VFAST Transactions on Software Engineering
https://vfast.org/journals/index.php/VTSE@ 2026, ISSN(e): 2309-3978, ISSN(p): 2411-6246
Volume 14, Number 1, January-March 2026 pp:278-300

An Explainable Deep Learning Framework for Brain Tumor Detection Using MRI Images

Atta Ullah [1], **Nadeem Akhtar** [2], **Sidra Hameed** [3], **Habib Ullah Sajid** [4], **Humaira Noreen** [1], **Muhammad Hasnain** [3]

[1]Department of Computer Science, The Islamia University of Bahawalpur, Pakistan; [2]Associate Professor, Department of Information Technology Faculty of Computing Information Technology (FCIT) University of the Punjab, Lahore, Pakistan; [3]Department of Artificial Intelligence, The Islamia University of Bahawalpur, Pakistan; [4]Principal at Apex International School Lahore, Pakistan

Keywords: *Brain Tumor, Deep Learning, CNN, ResNet, Inception, Explainable AI, Grad-CAM.*

Journal Info:
Submitted:
January 28, 2026
Accepted:
March 20, 2026
Published:
March 28, 2026

Abstract Brain tumor detection using Magnetic Resonance Imaging (MRI) is a critical diagnostic procedure that demands high interpretability, accuracy, and efficiency. This paper presents a system of brain tumor classification that is based on interpretable deep learning model. The Convolutional Neural Network (CNN) used is a customized one trained on two publicly obtained Br35H dataset along with a four-class brain tumor MRI dataset. There are four-class brain tumor, glioma, meningioma, pituitary and no tumor. Image denoising, data augmentation, and normalization are applied using the methodology to enhance robustness and generalizability in models. Gradient-weighted Class Activation Mapping (Grad-CAM) and Local Interpretable Model-agnostic Explanations (LIME) are combined to help tackle the problem of deep learning model opacities. These devices visualise class-discriminating areas and important local superpixels enhancing clinical transparency. The proposed CNN has been experimentally demonstrated to achieve about 94% and 98% accuracy on the Br35H dataset and the multiclass brain tumor MRI dataset respectively. The accuracy, recognition, and the F1-scores are comparable across classes. The results indicate that the framework is capable of capturing the features, type of tumor, and generates interpretable visual data to be relevant to the clinical world. The paper presents a full and interpretable deep learning model in the diagnosis of brain tumors acquired through MRI. It helps close the high diagnostic accuracy/reliable model explainability gap.

***Correspondence author email address:** sidrahameed786786@gmail.com
DOI: 10.21015/vtse.v14i1.2335

1 Introduction

Brain tumors are highly complex, polymorphic, and lethal diseases that pose a significant challenge in neurology and oncology. Early and accurate imaging diagnosis is critical to improve patient prognosis. Nevertheless, translating these acquisitions into clinical settings involves time-consuming manual interpretation of MRI scans, which suffers from interobserver variability. Recent data demonstrate that primary brain and other central nervous system (CNS) tumors make up one of the leading global causes of disability-adjusted life years, with survival rates differing significantly based on tumor type and grade [1, 2, 4, 5]. Therefore, the development of automatic and reliable diagnostic systems is a promising approach to improving clinical efficiency and diagnostic accuracy.

In recent years, deep learning (DL) techniques such as Convolutional Neural Networks (CNNs), Resid-

ual Networks (ResNets) [7], and Recurrent Neural Networks(RNNs) have proven that they can lead to high-performance results for image analysis with a lot of success, including brain tumor classification and segmentation. Deep learning on MRI images using CNNs can achieve high classification accuracy with large datasets and advanced architectures [8, 9]. Further, explainable AI (XAI) techniques like Grad-CAM [88] have also been increasingly utilized to visualize regions of concentration in models, which leads to dealing with the issue of black-box nature in deep learning systems [10, 47, 47]. However, several limitations still exist, preventing these advancements from being widely integrated into medical practice, though they show promise for supporting clinical decision-making in DL frameworks.

However, there are areas within this domain that still demand extensive research before effective brain tumor detection systems can be fully implemented. The largest obstacles are diverse datasets, cross-domain shifts between scanners and institutions, and class imbalance between tumor/non-tumor samples. There is no visual inspection of a model's data generalization when it is absent, with only asymptotic procedures to achieve such targeting, which don't encompass formal methods [12–14]. The basic building blocks of safety, transparency, and consistency and the evidence supporting specialisation use would require such diagnostic AI tools to have been derived mathematically or through formality verification techniques. How can we achieve both high performance and systematic verification of AI-based diagnostic systems? One approach is combining deep learning methods with formal verification tools (e.g., TLA+).

AI and DL have revolutionized the diagnostic and prognostic pipelines in healthcare, especially on clinical imaging. The concepts utilized alongside artificial intelligence support radiologists in the identification and segmentation of anatomical and pathological abnormalities, alleviating diagnostic load and increasing global accuracy while eliminating the case to case essence of radiological examination & interpretation [15–18]. The task of neuro-oncology can enable such systems to provide faster, more consistent interpretation of MRI scans, which is essential for early treatment and intervention planning. Past research established that AI-assisted radiological processes can facilitate quicker diagnosis, reduce human error, and improve clinical outcomes, and are significantly more affordable for care delivery services [19–21]. Individualized patient advice can also be provided through the implementation of AI in clinical decision-support systems, thereby effectively linking radiologic evidence to individualized planning.

While significant progress has been made, interpretability, fairness, and reliability pose important barriers to the large-scale adoption of AI in healthcare. Clinicians are skeptical they will trust black-box algorithms that cannot give them explicit justifications and attributions of cause [22, 24, 25, 49]. This represents one of the largest constraints to clinical adoption, particularly in high-stakes medical decision-making pathways for which there are serious consequences of error, such as brain tumor diagnosis. Translating clinical AI-based diagnostics into practice presents ongoing hurdles such as data privacy, generalizability across institutions and regulatory oversight that renders the clinical pathway less than simple. Thus, we have to focus not also on data privacy alone but itself on XAI methods to achieving interpretability, security and absolute validation of such methodologies.

Not only do such approaches address barriers to adoption, but they can also be especially valuable during the diagnostic phase of patient care. For instance, LIME-based classifiers have been successfully employed to enhance the transparency of COVID-19 detection from X-ray images [3], thereby supporting clinicians and enabling better decision-making.

The accelerated development of medical imaging technologies has led to a uniquely high data generation rate, establishing brain MRI datasets as cornerstone components in modern healthcare Big Data ecosystems. For example, modern large hospital installations generate petabytes of MRI scan data each year, straining storage, processing, and analytics infrastructure [26, 28, 30]. Such large volumes of high-speed, diverse big data (e.g., multidimensional MRI scans) require the use of Big Data frameworks, including distributed storage, parallel computing platforms, and scalable analytics. Recent medical physics reviews have already shown that the combination of enterprise imaging and Big Data technologies can improve patient care, stream-

line workflows, and support large-scale image analytics powered by popular AI architectures [30–32]. As a great example, Big Data capabilities enable processing of large repositories of MRI scans, training deep neural network algorithms with thousands of volumes, and providing strong generalizable outputs.

The adoption of Big Data analytics especially in cloud and edge computing models is reshaping the design and deployment of MRI-based diagnostic frameworks, among others. Research has shown that the integration of cloud-based infrastructures along with deep learning models highly enhances diagnostic accuracy (15–20% more accurate) and processing time (as much as 60% faster) [29, 33, 34]. These methods also facilitate federated and distributed learning across organizational boundaries, allowing models to be trained together without the need for centralized data aggregation [27, 30, 35]. Nonetheless, these innovations bring new complications around vendor and MRI protocol diversity; governance and privacy for sensitive medical images; and the scalability of analytics frameworks (e.g., HDFS, Spark, distributed deep learning frameworks) [29, 31, 32]. It is important to address such issues to develop a scalable, robust and reliable brain tumor detection framework that would perform effectively in varying imaging conditions.

As of late, the application of deep learning techniques like Convolutional Neural Network (CNN), Residual network (ResNet) and Inception-based architecture has relegated the domain of analyzing medical images. Numerous architectures might learn automatically hierarchies of high-level features and abstractions over images, without the manual engineering of features. Similar experiments have demonstrated classification rates of 99.88% with the use of magnetic resonance imaging (MRI) to detect brain tumors, thereby validating the effectiveness of residual connections to solve vanishing gradient problems, as well as enhance depth performance at a network scale, through brain tumor detection [64–66]. Moreover, CNN architectures that combine both regional and residual style also have shown improvements in performance due to the use of boundary and texture information about the tumor subtypes with over 98% accuracy reached by Res-BRNet architecture of the hybrid from both domains have also been demonstrated to achieve better results in this domain as well as others [67, 68]. These findings demonstrate the extent to which the deep neural networks go or even beyond the capabilities of humans in case they are properly data-pretreated and hyperparameters configured with sufficient data.

Meanwhile, medical AI is being focused more on explainability and interpretability, and the purely black-box systems are no longer considered suitable to use in high stakes clinical scenarios. Namely, visual feedback systems are present in all their models predictions indicating the areas that have contributed the most to their predictions (e.g., Gradient weighted Class Activation Mapping, 2025) to enhance transparency and trust of clinicians in [69, 70]. Conducting explainable AI (XAI) in medical imaging tasks, it has been observed that although these models are excellent predictors despite the model type being CNNs, RNNs or Vision Transformers (ViTs) the eventual clinical implementation of such systems has historically been hampered by model interpretability tools that attempt to visualize what part of the input contributed to determining the criteria of decision making by the model using tools such as Grad-CAM [71], Local Interpretable Model Agnostic Explanations (LIME) [72], or Shapley Additive exPlanations (SHAP) [73]. One such technique applies ResNet50 in conjunction with Grad-CAM to localize affected regions of a brain tumor, achieving up to 98.52% accuracy while providing clinically useful localization of the affected portion. This highlights the necessity of explainability as a key requirement towards safe, medical-grade AI system development [74, 75].

Although performance and interpretability have improved significantly, implementing deep learning-based diagnostic systems in the clinic is difficult. Important challenges include domain shifts across imaging centers, limited labeled training data, class imbalance, and the need for continuous model validation and monitoring. It is well known that models that achieve high performance on carefully selected benchmark datasets often do not generalize well when they encounter clinical data, and interpretability tools like Grad-CAM can obfuscate the behavioral understanding of table anatomy/visual context even more by inaccurately attributing their relevance to visual artifacts or other non clinically relevant areas [76–

78].

Several recent architectures, including Inception-v4, DenseNet and self-supervised Vision Transformers, have been utilized for the task of brain tumor classification as well [79]; however, moving towards these types of models adds unnecessary complexity along with data and processing overhead [80]. Thus, a clinically actionable brain tumor detection system should focus on robust modeling (CNNs, ResNet, Inception), feasible data management techniques, utility-oriented explainability tools (Grad-CAM), and, if required, formal validation to ensure trustworthy and interpretable diagnostics.

The objective of this research is not only to propose a deep learning framework with explanation ability in accurate brain tumor detection using MRI images, but also to develop the same. State-of-the-art solutions to clinical artificial intelligence are often specialized otologies of excellent models for specific problems, albeit a dedicated pipeline providing accuracy specifications which incorporates extensive preprocessing and cross-cutting diverse deep learning architectures with explainable AI components to ensure sufficient assurance and interpretability. The key contributions of the paper are as follows:

- An explainable deep learning framework based on a custom convolutional neural network (CNN) is presented for automated brain tumor classification from MRI images, balancing diagnostic accuracy, efficiency, and clinical interpretability.
- Two publicly available datasets from Kaggle, the Br35H brain tumor MRI dataset and a four-class Brain Tumor MRI dataset (glioma, meningioma, pituitary, and no tumor), were utilized. Extensive preprocessing steps, including image cleaning, intensity normalization, data augmentation, and stratified splitting, were implemented to ensure high-quality inputs, mitigate class imbalance, and enhance model robustness.
- A custom CNN architecture was developed to capture local spatial and texture-level patterns characteristic of tumor and non-tumor regions in MRI scans. This architecture was adapted and trained on both binary (tumor vs. non-tumor) and multi-class tumor classification tasks to assess its generalizability.
- To address the black-box limitation of deep neural networks, Gradient-weighted Class Activation Mapping (Grad-CAM) and the Local Interpretable Model-agnostic Explanations (LIME) method were integrated to visualize salient image regions and superpixels that influence model predictions. This dual XAI strategy promotes transparency and supports human-in-the-loop validation of AI-assisted diagnoses.
- The proposed framework was rigorously evaluated using standard metrics (accuracy, precision, recall, and F1-score) on both datasets. The CNN achieved approximately 94% accuracy on Br35H and 98% accuracy on the multi-class Brain Tumor MRI dataset, demonstrating effectiveness for MRI-based brain tumor detection while providing clinically meaningful, interpretable explanations.

This research contributes an end-to-end, explainable, and reproducible framework for MRI-based brain tumor classification. By combining a custom CNN architecture with dual visual interpretability mechanisms based on Grad-CAM and LIME, the study advances the field toward more transparent, accountable, and trustworthy AI-assisted medical diagnostics.

2 Background

2.1 Brain Tumor Detection and MRI Imaging

Brain tumors originate when cells within the brain tissue begin to proliferate in an abnormal way. This proliferation can interfere with normal functioning of the brain, which may result in significant disability or death [2, 4]. Magnetic Resonance Imaging (MRI) is the gold standard noninvasive imaging modality for brain tumor diagnosis. Compared to Computed Tomography (CT) or Positron Emission Tomography (PET), it provides excellent soft tissue contrast, multi-planar imaging without ionizing radiation [5, 6]. MRT is crucial to detect and characterize the mass morphology, tumor edema, and infiltration margins. These characteristics are important for clinical diagnosis, therapy planning, and surgical intervention. In contrast, manual reading of MRI images is time-consuming, operator-dependent and subject to inter-observer variability that can jeopardize diagnostic accuracy. That is why there is a high demand for computer-aided diagnosis (CAD) and classification

systems based on Machine Learning (ML) and Deep Learning (DL). These systems play a vital role in increasing the accuracy, effectiveness, and reliability of neuroimaging procedures pertaining to the brain and other organs [39, 67].

Although MRI is widely used to diagnose brain tumors, accurate detection can still be problematic, as the human brain is a complex organ, and there may be variability in disease characteristics across different patients. Tumors tend to show considerable heterogeneity in shape, size, texture, and anatomical location, which, together with varying MRI acquisition protocols, complicate model performance and interpretations [40, 41]. In addition, the distributions of healthy and pathological tissues overlap, making feature extraction and segmentation challenging and often leading to misclassification [42, 43]. To mitigate these issues, they suggest that complex preprocessing pipelines are needed, with improved intensity normalization and skull stripping, followed by bias field correction to reduce image variability as input to DL algorithms used for classification/segmentation.

Deep learning has made remarkable progress in recent years, and for brain tumor detection, deep-learning-based methods have exhibited much higher accuracy and generalization capacity. The Br35H Brain Tumor MRI Dataset is publicly available and provides high-quality T1-weighted magnetic resonance imaging (MRI) scans with a balanced class distribution, making it ideal for improving deep learning architectures [12]. Convolutional Neural Network (CNN) models have outperformed classical machine learning classifiers such as Support Vector Machines (SVM) or Random Forests (RF), mainly through their ability to learn spatial and texture features of tumor morphology [36, 39]. Cross-validation, data augmentation, and other techniques help prevent overfitting and improve model generalization to unseen examples. Collectively, these advancements have established deep learning frameworks as an indispensable component of contemporary techniques for MRI-based brain tumor classification schemes, achieving superior diagnostic precision, computational efficiency, and robustness compared to traditional clinical assessment.

2.2 ML and DL in Medical Imaging

ML and DL algorithms have reached impressive performance in numerous facets of medical image analysis for pattern recognition and classification. Convolutional Neural Networks (CNNs) are being adopted to extract the multiscale spatial features from medical images and have demonstrated remarkable performance, especially in tumor classification of MRI [36, 56]. Architectures based on transfer learning, like ResNet50 and Inception V3, have subsequently been utilized to achieve improved diagnostic accuracy by using weights trained on large datasets outside the domain of breast histopathology (e.g. ImageNet). This is how models learn general visual characteristics that are then fine-tuned for specific medical fields [55, 63]. While the residual connections used in ResNet50 can contribute to preventing vanishing gradients when training deeper networks, InceptionV3 uses intensity-concise parallel convolutional modules for better multi-scale feature extraction. These meta-models have achieved state-of-the-art results in various medical imaging applications, including the classification of brain [10], breast and lung cancers [53].

Although this development, traditional DL models are still criticized for their interpretability issues, which limit their clinical applications. With advances in deep learning, computer-aided diagnostic systems are becoming the norm across many medical imaging modalities. Although CNNs are currently the dominant models for image-based classification, hybrid architectures combining CNNs with recurrent neural networks (RNNs), long short-term memory (LSTM) units, or transformers provide greater capacity to model the spatial and temporal dependencies present in medical data [2, 4]. These hybrid structures can also handle sequential MRI slices, track tumor progression, and temporally summarize multi-view image representations. For instance, Sagan-istyle networks introduce a mixture of residual-based and dense connectivity [56, 67], which prevents vanishing gradients and enables efficient feature propagation. The performance of hybrid and transfer learning-based networks for brain tumor segmentation is high enough, even though such models have been trained on a significantly smaller number of annotated datasets. l annotated datasets

There is increasing interest in leveraging multimodal modeling and privacy-preserving techniques to improve the clinical relevance of deep learning methods. Multi-modal learning facilitates the integration of various data

sources, including MRI, CT, PET, and patient metadata, as input modalities for multivariate models that can learn complementary representations, thereby enhancing diagnostic accuracy [33, 35]. Federated learning frameworks enable model training across multiple collaborating institutions while keeping patient-level data within the respective institutions, preserving patients' privacy while enabling the development of larger-scale datasets [94, 95]. In the contexts above, transfer learning remains a key resource because of the limited data available in medical domains; pre-trained models on natural image datasets can be adapted to solve target tasks. Sandwiching these results, they all illustrate the rise of ML and DL as foundational technologies for precision medicine, intelligent diagnostics, and explainable healthcare systems.

2.3 Explainable Artificial Intelligence (XAI) in Medical Diagnosis

Explainable Artificial Intelligence (XAI) is currently recognized as a key factor in building trust in AI-based healthcare systems. Explainable Artificial Intelligence (XAI) techniques aim to help clinicians understand and trust the outcomes of deep learning-based decisions by providing insight into the reasoning process that led to the conclusion. Out of the various methods proposed for XAI, Gradient-weighted Class Activation Graph Mapping (Grad-CAM) has gained particular attention because it can produce saliency heatmaps that depict important locations within an image associated with model-specific predictions [52, 54]. Grad-CAM has improved the interpretability of CNNs such as InceptionV3 and ResNet50 in terms of being able to demonstrate which regions are most affected by the tumor, as found on MRIs, indicating that the model focuses closely on what is clinically understood in brain tumor diagnosis [10, 85]. Grad-CAM can not only build confidence when integrated into diagnostic pipelines, but also allow for clinical validation as radiologists can visually assess whether the attention given by models is focused on regions of interest overlaid with abnormalities. This trade-off between deep learning capability compared to interpretability is one of the foundation elements of contemporary AI-assisted diagnostic platforms.

Recent XAI cracks in the healthcare sphere are different approaches to visualization, interpretability metrics [60], bias measurement [61], and human in the loop verification. These improvements can help close the "trust gap" between clinicians and AI-powered diagnostic tools. This practice might not be trusted by human experts in cases such as brain tumor grading, lesion segmentation and cancer prognosis [53, 54], where blackbox machine learning algorithms do not produce interpretable explanations. For example, by making both the visual and analytical comprehension of AI Systems' findings possible through XAI, it paves inter-conditions for explainable decisions with an audit trail capable of being audited in compliance with regulatory needs as the EU's AI Act and FDA guidance for Software as a Medical Device (SaMD) [30, 62]. This approach is particularly relevant when it comes to model evaluation with scalpel drops, and also when incorporating methods such as Grad-CAM and LIME into the mix makes a lot more sense: since not only does this allow for collaboration between clinician and model, but it can really help augment SOTA performance evaluations. Model focal points can be evaluated with the knowledge of clinicians, thus facilitating trustworthiness, accountability and reproducibility in AI-supported medical imaging processes.

2.4 Grad-CAM in Medical Imaging

Interpretable or Explainable Artificial Intelligence (XAI) is one of the major research fields focused on improving transparency in deep learning-based healthcare models (e.g., [54]). - Grad-CAM has been used as the main visualization tool among XAI methods for generating interpretable regions for model decision-making (e.g., [83, 88]. Grad-CAM calculates the gradient of the target score with respect to the convolutional feature maps. This generates an activation map that indicates which parts of the image drive the model's prediction. [52, 53]. As a result, researchers and clinicians can confirm that the model attends to the correct anatomical regions during inference. Since then, Grad-CAM has been applied in various convolutional architectures: VGG, ResNet, and Inception to visualize results for classification, object detection, and semantic segmentation tasks in medical imaging.

Grad-CAM is widely used to interpret deep learning models for diagnosing conditions such as brain tumours, Alzheimer's disease, and other neurological diseases [10, 57, 85]. Grad-CAM visualizations, for example, delineate tumor boundaries or regions of interest in an MRI scan,

providing radiologists with tangible evidence that corresponds to machine learning predictions. InceptionV3 approach & InceptionResNetV2 manufactured clinical trust through feature activation: tumor link verified by the recent studies [52, 67]. In addition, CNN-transformer architectures and ViTs can benefit from Grad-CAM techniques to visualize the hierarchical attention to provide better interpretation of space-level features learned [57, 63].

While Grad-CAM has its benefits, it certainly has important disadvantages. It can sometimes emphasize irrelevant areas or noise, particularly in radiographs with overlapping textures or low tumor contrast (as several studies have shown). Standard Grad-CAM explanations are also model-specific and likely do not generalize to other datasets [53, 54]. In response to these challenges, some researchers have proposed enhancements over Grad-CAM that provide a more robust gradient computation [84, 86]. Still, Grad-CAM is one of the most valuable feature techniques, and using it in this model confirms the accuracy of deep learning predictions.

2.5 Local Interpretable Model-agnostic Explanations (LIME) in Medical Imaging

Local Interpretable Model-agnostic Explanations (LIME) is a highly used method for explaining the predictions of any complex classifier by locally approximating the target with an interpretable model [87]. For example, in the case of image data, LIME breaks the image into superpixels, creates perturbations by turning on or off these superpixels across different combinations, and records changes in performance. Then, a sparse linear model is fit in that local neighborhood, assigning importance weights to each superpixel based on how much it contributed to predicting the class. This sample-specific novel explanation paradigm is very useful in medical imaging, where each patient must be justified to a doctor for an automated diagnosis. Going forward, LIME has been applied further to brain tumor MRI classification to enhance transparency and trust for clinicians [88].

Abraham et al. complemented a DenseNet169 backbone with LIME (DenseNet169-LIME-TumorNet), showing that, for glioma, meningioma, and pituitary tumors, LIME heat maps were always detectable on MRI and could classify a public brain tumor dataset with nearly 99% accuracy. [89]. According to Chel et al. [47], the superpixel explanations of CNNs trained to classify (as healthy, low-grade tumor, and high-grade tumor) brain MRI occasionally revealed that the model had been trained on non-tumor structures, which represented potential failure modes and dataset biases that would not be apparent purely from classifications/accuracy. Local explanations of brain tumor detection in similarity models enabled by deformable local explainers, i.e., DeepEBTDNet, indicate capable assistance to radiologists, confirming that predicted tumors are at anatomical locations where known abnormalities are expected for the engine modeling attempts to detect abnormality-related tumors and at their areas of origin and spread [90].

LIME and explainable neural networks are used on many medical imaging tasks, beyond neuro-oncology. Examples are retina fundus analysis helping retinoblastoma assessment and lung disease identification on chest images, where unambiguous interpretations help support diagnosis [91, 92]. The LIME studies demonstrate that the approach can consistently localize important clinical regions within images and provide a better understanding of CNN decision-making, even with complex models. Using these observations, this work applies both LIME and Grad-CAM for the first time to derive complementary perturbation-based explanations of brain tumor MRI classification. These modules enable a more comprehensive qualitative analysis of the CNN model's decision-making process across binary and multiclass challenges.

3 Related Work

Guluwadi et al. [10] developed an explainable deep learning framework based on Gradient- weighted Class Activation Mapping (Grad-CAM) to interpret how the model is making decisions for brain tumor detection in Magnetic Resonance Imaging (MRI). The method used a public multimodal MRI dataset and a fine-tuned ResNet50 with ImageNet pretraining for tumor identification. To improve data quality and minimize the risk of overfitting, standard preprocessing methods (normalization, resizing, and augmentation) were applied. Visualizing areas of importance using the Grad-CAM methodology provided an informal way to examine

which regions were influencing a model's predictions, thereby increasing interpretability and confidence in a clinical setting. The proposed explainable AI model outperformed CNN and VGG16 baselines, achieving 97.6% accuracy, suggesting that a combination of explainable AI techniques can boost performance while providing a transparent solution for neuroimaging applications.

Iftikhar et al. [44] proposed an explainable convolutional neural network (CNN) framework for MRI-based classification of glioma and meningioma cases. Image preprocessing involved skull stripping, data normalization, and augmentation. The optimized CNN used depth-wise separable convolutions to quickly learn spatial and context features. We applied Gradient-weighted Class Activation Mapping (Grad-CAM) and Local Interpretable Model-Agnostic Explanations (LIME) to help clarify the image features that contributed to model predictions. The framework achieved 98.2% performance on conventional CNN-based benchmarks, suggesting the profound role of interpretation in clinical neuro-oncology.

Asmita et al. [45] offered a detailed overview of the history of AI in neuro oncology, framing its narrative with comparisons to both traditional black-box models and explainable AI (XAI) paradigms. The authors of the review covering more than one hundred MRI based brain tumor detection, segmentation, and classification studies concluded that although deep learning architectures such as ResNet, DenseNet, or Inception achieve excellent diagnostic accuracies, they still suffer from low interpretability which limits their application in clinical practice. We classified the existing top XAI techniques including Grad-CAM; Grad-CAM++; SHAP and LIME based on their degree of visualisation power, as well as their position in the current Medical Image Analysis pipeline. The article highlighted the importance of human-centricity, transparency and guidelines compliance for AI to be successful, reporting improved trustworthiness and reproducibility of explainable predictions and thus enhanced clinician accountability in AI-driven brain tumor diagnostics.

Sarker et al. [46] propose explainable methods for MRI brain tumor classification using transfer learning. Nevertheless, the methods should perform region-wise preprocessing steps like denoising, resizing or histogram equalization to normalize intensities and close spectra. Three deep learning models were analysed (InceptionV3, VGG19 an ResNet50), and ResNet50 achieved the highest accuracy at 97.8%. Grad-CAM visualizations effectively indicated tumor-affected areas, thus confirming model decisions and enhancing interpretability. Combining transfer learning and XAI can increase the diagnostic confidence of practitioners, even in data-poor clinical conditions, as the results of our study show.

Chel et al. [47] proposed Res-BRNet, a deep residual decision-based regional convolutional neural network for brain tumor classification from MRI images. This method combines residual learning and region-based feature extraction to model both global and local entity features. During pre-processing of MRI images, the researchers used a combination of skull stripping, normalization, and histogram equalisation to improve image quality, along with a publicly available dataset that includes glioma, meningioma, and pituitary tumor images. The Res-BRNet architectural design effectively integrates regional attention modules and residual block modules to guarantee a detailed map structure while providing efficient computation. The results of the experiment show that Res-BRNet achieves 98.4% accuracy, outperforming baseline models CNN (93.2%), ResNet50 (96.5%), and InceptionV3 (97.1%). The outcomes reveal that adding residual and regional learning can mitigate limitations in boundary differentiation and enhance robustness in clinical diagnostics, underscoring the practical significance of the strategy.

Soewu et al. [48] proposed an explainable, multiclass brain tumor detection model based on the Xception convolutional neural network and implemented Gradient-weighted Class Activation Mapping (Grad-CAM) together with SHAP to derive class activation maps. Specifically, the model training and validation was performed over MRI data of glioma, meningioma, pituitary and no-tumor classes with learnings possessing high training (99.95%), validation (p) (99.08%) and test accuracies (98.78%). Additionally, Grad-CAM visualizations accurately localized tumor-relevant regions of the image to provide interpretability for the model predictions. Significantly, the study also acknowledged the high computational cost and lack of external validation as important limitations, pointing toward opportunities in

future research and improvement in this space.

Saw et al. In [49], a deep learning pipeline for tumor detection and segmentation is proposed through MRI and CT modalities. This framework integrates a Harmony Search Optimization (HSO) algorithm with a convolutional neural network ensemble and obtained a classification accuracy of 99.13%, compared to single-modality-based approaches that abstract multiple modalities. Because MR and CT images share some spatial and intensity properties, a hybrid combination achieves better generalization across the imaging modalities. Nonetheless, reliance on multimodal data and limited access to high-quality CT datasets pose limitations that can affect both the scalability and clinical use of the proposed pipeline.

Dhana et al. [50] proposed an explanation-forward saliency-based deep learning method for the classification of brain tumor MRIs and used various visualization techniques such as Grad-CAM, Grad-CAM++, Score-CAM and XRAI. Quantitatively, AIC & SIC were used to assess the interpretability of these saliency methods. These results show that Score-CAM and XRAI produce stable visual-attribution maps, whereas Grad-CAM shows instability under background noise. The study concluded that explainable AI is a significant step toward transparency, but the authors identified robustness and the replicability of explanations as major challenges to clinical adoption, (see Table. 1).

Kumar et al. To streamline the second approach, [51] proposed a hybrid deep learning framework for brain tumor detection based on magnetic resonance imaging (MRI) that combined a convolutional autoencoder with traditional machine learning classifiers. This study used the publicly available Figure share dataset, which contains T1-weighted contrast-enhanced images of glioma, meningioma, and pituitary tumors. As a key highlight, the proposed model achieved a classification accuracy of 96.47%, outperforming individual CNN and SVM models. The authors noted that these advancements have also not been without ongoing challenges, such as dataset imbalance, overfitting, and feature redundancy, which they suggest will require further research into the potential of data augmentation and transfer learning to improve real-world applicability and overcome the limitations of current methods.

Seetha et al. [37] explored a convolutional neural network (CNN)-based approach for binary brain tumor classification based on MRI data. The image dataset consists of 2,065 MRI images, which were split into training, validation, and test subsets. The model achieved 94.39% test accuracy, reflecting the ability of convolutional layers to extract tumor-specific features. Yet, the authors noted that binary classification limits its clinical use in multi-class diagnosis, and scalability to larger datasets with greater diversity remains an obstacle.

Overall, the existing MRI-based brain tumor analysis literature reports promising classification performance, but these studies rely on isolated datasets or overly complex architectures and lack adequate explainability mechanisms. There are several works that employ deep CNNs or transfer-learning backbones on the Br35H and related datasets; however, they focus either on typical binary detection or on multi-class classification, without systematically incorporating complementary XAI techniques. Other methods achieve slightly higher accuracy with very deep models and ensemble approaches, which increase computational cost/complexity and decrease transparency, limiting their deployability in resource-limited clinical settings. On the other hand, the framework proposed in this study uses a compact yet interpretable CNN architecture, trained on two public datasets of MRI scans and enriched with dual explainable AI methods, Grad-CAM and LIME, yielding both competitive performance and clinically important visual explanations of model decisions.

4 Materials and Methods

Figure 1 presents a more streamlined, automated approach to brain tumor detection using deep learning and explainable AI. Our pipeline consists of Data Collection, Preprocessing, Data Augmentation, and Design and Interpretability Analysis of Classificative Models. We use two public MRI datasets available on Kaggle: the Br35H dataset for binary classification (tumor vs. non-tumor) and the Brain Tumor MRI dataset for four-class classification (glioma, meningioma, pituitary, no tumor). These datasets include a variety of MRI scans to train and test the model. And the framework has the potential to support clinical decision-making and enable earlier diagnosis and treatment planning by improving tumor

Table 1. Comparison of Brain Tumor detection studies.

Study	Year	Data Source	Methods	Accuracy	Limitation
[44]	2025	Custom MRI dataset (glioma, meningioma)	CNN + Grad-CAM + LIME	98.2%	Requires high computational resources; limited to binary and ternary classification tasks
[45]	2025	Survey of 100+ studies on MRI-based detection	Grad-CAM, SHAP, LIME (reviewed)	95%	Review study; lacks experimental validation; highlights interpretability gap in clinical AI models.
[48]	2025	MRI dataset (mult-class)	Xception + Grad-CAM + SHAP (Explainable CNN)	98.78%	High computational cost; lacks external dataset validation.
[46]	2025	MRI dataset from hospitals in Bangladesh	Transfer Learning (InceptionV3, VGG19, ResNet50) + Grad-CAM	97.8%,	Regional dataset with limited diversity; potential bias in demographic representation.
[10]	2024	Public MRI dataset (BMC Medical Imaging)	ResNet50 + Grad-CAM	97.6%	Limited dataset size and potential overfitting; interpretability depends on Grad-CAM visualization quality.
[47]	2024	Public brain tumor MRI dataset (Biomedicines)	Res-BRNet (Residual + Regional CNN)	98.4%	Requires fine-tuning on multi-modal MRI sequences; limited real-time inference evaluation.
[49]	2024	MRI + CT multimodal dataset	CNN + Harmony Search Optimization (HSO) ensemble	99.13%	Computationally expensive; limited multimodal dataset availability.
[50]	2024	MRI dataset (multi-class classification)	Grad-CAM, Grad-CAM++, Score-CAM, XRAI (saliency-based XAI)	73%	Visualization instability; explanations sensitive to noise and model bias.
[51]	2023	Figshare MRI dataset (glioma, meningioma, pituitary)	CNN + Autoencoder + ML classifiers	96.47%	Dataset imbalance and overfitting issues; limited clinical validation.
[37]	2023	2,065 MRI images (tumor vs non-tumor)	CNN (custom architecture)	94.39%	Restricted to binary classification; small dataset limits generalization.
This Study	2026	Br35H (binary) and Brain Tumor MRI (four class)	CNN, Grad-CAM	94% and 98%	Performance may be dataset-dependent; external validation on diverse real-world clinical datasets is required.

detection accuracy and interpretability in MRIs.

All MRI images are resized to a common spatial resolution, converted to a common color format during preprocessing, and denoised and intensity-normalized to minimize scanner-dependent variability. The training sets are augmented with controlled rotations, horizontal flips, translations, and zooming to increase dataset diversity and improve model generalization. In both binary and multi-class scenarios, stratified train-validation-test splits are used to maintain class balance while minimizing image leakage across the three partitions.

A custom Convolutional Neural Network (CNN) is developed as the primary classifier and trained from scratch using the preprocessed images from both datasets. The architecture consists of multiple convolutional and max-pooling layers, followed by fully connected layers incorporating batch normalization and dropout, which facilitate the extraction of discriminative spatial and texture features from tumor and non-tumor regions. Model parameters are optimized with the Adam optimizer at an initial learning rate of 3×10^{-4}. Training is regularized through early stopping and adaptive learning rate reduction based on validation performance to mitigate overfitting. Additionally, InceptionV3 and ResNet50 architectures are fine-tuned via transfer learning and serve as baseline models for comparison with the proposed CNN.

To address the inherent opacity of deep neural networks, explainable artificial intelligence (AI) techniques are applied to the trained CNN. Gradient-weighted Class Activation Mapping (Grad-CAM) is utilized to produce class-discriminative heatmaps that identify the most influential regions in the MRI scans for each prediction, thereby providing coarse localization of tumor structures. In addition, the Local Interpretable Model-agnostic Explanations (LIME) method generates perturbation-based, superpixel-level importance maps for selected cases, offering local surrogate explanations for individual predictions. The combined use of Grad-CAM and LIME visualizations enables radiological experts to qualitatively assess the CNN's decisions and enhances the clinical interpretability of the framework

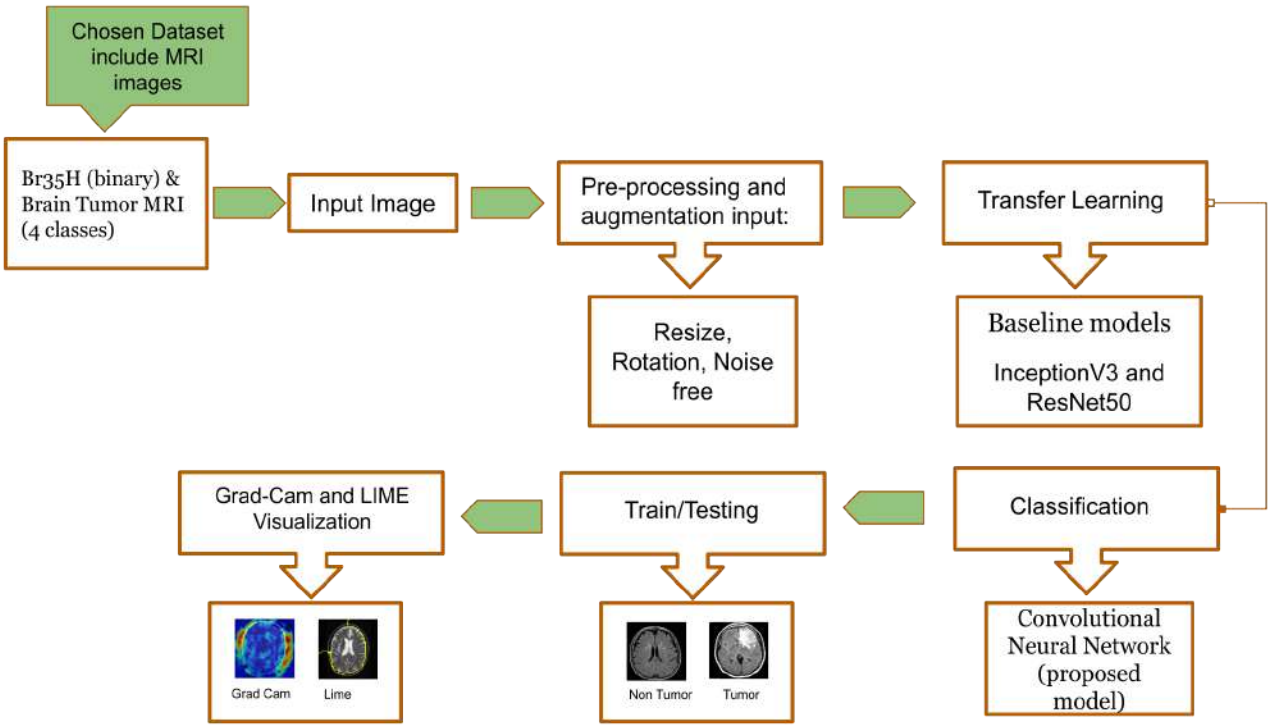

Figure 1. Brain Tumor Detection Framework.

for both binary and multi-class brain tumor classification tasks.

4.1 Dataset Description

This study utilized two publicly available brain tumor MRI datasets from the Kaggle platform. The first dataset, Br35H Brain Tumor Detection[1], contains 3,000 T1-weighted contrast-enhanced MRI images classified as either tumor or non-tumor. Each image represents an axial, coronal, or sagittal brain view, offering structural diversity that supports the development of robust deep learning models. All images are provided in standard formats (PNG or JPEG). For this study, we organized the images into training, validation, and testing subsets using stratified sampling to maintain class balance.

The second dataset, Brain Tumor MRI Dataset[2], consists of 7,153 MRI images divided into four categories: glioma, meningioma, pituitary, and no tumor. The images are stored in JPG format and are pre-partitioned into training (5,723 images, 80%) and testing (1,430 images, 20%) folders, with separate subdirectories for each tumor type. This multi-class dataset allows for evaluation of the proposed framework in a clinically relevant context, where the objective is both tumor detection and subtype classification.

For both datasets, we applied an identical preprocessing pipeline that included resizing to a fixed spatial resolution, denoising, and intensity normalization. Data augmentation techniques such as rotations, flips, translations, and zooming were applied to the training sets to enhance generalization. Figure 2 presents representative examples from both datasets, displaying non-tumor and tumor images from Br35H and the four classes from the Brain Tumor MRI dataset.

4.2 Data Preprocessing

A unified preprocessing pipeline was applied to both datasets to enhance image quality and improve model generalization, as illustrated in Figure 3. All MRI slices were resized to 128 $\times$ 128 pixels and converted to a consistent three-channel format suitable for convolutional neural network (CNN) input, while maintaining

[1] https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection

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

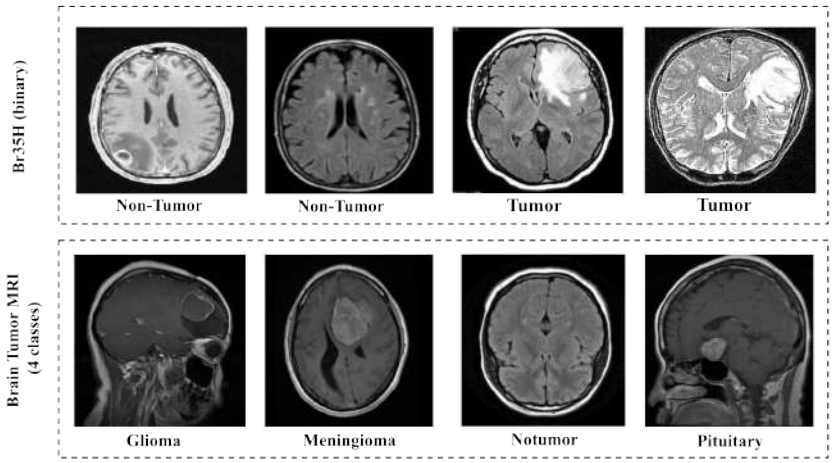

Figure 2. Sample images from the Both datasets.

the original single-intensity characteristics of brain MRI. Pixel intensities were normalized to the $[0, 1]$ range to reduce contrast variability across scans and to stabilize gradient-based optimization during training.

Data augmentation was applied to the training sets to increase the effective sample size and reduce overfitting. Augmentation operations included small rotations (up to 10°), horizontal flips, width and height shifts, shearing, and zooming. These transformations were selected to simulate plausible variations in patient positioning without altering tumor morphology. The `ImageDataGenerator` utility from TensorFlow/Keras was used to implement these transformations, perform on-the-fly rescaling, and maintain a consistent train–validation split (80–20 for each dataset). The preprocessed and augmented images were provided to the deep learning models as NumPy tensors, supporting efficient batch-wise training and reproducible experiments for both binary and multi-class classification tasks.

4.3 Model Architecture and Training

Three deep learning architectures were utilized to assess brain tumor classification performance: a custom Convolutional Neural Network (CNN) and two transfer-learning baselines, InceptionV3 and ResNet50. For the baseline models, convolutional backbones were initialized with ImageNet pre-trained weights. The original classification heads were replaced with fully connected layers tailored to the binary (Br35H) and four-class (Brain Tumor MRI) outputs. These models were then fine-tuned on preprocessed MRI images. This process established comparative benchmarks for the custom CNN.

The custom CNN was constructed as a compact yet expressive architecture. In our implementation, the CNN comprises three convolutional blocks with 32, 64, and 128 filters, respectively, each using 3×3 kernels followed by a dense layer with 256 units and a softmax output layer. Training is performed with a mini-batch size of 32 for up to 50 epochs, with early stopping based on validation accuracy and automatic learning-rate reduction when the validation metric plateaus. Each block is followed by batch normalization, a Rectified Linear Unit (ReLU) activation, and max pooling. This setup achieves progressive spatial downsampling and hierarchical feature extraction. The convolutional backbone is followed by fully connected layers, also with batch normalization and dropout. The network ends with a softmax output layer with either 2 or 4 neurons, depending on the task. This setup enables the network to capture discriminative local texture and structure for tumor and non-tumor regions. At the same time, it maintains controlled model complexity.

All models were trained using the Adam optimizer with an initial learning rate of 3×10^{-4}. Categorical cross-entropy loss and ℓ_2 weight regularization were applied to the dense layers. This helped reduce overfitting. Early stopping was based on validation accuracy.

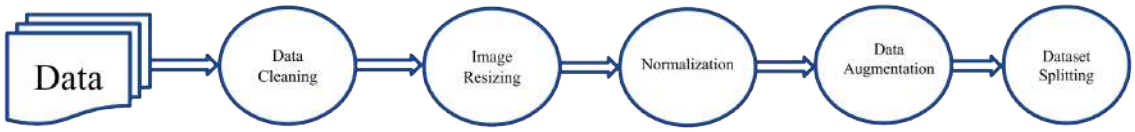

Figure 3. Data Preprocessing.

Adaptive learning rate reduction avoided unnecessary training epochs and accelerated convergence. Training used mini-batches generated by the data augmentation pipeline described earlier. Final model parameters were selected based on optimal validation performance. These were then used for subsequent quantitative evaluation and explainability analysis.

4.4 Evaluation Metrics

Model performance was assessed using four standard evaluation metrics: Accuracy, Precision, Recall, and F1-Score. These metrics are computed as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

where *TP*, *TN*, *FP*, and *FN* represent true positives, true negatives, false positives, and false negatives, respectively. These metrics provide a balanced evaluation of classification accuracy and robustness, particularly for imbalanced datasets.

4.5 Explainable AI Integration

To enhance interpretability and facilitate clinical adoption of the proposed framework, two complementary explainable AI techniques were integrated: Gradient Weighted Class Activation Mapping (Grad-CAM) and Local Interpretable Model-Agnostic Explanations (LIME). Grad-CAM was applied to the final convolutional layer of the trained convolutional neural network (CNN) to generate class-discriminative heatmaps that highlight image regions contributing most to each predicted class. These activation maps provide coarse localization of tumor-affected areas on MRI slices and enable clinicians to visually assess whether the network focuses on anatomically plausible regions.

Besides gradient-based explanations, LIME was used as a perturbation-based, model-agnostic model to get local surrogate explanations of a single prediction. On the few MRI slices of both the Br35H (binary classification) and Brain Tumor MRI (four-class classification) data sets, LIME divides the image into superpixels and perturbs them systematically to approximate their contribution to the probability prediction, producing superpixel-level importance masks. These masks determine areas that are positively correlated with the predicted class and offer a supplemental alternative view to Grad-CAM to highlight structures of influence.

The combined study of Grad-CAM heatmaps and LIME super pixel explanations can help radiologists to confirm that the CNN relies on its decisions on clinically significant tumor and non-tumor areas instead of focusing on the irrelevant background patterns or artifacts. The presented dual explainable AI approach increases the transparency, promotes the human-in-the-loop validation, and empowers the trust in the proposed AI-assisted brain tumor diagnosis pipeline both in binary and multi-class applications.

5 Results

In this study, we present an end-to-end approach to building a full-fledged, explainable deep learning framework for the classification of brain tumour from MRI scans. The proposed methodology combines systematic data preprocessing, a custom convolutional neural network (CNN) architecture, transfer learning baselines, and interpretable artificial intelligence (AI) methods to make accurate, well-explainable predictions on two publicly available datasets. We ran independent experiments on Br35H binary data and four-class Brain Tumor MRI data after preprocessing steps of resizing, normalization, and augmentation to assess both detection and subtype classification performance.

For each dataset, three deep learning models were

Table 2. Performance of Deep Learning Models on Br35H Brain Tumor Dataset (Binary Classification)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Proposed CNN	94.17	94.38	94.17	94.16
InceptionV3	77.20	73.05	77.20	69.13
ResNet50	70.51	64.37	70.51	60.69

Table 3. Performance of Deep Learning Models on Brain Tumor MRI Dataset (Four-Class Classification)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Proposed CNN (4 classes)	98.32	98.33	98.32	98.32
InceptionV3	78.24	84.25	78.28	77.93
ResNet50	96.64	97.79	96.68	98.21

trained and tested: the proposed CNN, InceptionV3, and ResNet50. For comparability, the models were trained with an identical train-validation-test protocol and optimized per layer using the Adam optimizer. The transfer learning models served as baselines, and the custom CNN was the focus due to its lightweight architecture and better compatibility with Grad-CAM and LIME interpretability techniques. To measure predictive performance and class-wise balance, accuracy, precision, recall, and F1-score were used for quantitative evaluation.

The performance of each model on the Br35H dataset is summarized in Table 2. Compared to baseline architectures, the suggested CNN surpasses them by a significant margin, with accuracy (94.17%), precision (94.38%), recall (94.17%), and F1-score (94.16%) all closely matched. For the four-class classification task on the Brain Tumor MRI dataset, as shown in Table 3, we obtained an accuracy of 98.32% with similarly high precision, recall, and F1-scores using our CNN. The results show that our proposed CNN performs well in both binary and multi-class settings while remaining computationally efficient. Beyond classification performance, Grad-CAM and LIME visualizations suggest that both methods qualitatively confirm that the driving features focus on clinically important regions of the tumor, supporting the transparency of our framework and its reliability in clinical application. Table 4 shows that the proposed CNN reaches between 97% and 99% F1-scores for all classes, confirming that high accuracy is always achieved regardless of whether it is glioma, meningioma, pituitary, or no-tumor.

5.1 Proposed CNN Performance

The proposed Convolutional Neural Network (CNN) yields strong, consistent results for both binary and multi-class brain tumor MRI classification tasks. The model obtains 94.17% accuracy on the Br35H dataset. Precision, recall, and F1-score are all around 94%. These findings demonstrate an even accuracy in discerning tumor versus non-tumor instances. No bias in favor of one class over the other. The performance demonstrates that the CNN effectively captures discriminative local texture and structural features. Its relatively small architecture also prevents overfitting.

A similar CNN architecture is altered for a four-class output layer on Brain Tumor MRI dataset. In addition, it reaches 98.32% accuracy while maintaining high precision, recall, and f1-score per class (glioma, meningioma, pituitary, and no-tumor). The classification report indicates that each class achieves F1 scores of 98%, 99%. This corroborates the network's capability to accurately differentiate between tumor subtypes and normal cases. The proposed CNN performs comparably to or outperforms state-of-the-art models such as InceptionV3 and ResNet50. It employs fewer parameters and produces better Grad-CAM and LIME explanations. A few recent state-of-the-art architectures achieve slightly better accuracies with more complex ensembles or multisource inputs. Nevertheless, the proposed CNN provides a practical trade-off between accuracy, interpretability, and computational efficiency. This strengthens its potential for embedding in clinically oriented, interpretable diagnostic pipelines.

Table 4. Class-wise performance of the proposed CNN on the Brain Tumor MRI dataset

Class	Precision (%)	Recall (%)	F1-Score (%)
Glioma	99.0	98.0	98.0
Meningioma	97.0	98.0	97.0
No tumor	99.0	98.0	99.0
Pituitary	99.0	99.0	99.0

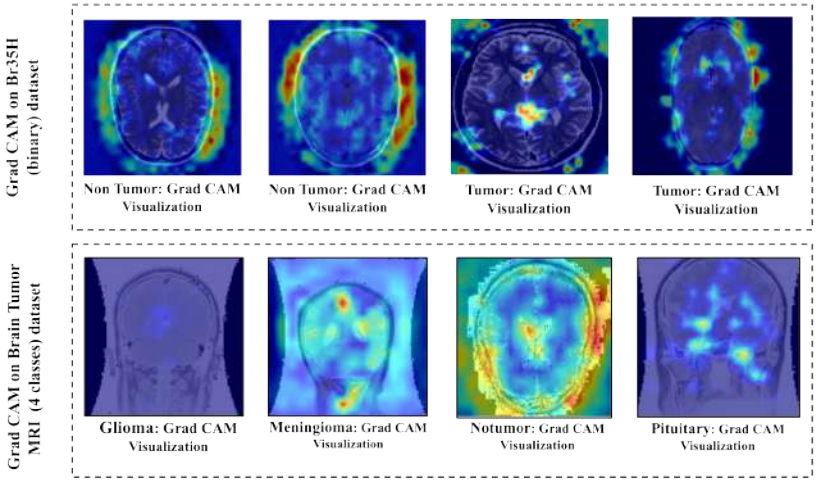

Figure 4. Grad-CAM visualizations of the proposed CNN on Br35H (binary) and Brain Tumor MRI (four-class) datasets

5.2 Grad-CAM and LIME Interpretations

Figure 4 presents Grad-CAM visualizations produced by the proposed convolutional neural network (CNN) for representative samples from both datasets. The upper panel depicts Br35H, where non-tumor and tumor cases demonstrate distinct activation patterns. The lower panel displays class-specific heatmaps for glioma, meningioma, no tumor, and pituitary cases from the Brain Tumor MRI dataset. In all instances, the Grad-CAM overlays highlight class-discriminative regions that contribute most significantly to the model's predictions.

For tumor cases, Grad-CAM consistently focuses on abnormal tissue, particularly at the lesion periphery, where texture and intensity patterns differ from the surrounding parenchyma. This local activation tends to coincide with tumor margins important to radiology implying that the CNN is based on clinically significant structures to formulate the predictions and not noise. Grad-CAM activation in non-tumor and no-tumor conditions is diffuse and weak and responses are only to normal anatomical landmarks, meaning that the model does not falsely detect pathology in structurally normal scans.

Figure 5 illustrate the LIME explanations of the same datasets with superpixel perturbations to find out the locally significant areas of the image. LIME (in Br35H) shows the neighboring superpixels in positive and negative cases around suspected tumors and on the less pathological areas, respectively, as a complement to coarser Grad-CAM heatmap. In the four-class Brain Tumor MRI case, LIME prioritizes superpixel configurations, which are based on the idea of cancer extent, shape and location of the tumor depending on glioma, meningioma, pituitary and no-tumors. Grad-CAM (gradient-based, feature-level) and LIME (perturbation-based, instance-level) can be used to gain complementary information and prove the fact that the CNN focuses on radiologically plausible tumor areas and thus enhances the interpretability and clinical plausibility of the framework.

5.3 Discussion

A comparison of InceptionV3, ResNet50, and the proposed convolutional neural network (CNN) reveals

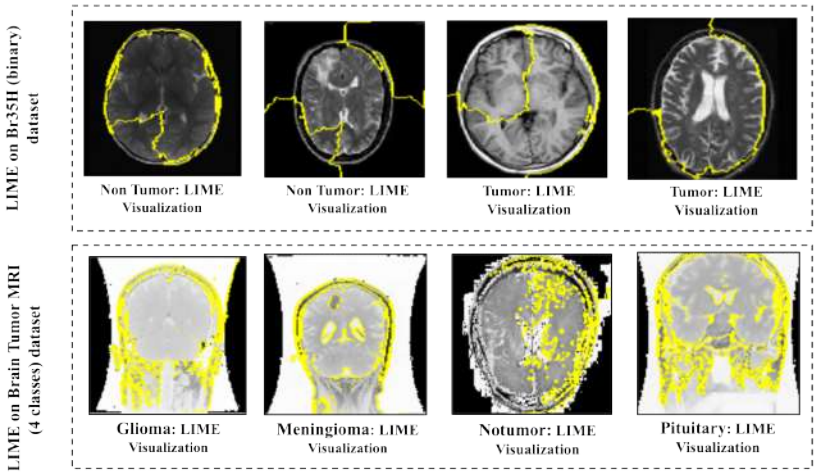

Figure 5. LIME-based superpixel explanations of the proposed CNN on Br35H (binary) and Brain Tumor MRI (four-class) datasets

significant differences in their performance for MRI-based brain tumor classification. The transfer learning baselines derived from ImageNet provided limited value on the Br35H dataset; while these models achieved stable results, they did not effectively capture the fine-grained intensity and image texture patterns characteristic of brain MRI, and the small size of the Br35H dataset further constrained their performance. In contrast, the custom-designed CNN, trained end-to-end on preprocessed MRI slices, achieved superior results, with an accuracy of 94.17% and closely aligned precision, recall, and F1-score metrics. The minimal variation among these metrics indicates that the CNN achieves a balanced trade-off between false positives and false negatives, demonstrating no substantial bias toward either category.

With the same CNN architecture, but the last softmax layer changed, the final result was a 98.32% accuracy on the multi-class Brain Tumor MRI dataset, and per-class F1-scores of 97 percent to 99% of glioma, meningioma, pituitary, and no-tumor. Such findings show that the model is very sensitive not only in the detection of the existence of tumors, but also in the differentiation between various types of tumours. The excellent performance on the binary and the four-class classification problems indicate that the CNN acquires generalizable tumor morphology representations so that it can be applied to the data sets that differ in terms of labeling schemes and acquisition procedures Table 5).

The suggested convolutional neural network (CNN) shows competitive performance, and increased usability in a variety of evaluation measures against previous research. As an example, Kumar et al. [51] obtained 96.47% accuracy using a hybrid CNN autoencoder on the Figshare multi-class dataset, albeit with overfitting and data asymmetry problems. The current CNN achieves an accuracy of 98.32% on the multi-class Brain Tumor MRI dataset and an accuracy of 94.17% on the binary Br35H dataset with using a simpler architecture without any hybrid block. On the same note, Seetha [37] trained a binary CNN classifier that had 94.39% accuracy on 2,065 MRI images, which again was restricted to one dataset. Compared to it, the proposed framework is more general and can accommodate binary and four class classification tasks and offers more interpretability with the help of Grad-CAM and LIME visualizations.

A hybrid Harmony Search Optimization (HSO)-CNN ensemble proposed by Saw et al. [49] achieved 99.13% accuracy, but this result required multimodal MRI and CT data as well as a more complex optimization pipeline, which substantially increased computational and implementation costs. The proposed CNN attains high diagnostic accuracy using only MRI data and a single, reusable architecture, thereby reducing resource consumption while maintaining explainability. Dhanalakshmi et al. [50] compared various saliency-based explainable artificial intelligence (XAI) techniques (Grad-CAM, Grad-CAM+, XraI) and concluded that expla-

Table 5. Comparative analysis with recent MRI brain tumor studies

Study (Year)	Dataset / Task	Model / Approach	Accuracy (%)
Iftikhar et al. [44]	Custom MRI dataset (glioma, meningioma)	Explainable CNN + Grad-CAM, LIME	98.2
Soewu et al.[48]	Multi-class MRI dataset (glioma, meningioma, pituitary)	Xception-based CNN + Grad-CAM, SHAP	98.78
Kumar et al. [51]	Figshare multi-class MRI dataset	Hybrid CNN autoencoder + ML classifiers	96.47
Saw et al. [49]	MRI + CT multimodal dataset	CNN + Harmony Search Optimization ensemble	99.13
Seetha [37]	2,065 MRI images (tumor vs non-tumor)	Custom CNN (binary classification)	94.39
Gulwadi [10]	Public MRI dataset (multi-type tumors)	ResNet50 + Grad-CAM	97.6
This study (2026)	Br35H (binary) and Brain Tumor MRI (four class)	Proposed CNN + Grad-CAM + Lime	**94% and 98%**

nation stability may vary significantly depending on the dataset. Building on this observation, the present study demonstrates that Grad-CAM, in combination with LIME strategies, generates clinically meaningful and repeatable attention maps for both binary and multi-class brain tumor MRI classification tasks. Similarly, Soewu et al. [48] achieved 98.78% accuracy with an Xception-based model using Grad-CAM and SHAP; however, their approach relies on a more computationally intensive backbone and does not include explicit cross-dataset evaluation as conducted in this study.

Overall, the results presented in Table 5 establish the proposed CNN as a robust, interpretable, and relatively low-cost model for MRI-based brain tumor localization and subtype classification. The framework demonstrates strong generalization across two heterogeneous public datasets and approaches state-of-the-art performance on the multi-class Brain Tumor MRI dataset. Additionally, it offers complementary Grad-CAM and LIME visualizations, although its numerical performance on Br35H is marginally lower than that of some highly optimized models. Such features make the framework a credible instrument of AI-assisted diagnosis, and provide a realistic compromise of functionality, transparency and clinical practicality in the analysis of medical images.

6 Conclusion

In this work, an explainable deep learning framework was introduced to detect and classify brain tumor types using MRI images. It provides a state-space admission control mixed-validation framework by integrating feature extraction from a convolutional neural network with an advanced data preprocessing pipeline and two explainable AI techniques, thereby optimizing diagnostic performance for both patients and treating departments while considering computational affordability. Results reported an overall accuracy of 94.17% on Br35H and 98.32% on the multi-class dataset with consistently high precision, recall, and F1-scores for all classes using two publicly available datasets to validate its method: Br35H (binary detection of tumors) and a four-class Brain Tumor MRI dataset comprising glioma, meningioma, pituitary, and no-tumor classes. Such quantitative performance, with very good generalization across both binary and multi-class scenarios, is supported by Grad-CAM and LIME visualizations, which further

confirm that the model focuses only on clinically informative regions, aiding human-in-the-loop validation of an automated prediction system and increasing clinician trust in predictions.

Author Contributions

Sidra Hameed: Conceptualization, Methodology, Conceptualized the study and led the research direction, Writing – Original Draft **Atta Ullah**: Main Contribution; Data Curation: Collected, cleaned, and organized datasets; Writing – Original Draft Preparation: Authored the initial manuscript draft.

Data Availability

The MRI datasets used in this study are publicly available on Kaggle. The Br35H Brain Tumor MRI dataset can be accessed at https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-dataset, and the four-class Brain Tumor MRI dataset (glioma, meningioma, pituitary, and no tumor) can be accessed at https://www.kaggle.com/datasets/sartajbhuvaji/brain-tumor-classification-mri. All preprocessing and training scripts are available from the corresponding author on reasonable request.

The MRI datasets used in this study are publicly available on Kaggle. The Br35H Brain Tumor MRI dataset can be accessed at

`https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-dataset` and the four-class Brain Tumor MRI dataset (glioma, meningioma, pituitary, and no tumor) can be accessed at

`https://www.kaggle.com/datasets/sartajbhuvaji/brain-tumor-classification-mri`. All preprocessing and training scripts are available from the corresponding author on reasonable request.

Declarations

Conflict of Interest

The authors declare that they have no conflict of interest.

Funding

This research received no external funding.

Ethical Approval

The study used a publicly available de-identified dataset (Br35H) obtained from Kaggle, and no institutional ethical approval was required.

Informed Consent

Not applicable, as all data were anonymized and publicly available.

References

[1] R. L. Siegel, T. B. Kratzer, A. N. Giaquinto, H. Sung, and A. Jemal, "Cancer statistics, 2025," *CA: A Cancer Journal for Clinicians*, vol. 75, no. 1, p. 10, 2025.

[2] Q. T. Ostrom, M. Price, C. Neff, G. Cioffi, K. A. Waite, C. Kruchko, and J. S. Barnholtz-Sloan, "CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2015–2019," *Neuro-Oncology*, vol. 24, no. 5, pp. v1–v95, 2022.

[3] M. Rana, N. A. Khan, and M. Hussain, "LIME-based Explainable AI classifier to detect COVID-19 pandemic through X-ray images," *VAWKUM Transactions on Computer Sciences*, vol. 14, no. 1, pp. 28–39, 2026.

[4] D. N. Louis, A. Perry, P. Wesseling, D. J. Brat, I. A. Cree, D. Figarella-Branger, C. Hawkins, H. K. Ng, S. M. Pfister, G. Reifenberger, *et al.*, "The 2021 WHO classification of tumors of the central nervous system: A summary," *Neuro-Oncology*, vol. 23, no. 8, pp. 1231–1251, 2021.

[5] M. Weller, M. van den Bent, K. Hopkins, *et al.*, "EANO guidelines on the diagnosis and treatment of adult astrocytic and oligodendroglial gliomas," *Lancet Oncology*, vol. 22, no. 10, pp. e391–e406, 2021.

[6] A. Chattopadhyay and M. Maitra, "MRI-based brain tumour image detection using CNN based deep learning method," *Neuroscience Informatics*, vol. 2, no. 4, p. 100060, 2022.

[7] D. Lamrani, B. Cherradi, O. El Gannour, M. A. Bouqentar, and L. Bahatti, "Brain tumor detection using MRI images and convolutional neural network," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 7, 2022.

[8] K. S. A. Kumar, A. Y. Prasad, and J. Metan, "A hybrid deep CNN-Cov-19-Res-Net transfer learning architecture for an enhanced brain tumor detection and classification scheme in medical image processing," *Biomedical Signal Processing and Control*, vol. 76, p. 103631, 2022.

[9] T. L. Prasanthi and N. Neelima, "Improvement of brain tumor categorization using deep learning: a comprehensive investigation and comparative analysis," *Procedia Computer Science*, vol. 233, pp. 703–712, 2024.

[10] S. Guluwadi, *et al.*, "Enhancing brain tumor detection in MRI images through explainable AI using Grad-CAM with ResNet 50," *BMC Medical Imaging*, vol. 24, no. 1, pp. 1–19, 2024.

[11] Y. H. Chel and L. L. Poh, "Brain tumor classification in MRI: Insights from LIME and Grad-CAM explainable AI techniques," *IEEE Access*, 2025.

[12] Y. Tian, "Trade-off analysis of classical machine learning and deep learning models for robust brain tumor detection: Benchmark study," *JMIR AI*, vol. 4, p. e76344, 2025.

[13] T. Berghout, "The neural frontier of future medical imaging: A review of deep learning for brain tumor detection," *Journal of Imaging*, vol. 11, no. 1, p. 2, 2024.

[14] M. Abou Ali, J. Charafeddine, F. Dornaika, and I. Arganda-Carreras, "Enhancing generalization and mitigating overfitting in deep learning for brain cancer diagnosis from MRI," *Applied Magnetic Resonance*, vol. 56, no. 3, pp. 359–394, 2025.

[15] V. Barra, "Convergence of man and machine artificial intelligence in human life," *Journal of Inclusive Methodology and Technology in Learning and Teaching*, vol. 3, no. 4, 2023.

[16] J. van der Laak, G. Litjens, and F. Ciompi, "Deep learning in histopathology: the path to the clinic," *Nature Medicine*, vol. 27, no. 5, pp. 775–784, 2021.

[17] J. Seah, Z. Brady, K. Ewert, and M. Law, "Artificial intelligence in medical imaging: implications for patient radiation safety," *The British Journal of Radiology*, vol. 94, no. 1126, p. 20210406, 2021.

[18] R. Najjar, "Digital frontiers in healthcare: Integrating mHealth, AI, and radiology for future medical," *A Comprehensive Overview of Telemedicine*, vol. 307, 2024.

[19] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on MRI," *Zeitschrift fuer Medizinische Physik*, vol. 29, no. 2, pp. 102–127, 2019.

[20] C. Mello-Thoms and C. A. B. Mello, "Clinical applications of artificial intelligence in radiology," *The British Journal of Radiology*, vol. 96, no. 1150, p. 20221031, 2023.

[21] R. A. El Arab, M. S. Abu-Mahfouz, F. H. Abuadas, H. Alzghoul, M. Almari, A. Ghannam, and M. M. Seweid, "Bridging the gap: from AI success in clinical trials to real-world healthcare implementation—a narrative review," in *Healthcare*, vol. 13, no. 7, p. 701, 2025.

[22] A. S. Tejani, Y. S. Ng, Y. Xi, and J. C. Rayan, "Understanding and mitigating bias in imaging artificial intelligence," *Radiographics*, vol. 44, no. 5, p. e230067, 2024.

[23] S. N. Saw, Y. Y. Yan, and K. H. Ng, "Current status and future directions of explainable artificial intelligence in medical imaging," *European Journal of Radiology*, vol. 183, p. 111884, 2025.

[24] E. Neri, G. Aghakhanyan, M. Zerunian, N. Gandolfo, R. Grassi, V. Miele, A. Giovagnoni, A. Laghi, and SIRM Expert Group on Artificial Intelligence, "Explainable AI in radiology: a white paper of the Italian Society of Medical and Interventional Radiology," *La Radiologia Medica*, vol. 128, no. 6, pp. 755–764, 2023.

[25] C. Sáez, P. Ferri, and J. M. García-Gómez, "Resilient artificial intelligence in health: synthesis and research agenda toward next-generation trustworthy clinical decision support," *Journal of Medical Internet Research*, vol. 26, p. e50295, 2024.

[26] A. Rahman, M. S. Hossain, G. Muhammad, D. Kundu, T. Debnath, M. Rahman, M. S. I. Khan, P. Tiwari, and S. S. Band, "Federated learning-based AI approaches in smart healthcare: concepts, taxonomies, challenges and open issues," *Cluster Computing*, vol. 26, no. 4, pp. 2271–2311, 2023.

[27] Y. Zhou, Y. Wu, Y. Su, J. Li, J. Cai, Y. You, J. Zhou, D. Guo, and X. Qu, "Cloud-magnetic resonance imaging system: In the era of 6G and artificial intelligence," *Magnetic Resonance Letters*, vol. 5, no. 1, p. 200138, 2025.

[28] G. Nadeem and M. I. Anis, "Ethical and regulatory consideration in AI-based medical imaging," in *AI for Medical Image Analysis: Reconciling Innovation and Ethical Considerations*, pp. 265–290, 2025.

[29] M. Y. Shakor and M. I. Khaleel, "Recent advances in big medical image data analysis through deep learning and cloud computing," *Electronics*, vol. 13, no. 24, p. 4860, 2024.

[30] M. Badawy, N. Ramadan, and H. A. Hefny, "Big data analytics in healthcare: data sources, tools, challenges, and opportunities," *Journal of Electrical Systems and Information Technology*, vol. 11, no. 1, p. 63, 2024.

[31] N. McCarthy, A. Dahlan, T. S. Cook, N. O'Hare, M.-L. Ryan, B. St John, A. Lawlor, and K. M. Curran, "Enterprise imaging and big data: A review from a medical physics perspective," *Physica Medica*, vol. 83, pp. 206–220, 2021.

[32] T. Geroski, D. Jakovljević, and N. Filipović, "Big data in multiscale modelling: from medical image processing to personalized models," *Journal of Big Data*, vol. 10, no. 1, p. 72, 2023.

[33] R. Egala and M. V. S. Sairam, "A review on medical image analysis using deep learning," *Engineering Proceedings*, vol. 66, no. 1, p. 7, 2024.

[34] K. Memon, N. Yahya, M. Z. Yusoff, R. Remli, A.-W. M. Mustapha, H. Hashim, S. S. A. Ali, and S. Siddiqui, "Edge computing for AI-based brain MRI applications: a critical evaluation of real-time classification and segmentation," *Sensors*, vol. 24, no. 21, p. 7091, 2024.

[35] F. Mohsen, H. Ali, N. El Hajj, and Z. Shah, "Artificial intelligence-based methods for fusion of electronic health records and imaging data," *Scientific Reports*, vol. 12, no. 1, p. 17981, 2022.

[36] N. Abiwinanda, M. Hanif, S. T. Hesaputra, A. Handayani, and T. R. Mengko, "Brain tumor classification using convolutional neural network," in *World Congress on Medical Physics and Biomedical Engineering 2018*, pp. 183–189, 2018.

[37] J. Seetha and S. S. Raja, "Brain tumor classification using convolutional neural networks," *Biomedical & Pharmacology Journal*, vol. 11, no. 3, p. 1457, 2018.

[38] M. M. Zahoor, S. H. Khan, T. J. Alahmadi, T. Alsahfi, A. S. A. Mazroa, H. A. Sakr, S. Alqahtani, A. Albanyan, and B. K. Alshemaimri, "Brain tumor MRI classification using a novel deep residual and regional CNN," *Biomedicines*, vol. 12, no. 7, p. 1395, 2024.

[39] B. V. Babu, S. Srinivasan, S. K. Mathivanan, Mahalakshmi, P. Jayagopal, and G. T. Dalu, "Detection and classification of brain tumor using hybrid deep learning models," *Scientific Reports*, vol. 13, no. 1, p. 23029, 2023.

[40] M. S. Pinto, R. Paolella, T. Billiet, P. Van Dyck, P.-J. Guns, B. Jeurissen, A. Ribbens, A. J. den Dekker, and J. Sijbers, "Harmonization of brain diffusion MRI: Concepts and methods," *Frontiers in Neuroscience*, vol. 14, p. 396, 2020.

[41] M. Bento, I. Fantini, J. Park, L. Rittner, and R. Frayne, "Deep learning in large and multi-site structural brain MR imaging datasets," *Frontiers in Neuroinformatics*, vol. 15, p. 805669, 2022.

[42] S. Krishnapriya and Y. Karuna, "A survey of deep learning for MRI brain tumor segmentation methods: Trends, challenges, and future directions," *Health and Technology*, vol. 13, no. 2, pp. 181–201, 2023.

[43] A. Al Noman and A. S. M. Arif, "Brain tumor recognition from MRI using deep learning with data balancing methods and its explainability with AI," in *International Conference on Image Processing and Capsule Networks*, pp. 523–538, 2023.

[44] S. Iftikhar, N. Anjum, A. B. Siddiqui, M. U. Rehman, and N. Ramzan, "Explainable CNN for brain tumor detection and classification through XAI based key features identification," *Brain Informatics*, vol. 12, no. 1, p. 10, 2025.

[45] Asmita and P. Mittal, "From black box AI to XAI in neuro-oncology: a survey on MRI-based tumor detection," *Discover Artificial Intelligence*, vol. 5, no. 1, p. 30, 2025.

[46] S. Sarker, "Transfer learning and explainable AI for brain tumor classification: A study using MRI data from Bangladesh," in *2024 6th International Conference on Sustainable Technologies for Industry 5.0 (STI)*, pp. 1–6, 2024.

[47] Y. H. Chel and L. L. Poh, "Brain tumor classification in MRI: Insights from LIME and Grad-CAM explainable AI techniques," *IEEE Access*, 2025.

[48] T. Soewu, D. Singh, M. Rakhra, G. S. Chakraborty, and A. Singh, "Convolutional neural networks for MRI-based brain tumor classification," in *2022 3rd International Conference on Computation, Automation and Knowledge Management (ICCAKM)*, pp. 1–7, 2022.

[49] S. N. Saw, Y. Y. Yan, and K. H. Ng, "Current status and future directions of explainable artificial intelligence in medical imaging," *European journal of radiology*, vol. 183, p. 111884, 2025.

[50] S. Dhanalakshmi and S. Arulselvi, "Using ResNet architecture with MRI for classification of brain images," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 39, no. 1, pp. 148–158, 2025.

[51] C. M. Kumar and J. S. Sankar, "Comparative analysis of convolutional neural networks for brain tumor detection: A study of VGG16, ResNet, Inception, and DenseNet models," in *2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, pp. 41–46, 2024.

[52] H. Mzoughi, I. Njeh, M. BenSlima, N. Farhat, and C. Mhiri, "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*, vol. 41, no. 4, pp. 2123–2142, 2025.

[53] M. W. Nadeem, M. A. Al Ghamdi, M. Hussain, M. A. Khan, K. M. Khan, S. H. Almotiri, and S. A. Butt, "Brain tumor analysis empowered with deep learning: A review, taxonomy, and future challenges," *Brain sciences*, vol. 10, no. 2, p. 118, 2020.

[54] 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.

[55] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud, and A. Hussain, "Interpreting black-box models: a review on explainable artificial intelligence," *Cognitive Computation*, vol. 16, no. 1, pp. 45–74, 2024.

[56] V. Chamola, V. Hassija, A. R. Sulthana, D. Ghosh, D. Dhingra, and B. Sikdar, "A review of trustworthy and explainable artificial intelligence (XAI)," *IEEe Access*, vol. 11, pp. 78994–79015, 2023.

[57] S. Hossain, A. Chakrabarty, T. R. Gadekallu, M. Alazab, and M. J. Piran, "Vision transformers, ensemble model, and transfer learning leveraging explainable AI for brain tumor detection and classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 3, pp. 1261–1272, 2023.

[58] R. Kumar, P. Singh, and A. Aggarwal, "Challenges and opportunities of deep learning in brain MRI analysis: An overview," *Artificial Intelligence in Medicine*, vol. 144, p. 102679, 2023, doi: 10.1016/j.artmed.2023.102679.

[59] X. Zhou, Y. Zhang, and J. Chen, "Limitations of Grad-CAM in medical imaging and prospects for model interpretability," *Pattern Recognition Letters*, vol. 178, pp. 50–62, 2025, doi: 10.1016/j.patrec.2024.12.008.

[60] R. Amin, M. A. Al Ghamdi, S. H. Almotiri, M. Alruily, *et al.*, "Healthcare techniques through deep learning: issues, challenges and opportunities," *IEEE Access*, vol. 9, pp. 98523–98541, 2021.

[61] S. S. Band, A. Yarahmadi, C.-C. Hsu, M. Biyari, M. Sookhak, R. Ameri, I. Dehzangi, A. T. Chronopoulos, and H.-W. Liang, "Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods," *Informatics in Medicine Unlocked*, vol. 40, p. 101286, 2023.

[62] J. Schmidt, N. M. Schutte, S. Buttigieg, D. Novillo-Ortiz, E. Sutherland, M. Anderson, B. de Witte, M. Peolsson, B. Unim, M. Pavlova, *et al.*, "Mapping the regulatory landscape for artificial intelligence in health within the European Union," *NPJ Digital Medicine*, vol. 7, no. 1, p. 229, 2024.

[63] M. Aamir, Z. Rahman, N. Choudhry, J. A. Bhutto, W. A. Abro, and Z. Zhu, "From CNNs to Transformers: A Review of Evolving Deep Learning Architectures for Brain Tumor Classification," *IEEE Access*, 2025.

[64] M. Rana and M. Bhushan, "Machine learning and deep learning approach for medical image analysis: diagnosis to detection," *Multimedia Tools and Applications*, vol. 82, no. 17, pp. 26731–26769, 2023.

[65] Y. Xu, R. Quan, W. Xu, Y. Huang, X. Chen, and F. Liu, "Advances in medical image segmentation: A comprehensive review of traditional, deep learning and hybrid approaches," *Bioengineering*, vol. 11, no. 10, p. 1034, 2024.

[66] A. Khalili Fakhrabadi, "A Hybrid Inception-Dilated-ResNet Architecture for Deep Medical Image Analysis," *Scientific Reports*, 2025. Available from: https://www.nature.com/articles/s41598-025-91322-3.

[67] M. M. Zahoor, S. H. Khan, T. J. Alahmadi, T. Alsahfi, A. S. A. Mazroa, H. A. Sakr, S. Alqahtani, A. Albanyan, and B. K. Alshemaimri, "Brain tumor MRI classification using a novel deep residual and regional CNN," *Biomedicines*, vol. 12, no. 7, p. 1395, 2024.

[68] V. V. S. Sasank and S. Venkateswarlu, "Hybridized deep neural network using adaptive rain optimizer algorithm for multi-grade brain tumor classification of MRI images," *Digital Technologies Research and Applications*, vol. 1, no. 1, pp. 13–30, 2022.

[69] H. Zhang and K. Ogasawara, "Grad-CAM-based explainable artificial intelligence related to medical text processing," *Bioengineering*, vol. 10, no. 9, p. 1070, 2023.

[70] Edge Impulse, "AI explainability with Grad-CAM: Visualizing neural network decisions," 2025. [Online]. Available: https://www.edgeimpulse.com/blog/ai-explainability-with-grad-cam-visualizing-neural-network-decisions. [Accessed: Oct. 2025].

[71] D. Minh, H. X. Wang, Y. F. Li, and T. N. Nguyen, "Explainable artificial intelligence: a comprehensive review," *Artificial Intelligence Review*, vol. 55, no. 5, pp. 3503–3568, 2022.

[72] D. R. Johnson, R. Srivats, A. Sharma, and V. Kalyanasundaram, "Explainable AI and deep learning for brain tumor classification: A comprehensive approach using Grad-CAM visualization," in *International Conference on Computer Vision and Image Processing*, pp. 160–174, 2024.

[73] I. D. Mienye, "A survey of explainable artificial intelligence in healthcare," *Artificial Intelligence in Medicine*, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352914824001448

[74] N. Subbarayudu, S. Kollem, and T. Venkatakrishnamoorthy, "Explainable MRI-Based Brain Tumor Detection Using ResNet50 with Grad-CAM-Driven Visualization," in *2025 2nd International Conference on Intelligent Systems for Cybersecurity (ISCS)*, pp. 1–5, 2025.

[75] A. Rath, "ResNet50-based deep learning model for accurate brain tumor segmentation and classification with Grad-CAM visualization," *Journal of Medical Imaging and Health Informatics*, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S3050475924001039

[76] Y. Zhang, "Challenges and opportunities in deploying deep learning models for medical imaging," *Frontiers in Materials*, vol. 10, p. 1583615, 2025. [Online]. Available: https://www.frontiersin.org/journals/materials/articles/10.3389/fmats.2025.1583615/full

[77] A. Khan, "Deep learning in medical imaging for disease diagnosis: Opportunities and challenges," *World Journal of Advanced Research and Reviews*, vol. 25, pp. 2522–2526, 2025. [Online]. Available: https://journalwjarr.com/sites/default/files/fulltext_pdf/WJARR-2025-0558.pdf

[78] J. Wang, "Robustness and interpretability of AI models in medical imaging under domain shifts and data variability," *Radiology: Artificial Intelligence*, 2025. [Online]. Available: https://pubs.rsna.org/doi/10.1148/ryai.250682

[79] A. Y. Jaffar, "Combining local and global feature extraction for brain tumor classification: A vision transformer and iResNet hybrid model," *Engineering, Technology & Applied Science Research*, 2024. [Online]. Available: https://etasr.com/index.php/ETASR/article/view/8271

[80] K. Suneetha, "Brain tumor detection using deep learning," *Journal of Theoretical and Applied Information Technology*, vol. 103, no. 11, 2025. [Online]. Available: http://www.jatit.org/volumes/Vol103No11/25Vol103No11.pdf

[81] J. Andreu-Perez, "Big data in multiscale modelling: From medical image processing to personalized models," *Journal of Big Data*, vol. 10, no. 1, p. 77, 2023, doi: 10.1186/s40537-023-00763-y.

[82] S. Hameed, M. Nauman, M. Hasnain, N. Akhtar, F. Hussain, and Z. Afzal, "An explainable deep learning framework for automated classification of ocular diseases in a big data environment," *VFAST Transactions on Software Engineering*, vol. 13, no. 3, pp. 258–278, 2025.

[83] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[84] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Applications of Computer Vision (WACV)*, 2018, pp. 839–847, doi: 10.1109/WACV.2018.00097.

[85] X. Zhou, Y. Zhang, and J. Chen, "Limitations of Grad-CAM in medical imaging and prospects for model interpretability," *Pattern Recognition Letters*, vol. 178, pp. 50–62, 2025, doi: 10.1016/j.patrec.2024.12.008.

[86] D. Tang, J. Chen, L. Ren, X. Wang, D. Li, and H. Zhang, "Reviewing CAM-based deep explainable methods in healthcare," *Applied Sciences*, vol. 14, no. 10, p. 4124, 2024.

[87] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should i trust you?" Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

[88] S. Hameed, M. Nauman, N. Akhtar, M. A. B. Fayyaz, and R. Nawaz, "Explainable AI-driven depression detection from social media using natural language processing and black box machine learning models," *Frontiers in Artificial Intelligence*, vol. 8, p. 1627078, 2025.

[89] L. A. Abraham, G. Palanisamy, and G. Veerapu, "Transparent brain tumor detection using DenseNet169 and LIME," *Scientific Reports*, vol. 15, no. 1, p. 28185, 2025.

[90] N. Ullah, M. Hassan, J. A. Khan, M. S. Anwar, and K. Aurangzeb, "Enhancing explainability in brain tumor detection: A novel DeepEBTDNet model with LIME on MRI images," *International Journal of Imaging Systems and Technology*, vol. 34, no. 1, p. e23012, 2024.

[91] B. T. da Costa, "Explainable AI in Medical Applications," Master's thesis, Universidade de Aveiro (Portugal), 2023.

[92] B. Aldughayfiq, F. Ashfaq, N. Z. Jhanjhi, and M. Humayun, "Explainable AI for retinoblastoma diagnosis: interpreting deep learning models with LIME and SHAP," *Diagnostics*, vol. 13, no. 11, p. 1932, 2023.

[93] A. Chakrabarty and S. Das, "Br35H brain tumor MRI dataset," Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection. [Accessed: Oct. 16, 2025].

[94] E. Afrihyiav, E. C. Chianumba, A. Y. Forkuo, O. Omotayo, O. O. Akomolafe, and A. Y. Mustapha, "Explainable AI in healthcare: visualizing black-box models for better decision-making," *Unpublished manuscript*, 2022.

[95] M. W. Rahman, M. R. Khan, and M. Nijim, "Privacy-Preserving Deep Learning for Disease Diagnosis in Medical Imaging: A Systematic Review," *IEEE Access*, 2025.