected and Conference proceedings.
Exclusions	EC1	Studies that not conducted based of Brain MRI.
	EC2	Review studies.
	EC3	Data from books, presentations, technical documentations, and website

2.4 Study Selection

The papers were selected using the inclusion and exclusion criteria defined in Table I. A bibliography of 343 papers was compiled from databases. We excluded duplicate studies, systematic reviews, scoping reviews, brain tumor diagnosis without explainable AI, and explainable AI without brain tumor diagnosis. Finally, 31 studies using Grad-Cam for brain tumor detection, segmentation, and classification were included and utilized in the systematic review.

3. Results

3.1 Detailed Taxonomy of Grad-CAM Variants

Figure 2 illustrates the development of Gradient-weighted Class Activation Mapping (Grad-CAM) techniques. In 2017, the original Grad-CAM was proposed and has since been continually refined to make the deep learning models more interpretable - which meant better visualization of areas on an image that were important to the model's prediction of a particular tumor type . Some key milestones include Grad-CAM++ in 2018, which enables more accurate localization, and Smooth Grad-CAM++ in 2019, which minimizes noise for better visual explanations. Gradient class activation mapping methods such as Score Grad-CAM (2020) and Layer Grad-CAM (2021) also represented an improvement with more flexibility in model interpretability. Now looking to the future: 2D Grad-CAM for 2024 leads us to more Innovations that produce models that have even more advanced and user-friendly interpretability tools. This development is a sign of the research community's commitment to making AI more interpretable and understandable.

Grad-CAM (Gradient-weighted Class Activation Mapping): Grad-CAM is frequently used in medical imaging to visually highlight parts of an image that were most important to the model's decision. For example, in brain tumor detection, Grad-CAM can identify regions in MRI scans that contributed to the model's prediction, offering a heatmap for better clinical validation [16].Gradient-weighted class activation mapping (or Grad-CAM) was proposed in 2016 as a framework for visually interpreting the decisions of CNNs [16]. It creates heatmaps for each image that show how much influence each region has on model

predictions by calculating gradients with respect to the final convolutional layer of the CNN model. The main application of Grad-CAM in neuro-oncology is that it helps to visually explain the output of the AI model by marking areas of MRI or CT scans that have a large impact on prediction [28], thus helping clinicians gain visual insight into the output of AI. However, Grad-CAM may not be sufficient for fine-grained, precise segmentation of tumor boundaries due to its coarse heatmaps. To overcome these limitations, Grad-CAM++ was developed, which leverages pixel-level importance and higher-order derivatives to refine this interpretation, enabling high-resolution heatmaps to be generated that are more fine-grained and accurate[17]. This makes it very suitable for early-stage tumor detection, which involves detecting smaller or scattered tumor regions. However, Grad-CAM++ requires more computational effort due to the higher order gradients.

Smooth Grad-CAM++ integrates the noise-reducing principles of SmoothGrad to deal with noise in medical imaging contexts, resulting in smoother, more stable heatmaps [19]. Although this method is effective in noisy data environments such as MRI, the technique is computationally intensive, which limits its practical application in real-time cases. Score-CAM, a gradient-free variant, increases stability by using model confidence scores instead of gradients. This strategy, in which regions are iteratively masked to assess their impact on model reliability, is particularly useful when gradient reliability is weakened but the masking processes incur significant computational cost [20].

Ablation-CAM is another method that performs neuron ablation in the last convolutional layer and measures its impact on model predictions [21]. This variant provides detailed explanations, determines which neurons contributed to the final decision, and improves the interpretability of complex models. However, the computational cost is too high because a large number of runs are required to be practical in real-time diagnosis. Layer-CAM addresses this limitation of Grad-CAM by extending it to multiple CNN layers for broad interpretability, capturing both low-level features (edges, textures) and high-level features (shapes, patterns). This makes layer-CAM advantageous for multi-layer analysis, such as detailed cases of brain tumor segmentation [22].

Numerical Grad-CAM, introduced by [23], improves the stability of Grad-CAM by computing numerical gradients instead of analytical gradients, which helps mitigate problems associated with gradient disappearance or explosion, resulting in greater precision and reliability in the diagnosis of brain tumors [23].

Finally, 2D Grad-CAM was used by Tehsin et al. designed for slice-level analysis in MRI and CT scans, providing computationally efficient heatmaps for individual 2D slices. Spatial attention modules are used to better focus on disease-relevant areas within each layer and thus enable efficient, targeted diagnostic insights [24]. However, this 2D technique may miss certain 3D contextual information required for larger or irregularly shaped tumors, which is still a factor in complete tumor analysis.

The **Table I** provides a comparative overview of Grad-CAM and its various enhancements over time, highlighting their distinct improvements, limitations, and best-suited applications. Starting with the original Grad-CAM in 2017, which laid the foundation for class activation mapping, each subsequent variant addresses specific limitations in interpretability, spatial precision, or computational efficiency.

Table I: Comparative analysis of Grad Cam and its variants.

Sr. No	Variant	Year	Key Improvement	Mathematical Basis	DL method used	Best Suited For	Limitations
1	Grad-CAM[16]	2017	Provides basic class activation mapping	Uses gradients from the last convolutional layer	CNNs	Common interpretation	1.Coarse heatmaps. 2.limited spatial precision. 3.Unsuitable for fine-grained tasks like segmentation.
2	Grad-CAM++[17]	2018	Provides fine-grained details with higher-order gradients	Uses pixel-wise weighted gradients	CNNs, Multi-Object Detection	Small, detailed regions	1.Higher computational cost due to pixel-wise gradient calculations. 2.less efficient in large images
3	Pyramid Grad-CAM[18]	2018	Provides enhanced multi-scale localization for complex regions	Uses multi-scale feature fusion from various CNN layers	CNNs, Multi-Scale Feature Extraction	Brain tumor localization and segmentation	1.High computational demand due to multi-scale processing.
4	Smooth Grad-CAM++[19]	2019	Provides noise reduction with multiple perturbed inputs	Uses averaged heatmaps from perturbed inputs	CNNs, Noise Reduction	Noisy environments (e.g., MRI scans)	1.Requires multiple perturbed inputs, which can be slow. 2.less suitable for time-sensitive settings.
5	Score-CAM[20]	2020	Provides gradient independent, stable outputs	Uses masking regions and model confidence scores	CNNs, Confidence Scoring	Situations with unstable gradients	1.Computationally expensive due to iterative masking, not ideal for high-resolution image analysis
6	Ablation-CAM[21]	2020	Provides Neuron ablation for targeted relevance	Systematic neuron ablation	CNNs, Neuron Ablation	Complex models with many layers	1.Requires multiple forward passes for ablation, making it impractical for real-time usage
7	Layer-CAM[22]	2021	Provides Multi-layered explanations for richer insights	Grad-CAM computed for multiple layers	CNNs, Layer-wise Feature Analysis	Detailed segmentation and hierarchical analysis	1.Complex heatmaps may complicate interpretation 2.increased computational load.
8	Numerical Grad-CAM[23]	2022	Provides enhanced stability and interpretation through numerical gradients	Numerical gradient computation via input perturbations	CNNs, Numerical Differentiation	Precise and reliable localization in brain tumor diagnosis	1.Sensitive to perturbation parameters. 2.high computational cost, requires tuning for accuracy.
9	2D Grad-CAM[24]	2024	Provides slice-level efficiency and spatial attention	Calculate gradient for each 2D slice and spatial attention modules	CNNs, Spatial Attention Mechanism	Fast and focused 2D slice interpretation	1.Limited to 2D images. 2.loses contextual information from 3D scans. 3.less effective in 3D analysis

For example, Grad-CAM++ (2018) proposed pixel-by-pixel gradients for more detailed localization, but required more computational effort. Pyramid Grad-CAM (2018) and Layer-CAM (2021) improve multiscale and multilayer functions, making them suitable for complex multi-object analyzes such as brain tumor segmentation. Instead of the Score-CAM (2020) and Ablation-CAM (2020) gradients, confidence values or neuron ablation are used, which improves interpretability in unstable gradient cases but requires high processing time.

Numerical Grad-CAM (2022) and 2D Grad-CAM (2024) are newer methods that offer greater stability and efficiency, with Numerical Grad-CAM increasing precision for medical imaging tasks and 2D Grad-CAM improving speed for slice-level interpretations Limitations in 3D contexts. Overall, each variant reflects an attempt to balance interpretability, computational cost, and suitability for specific tasks, and demonstrates progressive refinement of AI interpretability techniques.

3.2 Evaluation of Grad-CAM's Role in Brain Tumor Detection, Segmentation, and Classification

In this section we are going to review and compare the recent literature from 2017 to 2024 based of utilization of Grad cam based techniques in brain tumor detection, segmentation and classification.

Table II: Literature review and Comparative analysis of MRI-based Brain Tumor Detection Studies Using Explainable AI

Sr.No	Study	Year	Imaging Modality	ML/DL/TF	Task	XAI	Dataset	Limitations
1	Selvaraju et al.,[16]	2017	MRI	CNN	Detection	Grad Cam	PASCAL VOC 2007 and ImageNet	1.Gradient-CAM is based on the gradient flowing into the last convolutional layer of a CNN. 2.Limited to only.
2	Pereira et al., [25]	2018	MRI	CNN	Classification	Grad Cam	BRATS 2017	1.The model's performance is highly reliant on the availability of well-annotated MRI datasets. 2.Limited generalizability. 3.Risk of overfitting due to the small size of the dataset used for training.
3	Lee et al., [18]	2018	MRI	CNN	Detection	PG CAM	image dataset of 550 patients.	1.Precision and accuracy for tumor localization are still relatively low. 2.The model relies on weakly supervised learning using image-level annotations rather than pixel-wise. 3.The dataset used for training has a high class imbalance.
4	Ahmad et al.,[26]	2019	T2 Weighted MRI	DL-CNN	Detection and classification	CAM	41 subjects with IDH1 mutant genotype and 30 subjects with wild-type IDH1 genotype	1.Dependency on T2-weighted MRI. 2.Limited sample size. 3.The authors did not mention external validation on independent datasets.
5	Nateka et al.,[27]	2020	MRI	SimUnet, ResNet,DenseUnit With Grad Cam	Segmentation	Grad Cam	BRATS 2018	1.Validation thorough additional dataset is required. 2.Limited generalizability. 3.The study focuses on 2D models due to the challenge of analyzing and visualizing interpretability metrics in 3D models.
6	Pintelas et al.,[28]	2020	MRI	DT,NB,LR,NN,SVM,KNN	Classification	Grad-CAM	Dataset of 3064 head images with glioma, meningioma, and pituitary tumors.	1.The success of the framework relies heavily on the feature extraction process. 2.Trade-off between Accuracy and Interpretability. 3.Limited Generalizability.
7	Windisch et al.,[29]	2020	MRI T1-weighted, T2-weighted	Resnet50, and Bayesian neural network	Detection	Grad-CAM	TCGA-GBM	1.The model showed a tendency to focus on irrelevant areas. 2.The MRI dataset used for training came from different views.

								and acquisition sequences. 3.Limited Generalizability
8	Esmaili et al.,[30]	2021	MRI T1-weighted, T2-weighted, and FLAIR	DenseNet-121, GoogLeNet, and MobileNet,	Detection	Grad-CAM	TCGA CIA	1.A significant number of tumors were incorrectly classified. 2.Grad-CAM struggled with complex images containing multiple objects. 3.The study had limited patient data.
9	Saleem et al.,[31]	2021	MRI T1-weighted, T2-weighted, FLAIR, and T1ce	DMFNet, Guided Backpropagation (GB)	Segmentation	Grad-CAM	BraTS-2018	1.The model sometimes fails to predict specific tumor regions. 2.Limited generalizability. 3.The T1-weighted sequence was found to be the least useful in tumor segmentation compared to FLAIR, T1ce, and T2 sequences.
10	Jin et al., [32]	2022	MRI T1-weighted, T2-weighted, FLAIR, and T1ce	CNN,	Classification	Grad Cam and SHAP	BraTS 2020	1.Current XAI Algorithms Are Not Clinically Designed. 2.Most XAI algorithms struggle to highlight the importance of different image modalities, such as T1, T2, or FLAIR.
11	Zeineldin et al.,[33]	2022	MRI native T1W, Gadolinium T1Gd, T2W, and FLAIR	ResNet-50, 3D DeepSeg, Vanilla gradient, guided backpropagation,	Segmentation and Classification	Integrated gradients, Guided integrated gradients, SmoothGrad, Grad-CAM, and guided Grad-CAM.	BraTS 2019 and 2021	1.Lack of Interpretability of Deep Learning Models. 2.Current explainability methods struggle to effectively interpret complex multi-modal data. 3.Many existing XAI methods focused on 2D MRI slices.
12	Dasanayaka et al., [34]	2022	MRI and WSI	DenseNet, ResNet	Classification	Grad-Cam	MRI sequences and WSI	1.Source of dataset is not mentioned. 2.Limited generalizability. 3.The tool currently supports a limited number of imaging modalities.
13	Maqsood et al.,[35]	2022	MRI	MobileNetV2, Multiclass SVM	Detection	Grad Cam	Figshare and BraTS 2018	1.Time-consuming feature selection method. 2.The approach was only tested on 2D MRI images. 3.High computational resources are required for processing.

14	Marmolejo-Saucedo et al., [36]	2022	MRI T1-weighted, T2-weighted, FLAIR, and T1ce	CNN	Detection and classification	numGrad Cam	BRATS 2017 .	1.Limited generalizability 2.There is a need for further validation. 3.The model requires significant computational resources.
15	Hussain et al., [37]	2023	MRI	VGG-19, scratch VGG-19,	Detection and classification	CAM, Grad-CAM, and Grad-CAM++	Kaggle:Brain MRI Images for Brain Tumor Detection Kaggle:Brain Tumor Classification	1.Extensive computational resources and time is required 2.Pre-trained models rely heavily on external datasets like ImageNet for transfer learning
16	Rahman et al., [38]	2023	MRI	lightweight CNN with Grad-CAM	Detection	Grad-CAM	Figshare	1.Validation thorough additional dataset is required. 2.Limited generalizability 3.Dependence on MRI Imaging.
17	Yan et al., [39]	2023	MRI T1-weighted, T2-weighted, FLAIR, and T1ce	3D nnU-Net model	Segmentation and Classification	Grad-Cam	BraTS 2018	1.Required significant computational resources 2.Framework heavily relies on availability of datasets. 3.Limited generalizability
18	Taşci et al., [40]	2023	MRI	DenseNet201 and SVM	Classification	Grad-Cam	Kaggle: brain tumor dataset and Figshare brain tumor dataset	1.The study cannot address the model's real-world applicability. 2.The paper highlights a high computational load.
19	Özbay et al., [41]	2023	MRI	CNN based on mRMR,DarkNet53, EfficientNetB0, DenseNet201, SVM, KNN, and Ensemble algorithms	Detection	Grad-Cam	Br35H 2020	1.Limited generalizability 2.The proposed model focuses only on binary classification 3.Additional validation is required. 4.Computationally intensive model
20	Valerio Ponzi et al., [42]	2023	MRI T1-weighted, T2-weighted, FLAIR, and T1ce	U-Net Model,R-CNN, DeepLab V3,	Segmentation	FCN Grad-CAM, and SHAP	BRATS 2020	1.Moderate Survival Prediction Accuracy 2.No Exploration of Advanced Techniques 3.Validation through additional dataset is required.
21	Šefčík et al., [43]	2023	T1-weighted and T2-weighted, and FLAIR	DNN	Classification	LRP, Grad Cam	BraTS 2020	1.Limited generalizability 2.Model neglects other potentially relevant features in the MRI images that could contribute to classification
22	Sarah et al., [44]	2024	MRI	Kmeans++, SGLDM,VGG16,VGG19,ResNet50	Detection	Grad Cam	Kaggle's: Br35H 2020	1.Data Quality Sensitivity.The performance heavily depends

									on the quality the input MRI images. 2.The approach requires significant computational resources. 3.Limited generalizability
23	Nhlapho et al., [45]	2024	MRI	EfficientNetB0, DenseNet121, Xception	Detection	Grad-CAM and Grad-CAM++.	Kaggle:sartajbhuvaji	1.The study primarily focus on binary classification 2.Models like VGG16 and VGG19 are computationally expensive.	
24	Tehsin et al., [46]	2024	MRI	DaSAM Based on CNN	Detection	2D Grad Cam	Kaggle: Brain Tumor Classification (MRI) and Figshare brain tumor dataset.	1.Limited generalizability 2.The model focuses on classification rather than pixel tumor segmentation 3.The author mention an imbalance in data distribution	
25	M, M. M et al.,[47]	2024	MRI	ResNet50	Detection	Grad-CAM	Kaggle: Brain MRI Images for Brain Tumor Detection	1.Limited generalizability 2.ResNet50, 50 layers, require substantial computational resources for training and inference.	
26	Zeineldin et al.,[48]	2024	MRI T1-weighted, T2-weighted , FLAIR, and T1ce	CNN and Grad Cam	Segmentation	Grad-Cam	BraTS 2019 and FeTS2022 datasets.	1.The current implementation focuses on 2 axial MRI slices 2.The hybrid architecture requires significant computational resources.	
27	Rahman et al.,[49]	2024	MRI	CNN	Classification	SHAP and Grad Cam++	BraTS 2020	1.The dataset used is imbalanced. 2.Limited generalizability	
28	Saeed et al., [50]	2024	MRI	DeepLabV3, ResNet18, DarkNet53, MobileNetV2 ,SVM	Segmentation and Classification	Grad-CAM	Brats 2021	1.Limited generalizability Increased computational load, making 2.model resource intensive.	
29	Nazir et al., [51]	2024	MRI	Customised CNN	Detection	SHAP, LIME, and GRAD-CAM	Kaggle: BR35H 2020	1.Small dataset size. 2.Limited generalizability 3.Focuses on binary classification	
30	Mzoughi et al.,[52]	2024	MRI	ViT,CNN	Classification	Grad-CAM, LIME, and SHAP	Figshare Dataset	1.Limited generalizability 2.The ViT model is more accurate but has higher computational demands than CNN.	
31	Amin et al.,[53]	2024	MRI	Inception-v3, DenseNet201	Detection	LIME and Grad Cam	Kaggle:Brain Tumor MRI Dataset	1.Limited generalizability 2.The integration of PIDL and explainability	

Grad-CAM proposed by Selvaraju et al. [16] was a significant contribution to the field of XAI. It is a powerful tool for visually highlighting crucial decision areas in convolutional neural networks (CNN). Their study showed that Grad-CAM significantly improved human interpretability when applied to MRI data, outperforming other approaches such as guided backpropagation by 16.79%. However, it is limited to CNN models and largely depends on gradients in the final convolutional layer, which limits its general applicability. Using the BRATS 2017 dataset and Grad-CAM, Pereira et al. [25] performed tumor classification and achieved a high accuracy of 92.98%. Despite these results, the model's reliance on a small but well-annotated dataset raises concerns about overfitting and limited generalizability.

PG-CAM is another one developed by Lee et al. developed variant. [18], which achieved a remarkable accuracy of 95.6% in tumor detection. However, its precision in tumor localization was limited because it relied on weakly supervised learning with image-level annotations. Ahmad et al. [26] used the ResNet CNN model with CAM and achieved 86.7% accuracy on T2-weighted MRI images. The model used a small data set, resulting in limited performance and a lack of external validation. Similarly, Nateka et al. [27] used Grad-CAM with SimUnet, ResNet and DenseNet models to obtain comparable Dice scores for tumor segmentation. Despite its success, this study encountered problems with interpretability in 3D models. Many research studies used multimodal imaging and ensemble approaches to increase accuracy and interpretability. Windisch et al. [29] implemented ResNet-50 with Bayesian neural networks and achieved a category accuracy of 93%; However, the model often highlighted unimportant areas on the MRI. They study [30] how to implement DenseNet-121 and Grad-CAM to achieve good classification and localization accuracy. However, misclassifications occurred with complex images. Furthermore, Jin et al. [32] expressed concerns about the clinical effectiveness of 16 XAI algorithms and indicated that none met the criteria for clinical decision support. Recent research studies have also focused on lightweight structures. Rahman et al. [38] developed NeuroXAI++, a grade CAM with lightweight CNN, which achieved near-perfect accuracy, precision, and recall. However, generalizability remains an issue as the model was validated on a single data set. Other methods, such as those used by Hussain et al. [37] used pre-trained VGG-19 models with Grad-CAM++ to achieve 99.92% accuracy, but their dependence relies on external datasets such as ImageNet for transfer learning.

Nazir et al. [51] in their study implemented techniques such as SHAP and LIME integrated with Grad-CAM, which resulted in 100% accuracy in binary classification, but resulted in a lack of generalizability due to limited dataset sizes. The study [52] compared ViT with CNN models for multi-classification tasks and concluded that ViT was more accurate but required high computational resources. The Amin et al. [53] introduce an XAI-assisted MRI analysis framework that uses Physics-Informed Deep Learning (PIDL) with LIME and Grad-CAM for brain tumor diagnosis and achieves 96% accuracy. It improves interpretability but is limited by dataset generalization and computational complexity. The framework shows potential for broader medical applications in consumer health.

3.3 Impact of Grad-CAM on Model Interpretability and Clinical Decision-Making

Grad-CAM has made a great contribution with noticeable impact in enhancing model interpretability within medical imaging, and it is very helpful for clinical decision-making. Grad-CAM uses visual heatmaps to explain why a deep learning model made a particular decision. These heatmaps highlight the image areas that had the most impact on the model's decision. This method has helped to bridge the gap between AI models and clinical decision-making, making complex neural network predictions more transparent. Research indicates that Grad-CAM enhances clinicians' trust by highlighting the brain regions where the model identified signs of tumors. This has been particularly useful in applications such as brain tumor classification and grading, where precision and interpretability are essential for patient care [16] [25]. Moreover, Grad-CAM has demonstrated its significance in correlating model decisions with clinical decision-making, frequently indicating whether a model is erroneously concentrating on irrelevant parts of an image. This feedback is beneficial for researchers aiming to enhance diagnostic models, as it enables them to make modifications that ensure models focus on pertinent areas of interest, hence boosting reliability and accuracy in essential diagnostic applications [29] [32].

However, Grad-CAM does have its limitations. Its dependence on gradient information can occasionally create noise, causing the model to highlight areas that may be less clinically relevant. Additionally, it may struggle to interpret complex, multi-modal MRI data accurately. These limitations suggest that Grad-CAM is often most powerful when used in tandem with other interpretability tools to achieve a more holistic view [28] [34]. Overall, Grad-CAM has had a significant impact on AI prediction capabilities in healthcare, fostering a greater level of trust and transparency between clinicians and AI systems. Grad-CAM enhances the interpretability and clinical relevance of predictions, leading to safer and more informed AI integration in clinical workflows, which is crucial for advancing explainable AI in healthcare [45] [53].

3.4 Future Directions for Grad-CAM Research in Neuro-Oncology

Future research in neuro-oncology for Grad-CAM will focus on broadening its clinical importance and enhancing its role in facilitating diagnosis and treatment planning. An essential goal is to enhance the accuracy of Grad-CAM in order to provide more precise and detailed visual explanations in various medical imaging scenarios. Other methods like Grad-CAM++ and SmoothGrad are being studied to generate more detailed heatmaps, aiming to improve the accuracy in distinguishing between different types of brain tumors [37] [38]. Another area showing great potential is the implementation of Grad-CAM for three-dimensional (3D) MRI data, allowing for more comprehensive analysis of tumors based on volume. At present, the majority of neuro-oncology models rely on two-dimensional (2D) MRI slices. Researchers aim to provide clinicians with a more comprehensive understanding of tumor structure by expanding Grad-CAM to 3D, but face obstacles due to the significant computational resources needed [27] [33].

Merging Grad-CAM with other explainability techniques, such as SHAP and LIME, offers a fascinating area to improve the multidimensionality and robustness of model interpretations. Integrating these methods could yield richer insights into how the AI model reaches its conclusions, thereby enhancing the trust

clinicians place in these tools. By incorporating these techniques, Clinicians' trusts in these tools may be increased, which may provide deeper insights into the AI model's decision-making process [42] [52].

Finally, it is crucial to prioritize evaluating Grad-CAM models on broader and more varied datasets to guarantee their effectiveness across diverse clinical environments. Numerous ongoing research projects use particular datasets like BraTS or Figshare, which might not completely capture the diversity present in actual clinical scenarios. In the future, researchers must collaborate with bigger and more diverse datasets that encompass a wider array of scanner types, imaging protocols, and patient demographics. This increase would enhance the reliability and versatility of Grad-CAM-based models in various healthcare environments [16] [31]. Summarizing, future research aims to improve Grad-CAM's spatial precision, apply it to 3D MRI, integrate it with other XAI techniques, and enhance dataset variation to enhance the practicality and reliability of AI-driven neuro-oncology tools for clinical applications. These enhancements will ultimately aid in better, clear decision-making for diagnosing and planning treatment for brain tumors [39] [51].

4. Discussion

In this section, we provide a detailed discussion of the key findings related to machine learning (ML), deep learning (DL) and Grad-Cam for brain tumor detection, segmentation, and classification using MRI data. We also examine the role of Grad- Cam, dataset availability, and common limitations in the current literature.

4.1 Investigating the best current ML, DL techniques used in detection, segmentation and classification of brain tumor using MRI.

4.1.1 Detection Techniques

Detection is often the initial step in identifying the presence of a brain tumor within MRI scans. For this purpose, several ML and DL models have been explored, aiming for high sensitivity and early diagnosis capabilities.

- **Convolutional neural networks (CNNs):** CNNs have emerged as the preferred technique for tumor detection because of their superior performance in visual data analysis. Due to its ability to achieve high detection accuracy, architectures like ResNet50, VGG-19, and MobileNetV2 are commonly used [35] [37].
- **Region-Based CNNs (R-CNNs):** Region-based CNNs provide a more targeted approach for more accurate detection, especially in identifying the tumor location within larger MRI scans. R-CNN variants focus on specific regions that may contain tumors and are especially useful when detection needs a higher degree of spatial accuracy.
- **Physics-Informed Deep Learning (PIDL):** The main objective of PIDL is to integrate physics-based principles with deep learning algorithms. This approach helps to improve the overall detection accuracy by guiding the model with MRI-specific physics [53].

4.1.2 Segmentation Techniques

Once the tumor has been detected, segmentation is the next step, which involves outlining the tumor regions within MRI scans. This process is crucial for analyzing tumor size, shape, and growth.

- **U-Net and Variants:** Due to its high precision and relatively low computational cost, U-Net has emerged as benchmark for medical image segmentation. It can segment multi-dimensional MRI data such as 3D U-Net, which is essential for volumetric analysis of brain tumors [33].
- **DeepLab and Fully Convolutional Networks (FCNs):** These networks are designed for pixel-wise segmentation, which is highly useful in brain tumor cases. DeepLab's use of extended convolutions enables the segmentation of fine details in MRI images, making it effective for distinguishing between different tissue types, such as tumor cores and peritumoral regions [32].
- **3D Segmentation Models:** Because MRI scans often include multiple image slices, 3D segmentation models – such as the 3D DeepSeg model – have become increasingly important. These models provide a more comprehensive view by segmenting across three dimensions, but require significant computational resources to function effectively [48].

4.1.3 Classification Techniques

The final step is classification, where the segmented tumor is analyzed to determine its type and grade. This information is essential for prognosis and treatment planning.

- **Support Vector Machines (SVM) with CNN Features:** SVMs are often used as a classifier after extracting CNN features. This combination leverages the powerful feature extraction capabilities of CNN with the efficient classification of SVM, as seen in multiclass tumor classification, achieving high accuracy [35].
- **Multilayer Perceptrons (MLPs):** Although less commonly used than CNNs, MLPs are occasionally used in conjunction with other feature extraction methods for classification tasks, particularly when a simpler model is required for lower computational cost.
- **Transfer Learning with Pre-trained Networks:** Models like VGG-19 trained on large image datasets can be optimized for brain tumor classification. Transfer learning allows these models to use pre-learned features, improving classification accuracy on smaller, domain-specific datasets [37]
- **Explainable AI (XAI) Models with Grad-CAM:** As interpretability becomes increasingly important in clinical contexts, Grad-CAM (Gradient-Weighted Class Activation Mapping) is widely used alongside classification networks. By creating heatmaps that highlight which areas influenced the model's

predictions, Grad-CAM helps clinicians validate the model's conclusions [42].

4.2 Investigating the XAI methods used in **detection, segmentation and classification of brain tumors using MRI**.

To thoroughly understand the role of explainable AI (XAI) in brain tumor detection, segmentation and classification using MRI, we examine the different XAI methods used in these phases. The main goal of integrating XAI is to improve the interpretability and trustworthiness of AI models and help clinicians better understand how models make predictions, which is critical for high-risk medical applications.

4.2.1 XAI in Detection of Brain Tumors

In tumor detection, XAI is used to clarify why a model flags certain regions within an MRI scan as potential tumors. This step is especially valuable for early diagnostics, where interpretability can support early treatment decisions.

- **Saliency Maps:** Saliency maps are a popular XAI method that highlights areas in an MRI scan that have the greatest influence on model predictions. By visualizing these areas, doctors can understand what the model is focusing on, thereby verifying that the model has correctly identified potential tumor regions. Saliency maps are often used in conjunction with convolutional neural networks (CNNs) and are easy to implement and interpret [35].
- **Grad-CAM (Gradient-weighted Class Activation Mapping):** Grad-CAM has become an essential tool for tumor detection because it creates heatmaps that show which regions contribute the most to the model's predictions. Grad-CAM is particularly effective on CNNs, making it ideal for brain tumor detection applications. By applying Grad-CAM, clinicians can examine model-predicted tumor areas and determine whether the highlighted regions are consistent with medical knowledge [53].

4.2.2 XAI in Segmentation of Tumor Regions

Segmentation focuses on delineating the exact boundaries of the tumor within MRI scans, a step that requires high precision and interpretability.

- **SHAP (SHapley Additive exPlanations)** SHAP (SHapley Additive exPlanations): SHAP values are often used in segmentation models to explain the contribution of each pixel to the prediction. In brain tumor segmentation, SHAP helps interpret the boundaries of the segmented region by explaining pixel-wise contributions. Due to its theoretical basis in game theory, SHAP is particularly suitable for interpretations with high accuracy, although it can be computationally intensive [33].
- **Integrated Gradients:** Integrated Gradients provide another method of pixel-level interpretability by showing how each pixel in an MRI scan contributes to the model's decision. This technique is advantageous in MRI segmentation, where the relevance of each voxel must be clearly identified to distinguish tumor boundaries from surrounding tissue. Built-in gradients can increase confidence in model predictions by displaying stepwise contributions from pixels or voxels [32].
- **Attention Mechanisms:** Some segmentation models use attention layers that assign greater focus to specific MRI regions based on their relevance, essentially creating internal "heatmaps." This attention-based segmentation approach is particularly useful in multimodal MRI, where different types of MR images provide different information, enabling transparent and interpretable segmentation results [48].

4.2.3 XAI in Classification of Tumor Types and Grades

Once a brain tumor is detected and segmented, the next step is to classify the tumor by type and grade. Here, XAI methods support clinical decision-making by explaining the features that lead to each classification outcome.

- **LIME (Local Interpretable Model-Agnostic Explanations):** LIME is a model-agnostic provides local interpretability by generating explanations for individual predictions. This involves sampling the input data and observing how small perturbations affect predictions. In tumor classification, LIME can show which MRI features contribute to the diagnosis of benign or malignant, thereby helping physicians verify whether the model uses medically relevant features in its classifications [42].
- **Grad-CAM for Classification Models:** Grad-CAM was originally designed to visualize CNNs and has proven useful in classification models by highlighting regions that the model considers important for specific tumor types or grades. For example, when distinguishing high-grade from low-grade tumors, Grad-CAM can show which regions are most significant on MRI, allowing doctors to confirm whether these match known tumor characteristics. The resulting heatmaps help increase physician confidence and aid in model evaluation [35].
- **Counterfactual Explanations:** Counterfactual explanations are a new XAI technique that generates alternative scenarios to illustrate how subtle changes in inputs can lead to different classifications. For brain tumor classification, for example, counterfactual calculations can show how the model's classification of a tumor could change through minor adjustments to certain features. This method provides insights into decision boundaries and helps clinicians assess the reliability and sensitivity of the model [53].

4.3 Studying the impact of code of paper is publically accessible.

The availability of code is critical to reproducibility and future study. According to our evaluation, the majority of the publications did not openly publish their code. As shown in Table III, only eight out of the 31 reviewed publications specifically acknowledged their code and were made publicly available. The lack of

open-source implementations hinders other researchers' capacity to reproduce or expand on the findings of these investigations, which is a typical difficulty in medical AI research. However, some recent studies have begun to share their code, indicating a favorable trend toward open science.

Table III: depicting the code availability count in research studies

Code Availability	Yes	No
Count	8	23

4.4 Investing the most widely used datasets in brain tumor detection, segmentation and classification.

Several well-established datasets have been consistently used for brain tumor analysis in the literature. Based on the provided data, the most widely used datasets in brain tumor detection, segmentation, and classification are:

BraTS (2017, 2018, 2019, 2020, 2021): This dataset appears frequently and is the most popular for brain tumor segmentation tasks.

TCGA (GBM, LGG, CIA): The Cancer Genome Atlas datasets are widely used for segmentation and classification applications.

Kaggle Brain Tumor datasets: Several studies use Kaggle datasets, including "Brain MRI Images for Brain Tumor Detection" and "Brain Tumor Classification (MRI)."

Figshare: Brain tumor detection dataset is used in study.

Other Datasets: Additional datasets, totaling five, contribute to the diversity of sources in brain tumor research.

These datasets, particularly BraTS and Kaggle, are fundamental to brain tumor analysis research.

Table IV: Brain Tumor Analysis Dataset Usage Summary

Datasets	BRATES 2017	BRATES 2018	BRATES 2019	BRATES 2020	BRATES 2021	BR35H	Kaggle	Figshare	TCGA	Others.
Count	2	4	2	6	2	2	7	4	2	5

4.5 Summarising the common limitations of existing studies.

Despite the progress in deep learning for brain tumor detection, segmentation, and classification, several challenges remain:

- **Generalization Issues Across Datasets:** Many studies rely on specific, limited datasets that may not generalize well to broader populations or MRI systems. This poses a challenge for clinical deployment, as models often perform inconsistently across different patient demographics and imaging techniques [35].
- **Computational Complexity:** Techniques such as SHAP and Integrated Gradients, while effective, are computationally demanding. This complexity makes them challenging to apply in real-time clinical settings, where quick and efficient analysis is often required [33].
- **Interpretability vs. Accuracy Trade-off:** While simpler models are more interpretable, they often sacrifice accuracy compared to complex, black-box models. This trade-off remains a key limitation, as models that balance both accuracy and interpretability are challenging to develop [32].
- **Reliability of XAI Techniques:** XAI methods such as Grad-CAM and LIME may not always provide consistent or accurate explanations. This inconsistency can create issues for clinical decision-making, as the explanations may not fully align with medical knowledge or accurately reflect the model's reasoning [42].
- **Lack of Clinical Validation:** Few studies conduct rigorous clinical validation of their XAI methods. Consequently, the clinical utility and reliability of these methods remain largely theoretical, underscoring the need for extensive testing in real-world healthcare environments [53].
- **Limited Integration of Multimodal Data:** Most current studies focus on single-modality MRI, overlooking the potential of integrating multimodal imaging (e.g., combining MRI with PET or CT). This limitation restricts the scope of XAI applications, as combining different imaging types could improve diagnostic accuracy and interpretability [48].
- **Challenges in Fine-tuning for Specific Clinical Applications:** Developing XAI models that are finely tuned for particular diagnostic tasks, such as differentiating between tumor grades or types, remains an ongoing challenge. Often, these methods need further customization to meet the nuanced needs of specific clinical applications in neuro-oncology [53].

5. Conclusions

In summary, the application of explainable AI (XAI) techniques such as Grad-CAM, SHAP and LIME represents a significant advance in the field of brain tumor diagnostics. These methods help clarify AI-driven decisions and provide clinicians with insights into how models interpret MRI data. Research shows that XAI can improve interpretability and sometimes increase accuracy by highlighting critical tumor features [16] [33].

Despite this promise, important challenges remain. The generalizability of the model is a major limitation as most studies are based on specific datasets, making it difficult to reproduce results across different MRI systems and patient groups. The high computational cost of techniques such as SHAP also limits real-time application in clinical environments [42] [36]. Furthermore, although XAI improves interpretability, this often comes at the expense of accuracy and

potentially compromises diagnostic reliability. Furthermore, current research tends to focus on single-modality MRI and misses opportunities for improved precision through multimodal imaging. The integration of data from MRI, CT or PET could provide more comprehensive insights, but is still little researched [34]. Additionally, few models undergo the rigorous clinical testing required for safe and effective use in healthcare settings.

To move forward, research should focus on improving generalizability, computational efficiency and clinical validation, with the ultimate goal of making XAI a reliable tool in neuro-oncology. Solving these issues will be critical to enabling physicians to make transparent and accurate decisions based on AI-powered insights.

Declarations

Author Contribution

Krishan Kumar: the main author, took part in the conception, design, and implementation of the study; data collection and analysis; and drafting and revision of the manuscript. Dr. Kiran Jyoti: Supervisor, Right from the conception to the final version of the work—contributed to the design and methodology of the study; critical revisions and feedback on the paper were given; ensured overall quality and integrity of the work

Funding

No Funding

Ethical compliance

No

Declaration of competing interest

No conflict of interest in this work.

Data availability

Data can be shared as and when required by contacting the author.

References

1. Al-Galal, S.A.Y.; Alshaikhli, I.F.T.; Abdulrazzaq, M.M. MRI brain tumor medical images analysis using deep learning techniques:A systematic review. *Health Technol.* 2021, 11, 267–282.
2. Louis, D.N.; Perry, A.; Wesseling, P.; Brat, D.J.; Cree, I.A.; Figarella-Branger, D.; Hawkins, C.; Ng, H.K.; Pfister, S.M.;Reifenberger, G. The 2021 WHO classification of tumors of the central nervous system: A summary. *Neuro-Oncology* 2021, 23,1231–1251.
3. Nodirov, J.; Abdusalomov, A.B.; Whangbo, T.K. Attention 3D U-Net with Multiple Skip Connections for Segmentation of Brain Tumor Images. *Sensors* 2022, 22, 6501.
4. <https://www.cancer.net/cancer-types/brain-tumor/statistics>
5. C. Wang et al., “Phenotypic and genetic associations of quantitativemagnetic susceptibility in U.K. biobank brain imaging,” *Nature Neurosci.*,vol. 562, pp. 1–14, May 2022.
6. Z. Liu et al., “Deep learning based brain tumor segmentation: A survey,”*Complex Intell. Syst.*, vol. 9, no. 1, pp. 1001–1026, 2020.
7. M. A. Ottom, H. A. Rahman, and I. D. Dinov, “ZNet: Deep learningapproach for 2D MRI brain tumor segmentation,” *IEEE J. Transl. Eng. Health Med.*, vol. 10, May 2022, Art. no. 1800508.
8. H. S. Abdulbaqi, K. N. Mutter, M. Z. M. Jafri, and Z. A. Al-Khafaji, “Estimation of brain tumour volume using expanded computed tomographyscan images,” in *Proc. IEEE 23rd Iranian Conf. Biomed. Eng. 1st Int. Iranian Conf. Biomed. Eng.*, 2016, pp. 117–121.
9. A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI, *Inf. Fusion* 58 (2020) 82–115.
10. [10] M. Ahamed, A. Imran, Joint learning with local and global consistency for improved medical image segmentation, in: *Annual Conference on Medical Image Understanding and Analysis*, 2022.
11. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2010.
12. R. Strudel, R. Garcia, I. Laptev, C. Schmid, Segmenter: transformer for semantic segmentation, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021
13. Biratu, E.S.; Schwenker, F.; Ayano, Y.M.; Debelee, T.G. A Survey of Brain Tumor Segmentation and Classification Algorithms. *J. Imaging* 2021, 7, 179. <https://doi.org/10.3390/jimaging7090179>
14. Ali, S., Abuhmed, T., El-Sappagh, S., et al. (2023). Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, 99, 101805. <https://doi.org/10.1016/j.inffus.2023.101805>
15. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;372:n71.

16. Selvaraju, R. R., Cogswell, M., Das, A., et al. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *Proceedings of the IEEE International Conference on Computer Vision*.
17. Chattopadhyay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks. *arXiv preprint arXiv:1710.11063*.
18. Lee, S., Lee, J., Lee, J., Park, C.-K., & Yoon, S. (2018). Robust Tumor Localization with Pyramid Grad-CAM. *arXiv preprint arXiv:1805.11393*.
19. Omeiza, D., Speakman, S., Cintas, C., & Weldermariam, K. (2019). Smooth Grad-CAM++: An Enhanced Inference Level Visualization Technique for Deep Convolutional Neural Network Models. *arXiv preprint arXiv:1908.01224*.
20. Wang, H., Wang, Z., Du, M., et al. (2020). Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.
21. Desai, S., & Ramaswamy, H. G. (2020). Ablation-CAM: Visual Explanations for Deep Convolutional Network via Gradient-free Localization. *IEEE Winter Conference on Applications of Computer Vision (WACV)*.
22. Jiang, Z., Shen, L., Han, Z., et al. (2021). LayerCAM: Exploring Hierarchical Class Activation Maps for Localization. *IEEE Transactions on Image Processing*.
23. [23] Marmolejo-Saucedo, J. A., & Kose, U. (2022). Numerical grad-cam based explainable Convolutional neural network for brain tumor diagnosis. *Mobile Networks and Applications*. <https://doi.org/10.1007/s11036-022-02021-6>
24. Tehsin, S., Nasir, I. M., Damaševičius, R., & Maskeliūnas, R. (2024). DaSAM: Disease and spatial attention module-based explainable model for brain tumor detection. *Big Data and Cognitive Computing*, 8(9), 97. <https://doi.org/10.3390/bdcc8090097>
25. Pereira, S., Meier, R., Alves, V., Reyes, M., & Silva, C. A. (2018). Automatic brain tumor grading from MRI data using Convolutional neural networks and quality assessment. *Lecture Notes in Computer Science*, 106-114. https://doi.org/10.1007/978-3-030-02628-8_12
26. Ahmad, A., Sarkar, S., Shah, A., Gore, S., Santosh, V., Saini, J., & Ingalthaliker, M. (2019). Predictive and discriminative localization of IDH genotype in high grade gliomas using deep convolutional neural nets. 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), 25, 372-375. <https://doi.org/10.1109/isbi.2019.8759313>
27. Natekar, P., Kori, A., & Krishnamurthi, G. (2020). Demystifying brain tumor segmentation networks: Interpretability and uncertainty analysis. *Frontiers in Computational Neuroscience*, 14. <https://doi.org/10.3389/fncom.2020.00006>
28. Pintelas, E., Liaskos, M., Livieris, I. E., Kotsiantis, S., & Pintelas, P. (2020). Explainable machine learning framework for image classification problems: Case study on glioma cancer prediction. *Journal of Imaging*, 6(6), 37. <https://doi.org/10.3390/jimaging6060037>
29. Windisch, P., Weber, P., Fürweger, C., Ehret, F., Kufeld, M., Zwahlen, D., & Muacevic, A. (2020). Implementation of model explainability for a basic brain tumor detection using convolutional neural networks on MRI slices. *Neuroradiology*, 62(11), 1515-1518. <https://doi.org/10.1007/s00234-020-02465-1>
30. Esmaeili, M., Vettukattil, R., Banitalebi, H., Krogh, N. R., & Geitung, J. T. (2021). Explainable artificial intelligence for human-machine interaction in brain tumor localization. *Journal of Personalized Medicine*, 11(11), 1213. <https://doi.org/10.3390/jpm11111213>
31. Saleem, H., Shahid, A. R., & Raza, B. (2021). Visual interpretability in 3D brain tumor segmentation network. *Computers in Biology and Medicine*, 133, 104410. <https://doi.org/10.1016/j.combiomed.2021.104410>
32. Jin, W., Li, X., & Hamarneh, G. (2022). Evaluating explainable AI on a multi-modal medical imaging task: Can existing algorithms fulfill clinical requirements? *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11), 11945-11953. <https://doi.org/10.1609/aaai.v36i11.21452>
33. Zeineldin, R. A., Karar, M. E., Elshaer, Z., Coburger, ., Wirtz, C. R., Burgert, O., & Mathis-Ullrich, F. (2022). Explainability of deep neural networks for MRI analysis of brain tumors. *International Journal of Computer Assisted Radiology and Surgery*, 17(9), 1673-1683. <https://doi.org/10.1007/s11548-022-02619-x>
34. Dasanayaka, S., Shantha, V., Silva, S., Meedeniya, D., & Ambegoda, T. (2022). Interpretable machine learning for brain tumour analysis using MRI and whole slide images. *Software Impacts*, 13, 100340. <https://doi.org/10.1016/j.simpa.2022.100340>
35. Maqsood, S., Damaševičius, R., & Maskeliūnas, R. (2022). Multi-modal brain tumor detection using deep neural network and Multiclass SVM. *Medicina*, 58(8), 1090. <https://doi.org/10.3390/medicina58081090>
36. Marmolejo-Saucedo, J. A., & Kose, U. (2022). Numerical grad-cam based explainable Convolutional neural network for brain tumor diagnosis. *Mobile Networks and Applications*. <https://doi.org/10.1007/s11036-022-02021-6>
37. Hussain, T., & Shouno, H. (2023). Explainable deep learning approach for multi-class brain magnetic resonance imaging tumor classification and localization using gradient-weighted class activation mapping. *Information*, 14(12), 642. <https://doi.org/10.3390/info14120642>
38. Rahman, A., Karim, M. R., Chowdhury, P., Hossain, A., & Islam, M. M. (2023). NeuroXAI++: An efficient X-AI intensive brain cancer detection and localization. 2023 International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM). <https://doi.org/10.1109/ncim59001.2023.10212818>
39. Yan, F.; Chen, Y.; Xia, Y.; Wang, Z.; Xiao, R. An Explainable Brain Tumor Detection Framework for MRI Analysis. *Appl. Sci.* 2023, 13, 3438. <https://doi.org/10.3390/app13063438>
40. Taşçı, B. (2023). Attention deep feature extraction from brain MRIs in explainable mode: DGXINet. *Diagnostics*, 13(5), 859. <https://doi.org/10.3390/diagnostics13050859>
41. Özbay, F. A., & Özbay, E. (2023). Brain tumor detection with mrmr-based multimodal fusion of deep learning from MR images using Grad-CAM. *Iran Journal of Computer Science*, 6(3), 245-259. <https://doi.org/10.1007/s42044-023-00137-w>
42. Valerio Ponzi, & Giorgio De Magistris1. (n.d.). Exploring Brain Tumor Segmentation and Patient Survival: An Interpretable Model Approach. *CEUR-WS.org - CEUR Workshop Proceedings*

43. Šefčík, F., & Benesova, W. (2023). Improving a neural network model by explanation-guided training for glioma classification based on MRI data. *International Journal of Information Technology*, 15(5), 2593-2601. <https://doi.org/10.1007/s41870-023-01289-5>
44. Sarah, P., Krishnapriya, S., Saladi, S., Karuna, Y., & Bavirisetti, D. P. (2024). A novel approach to brain tumor detection using K-means++, SGLDM, ResNet50, and synthetic data augmentation. *Frontiers in Physiology*, 15. <https://doi.org/10.3389/fphys.2024.1342572>
45. Nhlapho, W., Atemkeng, M., Brima, Y., & Ndogmo, J. (2024). Bridging the gap: Exploring Interpretability in deep learning models for brain tumor detection and diagnosis from MRI images. *Information*, 15(4), 182. <https://doi.org/10.3390/info15040182>
46. Tehsin, S., Nasir, I. M., Damaševičius, R., & Maskeliūnas, R. (2024). DaSAM: Disease and spatial attention module-based explainable model for brain tumor detection. *Big Data and Cognitive Computing*, 8(9), 97. <https://doi.org/10.3390/bdcc8090097>
47. M, M. M., T. R, M., V, V. K., & Guluwadi, S. (2024). Enhancing brain tumor detection in MRI images through explainable AI using Grad-CAM with Resnet 50. *BMC Medical Imaging*, 24(1). <https://doi.org/10.1186/s12880-024-01292-7>
48. Zeineldin, R. A., Karar, M. E., Elshaer, Z., Coburger, J., Wirtz, C. R., Burgert, O., & Mathis-Ullrich, F. (2024). Explainable hybrid vision transformers and convolutional network for multimodal glioma segmentation in brain MRI. *Scientific Reports*, 14(1). <https://doi.org/10.1038/s41598-024-54186-7>
49. Rahman, M. A., Masum, M. I., Hasib, K. M., Mridha, M. F., Alfarhood, S., Safran, M., & Che, D. (2024). GliomaCNN: An effective lightweight CNN model in assessment of classifying brain tumor from magnetic resonance images using explainable AI. *Computer Modeling in Engineering & Sciences*, 140(3), 2425-2448. <https://doi.org/10.32604/cmescs.2024.050760>
50. Saeed, T., Khan, M. A., Hamza, A., Shabaz, M., Khan, W. Z., Alhayan, F., Jamel, L., & Baili, J. (2024). Neuro-XAI: Explainable deep learning framework based on deeplabV3+ and bayesian optimization for segmentation and classification of brain tumor in MRI scans. *Journal of Neuroscience Methods*, 410, 110247. <https://doi.org/10.1016/j.jneumeth.2024.110247>
51. Nazir, M. I., Akter, A., Wadud, M. A., & Uddin, M. A. (2024). Utilizing customized CNN for brain tumor prediction with explainable AI. <https://doi.org/10.2139/ssrn.4834282>
52. Mzoughi, H., Njeh, I., BenSlima, M., Farhat, N., & Mhiri, C. (2024). Vision transformers (Vit) and deep convolutional neural network (D-cnn)-based models for MRI brain primary tumors images multi-classification supported by explainable artificial intelligence (XAI). *The Visual Computer*. <https://doi.org/10.1007/s00371-024-03524-x>
53. Amin, A., Hasan, K., & Hossain, M. S. (2024). XAI-empowered MRI analysis for consumer electronic health. *IEEE Transactions on Consumer Electronics*, 1-1. <https://doi.org/10.1109/tce.2024.3443203>

Figures

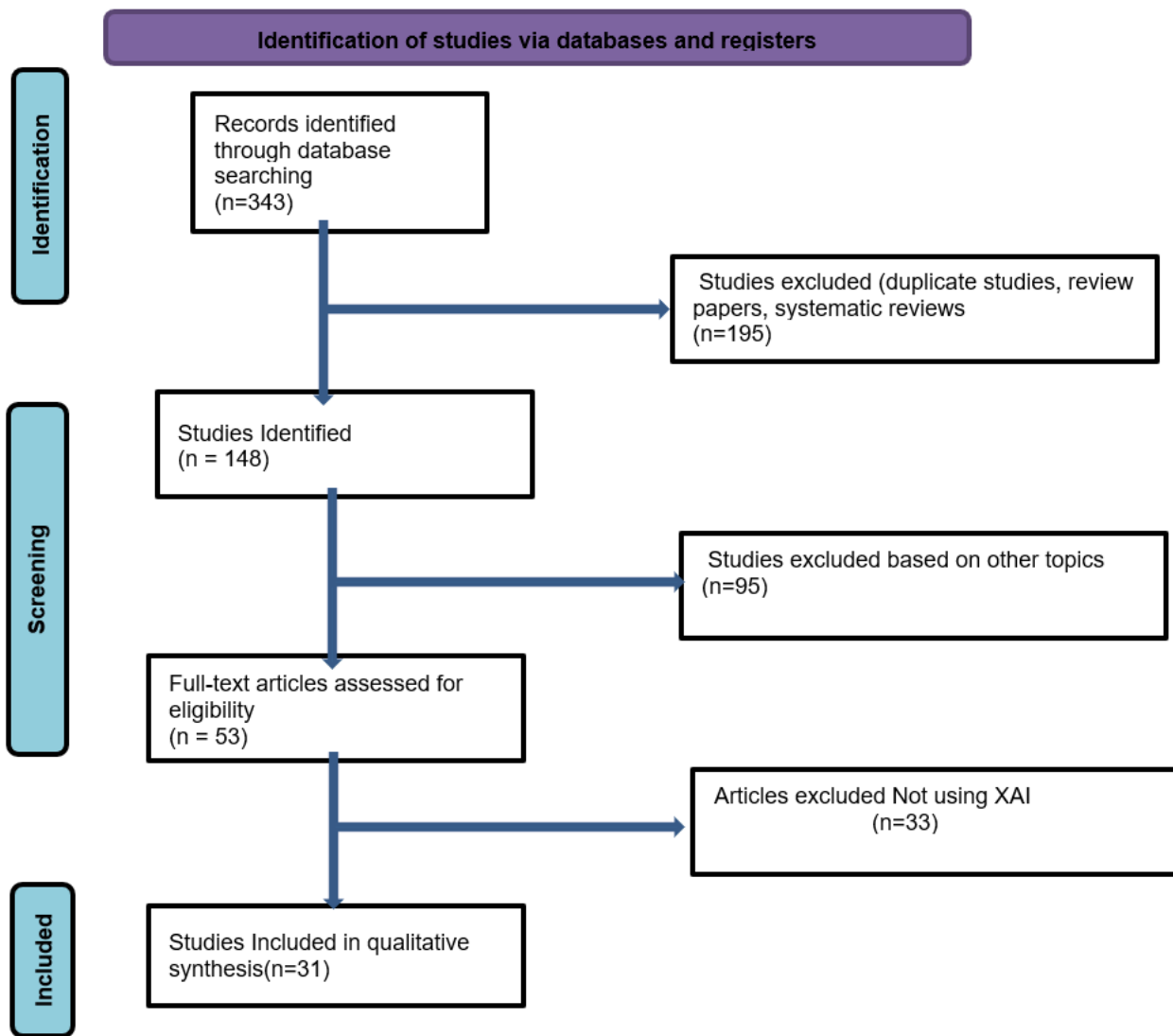


Figure 1

PRISMA flow demonstrating the procedure for choosing the most suitable 31 papers

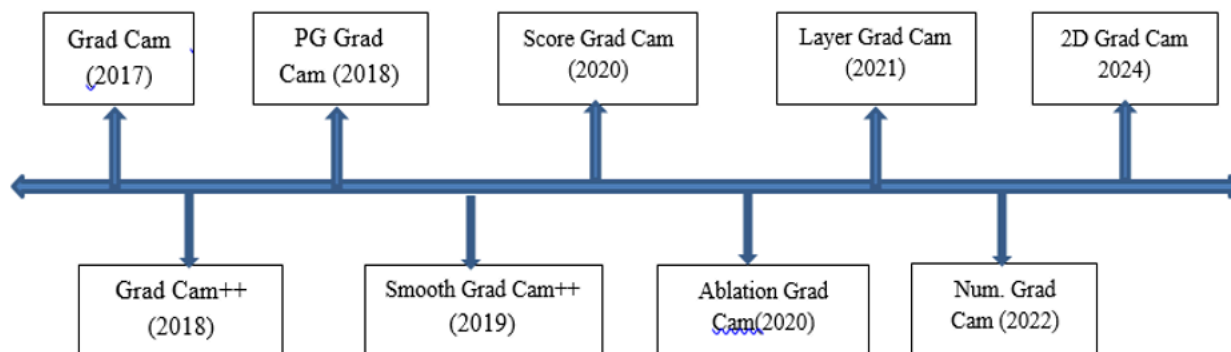


Figure 2

Evolution of Grad-CAM Techniques (2017–2024)

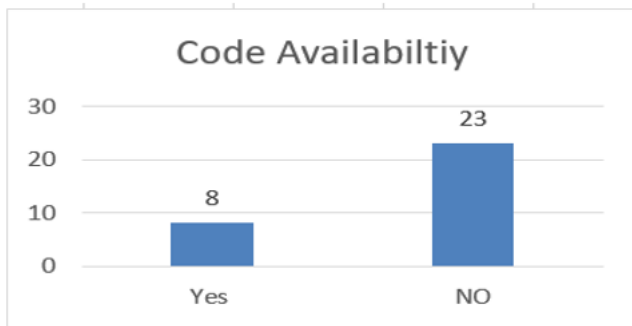


Figure 3

Proportion of Studies with Available Code for Reproducibility

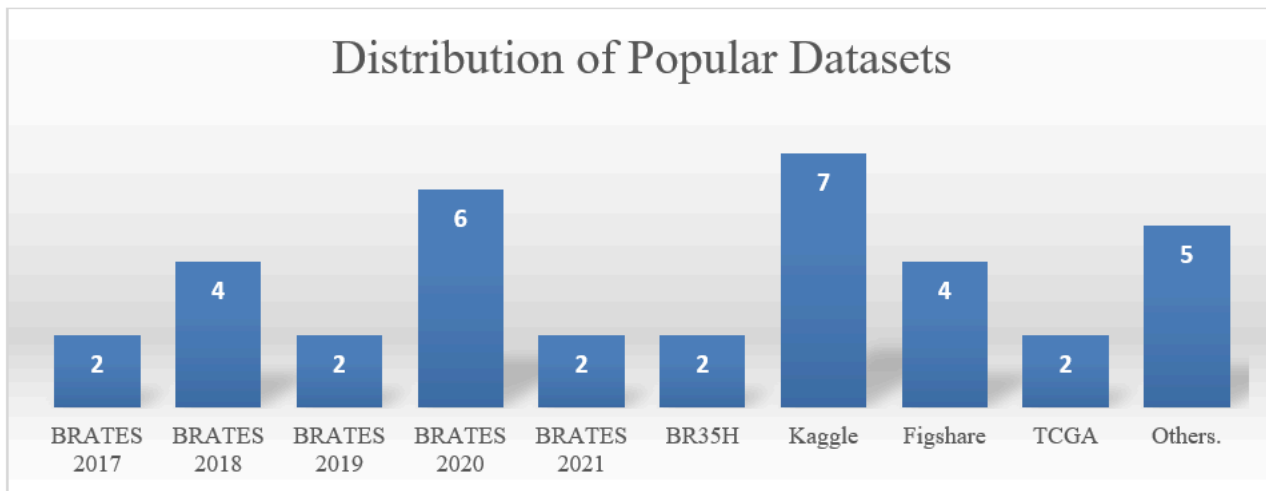


Figure 4

Distribution of Popular Datasets for Brain Tumor Detection, Segmentation, and Classification