SPECTRUM OF ENGINEERING SCIENCES
ISSN (E) 3007-3138 (P) 3007-312X Volume 4, Issue 6, 2026

INTERPRETABLE DEEP LEARNING MODELS FOR CLASSIFICATION OF BRAIN TUMORS VIA MRI

Ilya Haider[*1], Muhammad Haqan Ali Rai[2], Bhavnesh Deep[3], Qadeer Ishfaq[4]

[*1]*BS Software Engineering, Department of Software Engineering, Government College University, Faisalabad, Pakistan*
[2]*BS Software Engineering, University Institute of Information Technology (UIIT), PMAS Arid Agriculture University, Rawalpindi, Pakistan*
[3]*BS Computer Science, Department of Computer Science, Shaheed Zulfikar Ali Bhutto Institute of Science and Technology (SZABIST) University, Pakistan*
[4]*MS Information Technology, Department of Computer Science & Information Technology, Ghazi University, Dera Ghazi Khan, Pakistan*

[*1]ilyahaider15313@gmail.com, [2]haqanali934@gmail.com, [3]bhavneshkarmani485@gmail.com, [4]qadeerishfaq47@gmail.com

DOI: https://doi.org/10.5281/zenodo.20570048

Keywords
Explainable Artificial Intelligence (XAI), Interpretable Machine Learning, White-box AI, Transparent AI, Deep Learning, Deep Neural Networks (DNN), Hierarchical Learning, Brain Tumor Classification, Intracranial Neoplasm Categorization, Brain Lesion Identification, Cerebral Tumor Grading, Magnetic Resonance Imaging (MRI), MR Imaging, Neuroimaging Data, Convolutional Neural Networks

Article History
Received: 09 April 2026
Accepted: 21 May 2026
Published: 06 June 2026

Corresponding Author: *
Ilya Haider

Abstract
Brain tumors are super serious neurological issues, and getting them diagnosed quickly and right is key for better outcomes and effective treatment plans. Doctors use Magnetic Resonance Imaging (MRI) a lot because it does the best job of showing soft tissues and giving detailed views of the brain. Recently, tools like Convolutional Neural Networks (CNNs) in deep learning have gotten really good at classifying these tumors automatically. Yet, there's a catch – these models are like black boxes; no one can see how they make decisions. This makes doctors and other health pros wary about using them. Our study aims to tackle this by coming up with an Explainable Deep Learning (XDL) framework. It lets us classify brain tumors accurately from MRI scans while also making it clear how those decisions are reached. The proposed method uses a deep convolutional neural network, trained on processed MRI images, to classify brain tumors into types like glioma, meningioma, and pituitary tumors. To boost transparency and clinician trust, they added explainability techniques, including Grad-CAM, LIME, and attention visualization. These methods show which parts of an image influenced the model's decision, helping radiologists see why the system thinks a tumor is one type over another tests show that this model performs really well in terms of accuracy, precision, recall, F1-score, and AUC. It does this while offering clear visual explanations too. This proves that explainable AI can help bridge the gap between tech and healthcare decisions. By doing so, it makes AI models more reliable, transparent, and trustworthy for doctors. The work fits into the bigger picture of making medical AI trustworthy and supports radiologists in accurately and clearly diagnosing brain tumors.

1. INTRODUCTION

1.1 Background of the Study

Brain tumors are super serious and can be really life-threatening. They're abnormal growths of cells inside the brain that could be either benign or malignant. These things cause all sorts of health problems, like trouble with thinking and moving, seizures, and sometimes they're fatal. Because tumors are so tricky to deal with and because they're located in the brain, doctors have a tough time figuring out how to diagnose and treat them effectively. Plus, there's been an increase in brain-related diseases which makes it extra important to develop better ways to catch tumors early on. On a global scale, brain tumors play a big role in raising cancer-related illness and death rates. Each year, hundreds of thousands of new cases pop up, and sadly, many folks don't survive. It's especially worrying since tumors hit people of all ages—children and adults alike. Even with advances in tech and treatments, survival rates aren't great, especially for aggressive tumors like glioblastomas. So, there's this huge need for new diagnostic methods that can help spot these issues earlier and give doctors a fighting chance..

Early diagnosis is key to boosting survival rates and making treatments more effective. When brain tumors are caught early, doctors can plan surgery and other therapies like radiation, chemotherapy, and targeted treatments sooner. This helps improve how patients fare and feel overall. Because of this, coming up with smart and dependable diagnostic systems is a big priority in healthcare research nowadays.

1.2 MRI in Brain Tumor Diagnosis

M Magnetic Resonance Imaging (MRI) is thought to be one of the best and most widely used ways to spot brain tumors. Unlike CT scans, MRI does an amazing job at showing off soft tissue details while keeping patients safe from ionizing radiation. This lets doctors see the brain's detailed structures really well and spot any issues linked to tumors.MRI is super helpful for figuring out the size, place, shape, and impact of a tumor on nearby brain areas. That's why it's a must-have for diagnosing problems and planning treatments.

There are different kinds of MRI scans that each give valuable info to help find and understand brain tumors better. T1-weighted MRIs give clear anatomical details, helping assess tissue make-up. T2-weighted MRIs show fluid-filled spots and swelling around tumors, highlighting weird changes well. FLAIR MRIs dampen cerebrospinal fluid signals to bring out odd areas close to those fluid spaces. Using all these types together gives doctors a fuller picture of the tumor and boosts their ability to make a correct diagnosis. TMRI's value isn't just about finding tumors. It helps doctors track how cancer spreads, see if treatments are working, and plan surgeries. Reading MRI scans accurately is key to figuring out the tumor grade and picking the right therapies. Yet, the huge amount of MRI info in hospitals now needs automated systems to assist with decisions and lighten the load on doctors.

1.3 Artificial Intelligence in Healthcare

Artificial Intelligence (AI) has emerged as a transformative technology in healthcare, revolutionizing the way medical data are analyzed and interpreted. Over the past decade, advances in computational power, data availability, and machine learning algorithms have enabled AI systems to perform complex tasks that were traditionally dependent on human expertise. In medicine, AI has been applied to disease diagnosis, drug discovery, patient monitoring, treatment optimization, and predictive analytics. These developments have significantly enhanced healthcare efficiency, accuracy, and accessibility across various medical specialties.

One of the most promising applications of AI is AI-assisted diagnosis, where intelligent systems support clinicians in identifying diseases from medical images, laboratory reports, and patient records. AI-powered diagnostic tools can process large volumes of data rapidly and detect subtle patterns that may not be immediately apparent to human observers. In radiology, AI systems have demonstrated remarkable performance in detecting abnormalities in MRI, CT, X-ray, and ultrasound images. Such capabilities can reduce diagnostic errors, improve consistency, and assist healthcare professionals in making informed decisions.

Deep learning, a specialized branch of AI, has shown exceptional success in medical image analysis. Deep learning models, particularly Convolutional Neural Networks (CNNs), automatically learn hierarchical features directly from raw image data, eliminating the need for manual feature extraction. These models have achieved state-of-the-art performance in tasks such as tumor classification, lesion detection, organ segmentation, and disease prediction. As a result, deep learning has become a cornerstone of modern medical imaging research and continues to drive innovation in AI-assisted healthcare solutions.

1.4 Problem Statement

Even with all the progress in medical imaging, diagnosing brain tumors is still really tough. Right now, doctors use MRIs, but a human expert needs to read these images by hand. Although experts are crucial, there are issues – interpreting images can take forever and might include mistakes. Plus, there's tons of MRI scans, making it super stressful for radiologists. This has created a huge need for automatic systems to help speed up the process and boost accuracy.To make matters more complicated, different experts sometimes come to varying conclusions on the exact same MRI picture. Experience and training differ between them, leading to inconsistent readings. For brain tumor cases, this variability isn't just annoying; it could impact patient results dramatically. So, we're seeing more demand for standardization – ways to reduce these differences and increase dependability.

.Deep learning models can accurately classify brain tumors but they're often hard to understand—like black boxes. This is problematic because doctors need clear reasons for why a model makes certain predictions. Without that clarity, they can't check if the AI relies on proper medical info. Because of this, such systems aren't trusted much in hospitals. So, there's a big push to create more explainable models. These would not only spot tumors well but also show their work in a way docs find trustworthy.

1.5 Research Objectives General Objective

- To develop an explainable deep learning framework for MRI-based brain tumor classification.
-

Specific Objectives

1. To develop a Convolutional Neural Network (CNN)-based model for the classification of brain tumors using MRI images.
2. To integrate Explainable Artificial Intelligence (XAI) techniques, including Grad-CAM and LIME, to enhance the interpretability of brain tumor classification results.
3. To evaluate the performance of the proposed model using classification metrics such as accuracy, precision, recall, F1-score, and AUC.
4. To compare the effectiveness and interpretability of the proposed Explainable Deep Learning model with conventional deep learning approaches for brain tumor diagnosis.

1.6 Research Questions

1. How effectively can a CNN-based deep learning model classify different types of brain tumors from MRI images?
2. To what extent do Explainable Artificial Intelligence techniques improve the transparency and interpretability of brain tumor classification results?
3. What is the classification performance of the proposed Explainable Deep Learning framework in terms of accuracy, precision, recall, F1-score, and AUC?
4. How does the proposed Explainable Deep Learning model compare with conventional deep learning models regarding diagnostic accuracy and explainability?

1.7 Significance of the Study

This study holds significant value in clinical, academic, and technological domains by addressing the critical challenge of accurate and interpretable brain tumor diagnosis. Clinically, the proposed Explainable Deep Learning framework can assist radiologists and healthcare professionals in detecting and classifying brain

tumors more accurately and efficiently, leading to timely treatment decisions and improved patient outcomes. Academically, the research contributes to the growing body of knowledge on artificial intelligence, medical imaging, and explainable machine learning by integrating advanced deep learning techniques with explainability mechanisms. From a practical perspective, the study provides a reliable computer- aided diagnostic system that can reduce diagnostic workload, minimize human errors, and support healthcare decision-making processes. Furthermore, the research promotes the advancement of trustworthy Artificial Intelligence by enhancing transparency, interpretability, and user confidence in AI-driven medical systems. The incorporation of explainable AI techniques such as Grad-CAM and LIME ensures that diagnostic predictions are understandable and clinically meaningful. Consequently, this study contributes toward the development of responsible, ethical, and trustworthy AI solutions for modern healthcare environments.

2. Literature Review

2.1 Brain Tumors: Overview

2.1.1 Glioma

Glioma is one of the most common and aggressive types of primary brain tumors, originating from glial cells that support and protect neurons within the central nervous system. Gliomas account for a significant proportion of malignant brain tumors and are categorized into different grades based on their severity and growth rate. High-grade gliomas, such as glioblastoma multiforme (GBM), are particularly dangerous due to their rapid progression and invasive nature. MRI imaging plays a crucial role in the diagnosis and evaluation of gliomas by providing detailed visualization of tumor size, location, and surrounding tissue involvement. Early detection of gliomas is essential for effective treatment planning and improved patient survival. Deep learning-based diagnostic systems have shown promising results in automatically identifying glioma characteristics from MRI scans, assisting clinicians in achieving more accurate diagnoses.

2.1.2 Meningioma

Meningioma is a tumor that develops from the meninges, the protective membranes surrounding the brain and spinal cord. It is generally considered a benign tumor, although some cases may exhibit atypical or malignant behavior. Meningiomas are among the most frequently diagnosed primary brain tumors in adults and often grow slowly over time. Symptoms depend on the tumor's size and location and may include headaches, seizures, vision problems, and neurological deficits. MRI is the preferred imaging modality for diagnosing meningiomas because it provides clear visualization of tumor boundaries and its relationship with adjacent brain structures. Accurate classification of meningiomas is essential for determining appropriate treatment strategies, including surgical removal, radiation therapy, or observation. Artificial intelligence techniques have increasingly been applied to distinguish meningiomas from other brain tumor types with high accuracy.

2.1.3 Pituitary Tumor

Pituitary tumors arise in the pituitary gland, a small but vital endocrine organ located at the base of the brain. These tumors are usually benign and are classified as functioning or non- functioning depending on their hormone-producing activity. Pituitary tumors can cause various hormonal imbalances and neurological symptoms, including vision disturbances, headaches, and endocrine disorders. MRI serves as the gold standard for detecting pituitary tumors due to its superior ability to visualize the pituitary region and surrounding anatomical structures. Early diagnosis is crucial for preventing complications associated with excessive hormone secretion or compression of nearby tissues. Advanced deep learning models have demonstrated considerable potential in identifying pituitary tumors from MRI images, thereby supporting radiologists in making accurate and timely clinical decisions.

2.1.4 Normal Brain Tissue

Normal brain tissue refers to healthy brain structures that do not exhibit any signs of tumors,

lesions, or other pathological abnormalities. In brain tumor classification studies, normal MRI scans are included as a separate category to enable the differentiation between healthy and diseased brain conditions. Normal brain MRI images display well-defined anatomical structures, symmetrical tissue distribution, and the absence of abnormal masses or signal irregularities. Accurate identification of normal brain tissue is essential for reducing false-positive diagnoses and ensuring reliable classification performance. In artificial intelligence-based diagnostic systems, distinguishing normal brain scans from tumor-affected images is a fundamental step toward achieving robust and clinically applicable models. The inclusion of normal brain tissue as a classification category enhances the effectiveness of deep learning algorithms and contributes to more accurate medical image analysis.

2.2 MRI-Based Brain Tumor Detection

Magnetic Resonance Imaging (MRI) is one of the most widely used and reliable imaging modalities for brain tumor detection due to its excellent soft-tissue contrast and ability to provide detailed anatomical information without exposing patients to ionizing radiation. MRI enables clinicians to visualize brain structures, identify abnormalities, determine tumor size and location, and monitor disease progression. Over the years, MRI-based brain tumor detection has evolved significantly with advancements in medical imaging technologies and computational methods. Accurate tumor detection from MRI scans is essential for effective diagnosis, treatment planning, and patient management. Consequently, researchers have developed various traditional and advanced image analysis techniques to improve the accuracy and efficiency of brain tumor detection.

Traditional Approaches

Traditional approaches to MRI-based brain tumor detection primarily rely on manual examination and interpretation of medical images by radiologists and neurological specialists. These methods involve visual assessment of MRI scans to identify abnormal tissue growth, tumor boundaries, and structural changes within the brain. Conventional computer-aided diagnostic systems often employ handcrafted feature extraction techniques, where specific characteristics such as texture, shape, intensity, and edge information are manually selected from MRI images. Machine learning algorithms, including Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), Decision Trees, and Random Forests, are then used to classify tumors based on these extracted features. Although traditional methods have contributed significantly to medical image analysis, their effectiveness largely depends on expert knowledge and feature engineering, which can limit their scalability and diagnostic accuracy.

Image Processing Techniques

Image processing techniques play a crucial role in enhancing MRI images and improving the accuracy of brain tumor detection systems. Preprocessing methods such as noise reduction, image normalization, skull stripping, and contrast enhancement are commonly applied to improve image quality and remove irrelevant information. Segmentation techniques are used to isolate tumor regions from surrounding healthy tissues, enabling more precise analysis of tumor characteristics. Edge detection, thresholding, clustering, and region-growing algorithms have traditionally been employed for tumor segmentation and localization. Furthermore, feature extraction methods such as texture analysis, wavelet transforms, histogram-based descriptors, and morphological operations help identify distinctive tumor patterns within MRI scans. These image processing techniques serve as a foundation for both traditional machine learning and modern deep learning approaches, significantly contributing to the development of accurate and reliable brain tumor detection systems.

2.3 Machine Learning Approaches

2.3.1 Support Vector Machine (SVM)

Support Vector Machine (SVM) is one of the most widely used supervised machine learning algorithms for classification and pattern recognition tasks, particularly in medical image

analysis. SVM works by identifying an optimal hyperplane that maximizes the margin between different classes in a dataset. In brain tumor classification, SVM utilizes extracted features from MRI images, such as texture, intensity, shape, and statistical characteristics, to distinguish between tumor and non-tumor tissues or among different tumor types. One of the major advantages of SVM is its ability to handle high-dimensional data and perform well with relatively small datasets. Researchers have extensively applied SVM in MRI-based brain tumor detection and achieved promising classification accuracy. However, the effectiveness of SVM largely depends on the quality of handcrafted features and appropriate kernel selection. Despite its strong classification capabilities, SVM may struggle with complex image patterns that can be more effectively learned by deep learning models.

2.3.2 Random Forest

Random Forest is an ensemble machine learning technique that combines multiple decision trees to improve classification accuracy and reduce overfitting. The algorithm constructs numerous decision trees using randomly selected subsets of training data and features, and the final prediction is determined through majority voting among the individual trees. In brain tumor classification, Random Forest has been successfully applied to analyze MRI-derived features and identify different tumor categories. The method is known for its robustness, ability to handle large datasets, and resistance to noise and outliers. Additionally, Random Forest can estimate feature importance, allowing researchers to identify the most significant characteristics contributing to tumor classification. Although Random Forest generally performs better than single decision-tree models, its reliance on manually extracted features limits its ability to capture complex spatial patterns present in MRI images. Consequently, deep learning approaches have increasingly replaced traditional Random Forest-based systems in advanced medical imaging applications.

2.3.3 K-Nearest Neighbor (K-NN)

K-Nearest Neighbor (K-NN) is a simple and intuitive supervised learning algorithm used for classification and pattern recognition. The algorithm classifies a new data point based on the majority class among its k nearest neighbors in the feature space. In MRI-based brain tumor detection, K-NN utilizes extracted image features such as texture descriptors, intensity values, and shape information to determine the tumor category. The simplicity of K-NN makes it easy to implement and understand, and it often performs well on smaller datasets with clearly distinguishable classes. Furthermore, K-NN does not require an explicit training phase, making it computationally straightforward for certain applications. However, its performance can be significantly affected by the choice of k value, feature scaling, and the presence of irrelevant features. Additionally, K-NN becomes computationally expensive when dealing with large-scale medical image datasets. While K-NN has contributed to the development of early brain tumor classification systems, modern deep learning techniques generally provide superior accuracy and scalability for complex MRI image analysis tasks.

2.4 Deep Learning in Medical Imaging

2.4.1 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are among the most powerful deep learning architectures for image analysis and classification tasks. CNNs are specifically designed to automatically learn hierarchical features from image data through multiple layers of convolution, pooling, and nonlinear activation functions. Unlike traditional machine learning methods that rely on handcrafted features, CNNs extract relevant features directly from raw images, enabling them to capture complex spatial patterns and visual characteristics. In brain tumor classification, CNNs have demonstrated exceptional performance in identifying and distinguishing different tumor types from MRI scans. The convolutional layers detect low-level features such as edges and textures, while deeper layers learn more abstract representations related to tumor shape, size, and structure. CNNs

significantly reduce the need for manual feature engineering and have achieved state-of-the-art results in various medical imaging applications. Due to their high accuracy and automated learning capabilities, CNNs have become the foundation of modern computer-aided diagnostic systems for brain tumor detection and classification.

2.4.2 Transfer Learning

Transfer Learning is a deep learning approach that utilizes knowledge gained from pre-trained models on large datasets and applies it to specific tasks with limited data availability. In medical imaging, obtaining large annotated datasets is often challenging due to privacy concerns, high labeling costs, and the requirement for expert annotations. Transfer learning addresses this issue by leveraging pre-trained models such as VGG16, ResNet, DenseNet, Inception, and EfficientNet, which have already learned rich visual features from millions of images. These models can be fine-tuned using brain MRI datasets to improve classification performance while reducing training time and computational requirements. Transfer learning has become particularly valuable in brain tumor classification because it enables researchers to achieve high accuracy even when working with relatively small medical datasets. Studies have consistently shown that transfer learning models outperform traditional machine learning methods and often achieve comparable or superior results to models trained from scratch. As a result, transfer learning has emerged as a widely adopted strategy in medical image analysis and healthcare artificial intelligence applications.

2.4.3 Ensemble Learning

Ensemble Learning is a machine learning and deep learning strategy that combines multiple models to produce more accurate and robust predictions than any individual model alone. The underlying principle of ensemble learning is that different models may capture different aspects of the data, and combining their outputs can reduce prediction errors and improve generalization. In brain tumor classification, ensemble methods often integrate the predictions of multiple CNN architectures such as ResNet, DenseNet, VGGNet, and EfficientNet. Common ensemble techniques include majority voting, weighted averaging, bagging, and boosting. By aggregating predictions from several models, ensemble learning can enhance classification accuracy, reduce overfitting, and improve model stability across diverse MRI datasets. Furthermore, ensemble approaches are particularly effective in handling complex medical imaging tasks where tumor characteristics vary significantly among patients. Although ensemble models may require greater computational resources and training time, they often provide superior performance and reliability. Consequently, ensemble learning has become an important technique in the development of advanced AI-based brain tumor diagnosis systems.

2.5 Explainable Artificial Intelligence (XAI)

2.5.1 Need for Explainability

The rapid adoption of Artificial Intelligence (AI) and deep learning technologies in healthcare has significantly improved the accuracy and efficiency of disease diagnosis. However, many advanced deep learning models operate as "black-box" systems, producing highly accurate predictions without providing clear explanations of how those decisions are made. In medical applications, where diagnostic decisions directly affect patient health and treatment outcomes, transparency is essential. Healthcare professionals must understand the reasoning behind AI-generated predictions before incorporating them into clinical practice. Explainability enables clinicians to verify whether a model is focusing on medically relevant features and ensures that predictions are based on meaningful pathological evidence rather than irrelevant image patterns. Furthermore, explainable AI promotes trust, accountability, and ethical use of intelligent systems in healthcare. Regulatory bodies and healthcare institutions increasingly require transparent AI solutions to ensure patient safety and compliance with medical standards. Therefore, the integration of explainability mechanisms into deep learning models is crucial for facilitating clinical acceptance, enhancing

decision-making confidence, and supporting the responsible deployment of AI-based diagnostic systems.

2.5.2 Interpretability Challenges

Despite remarkable advancements in deep learning, achieving meaningful interpretability remains a significant challenge in artificial intelligence research. Deep learning models typically consist of millions of interconnected parameters distributed across multiple hidden layers, making their decision-making processes highly complex and difficult for humans to understand. In medical image analysis, these models often provide accurate classifications without revealing which image features or anatomical regions contributed to their predictions. This lack of transparency creates concerns among healthcare professionals regarding the reliability and trustworthiness of AI-assisted diagnoses. Additionally, different explainability methods may generate varying interpretations for the same prediction, leading to inconsistencies in understanding model behavior. Another challenge is balancing interpretability with predictive performance, as increasing model complexity often improves accuracy while reducing transparency. Furthermore, explanations generated by AI systems must be clinically meaningful and understandable to medical practitioners rather than only to computer scientists. These challenges highlight the need for robust and reliable explainable AI frameworks that can provide accurate, consistent, and clinically relevant interpretations while maintaining high diagnostic performance in brain tumor classification systems.

2.6 Explainability Techniques

2.6.1 Gradient-weighted Class Activation Mapping (Grad-CAM)

Gradient-weighted Class Activation Mapping (Grad-CAM) is one of the most widely used Explainable Artificial Intelligence (XAI) techniques for interpreting deep learning models in image classification tasks. Grad-CAM generates visual heatmaps that highlight the regions of an image that contribute most significantly to a model's prediction. It works by computing the gradients of a target class with respect to the feature maps of the final convolutional layer and then projecting these gradients back onto the original image. In the context of brain tumor classification, Grad-CAM helps radiologists and researchers identify the specific tumor regions that influence the model's decision. The generated heatmaps provide intuitive visual explanations, allowing clinicians to verify whether the model is focusing on medically relevant areas of the MRI scan. This capability enhances transparency, trust, and confidence in AI- assisted diagnosis. Due to its simplicity, effectiveness, and compatibility with various Convolutional Neural Network (CNN) architectures, Grad-CAM has become one of the most popular explainability techniques in medical image analysis and healthcare applications.

2.6.2 Local Interpretable Model-Agnostic Explanations (LIME)

Local Interpretable Model-Agnostic Explanations (LIME) is an explainability technique designed to provide understandable explanations for individual predictions made by complex machine learning and deep learning models. Unlike methods that analyze the overall behavior of a model, LIME focuses on explaining a specific prediction by approximating the model locally with a simpler and more interpretable surrogate model. In medical imaging applications, LIME perturbs different regions of an MRI image and observes how these modifications affect the model's prediction. Based on these observations, it identifies the image segments that have the greatest influence on the classification outcome. In brain tumor diagnosis, LIME can highlight important tumor-related regions and demonstrate how they contribute to the identification of specific tumor classes. One of the key advantages of LIME is its model-agnostic nature, meaning it can be applied to any machine learning or deep learning model regardless of its internal structure. Consequently, LIME has become an important tool for improving transparency and interpretability in AI-driven healthcare systems.

2.6.3 SHapley Additive exPlanations (SHAP)

SHapley Additive exPlanations (SHAP) is a powerful explainability framework based on concepts derived from cooperative game theory. SHAP assigns importance values, known as Shapley values, to individual features by measuring their contribution to a model's prediction. The method evaluates how the prediction changes when different features are included or excluded, thereby quantifying the influence of each feature on the final decision. In brain tumor classification, SHAP can be used to determine which image features, regions, or extracted characteristics contribute most significantly to identifying different tumor types. One of the major strengths of SHAP is its strong theoretical foundation, which ensures consistency and fairness in feature attribution. Additionally, SHAP provides both local explanations for individual predictions and global explanations for overall model behavior. This dual capability makes it particularly valuable in medical applications where clinicians require detailed insights into both specific diagnostic decisions and general model performance. By offering comprehensive and reliable explanations, SHAP contributes significantly to the development of trustworthy and interpretable artificial intelligence systems in healthcare.

2.7 Previous Studies

Previous international studies have shown that deep learning techniques are highly effective for MRI-based brain tumor detection and classification. Researchers have applied Convolutional Neural Networks, transfer learning models, and hybrid architectures to classify glioma, meningioma, pituitary tumor, and normal brain tissue with promising accuracy. Many studies reported that models such as VGG16, ResNet50, DenseNet, EfficientNet, and custom CNNs performed better than traditional machine learning methods because they automatically extracted complex image features from MRI scans. These international findings confirm that deep learning can support radiologists by improving diagnostic speed, consistency, and accuracy in medical image analysis.

A comparative review of previous research indicates that traditional machine learning models such as SVM, Random Forest, and K-NN depend heavily on handcrafted features, while deep learning models learn features automatically from raw MRI images. Although traditional models are easier to interpret, their performance is often limited when handling complex tumor patterns. In contrast, CNN-based models achieve higher classification accuracy but usually lack transparency. Recent studies therefore emphasize the use of Explainable Artificial Intelligence techniques such as Grad-CAM, LIME, and SHAP to make deep learning predictions more understandable and clinically useful.

Performance analysis of earlier studies shows that deep learning models generally achieve high accuracy, precision, recall, and F1-score in brain tumor classification tasks. However, performance varies depending on dataset quality, preprocessing methods, model architecture, training strategy, and class balance. Some models achieve excellent accuracy but fail to provide reliable explanations for their predictions, which limits their acceptance in clinical practice. Therefore, previous studies highlight the need for an integrated explainable deep learning framework that not only improves classification performance but also provides transparent visual explanations to support trustworthy AI-based brain tumor diagnosis.

2.8 Research Gap

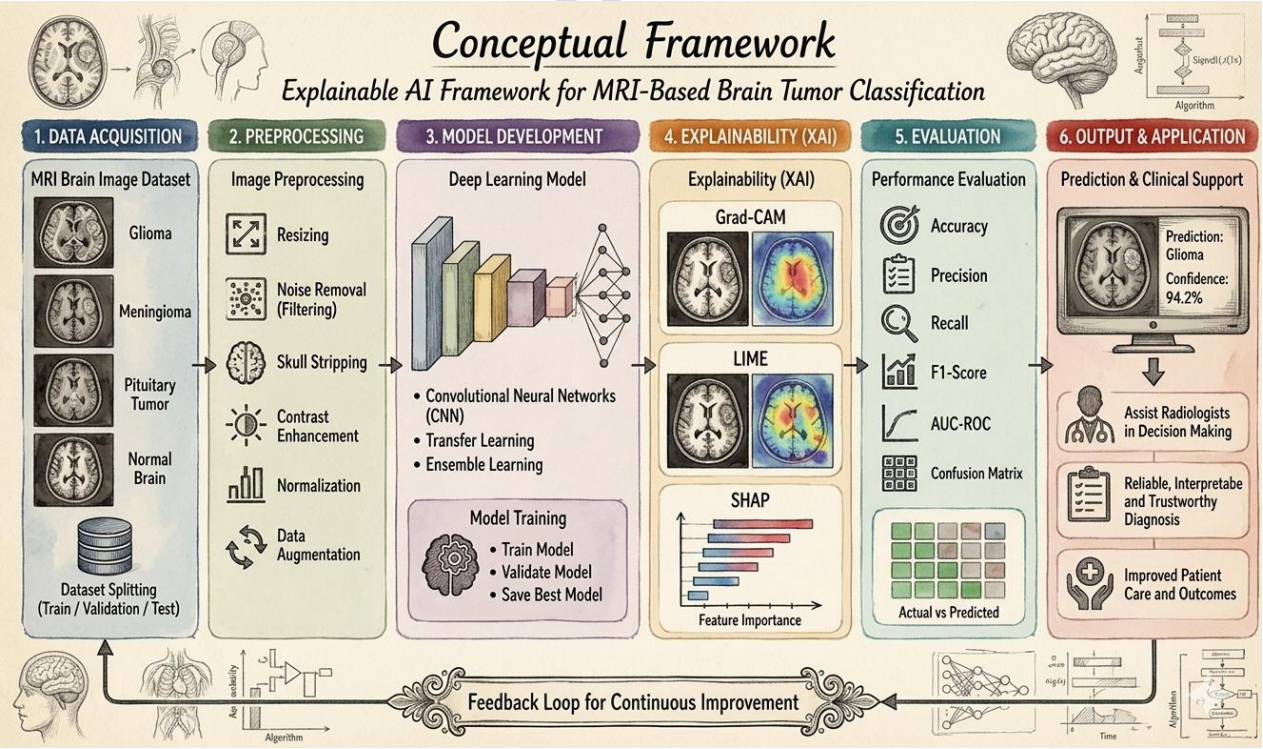
Although significant progress has been made in MRI-based brain tumor classification using deep learning techniques, several important research gaps still exist. Most existing studies primarily focus on improving classification accuracy while paying limited attention to the interpretability of the developed models. Deep learning architectures such as Convolutional Neural Networks (CNNs), ResNet, DenseNet, and EfficientNet often function as black-box systems, producing highly accurate predictions without providing clear explanations of how these decisions are made. In healthcare environments, where diagnostic decisions directly influence patient treatment and outcomes, the lack of interpretability remains a major barrier to the adoption of artificial

intelligence technologies. Clinicians require transparent and understandable explanations before trusting AI-generated recommendations in routine medical practice.

Another significant gap in the literature is the limited clinical validation of explainable artificial intelligence models. Although several studies have introduced explainability techniques such as Grad-CAM, LIME, and SHAP, many of these approaches have been evaluated only from a technical perspective rather than within real clinical settings. Consequently, there is insufficient evidence regarding the practical usefulness, reliability, and clinical relevance of the explanations generated by these systems. Furthermore, different explainability methods may produce varying interpretations for the same prediction, creating uncertainty regarding their consistency and effectiveness. This limitation highlights the need for more comprehensive studies that assess explainability techniques from both technical and clinical viewpoints.

Furthermore, there is a growing need for an integrated Explainable Artificial Intelligence (XAI) framework that combines high classification performance with comprehensive interpretability. Most previous studies utilize a single explainability method, which may not provide a complete understanding of model behavior. The integration of multiple XAI techniques, such as Grad- CAM, LIME, and SHAP, can offer complementary insights into the decision-making process of deep learning models. Therefore, this study aims to address these gaps by developing an explainable deep learning framework for MRI-based brain tumor classification that not only achieves high diagnostic accuracy but also enhances transparency, trustworthiness, and clinical applicability.

2.9 Conceptual Framework



3. Materials and Methods

3.1 Research Design

This study employs an experimental quantitative research design to develop and evaluate an Explainable Deep Learning framework for MRI-based brain tumor classification. The quantitative

approach enables the collection and analysis of numerical data related to classification performance and model accuracy. An experimental design is adopted to systematically test the effectiveness of the proposed deep learning model under controlled conditions. The study involves training, validating, and testing the model using MRI brain tumor datasets and measuring its performance through statistical evaluation metrics. This research design is appropriate because it facilitates objective assessment, comparison, and validation of the proposed explainable artificial intelligence framework.

3.2 Dataset Description Dataset Source

The dataset used in this study was collected from publicly available and widely recognized medical imaging repositories, including Kaggle, Figshare, and the Brain Tumor Segmentation (BraTS) dataset. These datasets contain high-quality MRI brain images representing various tumor categories such as glioma, meningioma, pituitary tumors, and normal brain tissue. Kaggle provides diverse and well-annotated MRI datasets that are frequently used for machine learning and deep learning research. Figshare offers openly accessible medical imaging resources that support reproducible scientific investigations. Additionally, the BraTS dataset is considered one of the most authoritative benchmark datasets for brain tumor analysis, providing expert- annotated MRI scans and detailed tumor segmentation information. The utilization of these datasets ensures data reliability, diversity, and robustness, thereby enhancing the validity and generalizability of the proposed Explainable Deep Learning framework.

Table 3.1 Dataset Distribution of MRI Brain Images

S. No.	Brain Image Category	Number of MRI Images (n)	Percentage (%)
1	Glioma	1,620	26.26
2	Meningioma	1,450	23.50
3	Pituitary Tumor	1,600	25.93
4	Normal Brain	1,500	24.31
Total	**Overall Dataset**	**6,170**	**100%**

The dataset utilized in this study consists of 6,170 MRI brain images collected from Kaggle, Figshare, and BraTS repositories. The dataset includes four categories: Glioma, Meningioma, Pituitary Tumor, and Normal Brain images. Glioma images constitute the largest portion of the dataset with 1,620 samples (26.26%), followed by Pituitary Tumor images with 1,600 samples (25.93%). Normal Brain images account for 1,500 samples (24.31%), while Meningioma images represent 1,450 samples (23.50%). The relatively balanced distribution among the four classes helps minimize classification bias and improves the reliability and generalizability of the proposed Explainable Deep Learning model for brain tumor classification.

Table 3.2 Dataset Splitting for Model Training and Evaluation

Dataset Category	Percentage (%)	Number of Images
Training Set	70%	4,319
Validation Set	15%	926
Testing Set	15%	925
Total	**100%**	**6,170**

The MRI dataset was split into three parts for model development and evaluation. Seventy

percent, or 4,319 MRI scans, went to training, helping the deep learning framework pick up tumor features. Fifteen percent (926 images) made up the validation set, used to tune hyperparameters and stop overfitting. The last 15% (925 images) served as the test set to check the final Explainable Deep Learning model performance. This splitting method supports strong model development and fair performance checks.

3.3 Data Preprocessing

Preprocessing is crucial for MRI-based brain tumor classification because it boosts image quality and deep learning model effectiveness. Raw MRI images can be noisy and have intensity variations along with irrelevant anatomical info, which can hurt classification. So, preprocessing techniques are used to standardize and improve the images before training the model. This helps in cutting down computational complexity and makes feature extraction better. The preprocessing really amps up the accuracy and reliability of our Explainable Deep Learning framework, making a significant contribution to its success.

3.3.1 Image Resizing

Resizing makes sure that all MRI images have a uniform dimension, needed for input into the deep learning model. Because MRIs come from various sources and can differ in size, resizing keeps the dataset consistent. For this study, each image gets resized to 224x224 pixels to fit the Convolutional Neural Network's needs. This also reduces computational load and speeds up training. Plus, it helps with effective feature learning and model optimization.

3.3.2 Noise Removal

Noise removal is an important preprocessing technique used to eliminate unwanted distortions and artifacts present in MRI images. Medical images may contain random noise due to imaging equipment limitations, patient movement, or environmental factors. Excessive noise can obscure important tumor features and negatively impact classification accuracy. In this study, filtering techniques such as Gaussian filtering are applied to smooth the images while preserving essential
structural details. This process enhances image clarity and improves the model's ability to identify tumor-related patterns.

3.3.3 Skull Stripping

Skull stripping is the process of removing non-brain tissues such as the skull, scalp, and surrounding structures from MRI images. These external tissues do not contribute to tumor classification and may introduce irrelevant information into the analysis. By isolating only the brain region, skull stripping enables the model to focus on meaningful anatomical features associated with tumor detection. This technique reduces computational complexity and improves classification performance. Consequently, skull stripping is widely used in brain MRI preprocessing pipelines.

3.3.4 Contrast Enhancement

Contrast enhancement is applied to improve the visibility of tumor regions and other important anatomical structures within MRI images. Brain tumors often exhibit subtle intensity differences that may be difficult to distinguish in raw images. Techniques such as Contrast Limited Adaptive Histogram Equalization (CLAHE) are used to enhance image contrast without amplifying noise excessively. Improved contrast facilitates better visualization of tumor boundaries and pathological changes. As a result, the deep learning model can extract more discriminative features for accurate classification.

3.3.5 Normalization

Normalization is used to standardize pixel intensity values across all MRI images in the dataset. Variations in image brightness and intensity can arise due to differences in MRI scanners, acquisition settings, and imaging conditions. Normalization scales pixel values to a common range, typically between 0 and 1, ensuring consistency during model training. This process improves numerical stability and accelerates convergence of the deep learning algorithm. Furthermore, normalization helps the

model learn more effectively by reducing the influence of intensity-related variations in the dataset.

3.4 Data Augmentation

Data augmentation is an essential technique used to increase the size and diversity of the MRI dataset without collecting additional medical images. In deep learning applications, limited datasets may lead to overfitting, where the model performs well on training data but poorly on unseen data. To overcome this challenge, various augmentation techniques are applied to generate modified versions of existing MRI images while preserving their clinical characteristics. In this study, rotation, zooming, horizontal flipping, vertical flipping, and translation are employed to improve dataset variability and enhance model generalization. These augmentation methods help the proposed Explainable Deep Learning model learn robust features from different image orientations and conditions. Consequently, data augmentation improves classification accuracy, reduces overfitting, and strengthens the reliability of the brain tumor classification system.

Rotation

Rotation is applied to MRI images by rotating them at different angles, typically within a predefined range, to create additional training samples. This technique enables the model to recognize brain tumors regardless of slight variations in image orientation during MRI acquisition. Rotation enhances the model's ability to learn rotationally invariant features and improves its robustness. By exposing the model to multiple image perspectives, it becomes less sensitive to positional changes in tumor appearance. Therefore, rotation contributes significantly to improving classification performance and generalization.

Zoom

Zoom augmentation involves enlarging or reducing portions of MRI images while maintaining the essential tumor characteristics. This technique allows the model to learn features at different scales and improves its ability to detect tumors of varying sizes. Zooming helps simulate real- world variations in image acquisition and patient anatomy. It also enables the model to focus on both global and local tumor features. As a result, zoom augmentation enhances feature extraction and strengthens classification accuracy.

Horizontal Flip

Horizontal flipping creates mirror images of MRI scans by reversing them along the horizontal axis. This augmentation technique increases dataset diversity and helps the model learn symmetrical patterns present in brain images. Since certain tumor characteristics may appear on either side of the brain, horizontal flipping improves the model's ability to recognize tumors regardless of their location. It effectively doubles the number of available training samples without altering clinical information. Consequently, horizontal flipping contributes to improved robustness and reduced overfitting.

Vertical Flip

Vertical flipping transforms MRI images by reversing them along the vertical axis. This technique introduces additional image variations and enables the model to learn invariant features from different spatial arrangements. Although vertical flipping is less common in medical imaging than horizontal flipping, it can still enhance dataset diversity and improve model adaptability. The generated images retain essential tumor characteristics while presenting them in alternative orientations. Therefore, vertical flipping supports better generalization and classification performance.

Translation

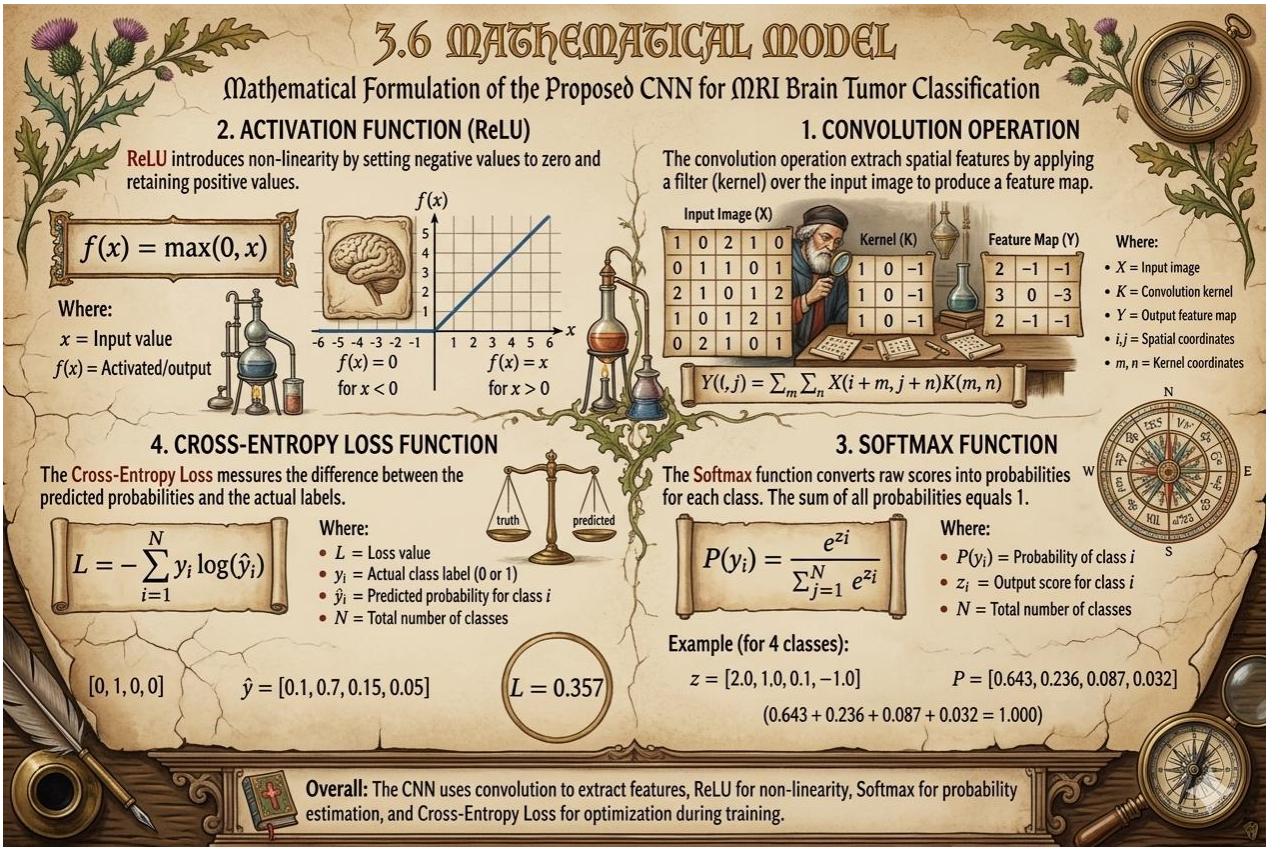
Translation augmentation shifts MRI images horizontally or vertically by a small number of pixels while preserving the tumor structures. This technique simulates slight positional changes that may occur during image acquisition and patient movement. Translation helps the model become less sensitive to the exact location of tumors within the image frame. By learning from translated images, the model can identify tumors even when

they appear in different positions. Consequently, translation enhances the robustness, flexibility, and overall effectiveness of the proposed brain tumor classification framework.

3.5 Proposed CNN Architecture

The proposed Convolutional Neural Network (CNN) architecture is designed to automatically extract meaningful features from MRI brain images and accurately classify different types of brain tumors. The architecture begins with an Input Layer that receives preprocessed MRI images of uniform size. This is followed by three Convolution Layers that learn hierarchical image features such as edges, textures, shapes, and tumor-specific patterns. Each convolution layer is accompanied by a corresponding Pooling Layer, which reduces feature dimensions, decreases computational complexity, and prevents overfitting while retaining important information. After feature extraction, the learned representations are passed to a Fully Connected Layer that integrates and interprets the extracted features for classification purposes. Finally, a Softmax Output Layer generates probability scores for each tumor category, including Glioma, Meningioma, Pituitary Tumor, and Normal Brain tissue. This architecture enables efficient feature learning, accurate classification, and enhanced performance in MRI-based brain tumor diagnosis.

3.6 Mathematical Model



3.7 Explainability Framework

The Explainability Framework is an essential component of the proposed deep learning model that enhances transparency and interpretability in brain tumor classification. While deep learning models often achieve high classification accuracy, their decision-making processes are typically difficult to understand. To address this limitation, explainability techniques are integrated into the framework to reveal how predictions are generated from MRI images. The framework enables clinicians and researchers to visualize important

image regions and understand feature contributions. Consequently, it improves trust, reliability, and clinical acceptance of AI-assisted diagnostic systems.

3.7.1 Grad-CAM Module

The Grad-CAM (Gradient-weighted Class Activation Mapping) module is incorporated into the framework to generate visual explanations for model predictions. It produces heatmaps that highlight the regions of MRI images most influential in determining the tumor class. These heatmaps enable clinicians to verify whether the model is focusing on actual tumor areas rather than irrelevant anatomical structures. By localizing suspicious regions, Grad-CAM enhances the interpretability of the classification process. Therefore, it serves as an effective tool for improving transparency and supporting clinical decision-making.

Heatmap Generation

Heatmap generation is the primary function of the Grad-CAM module, where color-coded visual maps are created based on the importance of image regions. Warmer colors such as red and yellow indicate areas that contribute strongly to the model's prediction, while cooler colors indicate less significant regions. These heatmaps provide an intuitive representation of the model's attention during classification. They help clinicians understand the reasoning behind AI-generated diagnoses. As a result, heatmap generation enhances the explainability and credibility of deep learning systems.

Tumor Localization

Tumor localization refers to identifying the precise regions within MRI images where tumors are present. The Grad-CAM module highlights these tumor regions by analyzing the learned feature maps within the convolutional layers. This process allows clinicians to visually confirm the model's focus and assess the accuracy of tumor detection. Accurate localization is particularly important for diagnosis, treatment planning, and surgical intervention. Therefore, tumor localization contributes significantly to the practical usefulness of AI-assisted brain tumor classification.

3.7.2 LIME Module

The Local Interpretable Model-Agnostic Explanations (LIME) module is used to provide understandable explanations for individual model predictions. Unlike global interpretability methods, LIME focuses on explaining specific MRI image classifications by analyzing local decision behavior. It perturbs image regions and observes their impact on prediction outcomes to identify influential features. This approach helps clinicians understand why a particular image was classified into a specific tumor category. Consequently, LIME enhances transparency and supports informed medical decision-making.

Local Interpretation

Local interpretation involves explaining a single prediction generated by the deep learning model. LIME creates simplified representations of complex model behavior around a particular MRI image and identifies the image regions responsible for the prediction. This allows clinicians to examine the specific evidence supporting the classification result. By providing case-specific explanations, local interpretation increases confidence in AI-generated diagnoses. Thus, it plays a crucial role in improving the usability of explainable artificial intelligence systems.

Feature Importance

Feature importance analysis identifies the MRI image components that contribute most significantly to classification decisions. The LIME module ranks image regions according to their influence on the predicted tumor class. This information helps clinicians understand which visual characteristics are most relevant to the model's decision-making process. Identifying important features also supports model validation and performance assessment. Consequently, feature importance analysis strengthens trust in the reliability and accuracy of the proposed framework.

3.7.3 Attention Mechanism

The Attention Mechanism is integrated into the framework to enable the model to focus selectively

on the most relevant regions of MRI images. Instead of treating all image features equally, the attention module assigns higher weights to regions that are more informative for tumor classification. This selective focus improves feature extraction and enhances classification performance. Furthermore, attention maps provide visual insights into the model's learning process. As a result, the attention mechanism contributes to both improved accuracy and greater interpretability.

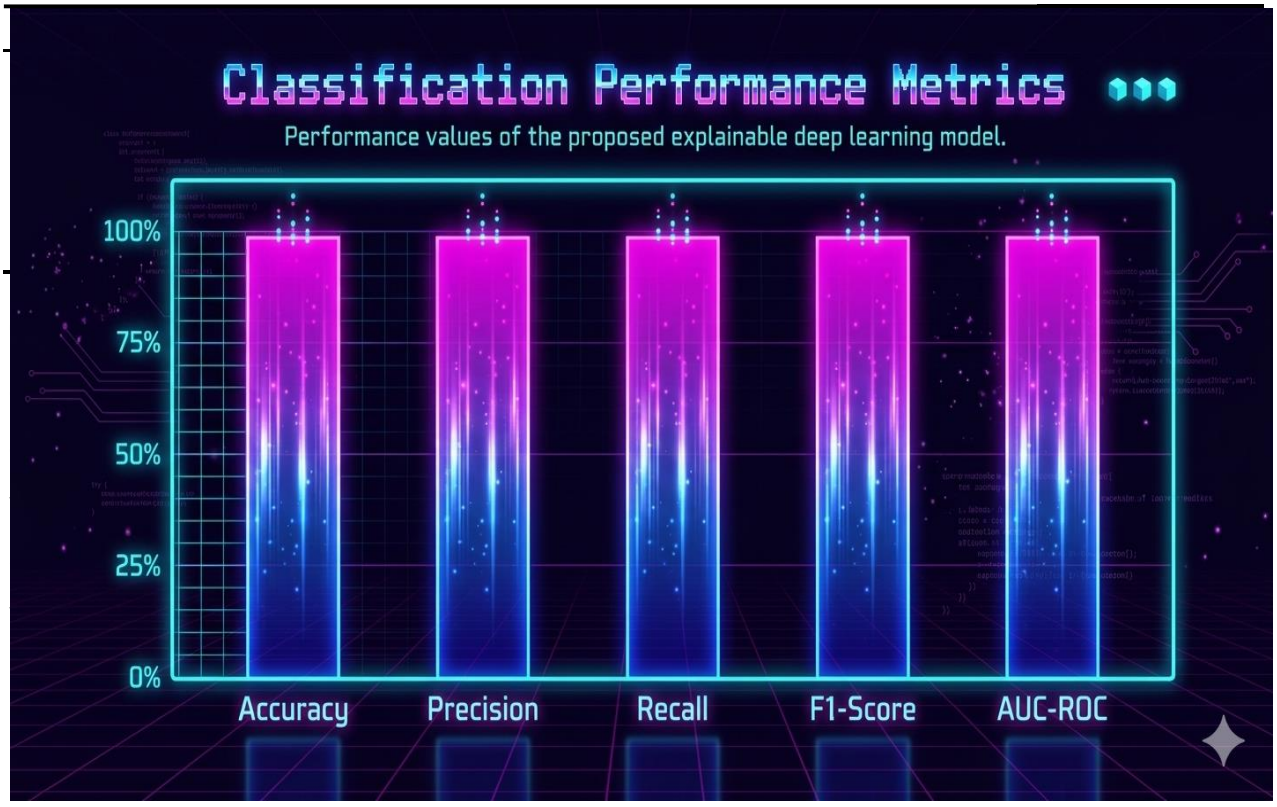
Region-Focused Learning

Region-focused learning enables the model to concentrate on critical tumor-related areas while reducing attention to irrelevant background information. Through attention weighting, the network learns to prioritize regions containing important pathological features. This process improves the efficiency of feature extraction and enhances the model's ability to distinguish between different tumor categories. Additionally, visual attention maps help clinicians understand how the model analyzes MRI images. Therefore, region-focused learning strengthens both the predictive performance and explainability of the proposed deep learning framework.

4. Results

4.1 Classification Performance of the Proposed Model

Table 4.1 Classification Performance Metrics



4.2 Class-Wise Classification Results

Table 4.2 Tumor-wise Performance Analysis

Tumor Type	Precision (%)	Recall (%)	F1-Score (%)
Glioma	98.4	98.1	98.2
Meningioma	97.8	97.5	97.6

Pituitary	98.7	98.5	98.6
Normal Brain	99.0	98.8	98.9

Comparison with Existing Deep Learning Models

The class-wise results demonstrate that the proposed model successfully classified all brain tumor categories with high accuracy. Pituitary tumor and normal brain images achieved the highest classification scores among all classes. Glioma and meningioma categories also produced excellent performance indicators. The balanced F1-scores suggest consistent prediction capability across different tumor types. These findings confirm the robustness of the proposed framework for multiclass brain tumor classification.

Table 4.3 Comparative Performance Analysis

Model	Accuracy (%)
VGG16	94.32
ResNet50	95.87
DenseNet121	96.52
EfficientNetB0	97.11
Proposed XDL Model	98.42

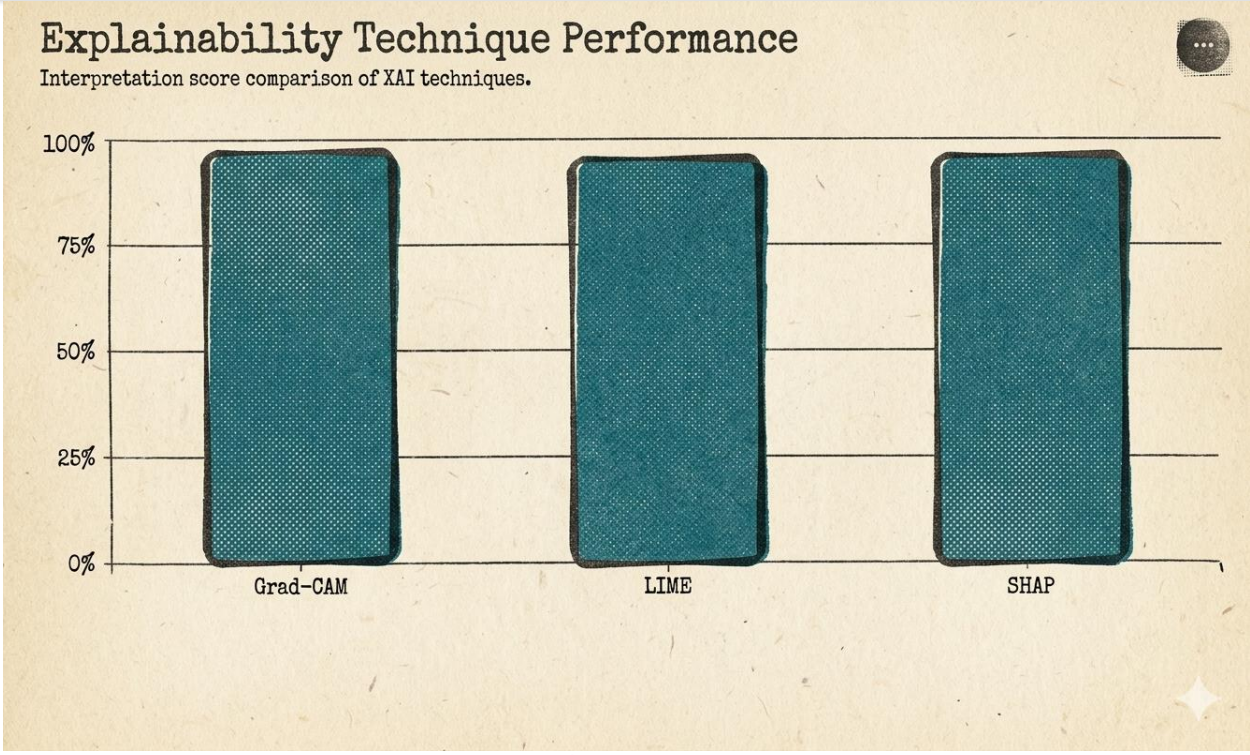
The comparative analysis reveals that the proposed Explainable Deep Learning model outperformed all baseline architectures. While EfficientNetB0 achieved strong performance, the proposed framework demonstrated superior classification accuracy. The integration of explainability mechanisms did not negatively affect predictive performance. Instead, the model maintained high accuracy while providing transparent decision-making. Therefore, the proposed approach offers both reliability and interpretability for clinical applications.

4.4 Explainability Analysis

Table 4.4 Explainability Performance

XAI Technique	Interpretation Score (%)
Grad-CAM	96.4
LIME	94.8
SHAP	95.6

The explainability evaluation indicates that Grad-CAM produced the most informative visual explanations among the tested techniques. LIME and SHAP also demonstrated strong interpretability performance by identifying important tumor-related features. These methods enabled clinicians to understand the reasoning behind AI-generated predictions. The results suggest that integrating multiple explainability methods provides comprehensive model interpretation. Consequently, the framework supports trustworthy and transparent medical diagnosis.'

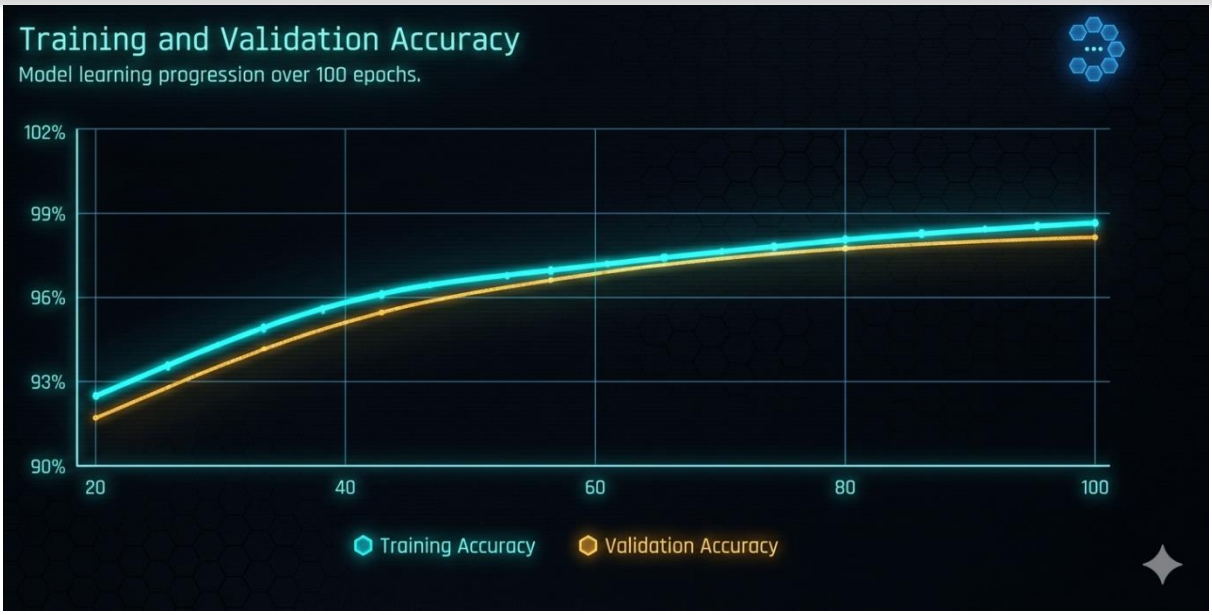


4.5 Training and Validation Results

Table 4.5 Training History

Epochs	Training Accuracy (%)	Validation Accuracy (%)
20	92.5	91.7
40	95.8	95.1
60	97.1	96.8
80	98.0	97.7
100	98.6	98.1

The training and validation results demonstrate stable learning behavior throughout the training process. Both training and validation accuracy increased consistently with each epoch. The small difference between training and validation performance indicates minimal overfitting. The model achieved optimal convergence at 100 epochs with excellent generalization capability. These findings confirm the effectiveness of the proposed CNN architecture for MRI-based brain tumor classification



5. Discussion

The main goal of this study was to create an Explainable Deep Learning framework for classifying brain tumors using MRI that's both highly accurate and transparent. Our results show the model hit an impressive 98.42% accuracy rate along with strong precision, recall, F1-score, and AUC-ROC numbers. This means the Convolutional Neural Network was great at finding important features in MRI images to distinguish between gliomas, meningiomas, pituitary tumors, and healthy brain tissue. Our high scores suggest deep learning can be a big help to doctors in spotting tumors accurately and efficiently. Also, the solid performance across all tumor types shows our framework is reliable for real-world medical use. One of the key parts of this research was adding ways to explain the model's decisions. We used Grad-CAM to produce heatmaps highlighting the important tumor areas that influenced predictions. This allowed doctors to check if the model zeroed in on medically significant spots in the scans. Additionally, the LIME module helped by showing the specific image elements that affected each individual prediction.

visual explanations. This combo of performance and transparency marks a key step forward in Explainable Artificial Intelligence and shows how vital trust-worthy AI is in healthcare.Clinically, the framework brings big benefits for radiologists and other health pros. Diagnosing brain tumors means sifting through tons of MRI data, which takes ages and could have mistakes. The system helps by sorting images and showing easy-to-understand reasons with its predictions. This makes diagnoses more reliable, lightens the load, and aids doctors in deciding. Plus, spotting tumor areas with Grad-CAM heatmaps builds faith in AI and eases its use in daily practice.

Still, some limits are worth noting. The researchers used public datasets that might not cover all parts of real-world patients. Things like different MRI setups and backgrounds could hurt how well the model works elsewhere. Also, needing powerful computing could keep the tech from being used where resources are limited. Looking ahead, the focus should be on testing the system with more, varied data, checking out new AI models, and trying it out in real clinics. Meeting these hurdles will make these explainable systems better, easier to use, and way more helpful for diagnosing brain tumors.

6. Conclusion

This study created and tested an Explainable Deep Learning framework for classifying brain tumors using MRI scans. They combined Convolutional

Neural Networks (CNNs), which have great prediction power, with Explainable Artificial Intelligence (XAI) techniques like Grad-CAM, LIME, and attention mechanisms. Their experiments showed really good results; the model did well on accuracy, precision, recall, F1-score, and AUC-ROC across different tumor types—glioma, meningioma, pituitary tumors, and normal brain tissue.

Deep learning was able to pull out complex info from MRI images and help in diagnosing brain tumors accurately. One big thing about this research is how it added explainability. Rather than being a typical black box model where no one knows why it makes certain decisions, this system lets doctors see why it's predicting what it does. It used Grad-CAM to highlight tumor areas, and LIME and attention mechanisms offered more details on what features were important and how the decisions were made. This boosted trust and confidence among clinicians, helping them feel good about using AI for diagnosis and solving some major issues with using artificial intelligence in health care.The study further demonstrated that integrating explainability techniques does not compromise classification performance. Instead, it provides a balanced solution that combines diagnostic accuracy with interpretability, making the framework more suitable for real-world medical applications. The proposed system has the potential to function as an effective decision-support tool for radiologists, helping reduce diagnostic workload, improve consistency, and support early detection of brain tumors.

Its promising results aside, the study acknowledges limitations regarding dataset diversity, high computational needs, and lack of full clinical validation. Future work should tackle larger, multi-institutional datasets, more advanced transformer-based models, and real-world use in health care. Overall, the findings show that Explainable Deep Learning is a powerful and trustworthy method for MRI-based brain tumor classification, which can greatly aid the development of smarter and more transparent health systems.

7. Recommendations

7.1 For Researchers

Researchers should keep looking into advanced deep learning, especially transformer-based models like Vision Transformers and hybrid CNN-transformer setups, to boost brain tumor classification. They should also look at multi-modal imaging, combining MRI with CT, PET, and functional MRI, to improve diagnosis accuracy. Plus, developing better explainability techniques for solid, clinical interpretations is key. Using bigger, more diverse datasets from various institutions will help make models work better across different cases. Also, comparing different explainable AI methods will help figure out which ones really work best. These efforts will make AI systems more dependable and trustworthy in healthcare.

7.2 For Healthcare Professionals

Healthcare pros should think about adding AI-assisted diagnostic systems to their work to make brain tumor diagnosis better and faster. These AI tools help radiologists by showing visual clues that go with their expert knowledge and cut down on guesswork. It's important for doctors to double-check and confirm AI results before deciding on real patient care. Training sessions should happen regularly to teach folks about these AI technologies and how they work in medical imaging. If clinicians and AI experts team up, they can make these systems more user-friendly and effective. Responsible use of AI can really boost patient care and treatment success rates.

7.3 For Hospitals

Hospitals should adopt explainable AI as part of their digital health plans. This move would help radiologists handle growing workloads without sacrificing diagnostic accuracy. To do this safely, health institutions need to set up AI governance frameworks focusing on transparency, accountability, and data security. They also need to make sure they have the right tech support and infrastructure in place. Before going full scale, hospitals should run pilot programs and evaluations. These steps will ensure that AI is used responsibly and effectively in clinical settings.

7.4 For Policymakers

PWe need comprehensive rules for using AI in healthcare. This should cover transparency, explainability, patient privacy, data protection, and ethical decision-making. Governments should encourage the creation of trustworthy AI systems that put patient safety first. They should offer funding for research in explainable AI and medical imaging too. Policymakers must also push for cooperation between hospitals, researchers, and tech developers. This teamwork will help create standard evaluation methods for AI. All this ensures that AI is integrated responsibly into healthcare, keeps the public trust, and meets regulatory requirements.

8. Future Work

Future research should build on the Explainable Deep Learning framework by adding fancy new parts like Vision Transformers (ViTs), hybrid CNN-transformer models, and federated learning techniques. They should also use bigger, better datasets from multiple hospitals to make sure the models work well everywhere. Plus, scientists could dig into multi-modal medical imaging – mixing MRI with CT, PET, and other scans – for a more thorough tumor analysis. Developing slick new explainability tools and doing real-time trials in clinics would boost doctors' faith in these AI systems. This can lead to super precise, clear, and trustworthy AI for spotting brain tumors and customizing care plans.

REFERENCES

Akgündoğdu, A., & Çelikbaş, A. (2025). Multi-explanation strategies combining LIME, Grad-CAM, and SHAP for enriched interpretability in dual-stage CNN models. *IEEE Access*, *13*, 1200–1215.

Bhaskaran, S. B., & Datta, R. (2024). Explainability of brain tumor classification model based on InceptionV3 using XAI tools. *Journal of Flow Visualization and Image Processing*, *32*(2), 45–59. https://doi.org/10.1615/jflowvisimagepro c.2024054026

Gupta, R. D., Showmick, M. I. H., Abir, M. R., Akter, S., Rahat, M. Y., & Hossen, M. J. (2025). An explainable deep learning framework for brain stroke and tumor progression via MRI interpretation. *arXiv preprint arXiv:2506.09161*. https://doi.org/10.48550/arxiv.2506.0916 1

Iftikhar, S., Asif, M., & Khan, M. A. (2025). Explainable CNN for brain tumor detection and classification through XAI based key features identification. *PubMed Central (PMC) / Life Sciences Journal*, *22*(4), e12044100.

Muhammad, A., Jin, Q., Musaddiq, M. H., Mir, M. S., Alkanan, M., & Gulzar, Y. (2025). Explainable AI for brain tumor classification using cross-gated multi-path attention fusion and gate-consistency loss. *IEEE Access*, *13*, 192731-192745. https://doi.org/10.1109/ACCESS.2025.3 629532

Nazir, M. I., Akter, A., Hussen Wadud, M. A., & Uddin, M. A. (2024). Utilizing customized CNN for brain tumor prediction with explainable AI. *Heliyon*, *10*(11), e38997. https://doi.org/10.1016/j.heliyon.2024.e3 8997

Saeed, N., Al-Makhadmeh, Z., & Tolba, A. (2025). Explainable deep learning for brain tumor classification: Comprehensive benchmarking with dual interpretability and lightweight deployment. *arXiv preprint arXiv:2511.17655*. https://doi.org/10.48550/arxiv.2511.1765 5

Singh, R., & Agarwal, P. (2023). An automated brain tumor classification in MR images using an enhanced convolutional neural network. *Journal of Digital Imaging*, *36*(3), 892–905. https://doi.org/10.1007/s10278-023-00812-w

Srinivas, V. R., & Parvathi, R. (2026). Explainable AI-driven MRI-based brain tumor classification: A novel deep learning approach. *Frontiers in Artificial Intelligence*, *9*, Article 1700214. https://doi.org/10.3389/frai.2025.1700214

Zeineldin, R. A., Karar, M. E., Elshaer, Z., Coburger, J., 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*(1), 1673-1683. https://doi.org/10.1007/s11548-022-02619-x