QUBAHAN JOURNAL OF MEDICAL SCIENCES
VOL. 3, Sep 2026
https://doi.org/10.48161/qjms.v3a88

EXPLAINABLE AI-BASED MULTI-SCALE FEATURE FUSION FOR ACCURATE TUMOUR DETECTION AND CLASSIFICATION IN RADIOLOGICAL IMAGES

Rizwan Hameed[1*], Noman latif[2], Sarthak Sengupta[3], and Aderonmu Kehinde Adebayo[4]

[1] Computer Science, School of Computing, Gold Campus, Superior University, Lahore, Pakistan;
[2] Department of Pulmonology, Lahore General Hospital, Lahore, Pakistan;
[3] ADMIRE Center for Advancing Digital Health, IIHMR Bangalore, Institute of Health Management Research, Bangalore, India.
[4] Kinesiology Health Education and Recreation, Faculty of Education, Obafemi Awolowo University, Ile-Ife, Nigeria;

* **Corresponding author:** su92-phcsw-f25-020@superior.edu.pk.

ABSTRACT: Detection and classification of brain tumors are essential in medical imaging, as early and accurate diagnosis aids in treatment planning and patient outcomes. Magnetic Resonance Imaging (MRI) is a common imaging technique used in medicine, especially for brain imaging. Magnetic Resonance Imaging (MRI) is a medical imaging technique used for brain scanning, which is time-consuming and traditionally highly dependent on expert interpretation. While many AI models have been developed for the automatic detection of tumors, recent achievements in Artificial Intelligence (AI) and deep learning have shown great potential; however, many available models lack interpretability and fail to leverage the multi-level features of images. For this reason, this study presents an Explainable AI (XAI) framework for accurate brain Tumor detection and classification using MRI images. The study employed the Brain Tumor MRI Dataset, which contains 7,200 MRI images belonging to four classes (Glioma, Meningioma, Pituitary tumor, and No tumor). To enhance the model's performance, the images underwent preprocessing techniques such as resizing, normalization, and data augmentation. A basic CNN model and a CNN with High-Quality Gradient Class Activation Maps (CNN-XAI) were developed, both using batch normalization, dropout, Early Stopping, and ReduceLROnPlateau. Four metrics, accuracy, precision, recall, and F1-score, along with ROC-AUC analysis, a confusion matrix, and an interpretability assessment, were used to evaluate performance. The experimental results showed that the CNN baseline model achieved the highest classification accuracy of 87.44%, with Precision, Recall, and F1-score values of 88.69%, 87.44%, and 87.13%, respectively. The proposed CNN-XAI framework achieved 82% accuracy and improved transparency and interpretability by localizing the tumor with Grad-CAM. Both models demonstrated good discriminative ability, with near-perfect performance for the pituitary and non-tumor classes in ROC-AUC analysis. Glioma tumors continued to be the most challenging category owing to their imaging heterogeneity. These results suggest that CNN-based deep learning models contribute to automated brain tumor classification, and that explainable AI techniques provide a stronger understanding of and trust in the clinical application of these models. The proposed framework demonstrates the potential of combining deep learning and explainability for reliable, interpretable AI-assisted diagnosis in radiological imaging systems.

Keywords: brain tumor classification, magnetic resonance imaging (MRI), deep learning, CNN, explainable artificial intelligence, Grad-CAM, medical image analysis.

I. INTRODUCTION

Brain tumors are some of the most dangerous neurological diseases due to their uncontrolled growth inside the cranial cavity. According to the report released by WHO (2025), cancer resulted in more than 10 million deaths in 2020, making it one of the top killers globally. This underscores the importance of efficient systems for early tumor detection. A few medical imaging techniques, such as Computed Tomography (CT) scans and Magnetic Resonance Imaging (MRI), are widely used to detect brain tumors because they provide detailed visualization of brain structures (Arora et al., 2022). Artificial Intelligence (AI) has become an important research area for automated tumor detection and classification. Convolutional Neural Networks (CNNs) and deep learning are also becoming important areas in AI. Although CNN-based models achieve promising performance, they still struggle to capture multi-scale contextual information and also lack transparency in decision-making (Nirob et al., 2025). In recent years, XAI and multi-scale feature fusion have proven to be effective approaches for enhancing the performance of tumor classification algorithms. Tools like Grad-CAM and SHAP help identify the parts of an image that influence predictions, thereby making the diagnoses made by radiological imaging systems more reliable and dependable (Biswas et al., 2024).

The accurate detection and classification of tumors in radiological images remain a significant concern despite advances in technology. The above problems arise from tumor heterogeneity, irregular lesion shapes, complex anatomy, and areas of low contrast (Raposo, 2024). The current computer-based diagnostic models primarily use CNNs, which are efficient at local spatial feature extraction. However, these CNNs are inefficient at capturing long-range dependencies and semantic information, which are essential for tumor detection. As a result, there are cases where these models exhibit poor generalization, incorrect classifications in complex radiology images, and inefficient feature representations. Most tumor detection models are opaque because they operate as black boxes, thereby reducing their practicality in clinical settings. Despite achieving high accuracy in classification, the explainability of such models remains relatively low, as they do not provide clinically meaningful information about the decision-making process. Conventional single-scale feature extraction approaches are insufficient to effectively capture both fine-grained texture patterns and high-level semantic tumor representations simultaneously. Such limitations impair the detection of small, irregular, and heterogeneous tumor regions. There is a critical need for an explainable AI-based multi-scale feature-fusion framework to improve tumor detection.

This study aims to develop an Explainable AI-based multi-scale feature-fusion framework for accurate tumor detection and classification in radiological images. The key goals of this study are to identify the problems encountered when detecting and classifying tumors in radiographic imaging systems. After that, a multi-scale feature fusion algorithm will be designed along with Explainable AI to analyze tumors. Lastly, the objective of this study is to assess the performance of the developed model compared with existing CNN-based tumor detection models. Objectives of this research include exploring the main issues associated with tumor detection and classification in radiology images; assessing the impact of multi-scale feature fusion on tumor detection and classification performance; evaluating the performance of the newly designed Explainable AI model in improving tumor detection performance; and comparing it with traditional CNN-based models. The following are the research hypotheses.

- H1: Multi-scale feature fusion significantly improves tumor detection accuracy in radiological images.
- H2: Explainable AI techniques significantly enhance the interpretability and transparency of tumor detection models.
- H3: An Explainable AI-based multi-scale feature fusion framework performs better than traditional CNN-based tumor detection models.
- H4: The integration of multi-scale feature extraction and Explainable AI significantly improves tumor classification performance in radiological imaging systems.

The importance of this research lies in its contribution to the development of intelligent, explainable health care systems that aid in tumor detection. The framework designed under Explainable AI-based Multi-Scale Feature Fusion attempts to overcome the above shortcomings. The integration of multi-scale feature fusion enables the framework to effectively capture low-level texture features and high-level semantic information from radiological images. This can improve the detection of small, irregular, and heterogeneous tumor images. Algorithms such as Grad-CAM and SHAP can visualize prediction outcomes so that radiologists can understand the reasoning behind AI-generated predictions (Gharaibeh, 2025). Such developments increase transparency and acceptance of AI-based health care systems. The results of this

research article could be beneficial for further advances in the development of intelligent health care solutions. This study discusses the development of a multi-scale feature fusion system, based on the principles of Explainable Artificial Intelligence, to perform tumor detection and classification using imaging techniques such as MRI and CT scans. This study primarily concerns medical imaging systems.

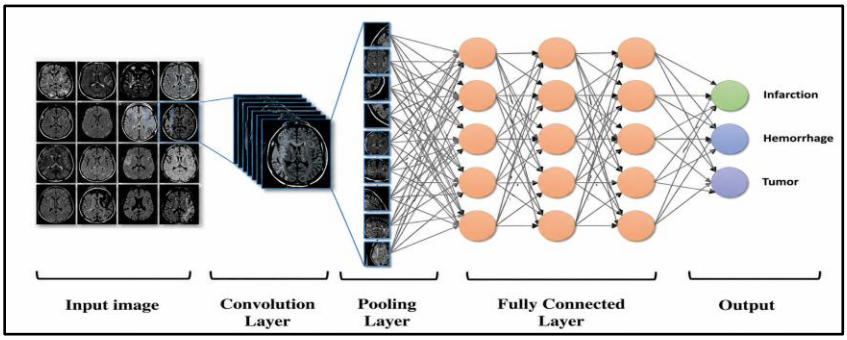
II. LITERATURE REVIEW

1. OVERVIEW OF RADIOLOGICAL IMAGING SYSTEMS

The study performed by Lother et al. (2023) indicated that imaging had been used quite effectively in the clinical assessment of tumors. Various imaging modalities include CT scans, MRIs, mammography, and PET scans. Some imaging modalities currently in vogue play a crucial role in identifying abnormalities. According to the results of the study by Abdullaeva et al. (2024), MRI is most effective for identifying CNS metastases and other abnormalities due to its excellent resolution. The gadolinium-enhanced T1 sequence is more sensitive than contrast-enhanced (Xi et al., 2022). The usual protocol for FLAIR scans is used to detect leptomeningeal diseases and cancer recurrence. The repeated administration of Gadolinium-based contrast agents can result in Nephrogenic Systemic Fibrosis and Gadolinium Deposition Disease in some individuals. As stated in the study, CT scan remains an essential imaging procedure in medicine. The CT-based diagnosis may have difficulty detecting hyperenhancing or heterogeneous metastatic lesions. Mammography and ultrasonography can be utilized to detect tumors; mammography can detect calcifications very well, while ultrasonography offers an advantage in cases of denser breasts. With recent developments in computer-aided diagnosis (CAD) and multimodality image fusion techniques, significant improvements have been achieved in tumor segmentation and classification.

2. TUMOUR DETECTION USING ARTIFICIAL INTELLIGENCE

AI has emerged as a transformative technology in medical imaging and tumor diagnosis due to its capability to analyze complex radiological data with high accuracy. Early AI-based tumor detection systems primarily utilized conventional machine learning algorithms combined with handcrafted feature extraction techniques. These approaches show limited performance because handcrafted features are insufficient to capture complex tumor patterns and heterogeneous lesion structures. Recent developments in deep learning, especially in CNNs, have significantly improved tumor detection performance by enabling automatic hierarchical feature extraction from radiological images. CNN-based models such as ResNet, U-Net, DenseNet, and EfficientNet have shown promising performance in tumor detection. Even though these AI-based tumor detection models exhibit high classification accuracy, they have been observed to be black-box models with poor interpretability. The general architecture of a CNN-based tumor detection system for radiological image analysis is illustrated in Figure 1. The model illustrates how image processing occurs through the sequential steps of image pre-processing, feature extraction, pooling, feature learning, and, finally, tumor classification in deep learning-based medical imaging systems.

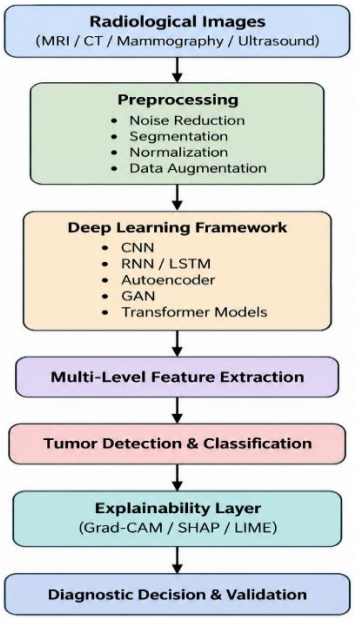


(Source: Baker et al., 2022).

FIGURE 1. CNN structures.

3. DEEP LEARNING IN MEDICAL IMAGING

According to Suganyadevi et al. (2021), the application of deep learning in the analysis of medical images is essential because these methods can extract feature representations from complex radiological datasets. Iqbal et al. (2023) stated that a deep learning architecture involves an input layer, a pooling layer, a convolutional layer, a fully connected layer, and an output layer. All these make deep learning effective in feature learning and classification. Convolutional neural networks (CNNs) are widely used because they can learn texture and spatial feature representations from both 2D and 3D medical images. There are different types of deep learning methods employed in medical imaging, such as Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, Deep Belief Networks (DBNs), and Generative Adversarial Networks (GANs) (Shaker & Xiong, 2023; Zaitoon et al., 2024). Here, both RNNs and LSTMs can learn temporal feature representations. Autoencoders are commonly applied for unsupervised feature extraction, dimensionality reduction, and denoising tasks. Figure 2 illustrates the general deep learning framework used in medical imaging systems for tumor detection and classification. GANs are highly useful for generating synthetic medical images and for data augmentation. Despite their many advantages, deep learning models still have limitations such as high computational requirements, risk of overfitting, dependence on labeled datasets, and limited interpretability.



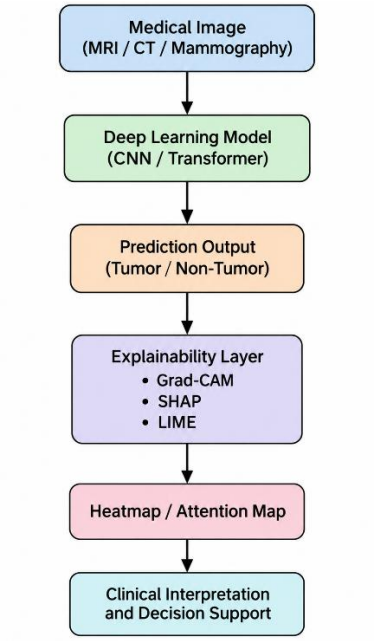
(Source: self-developed).

FIGURE 2. Deep learning framework for medical imaging and tumor detection.

4. EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI) IN HEALTHCARE

According to Abrantes & Pouria (2024), the importance of XAI stems from the growing interest in using deep learning-based tumor detection models in the healthcare system. The need for XAI stems from its ability to improve model transparency, reliability, accountability, and trustworthiness. XAI approaches can be classified into two categories: global and local explanations. Local explanations focus on understanding why a model makes a specific prediction for an individual input, whereas global explanations attempt to reveal overall behavior, logic, and decision-making patterns. Local explainability techniques are particularly important, as they help radiologists identify image regions that contribute significantly to disease prediction and tumor classification outcomes. According to Vilone & Longo (2021), existing XAI techniques can be further classified into ante-hoc and post-hoc approaches. Ante-hoc methods provide interpretability prior to the model's training phase, while post-hoc methods use explanations to understand the outputs generated after the model is trained. In the realm of post-hoc methods, both gradient-based and occlusion-based methods are commonly applied in medical imaging. For example, gradient-based methods such as Grad-CAM and Integrated Gradients generate heatmaps that highlight the areas of an image that play a crucial role in making predictions (Ennab & Mcheick, 2025). Figure 3 illustrates the general workflow of Explainable

Artificial Intelligence (XAI) techniques used in medical imaging systems for tumor detection and classification. The occlusion-based methods assess differences in predictions by masking different parts of the images.



(Source: self-developed).

FIGURE 3. General workflow of XAI in medical imaging systems.

5. MULTI-SCALE FEATURE FUSION TECHNIQUES

Multi-scale feature fusion methods have been widely used in recent years to improve the performance of tumor detection and classification in medical imaging systems. Many scholars have used CNNs, attention mechanisms, and transformer models to integrate low-level and high-level semantic features in radiographic images (Yu & Zhou, 2025). Multi-scale CNN models with residual connections, dense sampling, and pyramid structures have shown better results for lesion segmentation and classification. Apart from this, enhancements in the representation of spatial features and contextual learning were achieved using attention mechanisms and channel feature fusion techniques. The hybrid of CNNs and transformers has yielded positive outcomes due to their combination (Li et al., 2022).

6. CHALLENGES IN TUMOUR DETECTION AND CLASSIFICATION

While the use of deep learning techniques in tumor detection models, including CNNs, has made significant contributions to medical image analysis, current approaches have certain limitations. Limitations include poor feature-extraction capability, limited generalization, and an inability to capture the contextual information needed for effective classification. Limitations also include limited availability of labeled data, class imbalance ,an, and computational costsquera et al., 2024). Another limitation of AI-assisted tumor detection technology is explainability. Deep learning methods are often considered black-box models, limiting confidence in the reliability of applying AI techniques to medicine. Thus, there is a need for advanced XAI multi-scale feature fusion frameworks.

7. EXISTING TUMOUR DETECTION MODELS

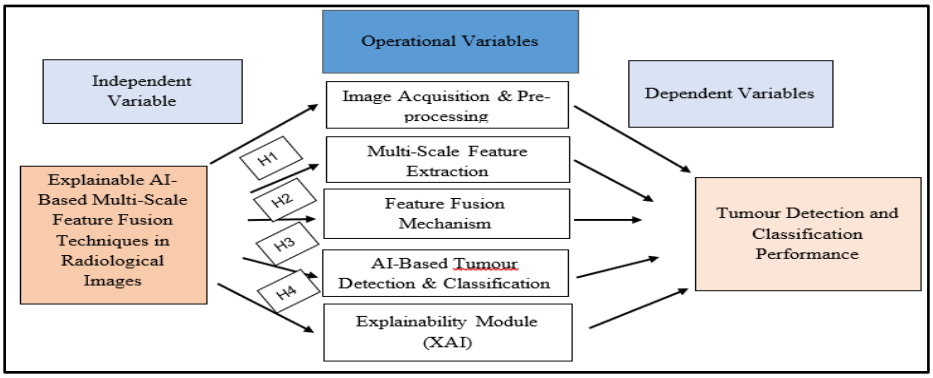
According to Abueed et al. (2025), existing techniques for detecting tumors include several deep learning models, such as CNNs, U-Net, ResNet, DenseNet, and transformer models, for automatic tumor detection in medical imaging. It has been observed that such techniques are highly efficient for feature extraction from tumors detected in MRI, CT scans, and mammograms. Attention models and multi-scale feature fusion approaches have been used to improve context learning and tumor localization accuracy in several studies.

But despite their successes, current systems continue to face challenges, including low interpretability, high computational complexity, feature redundancy, and poor generalization.

8. RESEARCH GAP IDENTIFICATION

Various machine learning frameworks based on deep architectures such as CNNs, U-Net, DenseNet, and transformers have shown excellent results in detecting different tumors in existing studies on medical image analysis. Although the frameworks mentioned above show high accuracy rates in tumor identification, various limitations remain. First, most studies have used deep learning-based methods for tumor detection and classification that rely mainly on local features, with inadequate attention to capturing the contextual or semantic multi-scale information necessary for accurate detection of irregular or heterogeneous tumors. Second, existing deep learning frameworks for detecting tumors mostly use single-scale feature representation techniques, which can be detrimental when detecting tumors in radiological images with low contrast and overlapping anatomical regions. An important gap in the existing literature is that relatively few studies focus on applying Explainable Artificial Intelligence (XAI) models to the development of tumor diagnosis systems. Most deep learning frameworks used in existing studies to detect different types of cancer and other diseases are treated as black-box models with low interpretability due to their complex, opaque architectures. Existing studies tend to focus solely on improving accuracy rather than addressing multiple aspects of AI applications simultaneously.

9. CONCEPTUAL FRAMEWORK DEVELOPMENT



(Source: Self-developed).

FIGURE 4. Conceptual framework.

III. METHODOLOGY

1. RADIOLOGICAL IMAGE DATASET COLLECTION

This research paper utilizes the open "Brain Tumor MRI Dataset" for tumor detection and classification in radiological imaging applications. This dataset contains 7200 human brain MRI images categorized into four classes: glioma, meningioma, pituitary tumor, and no tumor. The dataset is designed for deep learning-based medical image analysis and supports automated brain tumor classification tasks using MRI scans (Kaggle.com, 2021). The dataset comprises a number of training and testing subsets of an equal number of classes. The training subset consists of 1400 MRIs per class, while the testing dataset includes 400 MRIs per class to facilitate comparison. The selection of the dataset is attributed to its compatibility with multi-class classification tasks, the availability of sufficient data, and its relevance to Explainable AI-based studies in medical imaging.

2. DATA PREPROCESSING

Data preprocessing is carried out to enhance image quality and support the successful training of deep learning models. All images are initially resized to a uniform size for more efficient computation and processing. Various normalization approaches are used to standardize the pixel values across all images.

Techniques such as Gaussian and median filtering can be used to remove noise from images and achieve high-quality results. Additionally, various segmentation approaches are used to separate tumor areas from other regions. Finally, data augmentation techniques such as rotation, flipping, scaling, and zooming were applied to the images.

3. MULTI-SCALE FEATURE EXTRACTION FRAMEWORK

The proposed framework extracts multi-scale features hierarchically to identify low-, mid-, and high-level MRI representations for accurate tumor recognition (Huo et al., 2024). First, the low-level convolutional layers detect the basic features such as edges, textures, and intensity distribution in the MRI images. On the other hand, mid-level layers detect tumors' structural features, whereas high-level layers learn their semantic properties.

Where F_l represents the extracted feature map at layer l, W_l *denotes the convolutional kernel weights, Xl is the input feature map, bl* is the bias value, and σ is the nonlinear activation function (ReLU). The extracted hierarchical features are represented as follows.

Where F_{low} denotes low-level feature representations, Fmid denotes mid-level feature representations, and Fhigh denotes high-level semantic feature representations. This multi-scale hierarchical learning mechanism improves tumor boundary identification, texture representation, and spatial feature discrimination in MRI images.

4. FEATURE FUSION ARCHITECTURE

The feature fusion architecture combines multiple hierarchical feature representations obtained from different convolutional layers to generate robust tumor representations. Attention mechanisms are integrated to emphasize clinically relevant tumor regions and suppress irrelevant background features. This feature fusion strategy enhances discriminative tumor representation, improves classification stability, and supports accurate tumor localization and explainable AI-based interpretation through Grad-CAM visualization.

5. EXPLAINABLE AI INTEGRATION

The developed framework applies XAI methodologies to enhance the transparency and explainability of the brain tumor classification process. Deep learning methods are sometimes accused of being "black boxes" in the medical field, for example, making predictions without offering explanations. To address this problem, the Grad-CAM methodology has been incorporated into the CNN-based architecture. Using the Grad-CAM algorithmic approach helped identify the most critical MRI regions for classification. These heat maps were superimposed on the original MRI images to pinpoint tumor localization regions and enhance clinical understandability. Visualization helped the framework identify pathological parts of the image that seemed suspicious, as opposed to the other details that were not. The optimized CNN-XAI mechanism and the baseline CNN mechanism were used on the same dataset. This allowed CNN attention behavior to be analyzed and localization capabilities to be compared between models. The generated heatmaps enabled a more transparent, reliable AI diagnosis in medical imaging applications.

6. TUMOUR DETECTION AND CLASSIFICATION

The proposed framework was designed to automatically detect and classify brain tumors from MRI scans using Convolutional Neural Networks (CNNs). In the classification task, there were four classes: glioma, meningioma, pituitary tumor, and non-tumor. To ensure standardized model quality and performance, MRI images were preprocessed before training. All images were resized to 128 × 128 pixels and normalized to make the pixel intensity distribution of all images equal. The data was augmented to make the dataset more varied and less prone to overfitting using methods such as rotation, zooming, and horizontal flipping. Two CNN architectures were implemented. The first was a baseline CNN model with convolutional layers, max-pooling layers, flattening layers, and dense classification layers. The second one was an optimized CNN-XAI architecture, integrating batch normalization, dropout layers, and explainability. The two models adopted the Softmax activation function for multi-class classification. CNN models learned hierarchical features from

the MRI, from low-level edges to high-level tumor structures. The learned feature representations were then used to classify MRI images into their appropriate tumor classes.

7. Model Optimization

Optimization methods were used to improve convergence stability, minimize overfitting, and increase the model's capacity to generalize. In the initial testing, the CNN-XAI framework exhibited some instability in its validation performance and a slight tendency toward overfitting. Training stability was enhanced by using EarlyStopping and ReduceLROnPlateau. EarlyStopping monitored the validation loss and stopped training if no improvement was observed. It eliminated all unnecessary iterations, thereby preventing overfitting. The operation ReduceLROnPlateau prevents such problems by halving the learning rate once there is no improvement in the validation loss. Thus, the entire process becomes smoother as the weight values converge. To stabilize gradient propagation and improve training efficiency, batch normalization layers were added to the network. Also, to prevent overfitting, random neuron deactivation was applied during training using dropout layers. All these optimization methods contributed positively to enhancing the robustness and stability of the proposed framework.

8. VALIDATION AND INTERPRETABILITY ANALYSIS

Validation and interpretability analysis methods were combined to evaluate the clinical relevance, robustness, and reliability of the proposed framework. Various evaluation methods were employed, including confusion matrices, classification reports, ROC-AUC analysis, and Grad-CAM visualizations. The misclassification patterns across different ttumorclasses were aanalyzedusing confusion matrices. Performance metrics such as precision, recall, and F1 score were applied to evaluate the classification results on a per-class basis. The classification separability and discriminative power were evaluated using the ROC-AUC analysis. Interpretability modeling using Grad-CAM enabled visualization of CNN attention regions and the localization behavior of the tumor. The visualizations of the generated overlaid images showed that the framework focused on clinically relevant MRI regions for prediction generation. The techniques enhanced transparency and facilitated accurate AI-aided diagnosis.

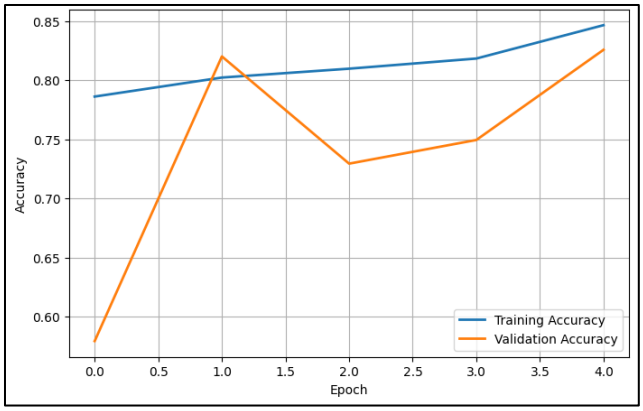
9. COMPARATIVE ANALYSIS FRAMEWORK

To compare the baseline CNN model with the optimized CNN-XAI framework, a comparative analysis framework was created. Comparisons were made using standard metrics such as accuracy, precision, recall, F1 score, ROC-AUC, and confusion matrices. The classification performance was evaluated using the baseline CNN (reference model), and the prediction capability and interpretability of the optimized CNN-XAI framework were evaluated. Comparative evaluation enabled analysis of model generalization, tumor localization behavior, and explainability performance. The framework enabled a thorough study of the classification accuracy–interpretability trade-off in medical imaging systems.

IV. PERFORMANCE EVALUATION

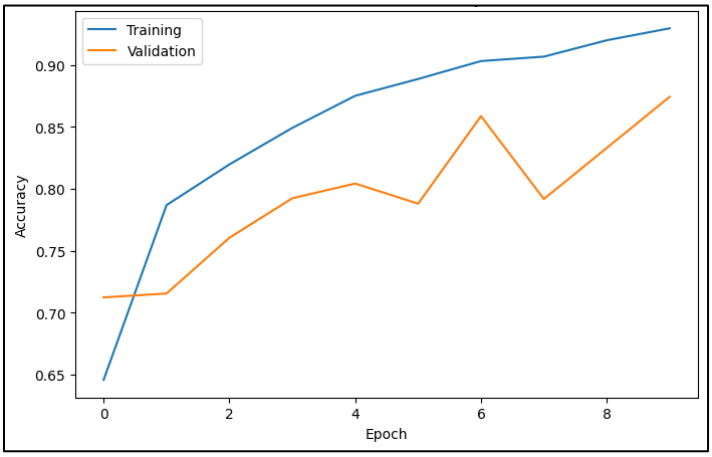
1. ACCURACY

To evaluate results, the overall accuracy of the tumor classification predictions was used as the primary metric. The experiments were performed on a multiclass MRI database comprising glioma, meningioma, pituitary tumor, and non-tumor brain images. Two CNN models were tested: a baseline CNN model and an optimized Explainable AI-integrated CNN framework.



(Source: Python).

FIGURE 5. Optimized XAI-CNN model accuracy.



(Source: Python).

FIGURE 6. Baseline CNN accuracy.

The baseline CNN achieved an overall accuracy of 87.44%, while the optimized CNN-XAI framework achieved 82%. The results showed that both models learned MRI features related to tumors and demonstrated good multi-class classification performance. However, the overall prediction performance of the optimized explainable AI framework was still worse than that of CNN. Because the baseline CNN has a simpler architecture and lower regularization complexity, its high accuracy may be attributed to this. For XAI, the model was additionally optimized with other regularization and optimization methods like batch normalization, dropout layers, EarlyStopping, and learning-rate scheduling. These additional parts increased interpretability and stability while slightly lowering classification performance. The training and validation accuracy graphs showed that the optimization techniques enhanced the model's convergence behavior and prevented severe overfitting. The optimized model exhibited more stable validation behavior with the implementation of learning-rate reduction and EarlyStopping techniques.

2. PRECISION

The percentage of correct predictions of tumor-positive cases among all positive predictions was measured as precision. High-precision values mean they are less likely to falsely predict a tumor and that their predictions are more likely to be accurate, which is crucial for medical diagnosis, because if they predict a tumor, unnecessary clinical procedures may be performed. The optimized CNN-XAI framework achieved a precision of 82.75%, while the baseline CNN framework has a precision of 88.69%. The analysis of the tumors revealed the highest Precision values in the pituitary group, as these tumors have the most distinct

MRI characteristics. Glioma ttumorsshowed relatively low accuracy due to overlapping structures on MRI and diffuse ttumorborders. The results of these studies showed that glioma classification remains one of the most difficult tasks in brain tumor detection. Despite these challenges, both frameworks performed well in precision across all tumor classes, demonstrating strong feature extraction and classification performance.

3. RECALL

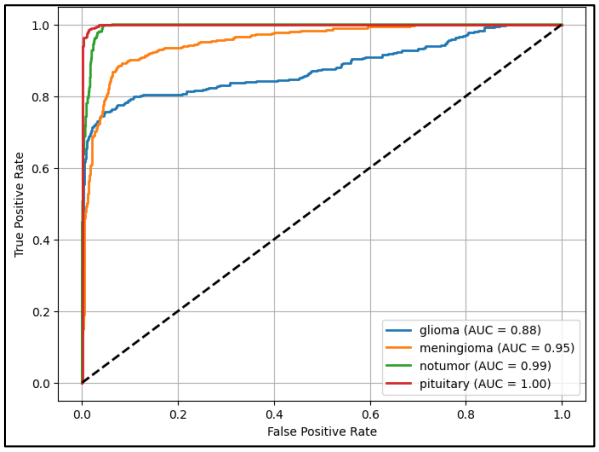
Recall was used to assess the framework's accuracy in identifying real tumor cases in MRI images. High recall values mean that the test has high sensitivity and few false negatives, which is crucial in medical applications, as false negatives could lead to an incorrect diagnosis and treatment for the patient. The CNN baseline achieved an overall recall of 87.44%, whereas the optimized CNN-XAI architecture achieved 82%. The best recall scores for the pituitary and non-tumor classes contributed to good class separability and feature learning. Recall rates for glioma tumors were poor, as some glioma cases were classified into other classes. Confusion matrix analysis revealed that misclassifications occurred most frequently among glioma cases across all tumor types. The results presented here highlight the challenges of glioma detection and the need for robust feature extraction and explanation techniques in medical imaging systems.

4. F1-SCORE

Precision and recall were evaluated using the F1-score, which offers a balanced measure. Medical image classification must be sensitive and reliable, making the F1 score an important measure for assessing the model's overall effectiveness. The overall F1-scores obtained by the baseline CNN and the optimized CNN-XAI framework were 81.17% and 87.13%, respectively. Both models yielded good sensitivity and reliable prediction results. The pituitary tumors had the highest F1-scores due to their unusual MRI morphology and a moderately low class overlap. Gliomas, on the other hand, had lower F1-scores due to tumor heterogeneity and overlapping MRI characteristics. The baseline CNN further demonstrated better classification and generalization performance by achieving a higher F1-score.

5. ROC-AUC ANALYSIS

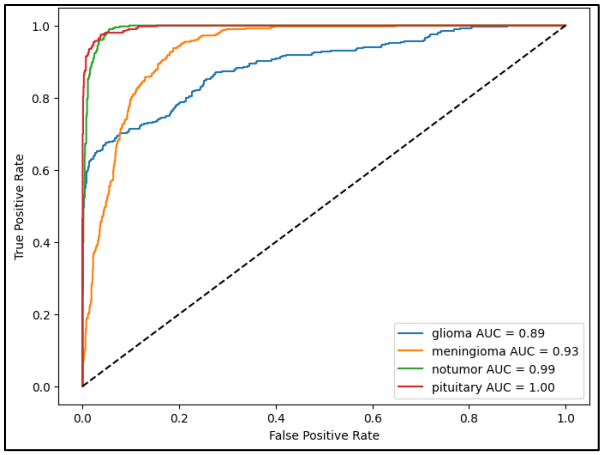
To assess the discriminative power of the proposed framework across all tumor categories, Receiver Operating Characteristic-Area Under the Curve (ROC-AUC) analysis was performed. The model's ability to separate positive and negative classes across various thresholds is evaluated using ROC-AUC analysis.



(Source: Python).

FIGURE 7. Baseline CNN ROC-AUC curve.

While the baseline CNN demonstrated good performance on the ROC-AUC metric, it performed very well across all tumor classes. AUC values were 0.8792 for gliomas, 0.9546 for meningiomas, 0.9940 for non-tumor images, and 0.9985 for pituitary tumors. Likewise, for the optimized CNN-XAI framework, high ROC-AUC performance was obtained, with the curves close to the upper-left corner of the graph, indicating high sensitivity and specificity.



(Source: Python).

FIGURE 8. Optimized XAI-CNN ROC-AUC curve.

The proposed CNN-XAI framework achieved good ROC-AUC performance across all tumor classes. The model achieved an AUC of ~0.89 for glioma, ~0.93 for meningioma, ~0.99 for non-tumor images, and ~1.00 for pituitary tumors. The Pituitary and non-tumor classes showed near-perfect AUC values due to the high separability of the MRI structures. The AUC values were lower in gliomas due to irregular tumor margins and differences in tumor structure. However, the results of the ROC-AUC experiments showed that both methods learned useful MRI tumor characteristics and exhibited good discriminative classification power.

6. COMPARATIVE ANALYSIS WITH EXISTING MODEL

The results of a comparative analysis showed that the CNN-XAI framework was significantly different from the baseline CNN framework. The baseline CNN outperformed the other networks in terms of accuracy, precision, recall, and F1-score. The optimized explainable AI framework, on the other hand, ensured greater interpretability and transparency through Grad-CAM visualizations and tumor localization analysis. The proposed framework yielded localized, clinically meaningful activation regions via Grad-CAM heatmaps. The baseline CNN, on the other hand, had more diffuse attention patterns spread across larger MRI regions. The results demonstrate that despite its reduced overall classification performance, integration with explainability enhanced localization performance. The study revealed a significant trade-off between classification accuracy and model interpretability in AI healthcare applications. While simpler CNN models may yield better numerical accuracy, explainable artificial intelligence may increase their transparency and reliability for physicians. In conclusion, this study demonstrated the potential of CNN models, combined with XAI methodologies, for developing algorithms for automatic brain tumor detection and classification.

V. RESULTS AND DISCUSSION

1. EXPERIMENTAL RESULTS

The experimental evaluation was conducted using a brain MRI dataset consisting of 7200 images, categorized into four classes: glioma, meningioma, pituitary, and non-tumor. Two deep learning models were implemented and evaluated: a baseline CNN model and the proposed explainable AI (XAI)- integrated CNN framework. The objective was to assess classification performance, interpretability, and clinical applicability. The baseline CNN achieved superior classification performance, with an accuracy of 87.44%, precision of 88.69%, recall of 87.44%, and an F1 score of 87.13%. The proposed CNN-XAI framework achieved an accuracy of 82%, a precision of 82.75%, a recall of 82%, and an F1 score of 81.17%. Although the proposed framework showed lower numerical performance, it provided enhanced interpretability through Grad-CAM visualizations. Moreover, the confusion matrix also showed that the most classifications were correctly predicted across all tumor categories. Some misclassifications still occurred between glioma and meningioma tumors due to similarities in their MRI characteristics. The pituitary tumors and non-tumor

images achieved the highest classification accuracy and the lowest misclassification rate, indicating strong feature separability. The effectiveness of both models was also confirmed using the ROC-AUC analysis. The proposed CNN XAI framework achieved approximately 0.89 AUC for glioma, 0.93 for meningioma, 0.99 for non-tumor, and nearly 1.00 for pituitary tumors. Also, the baseline CNN achieved AUC values of 0.8792, 0.9546, 0.9940, and 0.9985 for the glioma, meningioma, non-tumor, and pituitary classes, respectively. These findings demonstrate strong discriminative capability across all tumor categories. The experimental results also confirmed that both CNN-based frameworks were capable of accurately classifying brain tumors from MRI scans while maintaining high diagnostic reliability.

5.2 TUMOUR DETECTION PERFORMANCE

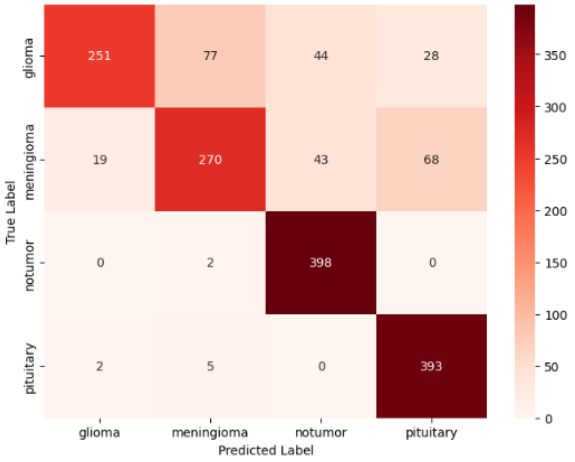


FIGURE 9. CNN-XAI confusion matrix.

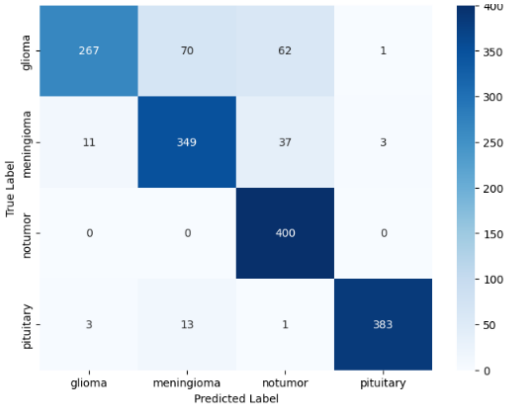


FIGURE 10. Baseline CNN confusion matrix.

Accuracy, precision, recall, F1-score, confusion matrix analysis, and ROC-AUC were used to evaluate tumor detection performance. The findings showed that the two models learned meaningful MRI features and could distinguish between tumor and non-tumor images. Pituitary tumors performed best amongst the four classes. The baseline CNN achieved a precision of 0.99, a recall of 0.96, and an F1 Score of 0.97 when classifying pituitary tumors. Similarly, the ROC-AUC value was close to 1.00, indicating near-perfect classification performance. This means that the image patterns of pituitary tumors, when acquired on an MRI scanner, are distinct and would be readily identified by CNN-based models.

The classification scores of the non-tumour group were also good. The model was able to differentiate between normal and tumor brain scans. This has clinical importance because false-positive results may translate to unindicated clinical tests and anxiety. The non-tumor class ROC-AUC score was high, indicating

strong discriminative ability and low false-positive rates. Conversely, glioma tumors were the most difficult to classify. The CNN on the baseline achieved a recall of 0.67 for glioma detection, meaning that not all gliomas were correctly classified as glioma, but rather as meningioma or other categories. This difficulty could be attributed to indistinct tumor margins, heterogeneous tissue properties, and overlapping MRI appearances.

Also, based on the confusion matrix, glioma had the most classification errors among the classes. Meningioma tumors demonstrated moderate-to-high classification performance and good precision and recall scores. Nonetheless, there was a certain overlap with glioma features, leading to eventual misclassification. These issues did not significantly affect overall tumor detection performance, indicating that CNN-based methods are effective for automatic brain tumor diagnosis. These results show that deep learning models have the potential to deliver promising tumor-detection performance and greatly assist radiologists in detecting brain abnormalities on MRI scans.

3. IMPACT OF MULTI-SCALE FEATURE FUSION

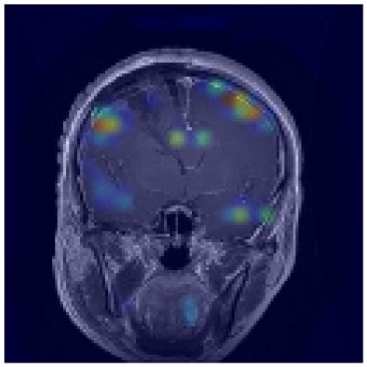


FIGURE 11. Grad cam for glioma.

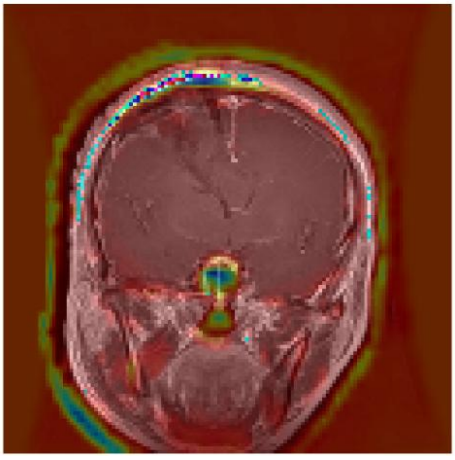


FIGURE 12. Baseline CNN tumor localization.

The proposed study aimed to use multi-scale feature learning to classify brain tumours. Although no explicit feature fusion architecture was used, the convolutional layers extracted hierarchical feature representations across multiple levels. These layers were trained on low-, mid-, and high-level MRI features, and the models were able to extract various tumor-related information. The first convolutional layers identified primordial image features such as edges, texture patterns, and intensity variations. Intermediate layers learned more intricate structural patterns, and deeper layers learned high-level tumor behavior associated with a particular tumor type. This hierarchical learning model has helped achieve effective feature representation and improved classification performance.

Grad-CAM visualizations were also helpful in the feature-learning process. The resulting heatmaps showed that the heat the model focused on was on clinically important areas of the tumor, not irrelevant background structures. CNN-XAI incorporated more local activation maps, indicating that deeper feature representations were provided, thus allowing for greater interpretability and tumor localization. The classification results showed that hierarchical feature extraction was sufficient to distinguish between the different tumor types. Classes with highly distinctive features, such as pituitary tumors, showed better classification performance, as these features are well represented by multiple CNN layers. On the other hand, glioma tumors were difficult to classify because their characteristics were less pronounced and highly variable.

4. DISCUSSION OF FINDINGS

The finding of the present research work is that deep learning models are a promising approach for automated brain tumor classification using MRI images. The results demonstrated that both the baseline CNN and the proposed CNN-XAI framework achieved high classification accuracy, indicating that the convolutional neural network algorithm can effectively extract meaningful tumor features. The Accuracy and F1-score obtained by the baseline CNN were higher, indicating that the CNN trained on the baseline dataset has better classification performance. However, the proposed CNN-XAI framework provided enhanced interpretability through Grad-CAM visualizations, enabling transparent analysis of the decision-making process. An important finding was that glioma tumors continued to be the most challenging type to classify. This is due to their poorly defined borders, internal anatomy, and overlapping MRI characteristics. However, pituitary tumors had near-perfect classification accuracy because their imaging features were easy to distinguish. These results were also supported by the strong discriminative performance of the confusion matrix and ROC-AUC analyses for both models. In the Grad-CAM visualizations, clinically relevant regions of the tumor were shown to be located within the proposed framework, which helped interpret the results and provided greater transparency in diagnostics. The effectiveness of these explainable AI techniques, such as Grad-CAM, in improving transparency and fostering trust in medical imaging systems reinforces the importance of such tools in clinical practice (Abrantes & Pouria, 2024; Ennab & Mcheick, 2025). Furthermore, the high classification accuracy achieved by both models is in line with earlier studies on the detection and analysis of medical and tumor images using CNN-based models (Anari et al., 2025). The findings suggest that in addition to classification accuracy, interpretability is an essential factor in AI applications in healthcare.

5. CLINICAL IMPLICATIONS

The suggested framework has significant clinical and radiological implications. The CNN models have demonstrated good accuracy in identifying brain tumors, highlighting their potential as decision-support systems for healthcare professionals. Automated tumor detection reduces diagnostic workload, enhances screening efficiency, and helps clinicians identify abnormalities at an early stage. Moreover, the incorporation of Grad-CAM, a key part of the model, boosts evidence and trustworthiness by visualizing the MRI region of influence that determines the model's prediction. This aligns with the aspirations for transparent and accountable AI in healthcare settings, where explainability and trust are critical for successful implementation in clinical practice (Abrantes & Pouria, 2024). Thus, the framework can help support more rational and sound clinical decision-making in healthcare.

VI. CONCLUSION AND FUTURE WORK

1. SUMMARY OF FINDINGS

In this study, an automated brain tumor classification system using MRI images was developed and evaluated based on a CNN and an Explainable AI (XAI) approach. The dataset used was a multiclass dataset containing MRI scans of glioma, meningioma, pituitary tumor, and non-tumor cases for research. Two deep learning models were applied: a baseline CNN and a CNN-XAI model incorporating optimization techniques and Grad-CAM visualization. The experimental findings demonstrated that both models achieved strong classification performance. The proposed CNN-XAI framework achieved an overall

accuracy of 82%, while the baseline CNN attained the highest accuracy of 87.44%. Precision, recall, F1-score, confusion matrices, and ROC-AUC evaluations demonstrated the efficacy of both models for brain tumor detection and classification. Apart from this, pituitary tumors and non-tumor images were determined with the highest reliability, while glioma tumor images were the third most difficult category due to the presence of similar MRI features and differences in tumor structures. One of the study's key findings was the successful application of explainable AI techniques. Using Grad-CAM visualizations, it was possible to identify the MRI regions that influenced the classification decision, thereby making the model transparent and interpretable.

2. CONTRIBUTIONS OF THE STUDY

One of the study's key findings was the successful application of explainable AI techniques. Using Grad-CAM visualizations, it was possible to identify the MRI regions that influenced the classification decision, thus making the model transparent and interpretable. Also, the research conducted a comparative evaluation between sa tandard CNN architecture and ean explainability-integratedCNN framework. By comparing the two models, this study showed that although there is a trade-off between classification and interpretability, some classification models remain clinically meaningful while maintaining comparable performance. Furthermore, the study aligns with ongoing research and development in deep learning for automated medical image analysis and intelligent decision-support systems.

3. LIMITATIONS

Despite achieving promising results, the study has several limitations. The experiments were conducted using a single publicly available MRI dataset, which may limit the applicability of the findings to other clinical environments and imaging conditions. The data were also limited to 4 classification categories and did not capture the full range of tumor types treated in hospital settings. Although hierarchical feature extraction was achieved through multiple CNN layers, an explicit multi-scale feature-fusion architecture was not implemented. Further classification performance might be improved by employing more advanced feature-fusion methods, particularly for the challenging glioma class. The other restriction is the use of a single explainability technique. Although Grad-CAM offered useful insights into interpretability, further XAI methods could augment the analysis of model behavior.

4. FUTURE RESEARCH DIRECTIONS

Future research should be conducted on larger, more heterogeneous MRI collections from multiple healthcare centers to enhance the model's robustness and generalizability. More complex multi-scale feature fusion, attention mechanisms, and hybrid deep learning models can be used to further enhance feature representation and classification performance. Moreover, other XAI techniques should be applied to further analyses, such as SHAP and attention-based visualizations, to gain better insight into the model's decisions. Moreover, optimization and comparison with other architectures, such as ResNet, DenseNet, and Vision Transformers, could yield insights into performance improvements. Finally, additional research into clinical application and in vivo validation, in collaboration with radiologists and practitioners, is required. This will help assess the effectiveness and usefulness of the XAI system and facilitate its acceptance in the day-to-day diagnostics process.

Author Contributions

The author conducted the conceptualization, methodology, data analysis, investigation, writing, review, editing, and final approval of the manuscript.

Funding

This research received no external funding.

Data Availability

The dataset will be available from the author upon reasonable request.

Conflicts of Interest

The author declares no conflict of interest.

REFERENCES

1. Abdullaeva, U., Pape, B., & Hirvonen, J. (2024). Diagnostic accuracy of MRI in detecting the perineural spread of head and neck tumors: a systematic review and meta-analysis. Diagnostics, 14(1), 113. https://doi.org/10.3390/diagnostics14010113
2. Abrantes, J., & Pouria Rouzrokh. (2024). Explaining explainability: The role of XAI in medical imaging. *European Journal of Radiology*, *173*, 111389–111389. https://doi.org/10.1016/j.ejrad.2024.111389
3. Abueed, O., Wang, Y., & Khasawneh, M. (2025). A Systematic Review of U-Net Optimizations: Advancing Tumor Segmentation in Medical Imaging. IET Image Processing, 19(1), e70203. https://doi.org/10.1049/ipr2.70203
4. Ali, A., Alghamdi, M., Marzuki, S. S., Tengku Din, T. A. D. A. A., Yamin, M. S., Alrashidi, M., ... & Ahmed, N. (2025). Exploring AI approaches for breast cancer detection and diagnosis: A review Article. Breast Cancer: Targets and Therapy, 927-947. https://doi.org/10.2147/BCTT.S550307
5. Alksas, A., Shehata, M., Saleh, G. A., Shaffie, A., Soliman, A., Ghazal, M., ... & El-Baz, A. (2021). A novel computer-aided diagnostic system for accurate detection and grading of liver tumors. Scientific reports, 11(1), 13148. https://doi.org/10.1038/s41598-021-916340
6. Anari, S., Sadeghi, S., Sheikhi, G., Ranjbarzadeh, R., & Bendechache, M. (2025). Explainable attention-based breast tumor segmentation using a combination of UNet, ResNet, DenseNet, and EfficientNet models. Scientific Reports, 15(1), 1027. https://doi.org/10.1038/s41598-024-84504-y
7. Arora, V., Sidhu, B. S., & Singh, K. (2022). Comparison of computed tomography and magnetic resonance imaging in evaluation of skull lesions. Egyptian Journal of Radiology and Nuclear Medicine, 53(1), 67. https://doi.org/10.1186/s43055-022-00745-9
8. Baker, M. R., Padmaja, D. L., Puviarasi, R., Mann, S., Panduro-Ramirez, J., Tiwari, M., & Samori, I. A. (2022). Implementing Critical Machine Learning (ML) Approaches for Generating Robust Discriminative Neuroimaging Representations Using Structural Equation Model (SEM). *Computational and Mathematical Methods in Medicine*, *2022*, 1–12. https://doi.org/10.1155/2022/6501975
9. Biswas, S., Mostafiz, R., Uddin, M. S., & Paul, B. K. (2024). XAI-FusionNet: Diabetic foot ulcer detection based on multi-scale feature fusion with explainable artificial intelligence. *Heliyon*, *10*(10), e31228. https://doi.org/10.1016/j.heliyon.2024.e31228
10. Ennab, M., & Mcheick, H. (2025). Advancing AI Interpretability in Medical Imaging: A Comparative Analysis of Pixel-Level Interpretability and Grad-CAM Models. *Machine Learning and Knowledge Extraction*, *7*(1), 12. https://doi.org/10.3390/make7010012
11. GHARAIBEH, N. (2025). Enhancing interpretability in brain tumor detection: Leveraging Grad-CAM and SHAP for explainable AI in MRI-based cancer diagnosis. Applied Computer Science, 21(3), 182-197. https://ph.pollub.pl/index.php/acs/article/download/7375/5202
12. Huo, X., Sun, G., Tian, S., Wang, Y., Yu, L., Long, J., ... & Li, A. (2024). HiFuse: Hierarchical multi-scale feature fusion network for medical image classification. Biomedical signal processing and control, 87, 105534. https://arxiv.org/pdf/2209.10218
13. Iqbal, S., N. Qureshi, A., Li, J., & Mahmood, T. (2023). On the Analysis of Medical Images Using Traditional Machine Learning Techniques and Convolutional Neural Networks. *Archives of Computational Methods in Engineering*, *30*. https://doi.org/10.1007/s11831-023-09899-9
14. Kaggle.com (2021). Brain Tumor MRI Dataset. https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset
15. Li, Z., Li, D., Xu, C., Wang, W., Hong, Q., Li, Q., & Tian, J. (2022, September). Tfcns: A cnn-transformer hybrid network for medical image segmentation. In International Conference on Artificial Neural Networks (pp. 781-792). Cham: Springer Nature Switzerland. https://arxiv.org/pdf/2207.03450
16. Lother, D., Robert, M., Elwood, E., Smith, S. M., Tunariu, N., Johnston, S., … Sharma, B. (2023). Imaging in metastatic breast cancer, CT, PET/CT, MRI, WB-DWI, CCA: review and new perspectives. *Cancer Imaging*, *23*(1). https://doi.org/10.1186/s40644-023-00557-8
17. Mosquera, C., Ferrer, L., Milone, D. H., Luna, D., & Ferrante, E. (2024). Class imbalance in medical image classification: towards better evaluation practices for discrimination and calibration performance. European Radiology, 34(12), 7895-7903.
18. Nirob, M. A. S., Bishshash, P., Khatun, T., Sharmin, S., Hasan, M. Z., & Arefin, M. S. (2025, February). Attention-based multi-scale fusion for brain tumor classification with explainable AI. In 2025 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 1- 6). IEEE. https://www.researchgate.net/profile/Md-Nirob/publication/392221177_Attention-Based_Multi-Scale_Fusion_for_Brain_Tumor_Classification_with_Explainable_AI/links/683aeab1c33afe388ac94446/Attention-Based-Multi-Scale-Fusion-for-Brain-Tumor-Classification-with-Explainable-AI.pdf
19. Raposo, H. (2024). Intelligent imaging: A systematic review of artificial intelligence techniques in disease detection, segmentation, and classification. Segmentation, and Classification (May 13, 2024). https://papers.ssrn.com/sol3/Delivery.cfm?abstractid=5342645

20. Shaker, A. M., & Xiong, S. (2023). Lung Image Classification Based on Long Short-Term Memory Recurrent Neural Network. *Journal of Physics: Conference Series*, *2467*(1), 012007. https://doi.org/10.1088/1742-6596/2467/1/012007
21. Suganyadevi, S., Seethalakshmi, V., & Balasamy, K. (2021). A review on deep learning in medical image analysis. *International Journal of Multimedia Information Retrieval*, *11*(1), 19–38. https://doi.org/10.1007/s13735-021-00218-1
22. Vilone, G., & Longo, L. (2021). A Quantitative Evaluation of Global, Rule-Based Explanations of Post-Hoc, Model-Agnostic Methods. *Frontiers in Artificial Intelligence*, *4*. https://doi.org/10.3389/frai.2021.717899
23. Wang, S., Mao, N., Duan, S., Li, Q., Li, R., Jiang, T., ... & Gu, Y. (2021). Radiomic analysis of contrast-enhanced mammography with different image types: classification of breast lesions. Frontiers in Oncology, 11, 600546. https://doi.org/10.3389/fonc.2021.600546
24. World Health Organization. (2025, February 3). Cancer. Retrieved from the World Health Organization website: https://www.who.int/news-room/fact-sheets/detail/cancer
25. Xi, X., Li, W., Li, B., Li, D., Tian, C., & Zhang, G. (2022). Modality-correlation embedding model for breast tumor diagnosis with mammography and ultrasound images. *Computers in Biology and Medicine*, *150*, 106130. https://www.sciencedirect.com/science/article/am/pii/S0010482522008381
26. Yaqub, M., Jinchao, F., Ahmed, S., Mehmood, A., Chuhan, I. S., Manan, M. A., & Pathan, M. S. (2023). DeepLabV3, IBCO-based ALCResNet: A fully automated classification and grading system for brain tumors. Alexandria *Engineering Journal*, *76*, 609-627. https://doi.org/10.1016/j.aej.2023.06.062
27. Yu, S., & Zhou, P. (2025). An optimized transformer model for efficient detection of thoracic diseases in chest X-rays with multi-scale feature fusion. *PLOS One*, *20*(5), e0323239. https://doi.org/10.1371/journal.pone.0323239
28. Zaitoon, R., Mohanty, S. N., Godavarthi, D., & Ramesh, J. V. N. (2024). SPBTGNS: Design an efficient model for survival prediction in brain tumor patients using a generative adversarial network with neural architecture search operations. IEEE Access, 12, 140847-140869. Digital Object Identifier 10.1109/ACCESS.2024.3430074