CAPITAL UNIVERSITY OF SCIENCE AND
TECHNOLOGY, ISLAMABAD

Explainability of MRI Analysis for Abnormal Brain Tissues

by

Shagufta Iftikhar

A thesis submitted in partial fulfillment for the
degree of Master of Science

in the

Faculty of Computing
Department of Computer Science

2024

This thesis is dedicated to my Family and my Teachers. I am deeply grateful to my beloved parents and siblings for their endless support and encouragement. I owe a special debt of gratitude to my supervisor, whose constant trust in me has helped me attain this crucial milestone.

CERTIFICATE OF APPROVAL

Explainability of MRI Analysis for Abnormal Brain Tissues

by

Shagufta Iftikhar

(MCS223003)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Tassawar Iqbal	COMSATS, Islamabad
(b)	Internal Examiner	Dr. Abdul Basit Siddiqui	CUST, Islamabad
(c)	Supervisor	Dr. Nadeem Anjum	CUST, Islamabad

Dr. Nadeem Anjum
Thesis Supervisor
September, 2024

Dr. Abdul Basit Siddiqui
Head
Dept. of Computer Science
September , 2024

Dr. M. Abdul Qadir
Dean
Faculty of Computing
September, 2024

Author's Declaration

I, **Shagufta Iftikhar** hereby state that my MS thesis titled "**Explainability of MRI Analysis for Abnormal Brain Tissues**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.

(Shagufta Iftikhar)

Registration No: MCS223003

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "**Explainability of MRI Analysis for Abnormal Brain Tissues**" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.

(Shagufta Iftikhar)
Registration No: MCS223003

List of Publications

It is certified that the following publication(s) have been made out of the research work that has been carried out for this thesis:-

1. **Explainable CNN for Brain Tumor Detection and Classification through XAI based Key Features Identification, MIT journal (submitted for publication)**

(Shagufta Iftikhar)
Registration No: MCS223003

Acknowledgement

First and foremost i am deeply grateful to Almighty Allah for blessing me with knowledge, strength, courage, and patience throughout my studies. I am also very grateful to my supervisor, **Dr.Nadeem Anjum**, for his close monitoring of the progress of this thesis, providing insights at every stage, and correcting the direction whenever necessary.

I would like to express my deepest gratitude to my dearest family members: my father, my mother, and my siblings, for their unconditional support during good and bad times. They have always encouraged me to stay motivated and achieve my goals.

Lastly, I would like to take a moment to acknowledge my efforts. Reflecting on my academic journey to complete this thesis, I am grateful to myself for maintaining a positive outlook and self-belief during challenging times. Balancing social life and household responsibilities, staying dedicated to my research, and remaining steadfast through every challenge have all been pivotal in overcoming obstacles and reaching this academic milestone. These efforts have truly been essential to my success.

(Shagufta Iftikhar)

Abstract

A tumor is an abnormal growth of cells that appears as a mass or lump in the body. Early detection of brain tumors is vital for enhancing patient survival rates. The existing manual diagnosis process is time-consuming and subjective, potentially leading to errors. To address this, several approaches have been developed.

Existing brain tumor classification model's complex structure often makes them difficult to interpret, leading to challenges in understanding decision-making processes of classification which can make models rely on unnecessary features or normal soft tissues. Additionally, Brain tumor classification models can have additional layers and perimeters which may produce imprecise results. So, there is a need to provide insights into model decision-making to make the system more robust and accurate with a reduced no of layers and make the model learn to focus on important features for classification which can be done by Explainable AI (XAI).

Our methodology is comprised of XAI-based CNN for achieving the key features for tumor classification that are the actual representation of tumor in MRI while reducing the number of layers. The effectiveness of the proposed model in detecting abnormal brain tissues within MRIs is demonstrated through XAI methods such as Grad-CAM, SHAP, and LIME. The proposed model achieved 99% accuracy on dataset 1 and 94% accuracy on an unseen dataset.

Contents

List of Figures

List of Tables

Abbreviations

ACC	Accuracy
AVG	Average
BT	Brain Tumor
CNN	Convolutional Neural Network
CPU	Central Processing Unit
DL	Deep Learning
F1	F1-Score
FP	False Positive
FN	False Negative
Grad-Cam	Gradient-weighted Class Activation Mapping
GPU	Graphics Processing Unit
Lime	Local Interpretable Model-agnostic Explanations
LRP	Layer-wise Relevance Propagation
MRI	Magnetic Resonance Imaging
ML	Machine Learning
ReLU	Rectified Linear Unit
Shap	Shapley Additive explanations
SVM	Support Vector Machine
TP	True Positive
TN	True Negative
XAI	Explainable AI

Chapter 1

Introduction

1.1 Tumor Background

A tumor is an abnormal growth of cells that manifests as a mass or lump within the body, representing a significant deviation from normal cellular behavior. Unlike healthy cells, which grow, divide, and die in an orderly manner, the cells in a tumor proliferate uncontrollably, often due to genetic mutations or environmental factors that disrupt regular cell cycle regulation. According to the World Health Organization (WHO), approximately 47,992 global cases of brain tumors are reported and 246,253 global deaths are caused by brain tumors. According to Pakistan Society of Neuro-oncology, a total of 2750 brain tumour cases are recorded in Pakistan [1]. Therefore, early detection of brain tumors is vital for enhancing patient survival rates.

Research has shown that various environmental factors, both external and internal, can contribute to the development of brain tumors. Among these factors, air pollution is one notable cause [2]. Genetic variation is another cause, with approximately 5–10 percent of cases occurring in individuals with a family history of the disease. Additionally, people exposed to radiation in their workplace have a higher likelihood of developing brain tumors [3]. Understanding the brain's working and structure of is key to identifying brain tumors and other neurological diseases [4].

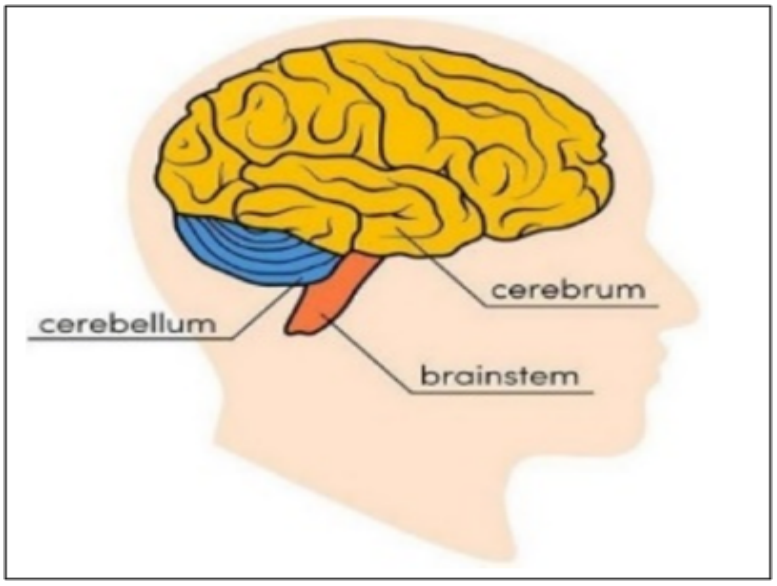


FIGURE 1.1: Brain Structure [2]

The brain has three main parts: the cerebrum, the brain stem, and the cerebellum as shown in above Figure 1.1.

1.1.1 Cerebrum

The cerebrum plays a crucial role in conscious thought, behavior, and movement. Comprising two hemispheres connected by the corpus callosum. The cerebrum is divided into four lobes.

1.1.2 Cerebellum

Positioned below the cerebrum near the skull's base, the cerebellum. It contains numerous Purkinje cells that receive signals from the cerebral cortex.

1.1.3 Brain Stem

The brain stem connects the spinal cord to the brain and includes the midbrain, pons, and medulla oblongata. It controls essential bodily functions and serves as a communication hub.

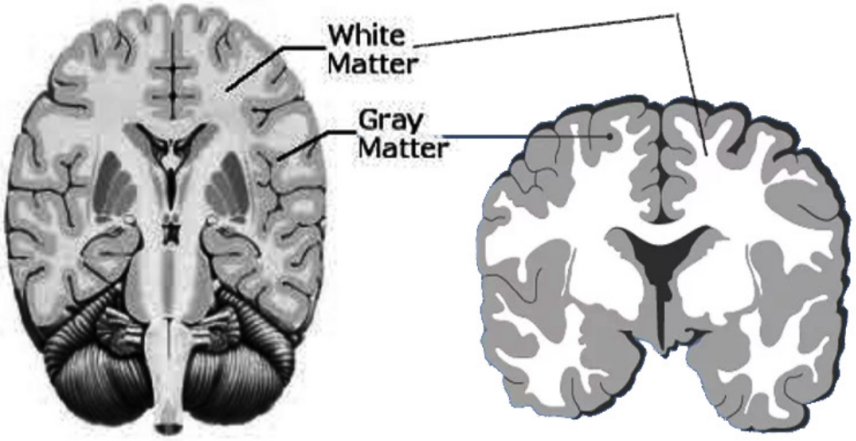


FIGURE 1.2: Brain Tissues

The brain is divided into two main types of tissue: grey matter and white matter as shown in Figure 1.2. The outer dark portion in color is a collection of neurons called grey matter. Grey matter consists of the cell bodies of neurons, which are the nerve cells for transmitting information throughout the brain. Grey matter is responsible for processing information, and it is found in the outer layer of the brain. This gray matter is sometimes referred to as the cerebral cortex [3]. White matter, on the other hand, is the innermost light portion made up of the nerve fibers, or axons, that connect different areas of the brain. These nerve fibers are covered in a fatty substance called myelin, which gives the white matter its characteristic white color. white matter is responsible for transmitting information between different areas of the brain and the rest of the body.

Symptoms of a brain tumor can differ significantly based on where it is located, its size, and its rate of growth. Common signs may include physical weakness, numbness, speech difficulties, vision changes, fatigue, memory problems, nausea, altered personality, and blurred vision [4, 5].

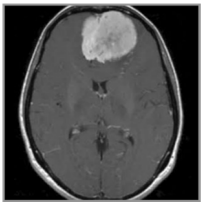
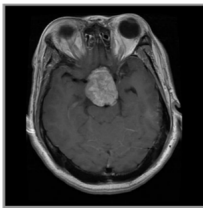
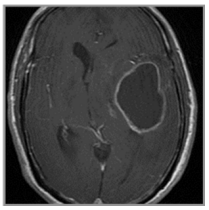
1.2 Types of Brain Tumors

Brain tumors are classified broadly into two categories: benign tumors, referred to as non-cancerous tumors, and malignant tumors, which are cancerous. Benign

brain tumors don't progress or disseminate [5]. Recurrence after removal through surgery of benign tumors is typically rare [6]. In contrast, malignant tumors are cancerous and tend to rapidly spread to other parts of the body. Without prompt and effective management, they can cause significant physiological dysfunction [5].

The two primary types of malignant tumors are those that originate within the brain (primary tumors) and those that develop elsewhere in the body and spread to the brain (secondary tumors) [7]. Additionally, meningiomas, pituitary tumors, and gliomas are subtypes of brain tumors that are highly prevalent [8] as shown in Table 1.1. According to the WHO, there exists four grade classification of brain tumors. Grades 1 and 2 refer to lower-grade (benign) tumors, such as meningioma and pituitary tumors [9]. In contrast, grades 3 and 4 indicate more severe (malignant) tumors, such as gliomas [10].

TABLE 1.1: Brain Tumor Types

Type	Image	Description
Meningioma		Extra-axial tumors develop adjacent to the meninges, which are the protective layers surrounding the brain and spinal cord.
Pituitary		An abnormal growth can form in the pituitary gland, a small organ at the base of the brain responsible for hormone production.
Glioma		Intra-axial tumors often exhibit thick, irregular borders that enhance around a central necrotic core with hemorrhagic features. These tumors arise from the brain's glial cells, which support neurons, and can emerge throughout the brain.

These types of tumors are characterized by essential features including location, shape, texture, and size. Their specific texture and location within the brain distinguish meningiomas and pituitary tumors. Conversely, gliomas are recognized by their distinct size and shape, as illustrated in Table 1.1 above [11]. Classifying tumors as meningiomas, gliomas, or pituitary tumors presents significant challenges due to their variability in size, shape, and severity [12].

Meningiomas often manifest with mild symptoms such as visual disturbances and morning migraines. Pituitary tumors may compress the optic nerve, leading to migraines, vision issues, and double vision. Gliomas induce various symptoms, including aphasia, cognitive decline, visual impairment or loss, challenges with walking or balance, and other associated manifestations [6].

These tumors are identified and classified by radiologists using different modalities of medical imaging. MRI and CT scans are commonly used to acquire different areas of the human body's detailed information. Due to the brain's sensitivity, MRI is preferred for assessing brain tumors because it is non-invasive. MRI generates detailed 3D images from multiple angles, offering comprehensive structural insights into the brain. Therefore, MRI is widely recognized as among the most effective modalities for automated analysis of medical images.[13–16].

Radiologists performing tasks on brain MRI focus on two primary objectives: (i) distinguishing between brain images with tumors and healthy brain images, and (ii) categorizing brain MRI scans that show tumors into different types [5].Manual detection and classification of brain tumors by radiologists are notably challenging due to significant diversity in tumor shapes and sizes within the same category, alongside the comparable appearances of various tumor types [8, 17].

Moreover, there is a shortage of experts in this specialized field. Manual identification and classification of brain tumors are laborious, time-consuming, and not easily reproducible when managing large volumes of MRI data. Errors in tumor type analysis can lead to serious consequences.

Additionally, categorizing brain tumors into multiple classes presents greater challenges over two class classification. Therefore, there is an urgent need for a reliable

automated system to assist radiologists in overcoming these challenges associated with manual brain tumor detection and classification [5].

To address this, several approaches for the segmentation, detection, and classification of brain tumors have been developed, utilizing ML and DL approaches. However, the superior performance of deep learning (DL) frameworks over traditional machine learning (ML) models in image classification and object detection tasks has highlighted their significance. This has motivated researchers to develop DL-based methods specifically for identification and classification of brain tumor.

1.3 Deep Learning Methods

Deep learning has significantly transformed medical image analysis, particularly in identifying and categorizing brain tumors. Using artificial neural networks with multiple hidden layers, deep learning enables systems to autonomously extract intricate patterns and features from medical images.

This advancement enhances detection accuracy and provides a deeper understanding of complex medical conditions, improving diagnostic processes in healthcare. The automated learning process boosts diagnostic efficiency and offers valuable insights to healthcare professionals, highlighting the potential of deep learning to revolutionize medical imaging and diagnostics.

In recent times, the research field has increasingly utilized multiple for brain tumor detection and classification based on deep learning (DL) approaches due to their enhanced performance. Figure 1.3 shows typical DL based brain tumor classification methodology. It is observed that deep learning models excel in classification tasks, surpassing traditional methods [5, 18–20].

However, their black box nature, characterized by intricate neural network layers, presents challenges in understanding how decisions are made inside these models, despite their exceptional performance. This lack of interpretability raises concerns, particularly in critical applications where understanding the decision-making process is essential.

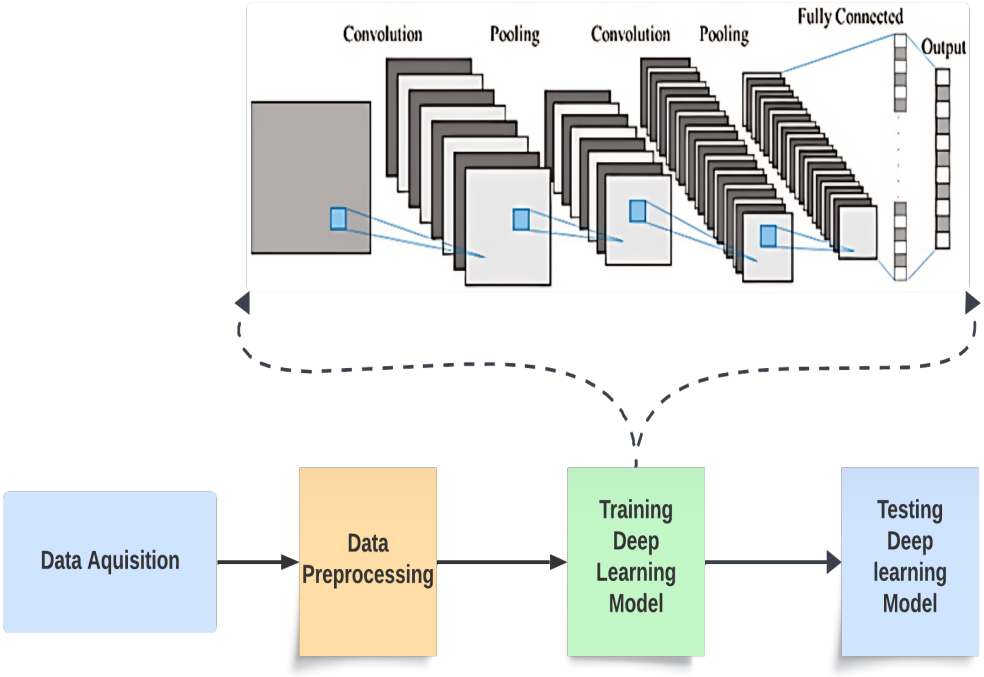


FIGURE 1.3: Typical CNN for Brain Tumor Classification

1.4 Deep learning methods with XAI

The complexity of deep neural networks can result in models that are difficult to interpret and explain. In medical settings, trust in the decision-making process of AI systems is paramount. Healthcare professionals and patients need to understand the reasoning behind a model's diagnosis or treatment recommendation [21–24]. Here comes Explainable AI, XAI that refers to the techniques and methods used to understand and interpret the decision-making process of these complex models. XAI aims to provide understanding that how DL models reach at their predictions or classification decisions, thereby enhancing transparency, and accountability.

XAI plays a crucial role in enhancing the interpretability and explainability of deep learning models, particularly in complex domains like healthcare. As deep learning models are highly effective at identifying patterns and making predictions, their "black box" nature often makes it difficult to understand how they arrive at

specific decisions. XAI aims to bridge this gap by providing methods and tools that make the workings of these models transparent and understandable to humans.

This transparency is essential for building trust, as stakeholders need to be confident in the model's decisions, especially in critical applications where errors can have significant consequences. XAI techniques, such as feature importance analysis, visualization of neural network layers, and generation of human-readable explanations, contribute to clarifying how deep learning models make decisions.

By enhancing the interpretability of these models, XAI not only enhances user trust but also aids in model debugging, improving overall performance and reliability. Furthermore, explainable models are better aligned with regulatory requirements, ensuring compliance and accountability in sensitive industries. Overall, XAI is indispensable for integrating deep learning models into real-world applications where transparency and trust are paramount.

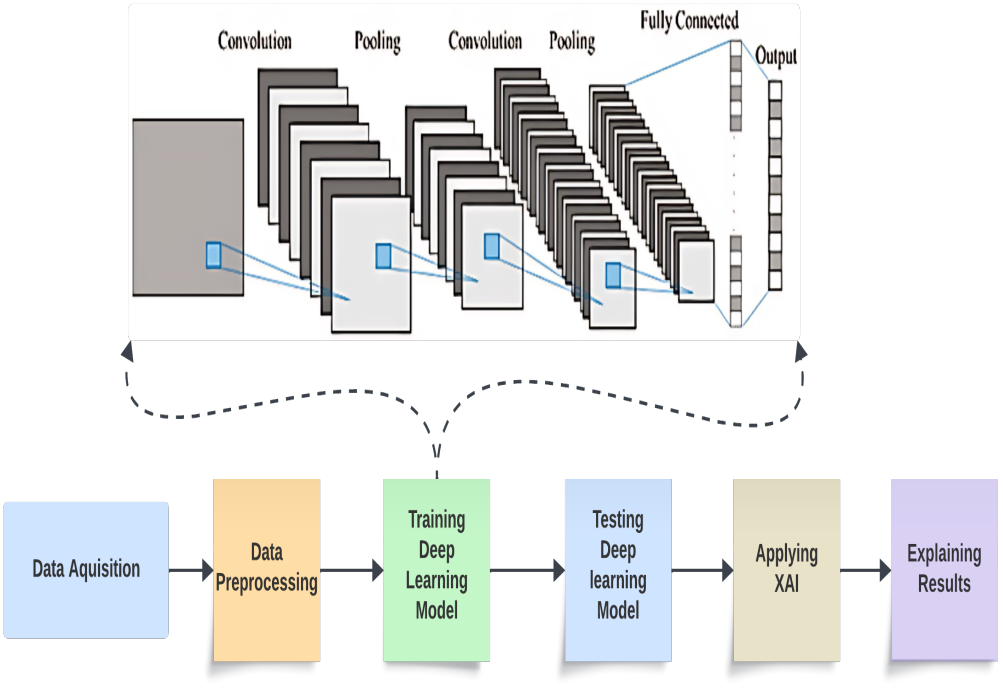


FIGURE 1.4: Typical CNN with XAI for Brain Tumor Classification

In recent years, several DL approaches with XAI have been developed. Figure 1.4 shows typical XAI based DL methodology for brain tumor classification.

However, all these previous studies that applied XAI have not provided proper visualization so they lack explainability. This lack of interpretation leads to model's complexity and models focus on unnecessary features, making the whole system questionable. Additionally, XAI is only being used for explainability not for enhancing the model's performance.

1.5 Motivation

It has been observed that DL models excel in classification tasks, surpassing traditional methods with remarkable accuracy. However from the literature survey of brain tumor classification using MRI, we found that deep learning models' complex structure often makes them difficult to interpret. This led to challenges in understanding the decision-making processes of classification which can make the model rely on unnecessary features. This lack of transparency is particularly problematic in medical settings, where trust and understanding of AI systems are crucial. Additionally, DL models can have layers or parameters that make no difference in output and may produce imprecise results. So, there is a need to present insights into the process of decision-making of the system for enhancing transparency. And to make the system more robust and accurate (focusing on tumorous features) with less complexity which can be done with XAI.

1.6 Problem Statement

Brain tumor classification model's complex structure often makes them difficult to interpret, leading to challenges in understanding decision-making processes of classification which can make models rely on unnecessary features or normal soft tissues. Additionally, Brain tumor classification models can have additional layers and parameters which may produce imprecise results. So, there is a need to provide insights into model decision-making to make the system more robust and accurate with a reduced no of layers and make the model learn to focus on important features (tumorous features) which can be done by Explainable AI (XAI).

1.7 Significance of the Solution

The classification decisions in previous techniques for brain tumor classification using MRI are not interpretable. And those previous studies that applied XAI have not provided proper visualization so they lack explainability. This lack of interpretation results in models focusing on unnecessary features or normal soft tissues, making the whole system questionable. Also, brain tumor classification models can have additional layers and parameters making them complex. Moreover, the previous studies did not use XAI to simplify the brain tumor classification models or to enhance the model's performance. So, our objective is to use XAI for developing an interpretable system that can improve the model's architecture and behavior, to improve the overall performance.

1.8 Research Questions

1. How to identify real contributing abnormal soft tissues from brain MRIs to improve the overall performance of the system?

2. What approaches and methodologies can provide insights into the model decision-making process to reduce the no of layers in brain tumor classification models?

1.9 Thesis Organization

The chapters in this thesis are organized as follows:

- Chapter 1 introduces the brain tumor and its subtypes including the motivation, problem statement of the proposed work and significance of the solution.

- Chapter 2 conducts a comprehensive literature review of brain tumor classification using deep learning approaches with and without XAI.

- Chapter 3 details the employed methodology, focusing on the use of XAI to make the brain tumor classification model less complex and interpretable.
- Chapter 4 presents the results and detailed discussion of the findings.
- Chapter 5 summarizes the conclusions drawn from the study's outcomes.

Chapter 2

Literature Review

Various researchers have proposed techniques for brain tumor classification. The inclusion criteria of our research for selecting literature review papers were based on their relevance to deep learning brain tumor classification using MR images. Moreover, papers on brain tumor classification with explainability are also selected. All these selected papers are from the years 2022 and above.

2.1 Survey of Existing Techniques

2.1.1 Deep Learning Techniques without Explainable AI

Alzahrani [25] proposed a transformer model combining global and local levels of attention mechanisms for classifying brain tumors. Their model got entire datasets global discriminative features and correlations using external attention. Then they achieved image patches local discriminative features and correlations through the self-attention mechanism. Additionally, the mixers of depth and point-wise convolution highlighted key areas of spatial and channel-wise features of Images. Lastly, final output feature maps were achieved by the use of a squeeze-and-excitation mechanism. Furthermore, they used data-augmentation techniques with ConvAttenMixer by using rescale and data augmentation layers.Data augmentation and

the simple classification head enhanced the performance of the model generally. For evaluation of the model, the dataset containing 7022 MRIs of the brain with 4 classes was used. Their model demonstrated an accuracy of approximately 97.94%.

In another study Gupta et al. [26], introduced an approach of 7-layered CNN to perform brain MRI classification in two classes as benign or malignant. They performed their investigation on a dataset containing 253 MRIs of the brain. For image preprocessing, the images were cropped to get the brain section only with the use of Edge Detection with Computer Vision (CV). The technique of data augmentation was used on the training dataset, using flipping, rotation, brightness, shear, and shift techniques to augment the amount of train data. They compared the performance of their CNN model with some pre-trained models including ResNet-50, VGG-16, and Inceptionv3, and some previous state-of-the-art models. Among these architectures, their proposed CNN model outperformed other architectures with 94% accuracy.

AlTahhan et al. [27] initially conducted classification using GoogleNet and AlexNet, two convolutional neural networks that were pre-trained and fine-tuned. Among these, AlexNet demonstrated superior performance. To further enhance the performance of fine-tuned AlexNet, they explored two hybrid approaches: one combining AlexNet with SVM and the other combining AlexNet with KNN. They worked with a limited dataset comprising 2880 brain MRIs (T1-weighted contrast-enhanced) and utilized K-Fold cross-validation to partition the data. The models, which included fine-tuned GoogleNet, fine-tuned AlexNet, AlexNet-SVM, and AlexNet-KNN, achieved accuracies of 88%, 85%, 95%, and 97%, respectively.

Rasheed et al. [28] introduced a methodology starting with image enhancement techniques, including sharpening with Gaussian blur and applying Adaptive Histogram Equalization using CLAHE. They subsequently employed a modified CNN model for classification. Their study utilized an MRI dataset consisting of 7023 images for model training. The performance of their approach was compared against pre-trained models such as VGG19, InceptionV3, VGG16, ResNet50, and MobileNetV2. During experimentation, the proposed method achieved a classification accuracy of 97.84%.

Özkaraca et al. [29] Firstly, basic CNN architecture, VGG16Net, and DenseNet architectural structures were investigated to know the effect of transfer learning approaches on the success rates of their classification problem. Their findings showed that the transfer learning approaches that they used did not provide the results as expected. So, they proposed a modified CNN architecture and used 10-fold cross-validation by which they achieved 96% accuracy. They used 7021 brain MRIs for model training.

Gómez-Guzmán et al. [30] proposed a generic convolutional neural network (CNN) model. They also performed a study on six pre-trained CNN models. They evaluated their model with the dataset containing 7023 MRIs. The dataset was preprocessed by resizing the images and adding labels, also augmentation on the dataset was applied by Zooming or Scaling, Rotating, and adding Brightness to MRIs. They evaluated their proposed CNN, Xception, MobileNetV2, ResNet50, InceptionV3, Inception, ResNetV2, and EfficientNetB0. Among all the listed models the best performance was of InceptionV3, as it achieved 97.12% average accuracy.

Islam et al. [31] employed transfer learning models such as DenseNet121, InceptionV3, MobileNet, and VGG19. They utilized a Kaggle dataset consisting of 7023 brain MRIs categorized into 4 classes. Image augmentation for class balance was conducted using Keras' ImageDataGenerator class. The study findings indicated that MobileNet achieved the highest accuracy of 99.60%.

Imam and Alam [32] investigated the influence of some loss functions, such as focal loss, and also the methods for data oversampling, including ADASYN and SMOTE. They also incorporated data augmentation in addressing the data imbalance issue. The dataset that they used was imbalanced with no of 4200 brain MRIs. They augmented samples of minority class with contrast, brightness, and sharpness alteration each with a random intensity of 80% to 120%. They evaluated VGG16-CNN, EfficientNetB0-CNN, EfficientNetB3-CNN, ResNet50-CNN, DenseNet201-CNN, MobileNet-CNN, GoogleNet-CNN, and XceptionNet-CNN. With augmentation, their proposed strategy, which combines VGG-16 and CNN, achieved an accuracy rate of 96% which was the highest among loss functions and oversampling methods.

Peng and Liao [33] introduced a convolutional neural network (CNN) approach for the classification of brain tumors from MRI images. Their dataset, sourced from Kaggle, consisted of 3264 MRIs categorized into four classes. Prior to input into the CNN architecture, the images were resized to 224 × 224 pixels. The CNN architecture comprised 24 layers designed for feature extraction. They reported an average classification accuracy of 94.4% on the testing set.

Shanjida et al. [34] developed a novel CNN-KNN architecture for the detection and classification of various types of tumors. They utilized a dataset of 2879 MRIs from Kaggle, categorized into four classes. As a preprocessing step, the images were resized and transformed into grayscale. The CNN was employed to extract different intensity-based features during the feature extraction phase. Subsequently, two machine learning classifiers, Softmax and KNN, were utilized in the classification step. KNN demonstrated superior accuracy compared to Softmax. Overall, their method achieved an accuracy of 95.7%.

To summarize the literature on existing techniques based on Deep Learning, Table 2.1 presents the numerous investigations that have been conducted on brain tumors in recent years.

TABLE 2.1: Analysis of Existing Techniques Based on Deep Learning

Method	Dataset	Accuracy	Limitation
Conv Atten Mixer [25]	7023 MRIs	97%	XAI not used
CNN-7 Layers [26]	253 MRIs	94%	XAI not used & only performed Binary Classification
AlexNet-KNN [27]	2880 MRIs	97%	XAI not used
Image Enhancement + Modified CNN [28]	7023 MRIs	97%	XAI not used
Modified CNN [29]	7023 MRIs	96%	XAI not used
InceptionV3 [30]	7023 MRIs	97%	XAI not used

TABLE 2.1: (Continued from Previous Page)

Method	Dataset	Accuracy	Limitation
MobileNet [31]	7023 MRIs	99%	XAI not used
VGG-CNN [32]	4200 MRIS	96%	XAI not used
CNN-24 Layers [33]	3264 MRIs	94%	XAI not used & achieved low accuracy
CNN-KNN [34]	2879 MRIs	95%	XAI not used & achieved low accuracy

2.1.2 Deep Learning Techniques with Explainable AI

Kumar et al. [35] proposed a framework for brain tumor classification using VGG19 and InceptionV3 CNNs, incorporating data augmentation techniques. They utilized a Kaggle dataset consisting of 253 MRI scans categorized into two classes. The MRI scans underwent preprocessing steps including normalization, thresholding, cropping, and resizing to 229 $\times$ 229 pixels.

For augmentation, they employed Keras' ImageDataGenerator class to apply rescaling, horizontal and vertical shifting, shear transformation, zooming, rotation, flipping, and brightness adjustments to the images. The weights of the pre-trained VGG19 and InceptionV3 models were frozen, and the model architecture was extended by adding a flatten layer, a dense layer, and a sigmoid activation function to the sequential model.

VizGradCAM was used for visualizing and interpreting the model's predictions. Their experiments demonstrated high accuracy rates of 98% for VGG19 and 96% for InceptionV3 CNNs.

Benyamina [36] performed binary classification of brain tumors using a pre-trained VGG16 model. MRI dataset from Kaggel containing 3000 MRI images of two classes was used for brain tumor classification. XAI method SHAP was used for model explainability.

Ahmed et al.[37] conducted a study using the VGG16 model to classify brain MRIs from a Kaggle dataset consisting of two classes: normal and tumor. The VGG16 model was trained and achieved a testing accuracy of 97.33%. To address the interpretability of DL models, Layer-wise Relevance Propagation (LRP) was applied to the VGG16. After training and making predictions with VGG16, LRP, an XAI method, was employed to provide explanations for the model's predictions. If the explanations provided by LRP were considered satisfactory, the trained model was deployed in the cloud, otherwise the model underwent retraining to enhance its performance.

Gaur et al. [38] introduced an explanation-driven deep learning model employing a CNN with a dual-input strategy. They utilized a dataset comprising 4 categories totaling 2,870 images. Initially, all 512 $\times$ 512 $\times$ 3 images were resized to 150 $\times$ 150 $\times$ 3 pixels. Gaussian noise with a mean of 0 and standard deviation of 100.5 was applied for data augmentation to enhance accuracy. Their CNN model consisted of six hidden layers with an output layer size of 1 $\times$ 4. To enhance accuracy, the researchers utilized two copies of the dataset in the CNN, processing one in the convolution layer and the other in the fully connected layer. They performed K-fold cross-validation with K=10 non-overlapping folds across 20 epochs, employing a batch size of 128. For interpretation of the model , they employed local interpretable model-agnostic explanations (LIME) and Shapley additive explanations (SHAP) to explain predictions related to brain tumors. Their trained CNN achieved 94.64% training accuracy and 85.37% test accuracy.

Mercaldo et al. [39] introduced convolutional neural networks along with class activation mapping to enhance explainability. They experimented with four distinct models: VGG16, ResNet50, Alex_Net, and MobileNet. Their approach was evaluated using a dataset consisting of 3000 brain MR images categorized into 2 classes, achieving a test accuracy of 99% with the ResNet50 model. Prediction explanations were provided using XAI method Grad-CAM.

To summarize the literature of existing techniques based on DL with XAI, below Table 2.2 presents numerous investigations that have been conducted on brain tumors in recent years.

TABLE 2.2: Analysis of Existing Techniques Based on Deep Learning with XAI

Method	Dataset (MRIs)	Accuracy	XAI Method	Limitation
VGG19 [35]	253	98%	Grad-Cam	Performed Binary Classification and explainable visuals not provided
VGG16 [36]	3000	N/A	Shap	Classification results not found
VGG16 [37]	3000	97%	LRP	Performed Binary Classification
CNN with dual-input [38]	2870	85%	Lime and Shap	Achieved low accuracy
Resnet50 [39]	3000	99%	Grad-Cam	Performed Binary Classification

2.2 Research Gap

From the literature survey of brain tumor classification using MRI, we found that the brain tumor classification models' complex structure often makes them difficult to interpret. The previous studies that applied XAI have not provided proper visualization so they lack explainability. This led to challenges in understanding the decision-making processes of the model which can make the model rely on unnecessary features or normal soft tissues. Additionally, brain tumor classification models can have additional layers and parameters that make no difference in output but make it complex, and may produce imprecise results. So, there is a need to present insights into the system's process of decision-making for enhancing transparency. And to make the system focusing on real contributing abnormal soft tissues with a reduced number of layers which can be done with XAI.

Chapter 3

Methodology

The current manual diagnostic process is slow and subjective, potentially leading to errors. Hence, there is a critical need for automated techniques that can provide accurate and efficient classifications. To address this, several approaches for segmenting, detecting, and classifying brain tumors have been implemented, utilizing ML and DL techniques. However, it is observed that deep learning models excel in classification tasks, surpassing traditional methods with remarkable accuracy [5, 18–20]. Yet, their black box nature, characterized by intricate neural network layers, presents challenges in understanding how decisions are made, despite their exceptional performance.

Deep neural networks can be complex, making them challenging to interpret and explain. In medical contexts, trust in AI decision-making is critical. Healthcare providers and patients require transparency about why a model makes specific diagnoses or treatment recommendations. Current challenges in understanding classification decisions can lead models to rely on unnecessary features or non-tumor tissues in MRI scans, resulting in inaccurate results for practical use. Also, brain tumor classification models are computationally intensive, with many layers and parameters, some of which may not affect positively the final output. Explainable AI (XAI) methods are used to interpret these models' decision-making processes. XAI aims to reveal how deep learning models make predictions or classifications, improving transparency and accountability.

Previous studies applying XAI have sometimes lacked proper visualization, limiting their interpretability and potentially leading models to focus on irrelevant features. These studies also haven't fully leveraged XAI to simplify brain tumor classification models or improve performance.

To address these aspects, our approach combines XAI with CNN to enhance system performance. Our methodology is described in below figure 3.1. We employed Grad-CAM to identify crucial tumor features in MRIs, ensuring accurate tumor classification. Additionally, XAI is used to reduce unnecessary layers of the model.

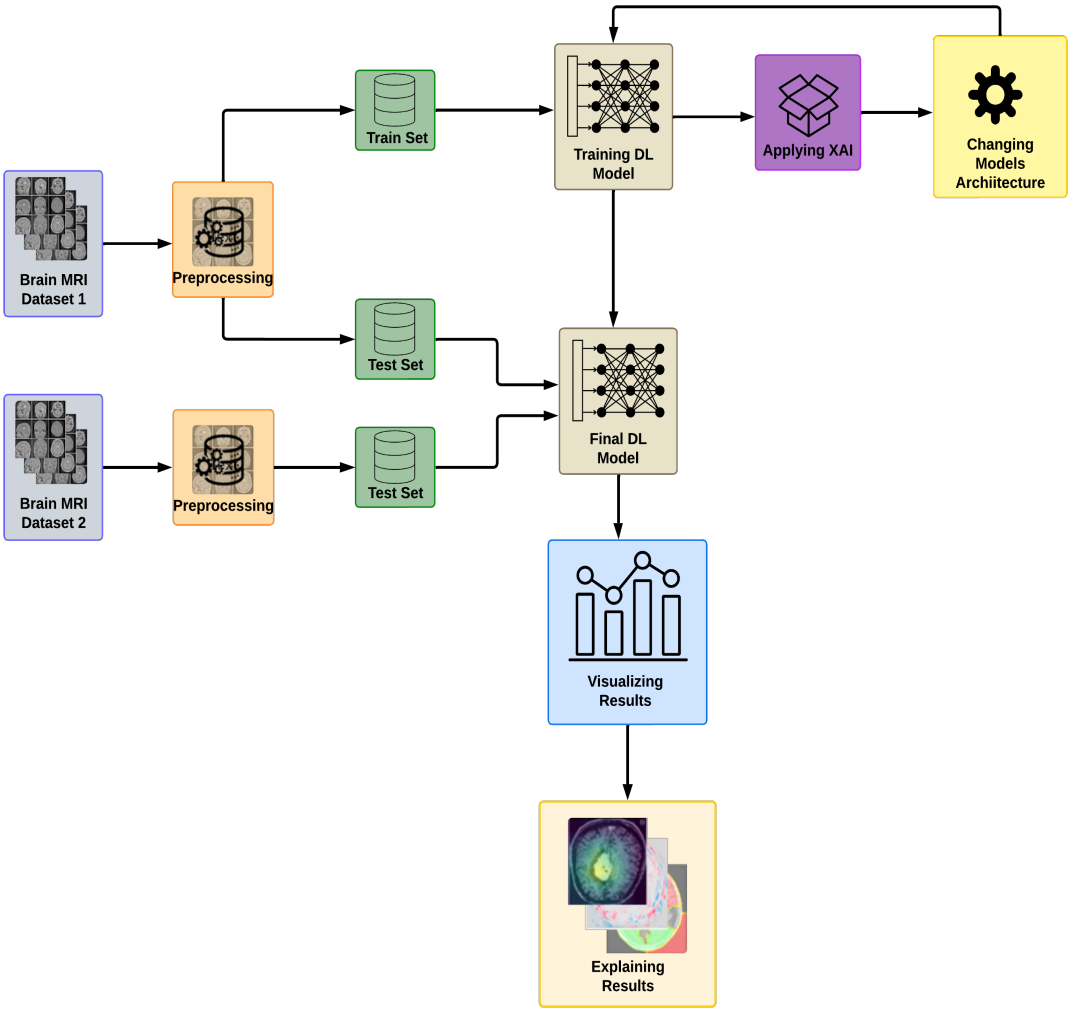


FIGURE 3.1: Proposed Methodology

In this process firstly we attained two publicly available datasets of brain MRIs. The MRIs were preprocessed to make it easy for deep learning models to understand the relevant features and to learn discriminative features in images. In a

preprocessing step, we performed cropping and resizing of the MRIs.

Then preprocessed dataset was divided into a train set, test set, and validation set. Then we trained our deep learning model on the train set. After training importance of all layers of the model was achieved through the XAI method Grad-Cam, and the layers that were below a certain threshold were removed from the model. Then the model was sent again in the training phase to get the final model.

This process was continued until a certain requirement was fulfilled. Afterward, testing was performed for model evaluation. For checking how well our model is generalized we tested it on a new dataset. To know why and how the model made the decisions in the testing phase, we visualized explanations made by XAI methods Grad-Cam, Shap, and Lime. These methods ensured the correctness of classification decisions. Lastly, a comparison of the proposed architecture with existing approaches was made.

3.1 Datesets

3.1.1 Dataset 1

The first brain MRI dataset is the Msoud dataset of brain tumors, constructed by Nickparvar (2021). Table 3.1 displays the specifics of the dataset, includes 7023 MR images of the human brain, available in grayscale with JPG formats. Expert radiologists carefully annotated these images to classify them based on the presence or absence of three types of tumors, resulting in four distinct classes in the dataset:

1. Glioma
2. Meningioma
3. No-tumor
4. Pituitary

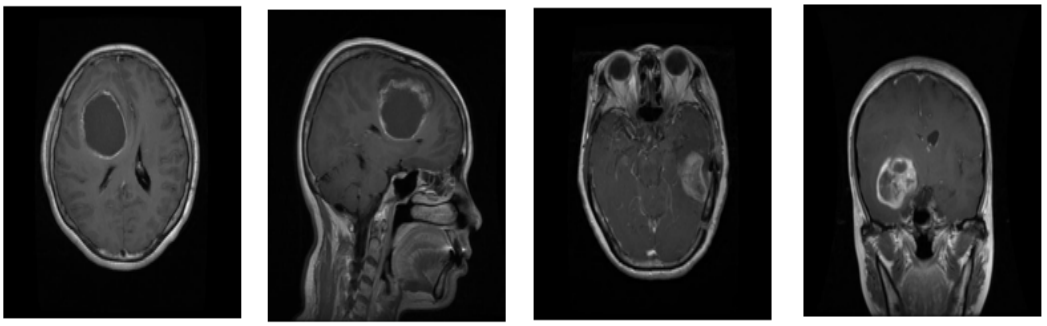


FIGURE 3.2: Glioma

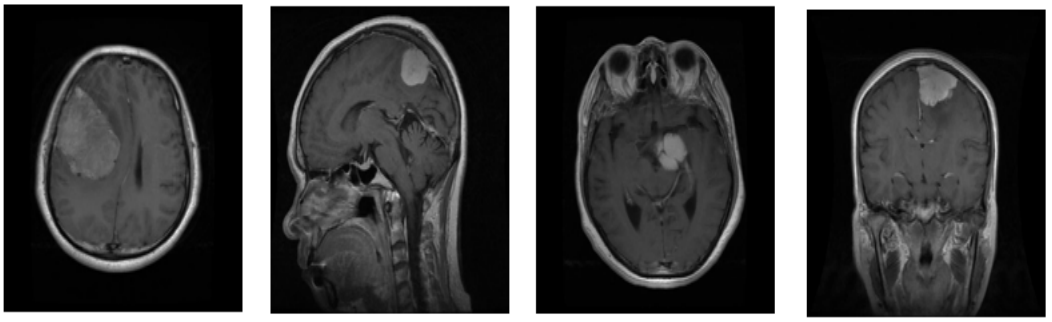


FIGURE 3.3: Meningioma

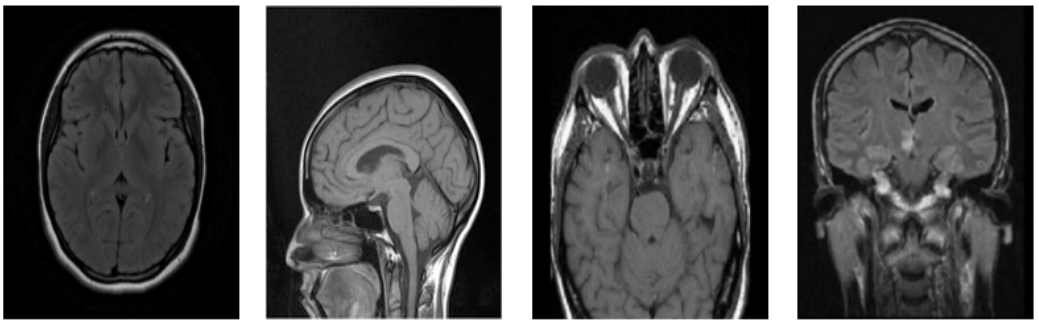


FIGURE 3.4: No Tumor

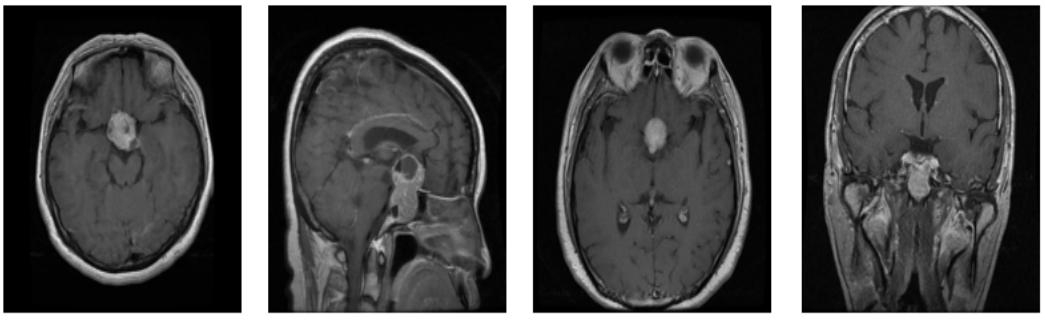


FIGURE 3.5: Pituitary

TABLE 3.1: Brain Tumor Dataset 1 Description

Tumor Type	Number of MRIs
Glioma	1621
Meningioma	1645
No Tumor	2000
Pituitary	1757

3.1.2 Dataset 2

The second brain MRI dataset is an open repository for detection and classification tasks of brain tumors. This dataset is used to test the model's generalizability. Table 3.2 displays the specifics of the dataset, which includes 3264 MRIs of the human brain, available in grayscale with JPG formats. Expert radiologists meticulously annotated these MR images to classify them based on the presence or absence of three types of tumors, resulting in four distinct classes in the dataset as mentioned below:

1. Glioma
2. Meningioma
3. No-tumor
4. Pituitary

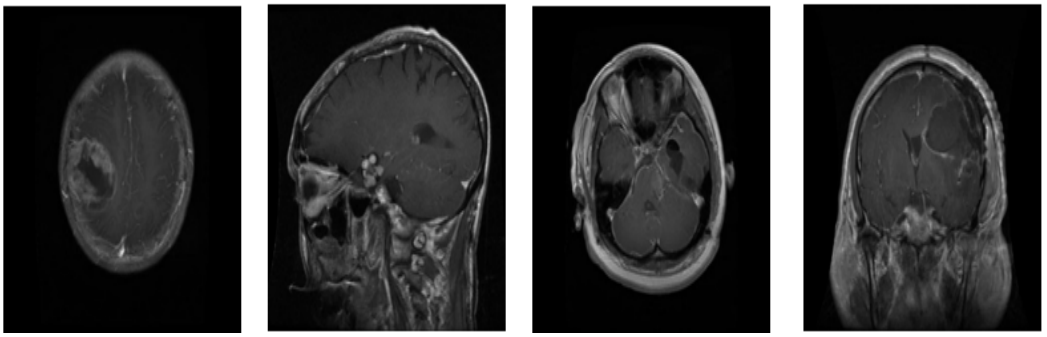


FIGURE 3.6: Glioma

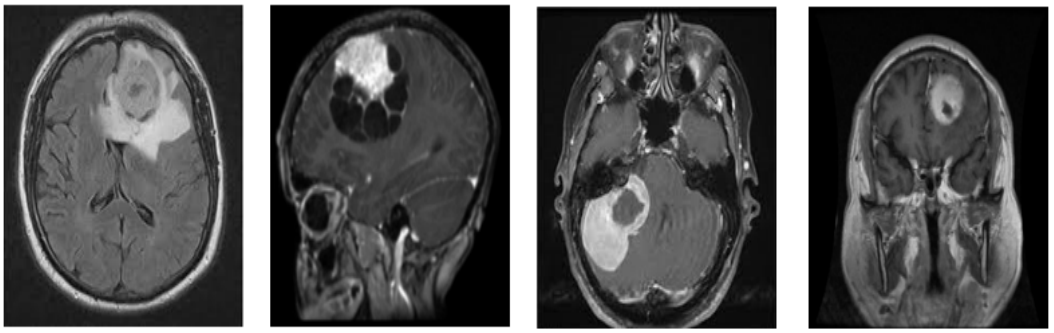


FIGURE 3.7: Meningioma

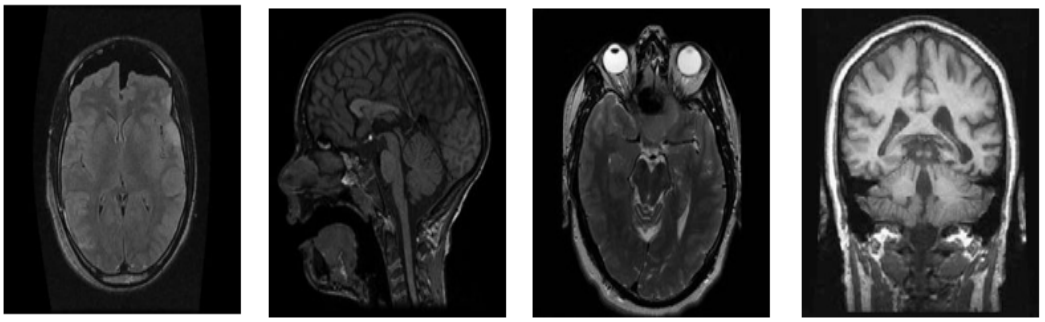


FIGURE 3.8: No Tumor

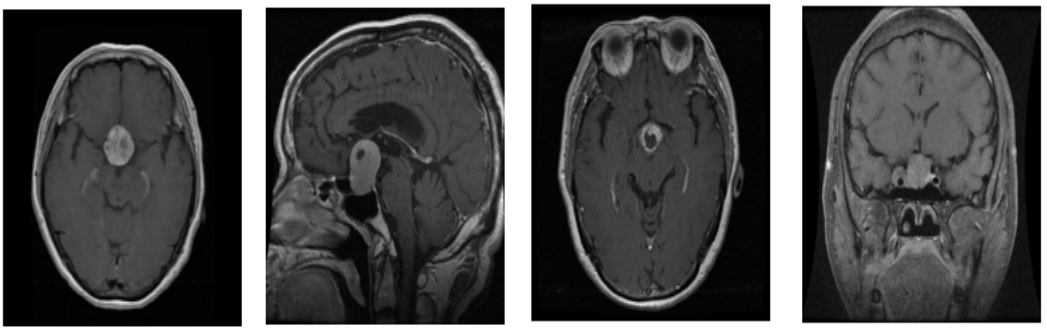


FIGURE 3.9: Pituitary

TABLE 3.2: Brain Tumor Dataset 2 Description

Tumor Type	Number of MRIs
Glioma	937
Meningioma	926
No Tumor	500
Pituitary	901

3.2 Data Pre-Processing

Data preprocessing is an important process of preparing the raw data for analysis or for training to use in machine learning or deep learning model [40]. This broadly refers to a set of operations and transformations that require the input data to be cleaned, formatted, and structured for the next analysis or model-building process. Some of these methods involve data cleaning, handling of missing values, handling of categorical data by encoding and the numerical data through normalization and scaling among others, and splitting of the data into training and test sets. Combined, these measures contribute toward the data quality, and increase its compatibility with the further analysis and the models' performance.

3.2.1 Cropping

Image cropping is important before the model training mainly because of the following reasons. Cropping assists in reducing the model's attention to unnecessary background or unwanted objects. It also helps to diminish the amounts of the data which results in shorter times of training and lesser expenses on computations. Moreover, a central idea is given to the generator by identifying critical regions which improves the model's accuracy since it is not compromised by the peripheral views. Finally, cropping ensures that aspect ratios and sizes are constants, which are common in various images, and one of the reasons that make the dataset uniform concerning deeper model training.

In this study, the first step in preprocessing was cropping in which firstly images were converted to grayscale. On the grayscale images, thresholding was applied to separate the object of interest from the background. Further morphological operations (erosion and dilation) to the threshold image were performed to remove noise. To crop the images, contours in the threshold image were found. Among all the contours, the largest contour in the image was selected. And the extreme points (top-left, top-right, bottom-left, bottom-right) of the largest contour were found. Finally, the original image using the extreme points was cropped and a few pixels to the cropped region were added.

3.2.2 Normalizing

The purpose of normalization is to scale the pixel values of the image to a common range, usually between 0 and 255, to improve the contrast and visibility of the image. This is particularly useful when working with images that have varying brightness or contrast levels. In this specific case, the normalization is done to ensure that the pixel values of the cropped image are within the range [0, 255], which is the typical range for 8-bit unsigned integer images. This is useful for displaying the image or for further processing steps that expect images with pixel values in this range.

3.2.3 Resizing the Images

Generally, deep learning models require larger input images and as a result, there are more parameters and computations which put pressure on the available hardware and thus slow down both training and inference. To overcome this problem resizing the input images to a smaller size is employed to reduce the computational cost. In this study, the original MRIs were in size $512 \times 512 \times 3$, therefore, all the images were resized to $224 \times 224 \times 3$. Not only does this resizing operation relieve the computational load but also improves the deep learning model's overall processing speed.

3.3 Proposed Model

In this research study, we have introduced a novel XAI-based CNN Model, which is specifically designed for the task of Multi-Class Image Classification focusing on important features [41]. This innovative model builds upon the foundation of the XAI that is a powerful structure for interpreting algorithms. However, we have extended its capabilities by incorporating an approach that effectively modifies the model's architecture and allows it to use important or tumorous features that should contribute to the classification task.

This research is primarily motivated by harnessing the strengths of the XAI architecture, which is renowned for its ability to interpret the black box nature of deep learning models, especially for medical image analysis like multi-class image classification of brain tumors. By integrating XAI with CNN, our approach leverages the power of focusing on contributing abnormal soft tissues while also enabling the model to have a reduced number of layers, making it a valuable tool for a wide range of image classification tasks. This hybrid model represents a significant advancement in the field of medical automated systems, offering reliability, trust, enhanced performance, and accuracy for multi-class image classification scenarios.

3.3.1 CNN

The architecture of CNNs is designed to mimic the visual processing of the human brain. Key components of CNNs include convolutional layers, activation functions, pooling layers, and fully connected layers [42]. Convolutional layers apply a series of convolutional filters to the input images, generating feature maps that highlight various aspects such as edges, textures, or patterns. These filters slide across the image and perform element-wise multiplications, which are then summed up to produce the feature maps. Activation functions, like ReLU (Rectified Linear Unit), are then applied to introduce non-linearity into the model, allowing it to learn more complex representations. Pooling layers follow the convolutional layers and serve to reduce the spatial dimensions of the feature maps, which helps in decreasing the computational load and mitigating the risk of overfitting. By retaining the most significant features and discarding irrelevant information, pooling layers ensure that the model captures the essence of the features.

After several rounds of convolution layers and pooling layers, the high-level, abstract features are passed through fully connected layers (FC layers). These layers combine all the features to make final predictions about the class of the input image. In context of multi-class classification like in our study, the output layer typically employs a softmax activation function to produce a probability distribution across the different classes, indicating the likelihood of each class being the correct one.

3.3.2 XAI

Explainable AI (XAI) architecture refers to the design and structure of AI systems that prioritize interpretability and transparency in their operation and outputs. This architecture integrates mechanisms that make the AI model's decision-making process understandable to humans, enabling users to gain insights into how and why certain decisions are made. It encompasses various techniques and approaches designed to make AI models more interpretable and transparent [43]. These can be broadly classified into several types, based on their methodology and application:

1. Post-Hoc Explainability techniques analyze and explain models after training. Feature importance identifies which features most influence the model's predictions, using methods like permutation importance and gradient-based importance. Saliency maps and heatmaps, visual tools in image-based models, highlight the parts of an image that most influenced the model's decision. LIME explains individual predictions by perturbing inputs and observing changes in outputs. SHAP uses concepts from cooperative game theory to provide a unified measure of feature importance by considering all possible feature combinations.

2. Intrinsic Explainability involves the models that are inherently interpretable. Decision trees use simple tree structures where decisions are based on feature splits, making them easy to understand. Linear and logistic regression models provide coefficients for each feature, indicating their weight and influence on the outcome. Rule-based models, like decision rules and association rule learning, use if-then rules for decision-making, which are also easy to interpret.

3. Model-Specific Explainability techniques are customized for particular types of models. For instance, attention mechanisms in models such as transformers illuminate which aspects of the input data the model emphasizes when making decisions. Feature visualization techniques, like activation maximization, display input patterns that maximize the activation of specific neurons or layers within a neural network.

4. Model-Agnostic Explainability methods are applicable to any AI model, regardless of its architecture. Partial dependence plots (PDPs) illustrate the impact of one or two features on the predicted outcome while averaging over the values of all other features. Surrogate models utilize straightforward, interpretable models such as decision trees or linear models to approximate the behavior of a more complex model.

5. Global Explainability techniques provide an overall understanding of the model's behavior. Feature importance ranking identifies and ranks features based on their overall impact on the model's predictions. Global surrogate models are simplified models that approximate the behavior of the entire complex model, offering an overview of how the model makes decisions.

6. Local Explainability methods provide insights into individual predictions. Counterfactual explanations describe how to change the input data to obtain a different desired output, helping to understand decision boundaries. Individual conditional expectation (ICE) plots show how predictions change when a single feature is varied, holding all other features constant, for individual instances.

7. Visualization Tools or techniques use visual aids to make the model's decision process more understandable. Heatmaps and saliency maps are visual representations that highlight important features or regions in the input data. Decision plots and graphs illustrate the sequence and impact of features on individual predictions.

3.4 XAI Based CNN Model

Our CNN model is designed for image classification tasks and consists of multiple layers that sequentially process the input data. The model begins with a convolutional layer that takes in 224x224x3 RGB images as input, applies 8 filters with a kernel size of 3x3, and then passes through a ReLU activation function. Next, a max pooling layer follows this layer that down-samples the feature maps

by a factor of 2 in both width and height. This pattern of convolutional and max pooling layers is repeated multiple times, with the number of filters increasing exponentially (16, 32, 64, 128, and 256) and the kernel size remaining constant at 3x3.

The padding is set to 'same' to maintain the spatial dimensions of the feature maps. After the convolutional and pooling layers, the model adds a batch normalization layer to normalize the activations, followed by an average pooling layer that down-samples the feature maps again. The output of the convolutional and pooling layers is then flattened into a 1D feature vector, which is fed into two dense layers with 512 neurons each and ReLU activation. Finally, the model ends with a dense layer that outputs a probability distribution over 4 classes using the softmax activation function.

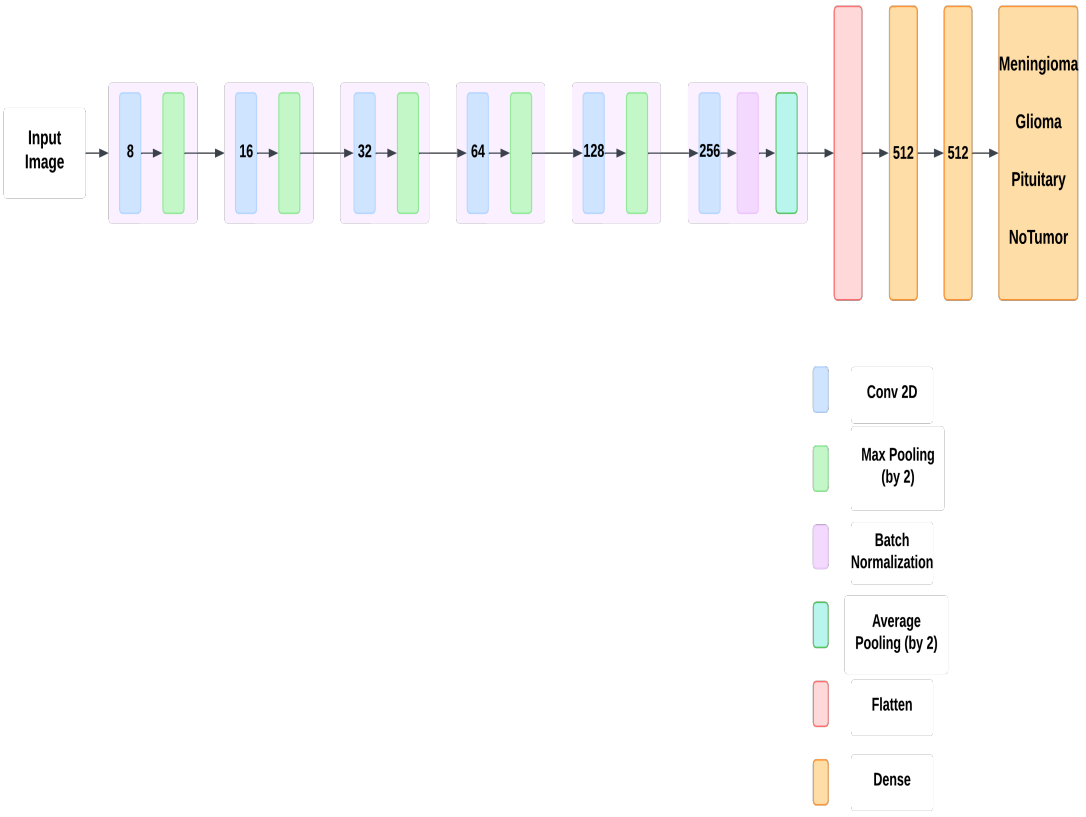


FIGURE 3.10: Proposed CNN Architecture

As brain tumor classification models complex structure often makes them difficult to interpret leading to challenges in understanding decision-making processes of

classification which can make models rely on unnecessary features or normal soft tissues.

Additionally, brain tumor classification models can have additional layers and parameters that make no difference in output but make it complex, and may produce imprecise results. To overcome these limitations, we used the XAI method Grad-Cam with CNN which made the model focus on important features with a reduced number of layers.

3.4.1 Gradient-Weighted Class Activation Mapping (Grad Cam)

Grad-CAM (Gradient-weighted Class Activation Mapping) is a technique used in the field of Explainable AI (XAI) for understanding the importance of different layers and features within a convolutional neural network (CNN). Grad-CAM helps to visualize which parts of an input image contribute the most to the model's predictions. By examining different layers of the network, you can gain insights into the hierarchical importance of features at various model depths. Grad-CAM [44] represents a progression from conventional CAM methods [45]. Unlike CAM, which relies on a CNN structured around global average pooling (GAP) [46], Grad-CAM is adaptable across various CNN architectures. In contrast to typical pooling methods, GAP compresses each 2D feature map into a 1D vector using an average function before applying classification via the softmax function. The GAP operation is mathematically defined as shown in Equation (1).

$$g_p = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_{p(ij)} \tag{3.1}$$

where g_p is the p^{th} 1D feature after GAP, H and W are the height and width of the 2D feature map, respectively, and g_p is the p^{th} convolved feature map at position (i,j). The attention map of the model is generated by combining convolved feature maps with the weights connecting the GAP layer to the output, described in Equation (2):

$$M^p_{(x,y)} = \sum_j w^p_j A^j_{(x,y)} \tag{3.2}$$

where $M^p_{(x,y)}$ is labeled as the class activation map for the category p, w^p_j is the weight of the j^{th} feature map, and $A^j_{(x,y)}$ is the j^{th} convolved feature map at position (x,y).

To overcome CAM's limitations, Grad-CAM adapts the approach to be applicable across different CNN architectures. Initially, it computes the output score for each category as follows:

$$S_p = \sum_{i=1}^{H} \sum_{j=1}^{W} w^p_j A^j_{(i,j)} \tag{3.3}$$

where S_p is the score for category p, H and W are the height and width of the feature map, w^p_j is the weight of the j^{th} feature map for category p, and A_j is the j^{th} feature map.

The gradient between the output score S_p and the feature map A_j is computed to acquire the category-specific positioning map. The gradients are then used to weigh the importance of each feature map, allowing for the generation of a class-discriminative localization map that highlights the regions most relevant to the predicted class.

$$\delta^p_j = \frac{\partial S_p}{\partial A_j} \tag{3.4}$$

where δ^j_p represents the weight of j^{th} feature map, and G_p is the normalized heat map for category p.

The advantage of Grad-CAM is its reliance on output results and feature maps, making it independent of the specific CNN structure. This flexibility allows Grad-CAM to be easily applied across different CNN architectures, enhancing its utility in various deep learning applications. Due to this we used Grad-Cam with the CNN model, for the purpose of explainability.

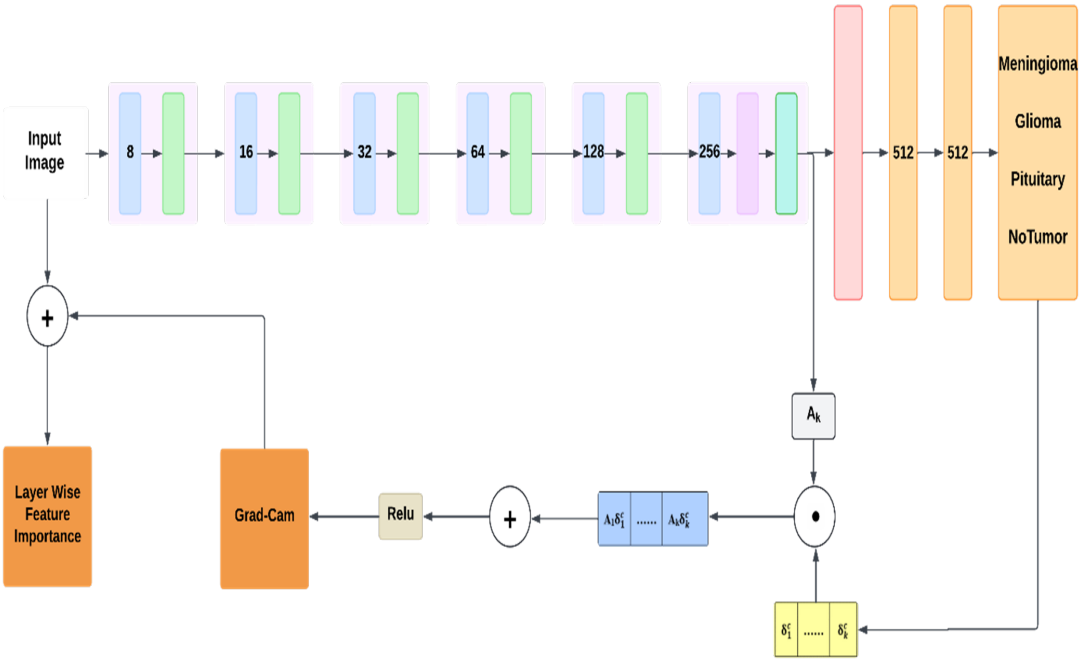


FIGURE 3.11: Proposed XAI Based CNN architecture

The Proposed XAI based CNN architecture is displayed in above Figure 3.11. In the forward pass, the images were passed through the CNN architecture to extract various features. After the forward pass, the network provided an output, which is the prediction of the tumor type. Once the prediction is made, a backward pass is performed to calculate the gradients of the output (specifically, the score for the predicted class) with respect to the feature maps of a chosen convolutional layer. The gradients were averaged over the width and height dimensions of the feature maps. The feature maps were then multiplied by their corresponding importance weights.

The weighted feature maps were summed to produce the Gradient activation mapping, highlighting the regions of the image that were most important for the classification. ReLU activation function was applied on this output to retain the features that positively influence the model's decision. This output was then combined with the input image to get the layer wise feature importance. These steps were applied across all layers to assess the feature importance at each layer. Through this process it is determined that which layers contribute most significantly to the model's decision.

3.4.2 Modification of Models Architecture

Based on the analysis achieved from Grad-Cam the CNN models' architecture was modified. Mean of all the layer's importance was calculated and saved. To remove the layers that are not contributing much in model's performance, a threshold was set and layers below that threshold were removed from the model. This threshold was the least mean value among all the layer's mean values. After removing those layers the model underwent the training phase again.

3.5 Interpretion of the Model

We employed these Explainable Artificial Intelligence (XAI) methods to interpret the model's decision-making processes and make them understandable and transparent.

3.5.1 SHapley Additive exPlanations (SHAP)

SHAP (Shapley Additive exPlanations) is a model-agnostic explainable artificial intelligence (XAI) technique. It assigns a value to each feature of an input instance for a given prediction, indicating its individual contribution to the model's output. This approach is grounded in Shapley values from game theory, which are employed to equitably distribute the total gains among participants in cooperative games [47].

SHAP works by estimating the contribution of each feature to the predicted output by simulating the absence of that feature. This is done by creating multiple versions of the input instance, each with a different feature missing or replaced with a reference value. The model is then evaluated on each of these modified instances, and the difference in the predicted output is used to estimate the contribution of the missing feature.

SHAP is used for both tabular data and image data. Explaining images with SHAP involves applying the same principles used in tabular data but adapted to

the peculiarities of image data. The core idea is still based on Shapley values, which allocate the contribution of each pixel (or group of pixels) to the final prediction made by an image classification model.

3.5.1.1 Algorithm for Shapley Values

1. Output: Shapley value corresponding to the j-th feature
2. Input: Epochs N, instance a, index of feature j, data matrix D, and the model f
   (a) From n = 1 to N:
      i. select an instance z from D matrix
      ii. Randomly permute and choose the feature values o
      iii. Arrange instance a: $a_o = \left(a_{(1)}, \ldots, a_{(j)}, \ldots, a_{(p)}\right)$
      iv. Arrange instance z: $z_o = \left(z_{(1)}, \ldots, z_{(j)}, \ldots, z_{(p)}\right)$
      v. Construct two new instances
         A. Including j: $a_{+j} = \left(a_{(1)}, \ldots, a_{(j-1)}, a_{(j)}, z_{(j+1)}, \ldots, z_{(p)}\right)$
         B. Without j: $a_{-j} = \left(a_{(1)}, \ldots, a_{(j-1)}, z_{(j)}, z_{(j+1)}, \ldots, z_{(p)}\right)$
      vi. Calculate marginal contribution: $\phi_j^m = \hat{f}(a_{+j}) - \hat{f}(a_{-j})$
3. Compute Shapley value as the average: $\phi_j(a) = \frac{1}{M} \sum_{m=1}^{M} \phi_j^m$

3.5.2 Lcoal Interpretable Model-agnostic Explanations

LIME (Local Interpretable Model-agnostic Explanations) explains the predictions of complex models by approximating the model locally around a specific instance with an interpretable model. For images, LIME generates explanations by perturbing the image, observing the changes in the model's predictions, and fitting a simple model to these observations [48].

LIME works by first segmenting the images into superpixels. Superpixels are contiguous regions of pixels that share similar characteristics. This reduces the

complexity of the image and allows LIME to handle a smaller number of features (superpixels) instead of individual pixels. After that LIME generates several perturbed versions of the original image. This is done by randomly turning superpixels on and off. Turning a superpixel off typically means replacing it with a baseline value, such as the average color or black pixels. Each perturbed image is thus a simplified version of the original image with certain superpixels removed or modified.

The original model is then used to predict the class probabilities for each perturbed image. This generates a set of prediction scores corresponding to the perturbed images. For each perturbed image, a binary vector is created indicating which superpixels are present (turned on) or absent (turned off). This matrix, along with the corresponding predictions, serves as the input for the local surrogate model.

LIME assigns weights to the perturbed images based on their similarity to the original image. A common approach is to use a kernel function to compute these weights, giving higher weights to perturbed images that are more similar to the original image. The proximity measure can be represented as:

$$w_i = \exp\left(-\frac{D(x, z_i)^2}{\sigma^2}\right) \tag{3.5}$$

where $D(x, z_i)$ is the distance between the original image x and the perturbed image z_i, and σ is a scaling parameter.

An interpretable model, such as a linear model or decision tree, is trained using the binary perturbation matrix and the weighted predictions. This model approximates the original complex model's behavior in the vicinity of the original image. The coefficients of the surrogate model explain.

In the case of a linear model, these coefficients indicate the importance of each superpixel in the model's prediction for the original image. Positive coefficients indicate that the presence of a superpixel increases the likelihood of the predicted class, while negative coefficients indicate the opposite.

3.6 Evaluation

Classifying brain tumor subtypes involves a systematic process where models are trained on a dataset, refined using a validation set to optimize hyperparameters, and finally evaluated on an independent test dataset to assess overall effectiveness.

3.6.1 Average Accuracy

Average accuracy, within the realm classification tasks, serves as a metric for assessing the comprehensive performance of a model across various classes or categories, especially in multi-class classification contexts.

The calculation of average accuracy involves a structured process as shown in equation 3.6. Initially for each j from 1 to N Accuracy j is calculated then accuracies for all N classes are added and divided by N to get the average accuracy, offering a holistic evaluation of the model's effectiveness across all classes. Where j is representing each class and N is the number of total classes.

$$\text{Average Accuracy} = \frac{1}{N} \sum_{j=1}^{N} \frac{\text{TP}_j + \text{TN}_j}{\text{TP}_j + \text{FP}_j + \text{TN}_j + \text{FN}_j} \tag{3.6}$$

In essence, average accuracy provides a unified perspective on how effectively a model distinguishes between different classes, thus proving to be a critical metric for evaluating the overall performance of a classification model in multi-class scenarios.

3.6.2 Average F1-score

The average F-score, commonly referred to as F1-score or simply F-score, is a statistical metric utilized for the precision and recall of a classification model. Calculated as the harmonic mean of precision and recall, it strikes a balance between these two metrics and provides a comprehensive assessment of model accuracy as shown in equation 3.7. The F1 score is particularly valuable in assessing overall

model performance, especially in scenarios where a single accuracy score may not fully capture effectiveness.

$$\text{Average F1-score} = \frac{1}{N}\sum_{j=1}^{N} F1\_score_j \tag{3.7}$$

$$F1\_score = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.8}$$

This metric serves as a critical tool in the thorough evaluation of classification models, offering insights into their ability to achieve a balanced trade-off between precision and recall. By considering both false positives and false negatives, the F1-score provides a nuanced perspective on the model's holistic performance. This makes it particularly relevant in fields where the consequences of misclassifications vary, emphasizing the need for a comprehensive evaluation approach.

3.6.3 Average Precision

Average precision in multi-class classification evaluates how accurately a model identifies instances belonging to each class. It extends the concept of precision from binary classification to multiple classes by calculating precision individually for each class. For each class j, precision is computed as the ratio of true positive predictions to the total instances predicted as belonging to that class (true positives plus false positives). The average precision across all classes is then calculated by taking the mean of these precision values.

$$\text{Average Precision} = \frac{1}{N}\sum_{j=1}^{N} \text{Precision}_j \tag{3.9}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3.10}$$

Higher average precision indicates that the model effectively minimizes false positives while correctly identifying instances of each class.

3.6.4 Average Recall

Average recall in multi-class classification evaluates how well a model identifies all instances belonging to each class across the entire dataset. Unlike precision, which focuses on the accuracy of positive predictions, recall measures the model's completeness in capturing all positive instances. For each class j, recall is calculated as the ratio of true positive predictions to the total number of actual instances of that class (true positives plus false negatives). The average recall for multi-class classification is then determined by averaging these recall values across all classes.

$$\text{Average Recall} = \frac{1}{N} \sum_{j=1}^{N} \text{Recall}_j \tag{3.11}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.12}$$

This metric is essential for assessing the model's ability to detect instances of each class, ensuring that no positive instances are overlooked. Higher average recall indicates that the model effectively minimizes false negatives while correctly identifying most or all instances of each class. In practical applications such as medical diagnostics or document categorization, average recall provides insights into the model's comprehensive performance across diverse categories.

Chapter 4

Implementation and Assesment of the Proposed Methodology

A comprehensive explanation of the proposed methodology is provided in Chapter 3. This chapter is devoted to the experimentation carried out for the assessment of the proposed methodology by providing the results obtained from its application on the mentioned datasets.

4.1 Tools and Technologies

4.1.1 Kaggle Jupyter Notebook

Kaggle Jupyter Notebook is an online platform that lets you write and execute Python code. It's a favored tool because of its simplicity and the numerous features it offers for data science, such as a code editor, terminal, and debugger.

4.1.2 Python Programming Language

Python is a versatile and widely used programming language. A significant advantage of Python is its rich library ecosystem, providing powerful tools and

frameworks essential for machine learning projects. Prominent machine learning libraries such as TensorFlow and Keras are built on Python. TensorFlow, a widely adopted open-source framework, is known for its flexibility and scalability in developing and training deep learning models. In contrast, Keras functions as a high-level neural networks API, streamlining the process of constructing and training models.

4.1.3 TensorFlow and Keras Libraries

TensorFlow and Keras are standout libraries in the field of deep learning. TensorFlow acts as a foundational, low-level library, providing the tools necessary for building complex deep-learning models. It is praised for its versatility and robustness, offering a wide range of tools for implementing various machine learning and deep learning algorithms. In contrast, Keras functions as a high-level library that serves as an abstraction layer over TensorFlow, making the design and training of deep learning models more straightforward.

Keras is especially appreciated for its user-friendly interface, which caters to users with varying levels of deep learning expertise. Its high-level approach offers a more intuitive and efficient development experience, allowing practitioners to concentrate on model architecture and experimentation without needing to manage the complexities of low-level implementation.

4.1.4 XAI Libraries

SHAP and LIME are standout libraries in the field of model explainability. SHAP, rooted in Shapley values, provides a comprehensive, low-level approach to measuring global feature importance by fairly distributing contributions among features. It is praised for its consistency and robustness, offering a reliable way to understand various machine-learning models. In contrast, LIME functions as a high-level tool that simplifies the explanation of individual predictions by approximating the model locally with simpler, interpretable models.

LIME's focus on local interpretability offers a more intuitive and efficient way to explain specific predictions of the model, allowing practitioners to be focused on understanding model's behavior without dealing with the complexities of global feature attribution.

4.1.5 Kaggle GPU

Kaggle GPU is an essential service that provides access to Graphics Processing Units (GPUs) specifically designed for machine learning (ML) and deep kearning (DL) experiments. GPUs are known for their superior speed compared to Central Processing Units (CPUs), making them especially beneficial for accelerating the training of deep learning models. This service is crucial for data scientists and researchers who need efficient, high-performance computing resources.

4.2 Dataset

In this study, two datasets are used. The first is the Msoud dataset, a significant and publicly available collection of brain MRIs labeled for detection and classification tasks. This dataset includes 7,023 images, each carefully reviewed and labeled by expert radiologists, as shown in below Figure 4.1. The labels indicate the presence or absence of three distinct types of brain tumors and a normal brain type. The large size of the dataset and the expertise behind the labeling makes it a valuable resource for developing and evaluating ML and DL models for detecting and classifying brain tumors in medical imaging. The types of brain tumors included are:

1. Glioma
2. Meningioma
3. No-tumor
4. Pituitary

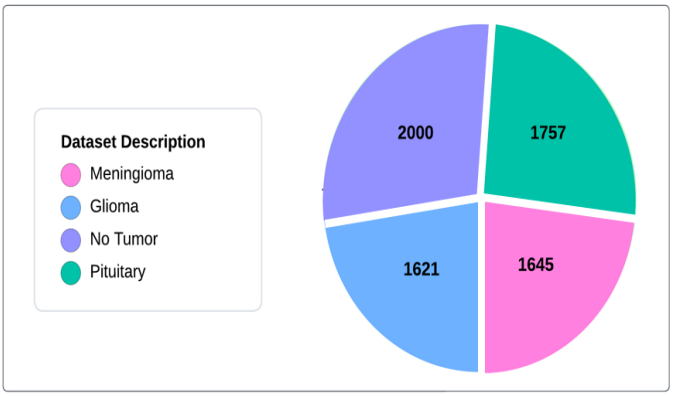


FIGURE 4.1: Dataset 1

The second dataset is a significant and openly available repository of brain MRIs with labels for detection and classification. This dataset encompasses 3264 images as shown in figure 4.2. The labels specifically focus on identifying the presence or absence of 3 distinct types of brain tumors and one Normal brain type. The types of brain tumors are as follows:

1. Glioma
2. Meningioma
3. No-tumor
4. Pituitary

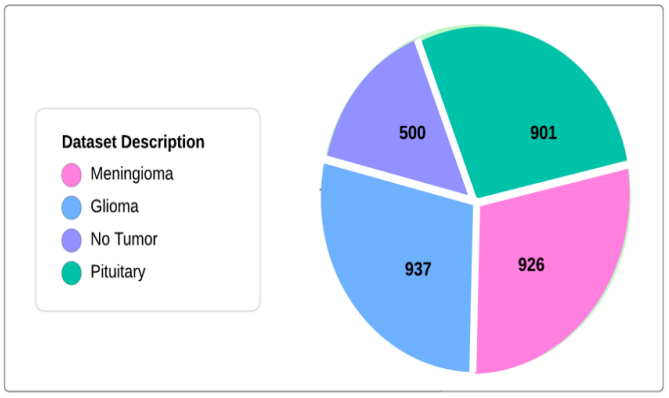


FIGURE 4.2: Dataset 2

Both datasets lack patient-specific clinical or medical details, making them ethically appropriate for use in this study. The absence of confidential or private

information ensures compliance with ethical guidelines, providing a secure foundation for analysis and research. The dataset 1 training set, validation set, and testing set consist of 5618, 702, and 702 images with the ratio of 80%,10%, and 10%, and dataset 2 was only used to test the model.

The images in both datasets initially varied in size and were not consistently focused on the brain section, presenting challenges for effective analysis. To address these issues, both datasets underwent a thorough preprocessing procedure. This involved cropping the images to ensure that only the brain section was retained, thereby removing any extraneous parts of the images. Following cropping, the images were normalized to standardize the pixel intensity values, which helps improve the performance of machine learning algorithms by ensuring uniformity across the dataset. Finally, the images were resized to the dimensions of 224x224 pixels. This resizing step not only standardizes the input size for models but also helps in optimizing computational efficiency and ensuring compatibility with widely used neural network architectures that often require specific input dimensions.

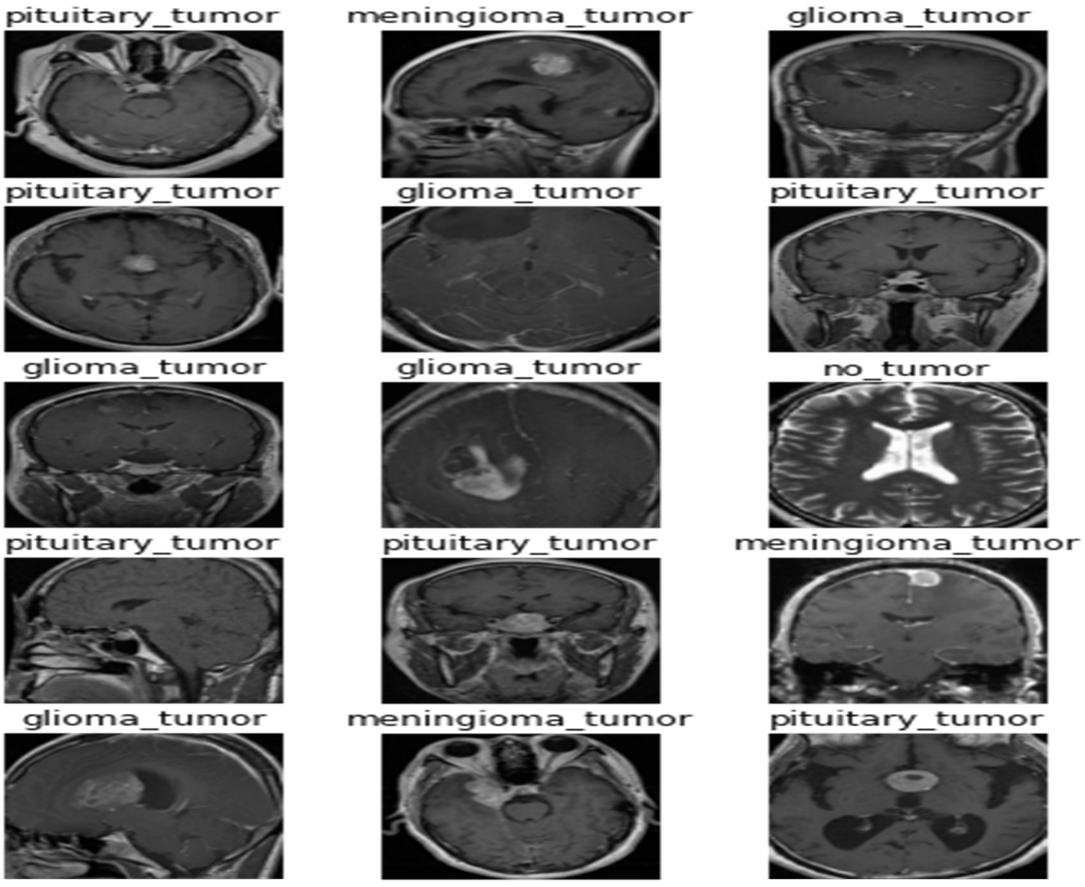


FIGURE 4.3: Preprocessed MRIs

4.3 Implementation of Model

The XAI-based CNN model was implemented in TensorFlow and Keras libraries. The model consists of a CNN and a Grad Cam. The CNN is a series of convolutional layers that extract features from the input image. The Grad Cam is an XAI method that achieves the feature and layer importance of the trained CNN model. Together, the combination of CNN and Grad-CAM ensures not only high performance in image classification tasks but also provides an understanding of the model's internal workings, enhancing trust and reliability in the model's predictions.

The model starts with a convolutional layer that processes 224x224x3 RGB images (MRIs) using 8 filters with kernel 3x3 and then passes it through a ReLU activation function. Next, a max pooling layer follows this layer that reduces the feature map size by half. This pattern of convolutional layers and max pooling layers repeats, with the filter count increasing exponentially (16, 32, 64, 128, 256) while maintaining a 3x3 kernel size and 'same' padding. After these layers, a batch normalization layer normalizes the activations, followed by an average pooling layer. The output is flattened into a 1D feature vector and passed through two dense layers with 512 neurons each and ReLU activation. The model concludes with a dense layer that outputs a probability distribution over 4 classes using softmax activation.

Then it follows the Grad Cam method, in which the gradients of the target class score concerning the feature maps of each layer were calculated. Then average of these gradients to obtain weights were performed that reflect the importance of each feature map for the target class score. Lastly, the Grad Cam map by performing a weighted sum of the feature maps is generated, followed by ReLU activation to highlight important features of each layer.

By comparing Grad-CAM maps from different layers, the most influential layers and feature maps were determined. A threshold was set to remove those layers that were not capturing important features and were below the threshold, through this the model was refined and its performance was improved.

TABLE 4.1: Model Parameter

Parameter	**Value**
Convolution Layer	5
Convolution Layer Filter	8,16, 64,128, 256
Convolution Layer Kernel Size	3
Convolution Layer Activation Function	Relu
Pooling Layer	5
Pooling Method	Max Pooling, Average Pooling
Pooling Size	2
Optimizer	Adam
Loss Function	Sparse Categorical Cross entropy
Batch Size	40
Epochs	40

4.4 Evaluating Model Performance

To evaluate the performance of our proposed XAI-based model against the initial model without XAI, we utilized the XAI Grad-CAM method. Grad-CAM allowed us to visualize the critical regions of the brain MRIs that the model focused on when making predictions. The Grad-Cam results of the Meningioma, Glioma, and Pituitary class are shown in below Figure 4.4, Figure 4.5, and Figure 4.6 respectively.

By comparing these visualizations, we observed that our proposed XAI-based model significantly improved its ability to accurately highlight and focus on abnormal brain tissues. This enhanced focus led to the accurate and explainable classification of brain tumor subtypes compared to the initial model without XAI. The results demonstrate the effectiveness of incorporating explainable AI techniques,not only to improve the accuracy of the model but also in providing valuable insights into its decision-making process, thereby increasing trust and reliability in its predictions.

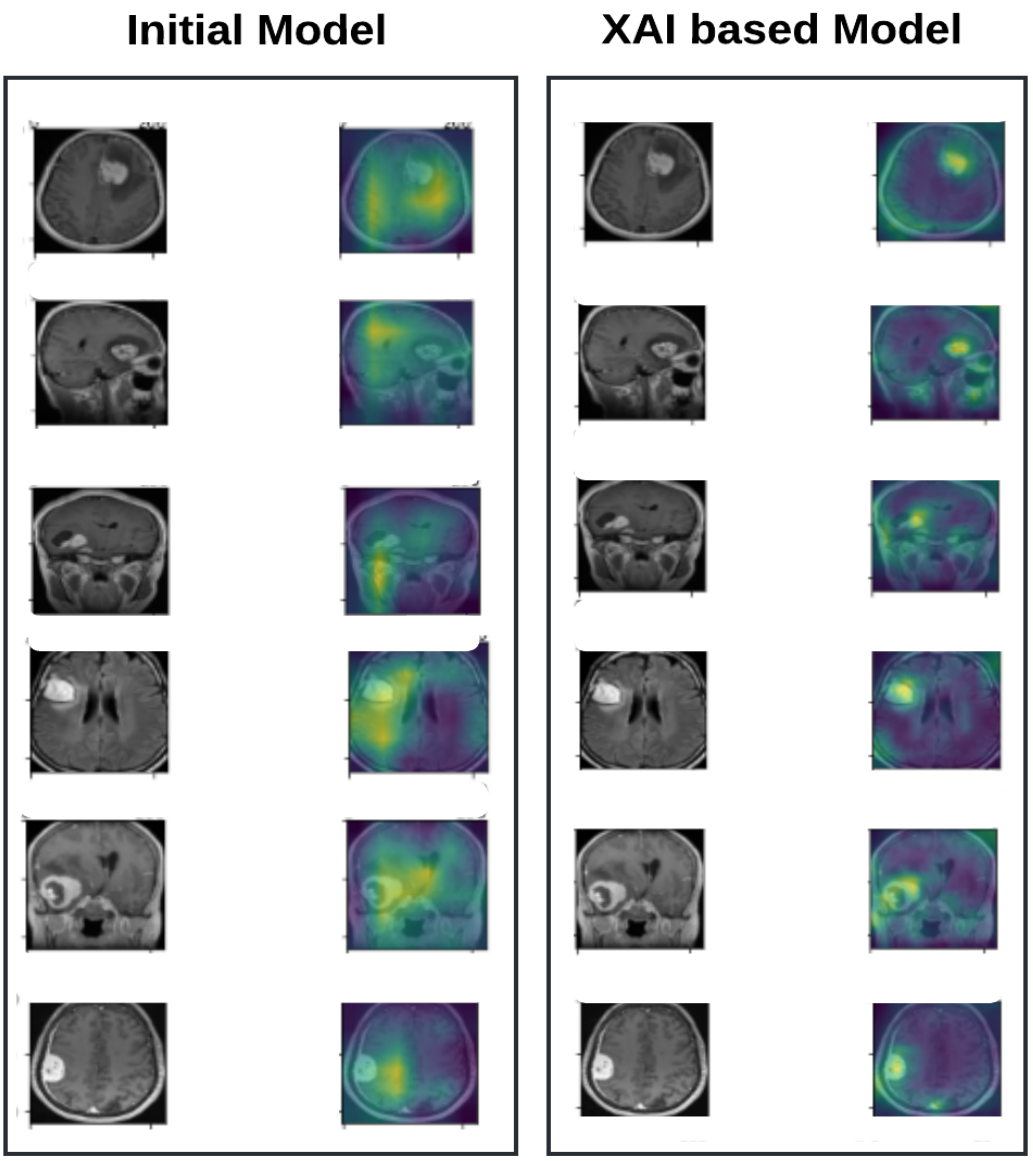


FIGURE 4.4: Meningioma Class Results of Initial and XAI Based Model

The XAI method Grad-CAM results provide compelling evidence that our proposed XAI-based model relies on crucial features related to abnormal brain tissues or tumorous features. This capability enhances the model's confidence in its MRIs classification decisions by emphasizing the regions of brain MRI scans that are most indicative of specific tumor subtypes. By highlighting these critical features, the model not only improves its accuracy but also strengthens its interpretability. This insight into where the model focuses its attention underscores the effectiveness of our proposed approach in leveraging explainable AI techniques to enhance both the performance and transparency of medical image classification tasks.

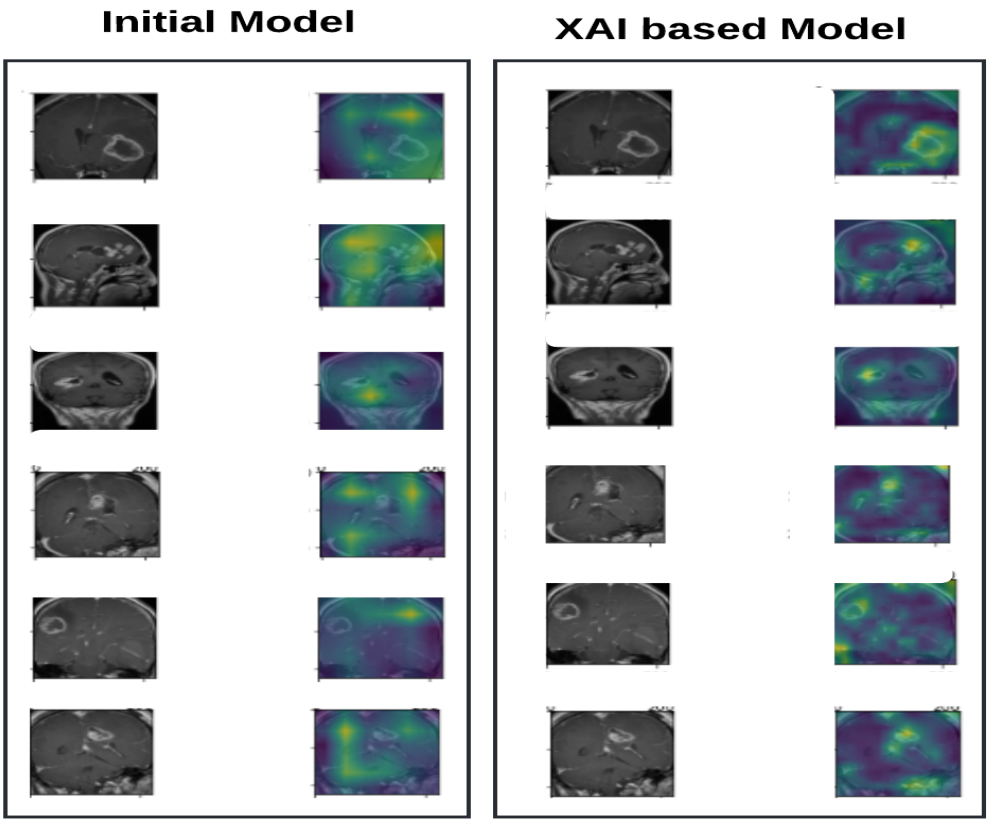


FIGURE 4.5: Glioma Class Results of Initial and XAI Based Model

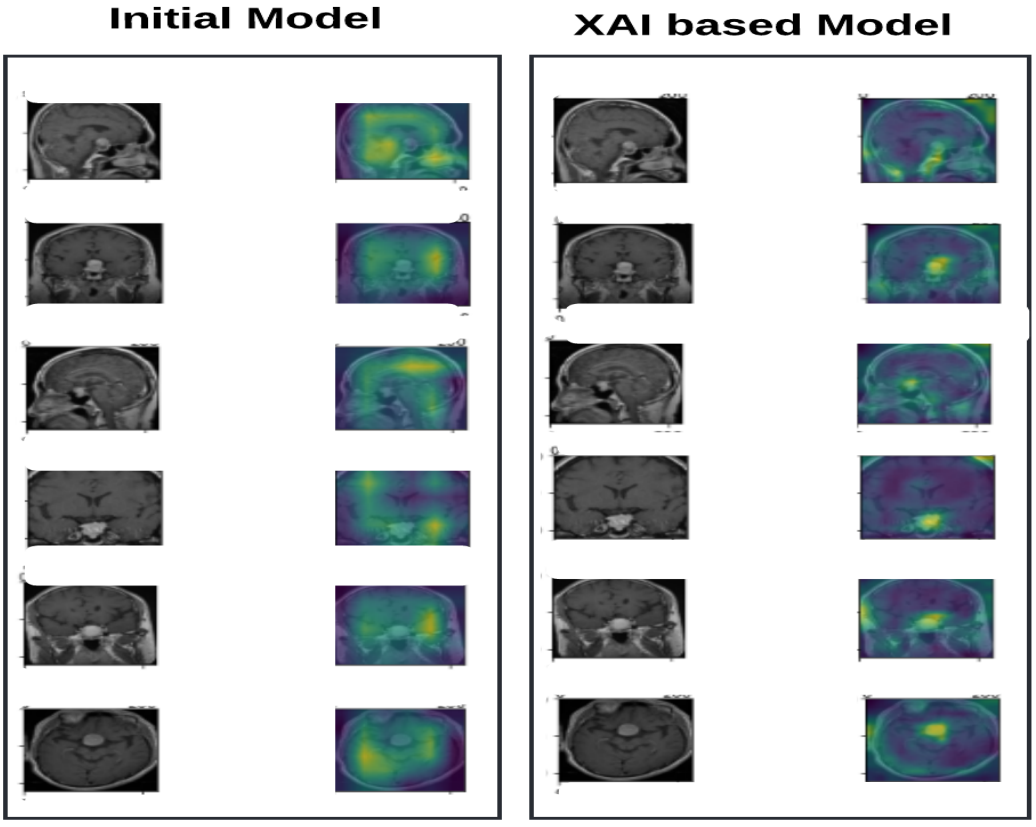


FIGURE 4.6: Pituitary Class Results of Initial and XAI Based Model

4.5 Explaining Model

The interpretability of our proposed model is achieved through the implementation of XAI methods such as SHAP and LIME. These methods enable us to analyze how the model makes decisions and identify the features it considers most influential in classifying brain MRI scans. By applying SHAP and LIME, we observed that our model effectively detects abnormal brain tissues within the MRIs, providing clear insights into its decision-making process. The results obtained from SHAP and LIME are illustrated in the figures below, demonstrating the specific features and regions of the brain scans that significantly influence the model's predictions.

4.5.1 Results Achieved by Shap

In SHAP, the visual representation use colors to denote the level of contribution of features to the model's predictions. Features highlighted in red indicate significant positive contributions to the output prediction. These features are crucial in influencing the model's decision-making process and are considered highly relevant in outcome, such as identifying abnormal brain tissues in MRIs. On the other hand, features highlighted in blue suggest minimal contributions to the prediction. These features are less influential and do not heavily impact the model's decision.

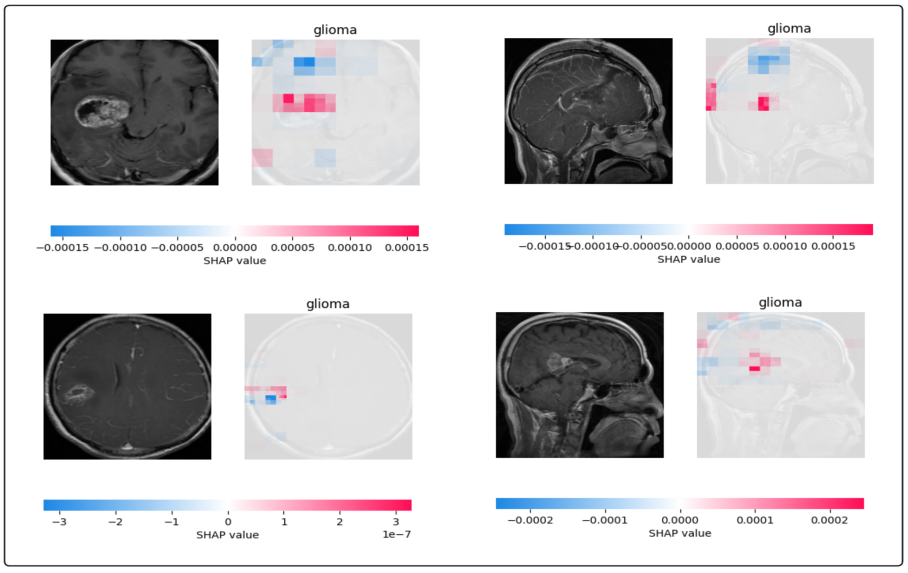


FIGURE 4.7: Glioma Class Shap Explaination

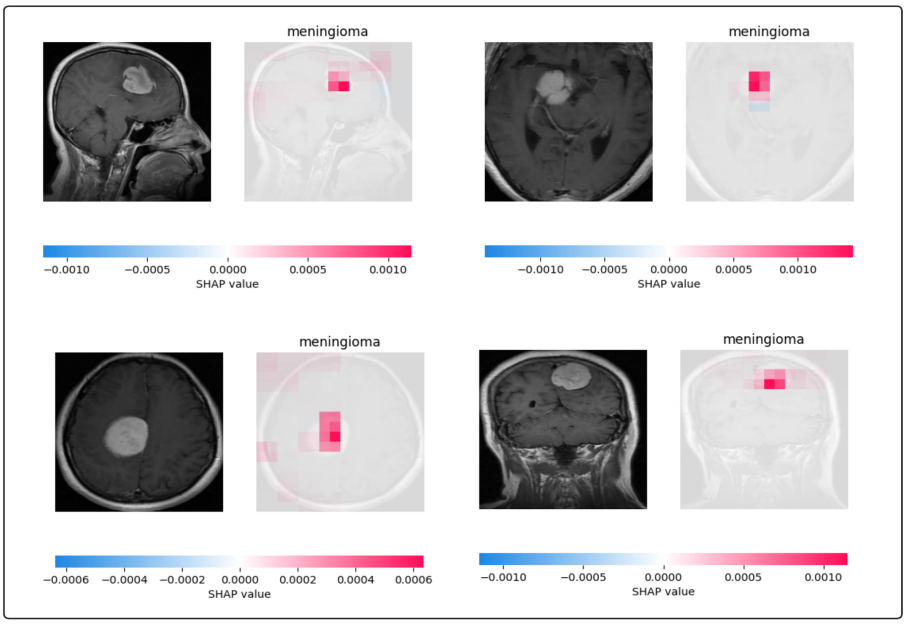


FIGURE 4.8: Meningioma Class Shap Explaination

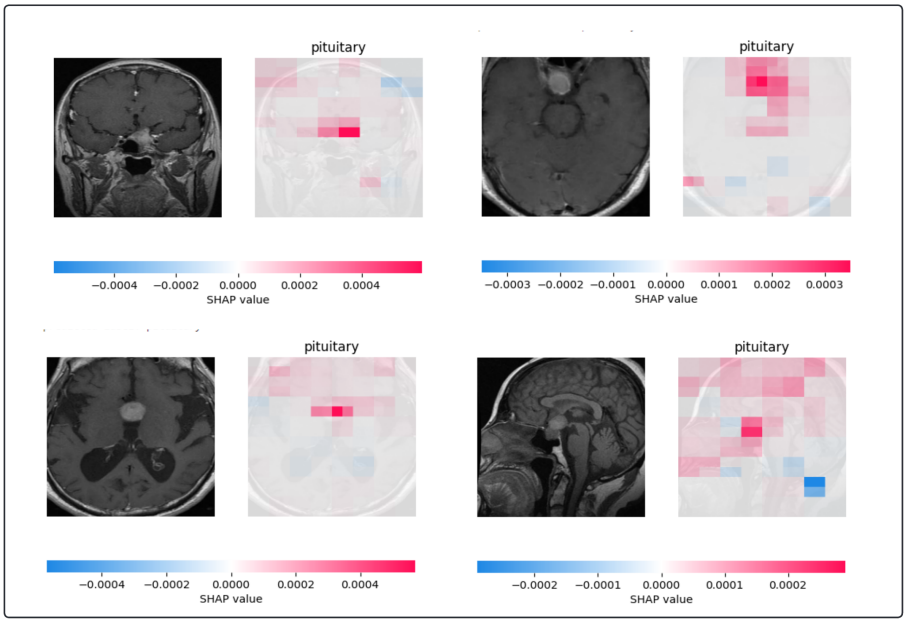


FIGURE 4.9: Pituitary Class Shap Explaination

4.5.2 Results Achieved by Lime

In LIME (Local Interpretable Model-agnostic Explanations), the visual representation employs a color gradient to signify the level of importance of features for a

specific prediction or instance. Features highlighted in green represent areas where the model's prediction is positively influenced. These features are deemed significant contributors to the model's decision-making process, exerting a substantial influence on the outcome, such as identifying abnormal brain tissues in MRI scans. Conversely, features highlighted in red indicate areas where the model's prediction is negatively influenced or where the feature is not influential. These features have minimal impact on the model's decision for the specific instance being analyzed.

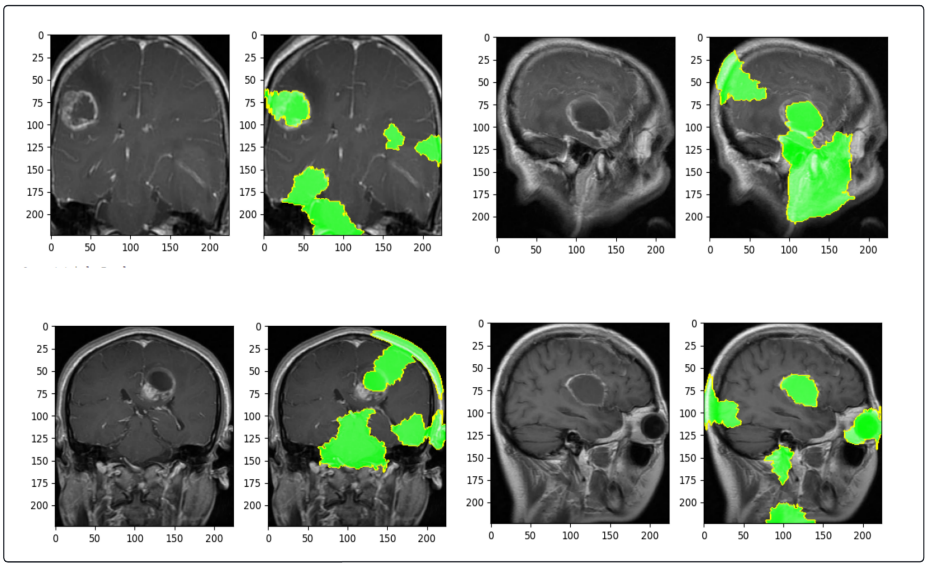


FIGURE 4.10: Glioma Class Lime Explaination

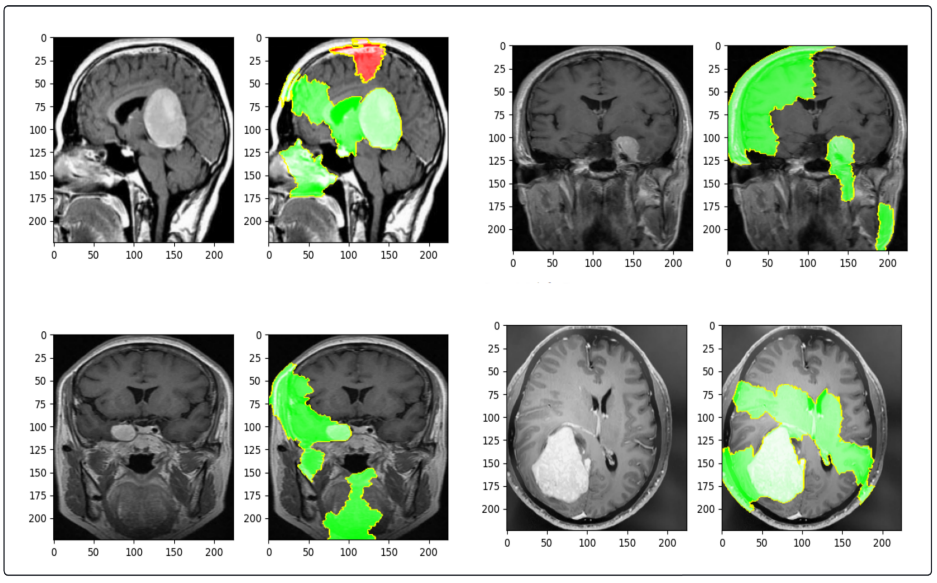


FIGURE 4.11: Meningioma Class Lime Explaination

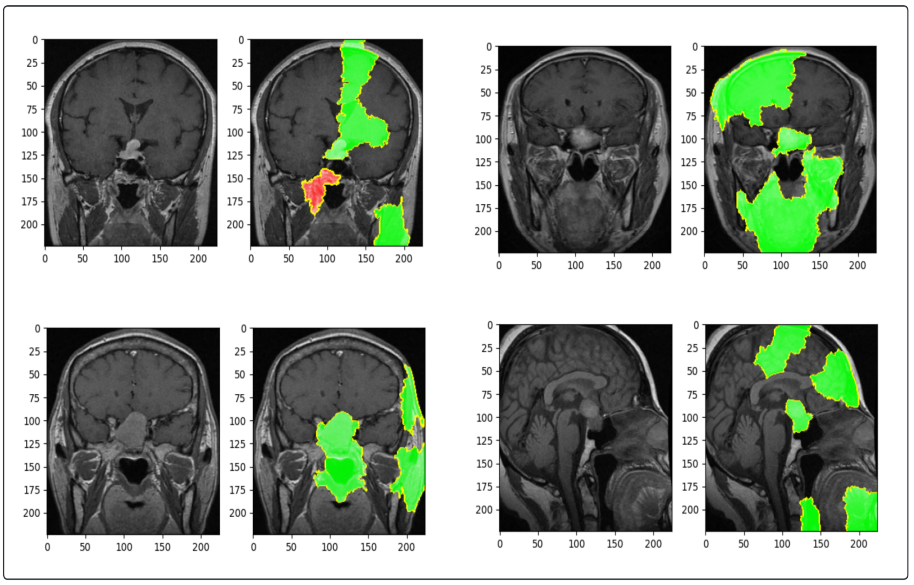


FIGURE 4.12: Pituitary Class Lime Explaination

4.6 Analysis of Classification Result

To evaluate the classification performance of our model, we employed four key metrics: Average Accuracy, Average F1 score, Average Precision, and Average Recall. These metrics offer a thorough insight into the model's ability to distinguish between different classes.

On Dataset 1, our model demonstrated exceptional performance, achieving an impressive average Accuracy of 99.21% and an average F1-Score of 99.20%. These results highlight the model's high level of precision and recall, which are critical for accurately classifying brain tumor subtypes. This indicates that the model is effectively learning the distinguishing features of each subtype, leading to enhanced performance in correctly identifying each category.

TABLE 4.2: Comparison of Proposed Model on Dataset 1

	Accuracy	F1-Score	Precision	Recall
Initial Model	98.95%	98.93%	98.96%	98.90%
Proposed Model	99.21%	99.20%	99.22%	99.18%

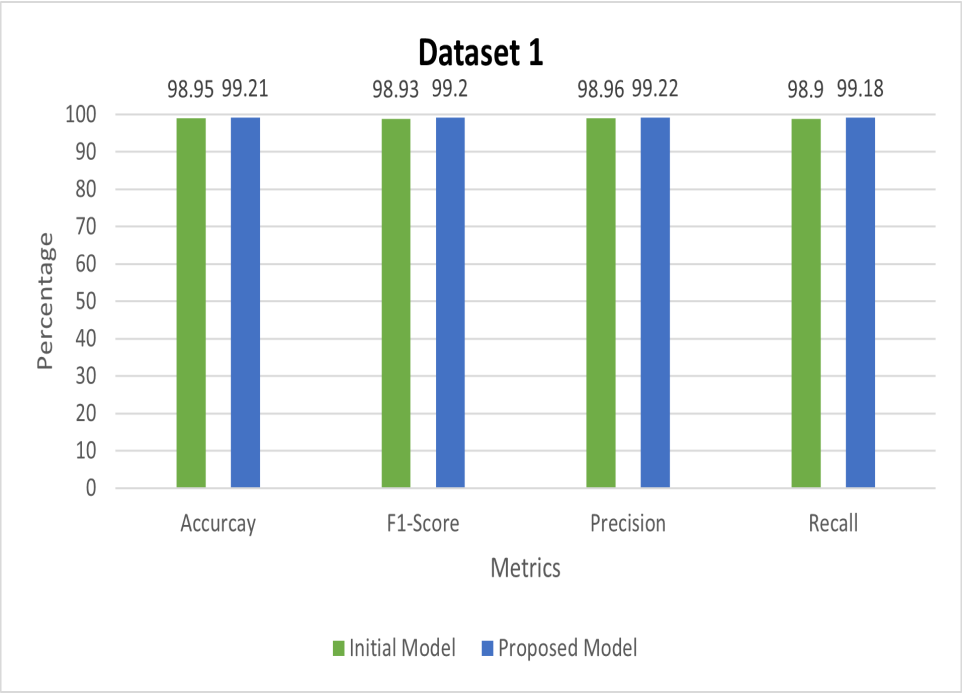


FIGURE 4.13: Initial and Proposed Model on Evaluation Metrics for Dataset 1

To assess the model's generalization ability, we evaluated its performance on a new dataset that the model had not encountered during training. This evaluation ensures that the model can effectively apply its learned knowledge to unseen data, which is crucial for real-world applications. For this purpose, we utilized the same comprehensive metrics: Average Accuracy, Average F1-Score, Average Precision, and Average Recall. Our proposed model achieved an impressive Average Accuracy of 94.72% and an Average F1-Score of 94.63% on the new dataset.

These results indicate that the model has successfully learned the distinctive features that are critical for accurate brain tumor classification. The ability to generalize well to unseen data showcases the robustness and reliability of the model, making it a valuable tool for practical medical applications where new and diverse data are frequently encountered.

TABLE 4.3: Comparison of Proposed Model on Dataset 2

	Accuracy	F1-Score	Precision	Recall
Initial Model	93.33%	93.24%	93.35%	93.83%
Proposed Model	94.72%	94.63%	94.70%	95.10%

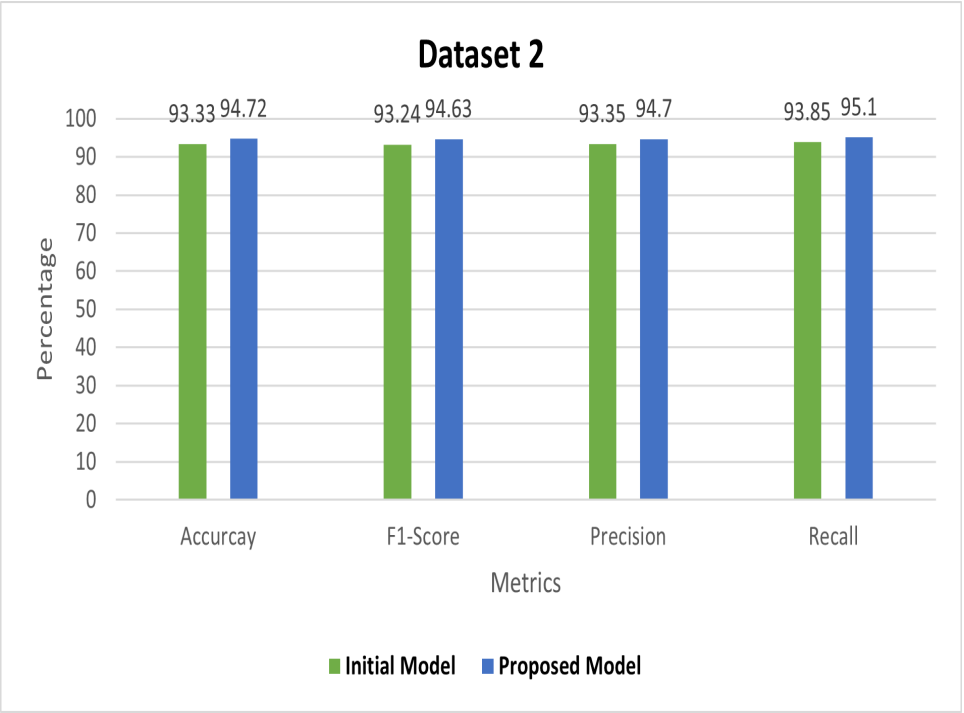


FIGURE 4.14: Initial and Proposed Model on Evaluation Metrics for Dataset 2

To visualize the performance of the model on brain tumor classification we created learning curves for both the initial and proposed models shown below in Figure 4.15. These curves reveal how the models' accuracy evolves over successive epochs.

The validation curve is particularly effective for depicting the learning behavior of the models. It provides insight into how the model's accuracy changes with different hyperparameters, helping us understand its performance and learning dynamics. The validation accuracy curve highlights classification performance, helping practitioners make informed decisions on model training and hyperparameter tuning, to prevent overfitting or enhance generalization.

The Initial Model, starts with a lower accuracy of around 0.5 at 5 epochs, increases significantly until around 15 epochs, reaching approximately 0.85, and then plateaus, showing minimal improvement and stabilizing around 0.9879. In contrast, the Proposed Model, starts at a higher initial accuracy of about 0.65, increases sharply until around 10 epochs, and then gradually improves, surpassing the Initial Model's accuracy at earlier epochs, and stabilizing around 0.9947. This suggests that the Proposed Model consistently outperforms the Initial Model across all epochs.

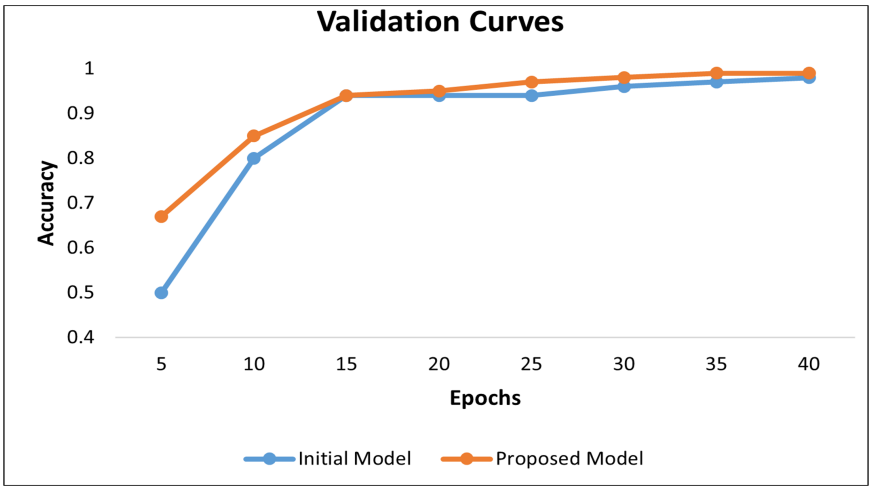


FIGURE 4.15: Validation Learning Curves of Initial and Proposed Model

4.7 Discussion of Results

The explainability results of the proposed model achieved using Grad-Cam, SHAP and LIME, provided insights into the specific features that influenced the model's predictions. By comparing the Grad-CAM visualizations of the initial and proposed models, we observed that our XAI-based model significantly improved its ability to accurately focus on abnormal brain tissues compared to the initial model. SHAP's visual representations showed that features with significant contributions to the predictions of the model were primarily associated with abnormal brain tissues. LIME's localized interpretations confirmed that the model's decisions were influenced by important features, leading to accurate and explainable predictions.

In terms of classification performance, the proposed XAI-based model significantly outperformed the initial model, achieving an accuracy of 99.47% on dataset 1. The generalization ability of the proposed model was also tested on an unseen dataset, where it maintained a strong performance with an accuracy of 94.72%.

The use of XAI techniques especially Grad-CAM, significantly improved the model's ability to focus on and highlight abnormal brain tissues by guiding it to prioritize the most relevant features during training. This refinement was only due to the targeted approach provided by XAI, which helped the model better identify and emphasize critical regions in the input images.

Alongside Grad-CAM, the XAI methods SHAP and LIME were crucial in validating the model's decisions and enhancing its interpretability. SHAP provided a global view of feature importance, confirming that the model focused on relevant abnormal brain tissues, while LIME offered localized interpretations, ensuring consistency and explainability in the model's predictions. This combination of global and local insights ensured the model was accurate, interpretable, and was not relying on irrelevant features or overfitting the data. This approach not only enhanced the model's performance but also made it a valuable tool in the clinical diagnosis of brain tumors, where understanding the reasoning behind a prediction is as important as the prediction itself.

4.8 Comparison with Existing Approaches

To further validate our results, we compared the performance of the proposed XAI based CNN model with different related studies of brain tumor classification. This comparison is displayed in below Table 4.4. The proposed XAI-based CNN demonstrates superior performance across all metrics compared to the other methods. With an accuracy of 99.21%, precision of 99.22%, recall of 99.18%, and F1-score of 99.20%, it provides more accurate and reliable classification of brain tumors, highlighting the effectiveness of integrating XAI techniques in CNN architectures.

Table 4.4: Comparison of Existing Techniques with Proposed Method

Ref	Method	MRIs	Accuracy	Precision	Recall	F1-Score
[28]	Image Enhancement + Modified CNN	7023	97.84%	97.85%	97.85%	97.90%
[29]	Modified CNN	7023	96%	96%	96%	96.50%
[30]	InceptionV3	7023	97.12%	97.97%	96.59%	97%
Our Work	**XAI-CNN**	**7023**	**99.21%**	**99.22%**	**99.18%**	**99.20%**

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In conclusion, our XAI-based model for brain tumor classification demonstrates significant promise in accurately identifying and classifying different subtypes of brain tumor MRIs by leveraging the power of XAI to focus on critical features in the images.
Key findings and contributions of our model include:

1. Identifying real contributing abnormal tissues from brain MRIs to improve the overall performance of brain tumor classification.

2. Achieving an interpretable model that provides transparency into decision-making process making it superior to existing non-interpretable approaches.

3. Providing insights into the model decision-making process to reduce the number of layers in brain tumor classification models.

4. Achieving outstanding average accuracy, F1-score, precision, and recall of 99.21% on Dataset 1 and 94.72% on Dataset 2.

5. Outperforming benchmark techniques by providing transparency and classification based on crucial abnormal features with a reduced number of layers.

The incorporation of XAI in the CNN architecture signifies an important advancement, enabling the model to make more informed decisions about feature emphasis. This enhancement improves classification accuracy and produces highly transparent and reliable models. And also boosts user trust, aids in debugging, and aligns with regulatory requirements, making it suitable for integrating models into real-world applications where transparency and trust are crucial. Our dataset preprocessing strategies, Grad-CAM-based model, and use of XAI methods like SHAP and LIME for model interpretability collectively build a model that makes its classification decisions on actual abnormal brain tissues.

The efficiency of our methodology is exemplified by its ability to classify MRIs based on actual tumor tissues without the need for complex feature extraction techniques, making it applicable to a broader range of medical imaging datasets and scenarios. The XAI-based model's ability to seamlessly blend classification and explainability within a unified framework further enhances its suitability for diverse medical image analysis applications.

5.2 Future Work

In the future, there are several promising directions to explore:

1. Incorporate more diverse brain tumor MRI datasets to improve the model's ability to generalize to new, unseen data across different populations and tumor subtypes.

2. Develop and test real-time application capabilities for clinical settings, ensuring the model can provide instant, interpretable results.

3. Explore the integration of other imaging modalities, such as CT or PET scans, to create a more comprehensive diagnostic tool.

4. Investigate more advanced XAI techniques to further refine feature analysis and improve model interpretability.

Bibliography

[1] Pakistan Brain Tumour Consortium, S. Enam, M. Shah, M. Bajwa, M. Khalid, S. Bakhshi, E. Baig, I. Altaf, A. Laghari, S. bin Anis, N. Akhunzada, M. Raghib, J. Gilani, N. Jawed, and S. Siddiqi, "The pakistan brain tumour epidemiology study," *Journal of the Pakistan Medical Association*, vol. 72, no. 11, pp. S4–S11, 2022.

[2] Q. T. Ostrom, N. Patil, G. Cioffi, K. 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 2013–2017," *Neuro-oncology*, vol. 22, no. Supplement_1, pp. iv1–iv96, 2020.

[3] R. Asad, S. U. Rehman, A. Imran, J. Li, A. Almuhaimeed, and A. Alzahrani, "Computer-aided early melanoma brain-tumor detection using deep-learning approach," *Biomedicines*, vol. 11, no. 1, p. 184, 2023.

[4] M. K. Kumar, D. S. N. Sreeja, S. Sadiq, D. Manisha, A. Jain, and B. Madhu, "Automated brain tumour classification using deep learning technique," in *E3S Web of Conferences*, vol. 430. EDP Sciences, 2023, p. 01032.

[5] N. Ullah, A. Javed, A. Alhazmi, S. M. Hasnain, A. Tahir, and R. Ashraf, "Tumordetnet: A unified deep learning model for brain tumor detection and classification," *Plos one*, vol. 18, no. 9, p. e0291200, 2023.

[6] Z. Rasheed, Y.-K. Ma, I. Ullah, Y. Y. Ghadi, M. Z. Khan, M. A. Khan, A. Abdusalomov, F. Alqahtani, and A. M. Shehata, "Brain tumor classification from mri using image enhancement and convolutional neural network techniques," *Brain Sciences*, vol. 13, no. 9, p. 1320, 2023.

[7] Q. T. Ostrom, C. McCulloh, Y. Chen, K. Devine, Y. Wolinsky, P. Davitkov, S. Robbins, R. Cherukuri, A. Patel, R. Gupta *et al.*, "Family history of cancer in benign brain tumor subtypes versus gliomas," *Frontiers in oncology*, vol. 2, p. 19, 2012.

[8] N. Ullah, J. A. Khan, M. S. Khan, W. Khan, I. Hassan, M. Obayya, N. Negm, and A. S. Salama, "An effective approach to detect and identify brain tumors using transfer learning," *Applied Sciences*, vol. 12, no. 11, p. 5645, 2022.

[9] Z. Atha and J. Chaki, "Ssbtcnet: semi-supervised brain tumor classification network," *IEEE Access*, 2023.

[10] S. Saeedi, S. Rezayi, H. Keshavarz, and S. R. Niakan Kalhori, "Mri-based brain tumor detection using convolutional deep learning methods and chosen machine learning techniques," *BMC Medical Informatics and Decision Making*, vol. 23, no. 1, p. 16, 2023.

[11] B. V. Isunuri and J. Kakarla, "Ensemble coupled convolution network for three-class brain tumor grade classification," *Multimedia Tools and Applications*, pp. 1–17, 2023.

[12] M. W. Nadeem, M. A. A. 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.

[13] K. G. Khambhata and S. R. Panchal, "Multiclass classification of brain tumor in mr images," *Int. J. Innov. Res. Comput. Commun. Eng*, vol. 4, no. 5, pp. 8982–8992, 2016.

[14] E. I. Zacharaki, S. Wang, S. Chawla, D. Soo Yoo, R. Wolf, E. R. Melhem, and C. Davatzikos, "Classification of brain tumor type and grade using mri texture and shape in a machine learning scheme," *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 62, no. 6, pp. 1609–1618, 2009.

[15] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[16] L. Singh, G. Chetty, and D. Sharma, "A novel machine learning approach for detecting the brain abnormalities from mri structural images," in *Pattern Recognition in Bioinformatics: 7th IAPR International Conference, PRIB 2012, Tokyo, Japan, November 8-10, 2012. Proceedings 7*. Springer, 2012, pp. 94–105.

[17] Y. Yang, L.-F. Yan, X. Zhang, Y. Han, H.-Y. Nan, Y.-C. Hu, B. Hu, S.-L. Yan, J. Zhang, D.-L. Cheng *et al.*, "Glioma grading on conventional mr images: a deep learning study with transfer learning," *Frontiers in neuroscience*, vol. 12, p. 804, 2018.

[18] E. U. Haq, H. Jianjun, K. Li, H. U. Haq, and T. Zhang, "An mri-based deep learning approach for efficient classification of brain tumors," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–22, 2023.

[19] C. Srinivas, N. P. KS, M. Zakariah, Y. A. Alothaibi, K. Shaukat, B. Partibane, and H. Awal, "Deep transfer learning approaches in performance analysis of brain tumor classification using mri images," *Journal of Healthcare Engineering*, vol. 2022, no. 1, p. 3264367, 2022.

[20] J. S. Paul, A. J. Plassard, B. A. Landman, and D. Fabbri, "Deep learning for brain tumor classification," in *Medical Imaging 2017: Biomedical Applications in Molecular, Structural, and Functional Imaging*, vol. 10137. SPIE, 2017, pp. 253–268.

[21] X. Li, H. Xiong, X. Li, X. Wu, X. Zhang, J. Liu, J. Bian, and D. Dou, "Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond," *Knowledge and Information Systems*, vol. 64, no. 12, pp. 3197–3234, 2022.

[22] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[23] D. Radecic, "Lime: How to interpret machine learning models with python," *Explainable Machine Learning at Your Fingertips*, 2020.

[24] V. Buhrmester, D. Münch, and M. Arens, "Analysis of explainers of black box deep neural networks for computer vision: A survey," *Machine Learning and Knowledge Extraction*, vol. 3, no. 4, pp. 966–989, 2021.

[25] S. M. Alzahrani, "Convattenmixer: Brain tumor detection and type classification using convolutional mixer with external and self-attention mechanisms," *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 10, p. 101810, 2023.

[26] I. Gupta, S. Singh, S. Gupta, and R. Nayak, "Classification of brain tumours in mri images using a convolutional neural network." *Current Medical Imaging*, 2023.

[27] F. E. AlTahhan, G. A. Khouqeer, S. Saadi, A. Elgarayhi, and M. Sallah, "Refined automatic brain tumor classification using hybrid convolutional neural networks for mri scans," *Diagnostics*, vol. 13, no. 5, p. 864, 2023.

[28] Z. Rasheed, Y.-K. Ma, I. Ullah, Y. Y. Ghadi, M. Z. Khan, M. A. Khan, A. Abdusalomov, F. Alqahtani, and A. M. Shehata, "Brain tumor classification from mri using image enhancement and convolutional neural network techniques," *Brain Sciences*, vol. 13, no. 9, p. 1320, 2023.

[29] O. Özkaraca, O. İ. Bağrıaçık, H. Gürüler, F. Khan, J. Hussain, J. Khan, and U. e. Laila, "Multiple brain tumor classification with dense cnn architecture using brain mri images," *Life*, vol. 13, no. 2, p. 349, 2023.

[30] M. A. Gómez-Guzmán, L. Jiménez-Beristaín, E. E. García-Guerrero, O. R. López-Bonilla, U. J. Tamayo-Perez, J. J. Esqueda-Elizondo, K. Palomino-Vizcaino, and E. Inzunza-González, "Classifying brain tumors on magnetic resonance imaging by using convolutional neural networks," *Electronics*, vol. 12, no. 4, p. 955, 2023.

[31] M. M. Islam, P. Barua, M. Rahman, T. Ahammed, L. Akter, and J. Uddin, "Transfer learning architectures with fine-tuning for brain tumor classification

using magnetic resonance imaging," *Healthcare Analytics*, vol. 4, p. 100270, 2023.

[32] R. Imam and M. T. Alam, "Optimizing brain tumor classification: A comprehensive study on transfer learning and imbalance handling in deep learning models," in *International Workshop on Epistemic Uncertainty in Artificial Intelligence*. Springer, 2023, pp. 74–88.

[33] C.-C. Peng and B.-H. Liao, "Classify brain tumors from mri images: Deep learning-based approach," in *2023 IEEE 5th Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS)*. IEEE, 2023, pp. 5–8.

[34] S. Shanjida, M. S. Islam, and M. Mohiuddin, "Mri-image based brain tumor detection and classification using cnn-knn," in *2022 IEEE IAS global conference on emerging technologies (GlobConET)*. IEEE, 2022, pp. 900–905.

[35] K. V. Kumar, M. Baid, and K. Menon, "Brain tumor classification using transfer learning on augmented data and visual explanation using grad-cam," in *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE, 2023, pp. 965–971.

[36] H. Benyamina, A. S. Mubarak, and F. Al-Turjman, "Explainable convolutional neural network for brain tumor classification via mri images," in *2022 International Conference on Artificial Intelligence of Things and Crowdsensing (AIoTCs)*. IEEE, 2022, pp. 266–272.

[37] F. Ahmed, M. Asif, M. Saleem, U. F. Mushtaq, and M. Imran, "Identification and prediction of brain tumor using vgg-16 empowered with explainable artificial intelligence," *International Journal of Computational and Innovative Sciences*, vol. 2, no. 2, pp. 24–33, 2023.

[38] L. Gaur, M. Bhandari, T. Razdan, S. Mallik, and Z. Zhao, "Explanation-driven deep learning model for prediction of brain tumour status using mri image data," *Frontiers in genetics*, vol. 13, p. 822666, 2022.

[39] F. Mercaldo, L. Brunese, F. Martinelli, A. Santone, and M. Cesarelli, "Explainable convolutional neural networks for brain cancer detection and localisation," *Sensors*, vol. 23, no. 17, p. 7614, 2023.

[40] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 91–99, 2022.

[41] M. Nawaz, A. A. Sewissy, and T. H. A. Soliman, "Multi-class breast cancer classification using deep learning convolutional neural network," *Int. J. Adv. Comput. Sci. Appl*, vol. 9, no. 6, pp. 316–332, 2018.

[42] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into imaging*, vol. 9, pp. 611–629, 2018.

[43] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information fusion*, vol. 58, pp. 82–115, 2020.

[44] 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 *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[45] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[46] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[47] Y. Gebreyesus, D. Dalton, S. Nixon, D. De Chiara, and M. Chinnici, "Machine learning for data center optimizations: feature selection using shapley additive explanation (shap)," *Future Internet*, vol. 15, no. 3, p. 88, 2023.

[48] N. B. Kumarakulasinghe, T. Blomberg, J. Liu, A. S. Leao, and P. Papapetrou, "Evaluating local interpretable model-agnostic explanations on clinical machine learning classification models," in *2020 IEEE 33rd international symposium on computer-based medical systems (CBMS)*. IEEE, 2020, pp. 7–12.

Verification of Results

The classification results presented in Section 4.6 are rigorously validated and discussed in detail in this section. For evaluating the model, we utilized Average Accuracy, Average F1-Score, Average Precision, and Average Recall. Our proposed model achieved an impressive average Accuracy of 99.21% and an average F1-Score of 99.20%. A thorough examination of the achieved outcomes is provided, including a breakdown of the performance metrics such as accuracy, precision, recall, and F1-score across different classes. The consistency of these results with the previously reported outcomes is confirmed through repeated trials and cross-validation.

1. **Results with Dataset 1**:

```
scores = o_model.evaluate(test_ds1)
```

```
19/19 [==============================] - 10s 233ms/step - loss: 0.0412 - accuracy: 0.9895
```

```
scores = n_model.evaluate(test_ds1)
```

```
19/19 [==============================] - 4s 228ms/step - loss: 0.0358 - accuracy: 0.9921
```

FIGURE 1: Test Accuracy for Initial and Proposed Model

```
predictions_1=[]
class_label_1=[]
predictions_2=[]
class_label_2=[]
for images, labels in test_ds1:
    batch_predictions = o_model.predict(images)
    predictions_1.extend(batch_predictions)
    class_label_1.extend(labels)
for images, labels in test_ds1:
    batch_predictions = n_model.predict(images)
    predictions_2.extend(batch_predictions)
    class_label_2.extend(labels)
# Convert the list of predictions to a numpy array
predictions_1 = np.array(predictions_1)
class_label_1 = np.array(class_label_1)
predictions_2 = np.array(predictions_2)
class_label_2 = np.array(class_label_2)
```

```
pred_label_1 = np.argmax(predictions_1, axis=-1)
pred_label_2 = np.argmax(predictions_2, axis=-1)
```

FIGURE 2: Code for Other Metrics

```
from sklearn.metrics import precision_score, recall_score, f1_score
f1=f1_score(class_label_1, pred_label_1, average="macro")
print("f1_score:",f1)
precision = precision_score(class_label_1, pred_label_1,average="macro")
recall = recall_score(class_label_1, pred_label_1,average="macro")
print("Precision:", precision)
print("Recall:", recall)
```

```
f1_score: 0.9893132172025141
Precision: 0.9896751687891138
Recall: 0.989081128234952
```

FIGURE 3: Initial Model Results

```
from sklearn.metrics import precision_score, recall_score, f1_score
f1=f1_score(class_label_2, pred_label_2, average="macro")
print("f1_score:",f1)
precision = precision_score(class_label_2, pred_label_2, average="macro")
recall = recall_score(class_label_2, pred_label_2, average="macro")
print("Precision:", precision)
print("Recall:", recall)
```

```
f1_score: 0.9920443276837125
Precision: 0.9922710561497327
Recall: 0.9918995250064455
```

FIGURE 4: Proposed Model Results

2. **Results with Dataset 2**:

```
scores = o_model.evaluate(test_ds)
```

```
9/9 [==============================] - 8s 189ms/step - loss: 0.5166 - accuracy: 0.9333
```

```
scores = n_model.evaluate(test_ds)
```

```
9/9 [==============================] - 2s 213ms/step - loss: 0.6440 - accuracy: 0.9472
```

FIGURE 5: Test Accuracy for Initial and Proposed Model

```
predictions_1=[]
class_label_1=[]
predictions_2=[]
class_label_2=[]
for images, labels in test_ds:
    batch_predictions = o_model.predict(images)
    predictions_1.extend(batch_predictions)
    class_label_1.extend(labels)
for images, labels in test_ds:
    batch_predictions = n_model.predict(images)
    predictions_2.extend(batch_predictions)
    class_label_2.extend(labels)
# Convert the list of predictions to a numpy array
predictions_1 = np.array(predictions_1)
class_label_1 = np.array(class_label_1)
predictions_2 = np.array(predictions_2)
class_label_2 = np.array(class_label_2)
```

```
pred_label_1 = np.argmax(predictions_1, axis=-1)
pred_label_2 = np.argmax(predictions_2, axis=-1)
```

FIGURE 6: Code for Other Metrics

```
from sklearn.metrics import precision_score, recall_score, f1_score
f1=f1_score(class_label_1, pred_label_1, average="macro")
print("f1_score:",f1)
precision = precision_score(class_label_1, pred_label_1,average="macro")
recall = recall_score(class_label_1, pred_label_1,average="macro")
print("Precision:", precision)
print("Recall:", recall)
```

```
f1_score: 0.9324865210127704
Precision: 0.9335065835065834
Recall: 0.9383748221906116
```

FIGURE 7: Initial Model Results

```
from sklearn.metrics import precision_score, recall_score, f1_score
f1=f1_score(class_label_2, pred_label_2, average="macro")
print("f1_score:",f1)
precision = precision_score(class_label_2, pred_label_2, average="macro")
recall = recall_score(class_label_2, pred_label_2, average="macro")
print("Precision:", precision)
print("Recall:", recall)
```

```
f1_score: 0.9463383104976026
Precision: 0.9470363275448022
Recall: 0.9510218112849691
```

FIGURE 8: Proposed Model Results