Applied Computer Science 2025, vol. 21, no. 3, pp. 182–197
https://doi.org/10.35784/acs_7375

Submitted: 2025-03-19 | Revised: 2025-05-06 | Accepted: 2025-05-18

CC-BY 4.0

Keywords: brain MRI classification, Xception architecture, explainable AI, grad-CAM, SHAP interpretability applications

Nasr GHARAIBEH [1*]
[1] Al-Balqa Applied University, Jordan, nas@bau.edu.jo
[*] Corresponding author: nas@bau.edu.jo

Enhancing interpretability in brain tumor detection: Leveraging Grad-CAM and SHAP for explainable AI in MRI-based cancer diagnosis

Abstract

This study aims to improve the interpretability of brain tumour detection by using explainable AI techniques, namely Grad-CAM and SHAP, alongside an Xception-based convolutional neural network (CNN). The model classifies brain MRI images into four categories — glioma, meningioma, pituitary tumour and non-tumour — ensuring transparency and reliability for potential clinical applications. An Xception-based CNN was trained using a labelled dataset of brain MRI images. Grad-CAM then provided region-based visual explanations by highlighting the areas of the MRI scans that were most important for tumour classification. SHAP quantified feature importance, offering a detailed understanding of model decisions. These complementary methods enhance model transparency and address potential biases. The model achieved accuracies of 99.95%, 99.08%, and 98.78% on the training, validation, and test sets, respectively. Grad-CAM effectively identified regions that were significant for different tumour types, while SHAP analysis provided insights into the importance of individual features. Together, these approaches confirmed the reliability and interpretability of the model, overcoming key challenges in AI-driven medical diagnostics. Integrating Grad-CAM and SHAP with a high-performing CNN model enhances the interpretability and trustworthiness of brain tumour detection systems. The findings underscore the potential of explainable AI to improve diagnostic accuracy and encourage the adoption of AI technologies in clinical practice.

1. INTRODUCTION

An MRI scan is necessary for investigating and determining various tumour types, especially gliomas, meningiomas and pituitary adenomas (Abd-Ellah et al., 2019; Khan et al., 2022). MRI is preferred because it provides high-resolution imaging of soft tissues, which is crucial for identifying subtle variations in brain anatomy that may indicate the presence of tumours (Nickparvar, 2024). As gliomas arise from glial cells, they present with an infiltrating nature that is very difficult to identify (Shamshad et al., 2024). In contrast, meningiomas tend to be located on the surface of the brain and are usually well-defined on an MRI scan, making them easily identifiable and evaluable for surgical treatment (Boitor et al., 2023). MRI is useful for delineating the extent of tumours arising from glial cells and informing surgeons of the best way to treat patients. It is also useful for examining pituitary adenomas, which can affect hormonal balance and cause neurological symptoms (Raghunath Mutkule, 2023). At the same time, it provides clear images of the sella turcica, where these tumours most often reside (Ahmed et al., 2023).

Despite advances in MRI technology, MRI scans are often interpreted in a complex, expert-based manner (Guo & Dou, 2023). This further complicates matters by increasing reliance on AI and ML models to aid diagnosis (Nhlapho et al., 2024). While such technologies often appear promising, they are frequently 'black boxes' that are almost impossible to operate, with decision-making processes that are not transparent to users (Babu Vimala et al., 2023). Interpretability is a persistent issue in clinical settings, as patients need to understand the rationale behind diagnoses in order to foster trust and facilitate effective treatment planning (Abunasser et al., 2023). Ghassemi et al. (2021) note that, although promising, most of these technologies remain largely opaque and function as black boxes. Their decision-making processes ultimately need to be more transparent to users. This lack of interpretability poses a significant challenge in clinical situations, as

patients must understand the diagnostic rationale in order to establish trust and facilitate effective treatment planning. If clinicians cannot determine how the model arrived at its conclusion, they may be reluctant to act on AI-generated responses to diagnosis, making them sceptical about adopting such technologies into practice (Jinsakul et al., 2019). They may then be reluctant to act based on AI-generated diagnostic responses, which makes them sceptical about adopting such technologies into practice (Jinsakul et al., 2019).

ResNet and Xception are regarded as key drivers of tumour detection alongside other CNNs, as they can sense subtle features in medical imaging (Ejiyi et al., 2023). They have demonstrated high accuracy in numerous cancer classification tasks (Mandiya et al., 2024). The most common visualisations have been created using Grad-CAM, which highlights the regions most influential in the ultimate predictions made by CNNs (Noreen et al., 2020). Some studies on Grad-CAM have noted that it can be relied upon even for deeper neural networks (Qiu et al., 2023). Furthermore, integrating Grad-CAM with other explainable AI techniques, such as Shapley values, in brain tumour detection tasks can provide region-specific and feature-level insights, thereby improving clinical decision-making and model transparency (Ahmed et al., 2024). Ahmed et al. (2024) have therefore emphasised the importance of explainability when deploying AI models in sensitive medical applications.

The following methods of XAI will therefore be integrated to address the interpretability challenges of AI in medical imaging: SHAP and Grad-CAM (Trivedi et al., 2021). SHAP provides insights into the contributions of individual features to model predictions, enabling clinicians to understand the MRI data that led to a particular diagnosis (Viswan et al., 2023). This is particularly important in high-stakes medical settings where the consequences of a misdiagnosis could be severe (Ejiyi et al., 2023). Grad-CAM, on the other hand, generates visual explanations that highlight which parts of MRI images most significantly influenced the model's decisions, providing more intuitive insights into the reasoning behind AI. Together, these techniques have the potential to make AI models more transparent, thereby fostering greater trust among healthcare professionals and patients in general (Nhlapho et al., 2024).

The challenges of interpretability in AI models extend beyond technical limitations to encompass ethical considerations (Khan et al., 2022). Several questions regarding accountability arise when AI is deployed in healthcare, including the potential for bias in decision-making processes (Mandiya et al., 2024). For instance, an AI model trained on a non-diverse dataset will produce biased predictions that cannot be generalised across patient populations (Awaluddin et al., 2023). This underlines the importance of extensive validation and continuous monitoring of the AI system's performance to ensure unbiased operation across different demographic groups (Nickparvar, 2024). Furthermore, ethical issues become more pronounced when AI is used for absolute health decision-making, creating a need for a framework to understand and mitigate the risks associated with diagnostic automation (Noreen et al., 2020).

The primary objective of this study is to achieve high accuracy in tumour classification while ensuring that the prediction results are interpretable and transparent. To this end, this work incorporates the advanced techniques of Grad-CAM and SHAP to provide clinicians with clear, visual and feature-wise interpretations of predictions, thereby facilitating a better understanding of, and trust in, the model's AI-based diagnostic process. This interpretability is crucial for fostering confidence in the clinical adoption of AI. It is important to establish a proper connection between technical performance and practical utility to ensure that AI is not used in sensitive medical applications such as brain tumour diagnosis.

2. LITERATURE REVIEW

Integrating deep learning (DL) into medical imaging has significantly improved the ability to detect brain tumours using MRI scans. However, despite the high performance of convolutional neural networks (CNNs), their lack of transparency continues to hinder their adoption in clinical practice. Consequently, there has been an increase in interest in Explainable Artificial Intelligence (XAI), which aims to make these models more interpretable for clinicians. The most widely used XAI methods include Gradient-weighted Class Activation Mapping (Grad-CAM), Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). This literature review examines and compares recent studies employing these techniques for MRI-based brain tumour detection, focusing on their CNN architectures, datasets, explainability methods, performance metrics and interpretability outcomes.

Several studies have proposed custom CNN architectures designed explicitly for brain tumour classification. Nazir et al. (2024), for example, developed a tailored CNN using the BR35H dataset, which

comprises over 3,000 brain MRIs, and combined it with Grad-CAM, SHAP and LIME. Their model achieved validation accuracy of 98.67% and an F1 score of 98.5%, demonstrating high predictive performance alongside detailed visual and feature-based interpretability (Nazir et al., 2024). Similarly, Rahman et al. (2024) introduced the Glioma CNN: a lightweight CNN architecture designed to differentiate between low-grade and high-grade gliomas using the BraTS2020 dataset. They used Grad-CAM++ and SHAP to generate visual and feature-level explanations, achieving an impressive accuracy of 99.15% (Rahman et al., 2024).

Pre-trained architectures such as EfficientNet have also been adapted for brain tumour detection, incorporating integrated explainable artificial intelligence (XAI) methods. T. R. et al. (2024) used the EfficientNetB0 model to classify four types of brain tumour (glioma, meningioma, pituitary tumour and no cancer). They applied Grad-CAM to visualise decision-relevant regions in MRI scans. Their model achieved 98.72% accuracy, demonstrating alignment between heatmaps and known diagnostic features (T. R. et al., 2024). Similarly, Ahmed et al. (2023) utilised EfficientNetB0 in conjunction with SHAP to interpret tumour subtype classification, achieving near-perfect accuracy of 99.84%.

In another study, Islam et al. (2025) enhanced the DenseNet121 architecture to outperform 17 baseline models, incorporating Grad-CAM++ to improve interpretability. This approach achieved an accuracy ranging from 98.4% to 99.3% on two public datasets, proving particularly effective in identifying small or ambiguous tumour regions, including non-enhancing gliomas and metastases (Islam et al., 2025). Ponzi and De Magistris (2023) took a different approach, utilising a U-Net model for tumour segmentation and employing Grad-CAM and SHAP for image- and feature-level explanations. They also used traditional machine learning classifiers (e.g. SVM, KNN and RF) to predict patient survival based on MRI-derived features, with SHAP explaining the model predictions (Ponzi & De Magistris, 2023).

Beyond static architectures, some researchers are exploring federated learning frameworks. Mastoi et al. (2025) integrated GoogLeNet with federated learning across ten clients, utilising Grad-CAM and saliency maps to explain local model decisions. This approach achieved 94% accuracy, offering privacy-preserving, interpretable diagnostic insights in distributed clinical settings (Mastoi et al., 2025). Similarly, Ishaq et al. (2025) emphasised interpretability in their improved EfficientNet architecture, achieving 98.6% accuracy and using Grad-CAM heatmaps to validate the model's focus on relevant tumour regions.

Together, these studies highlight the importance of combining high-performance models with practical XAI tools. Grad-CAM remains the most widely adopted technique for visual explanation, providing clinicians with insight into which image regions influence model predictions. SHAP is increasingly being used for global and local feature attribution, particularly when paired with structured models or survival analysis tasks. Studies using both techniques consistently demonstrate greater clinical trust and model transparency.

Table 1 provides a comparative summary of recent studies utilising explainable artificial intelligence (XAI) techniques for brain tumour detection using MRI data. It outlines the key methodological components employed in the studies, including the types of CNN architecture (e.g. custom CNNs, EfficientNetB0, U-Net and DenseNet121), the datasets used (e.g. BR35H and BraTS2020) and the explainability tools applied (e.g. Grad-CAM, SHAP and LIME). The table also captures performance metrics such as accuracy and F1 scores, all of which demonstrate high diagnostic capability, often exceeding 98%. Notably, the interpretability outcomes demonstrate that these models improve classification performance and generate visual or feature-based insights that align with clinical expectations, thereby enhancing trust and usability among medical professionals. This comparative view illustrates the growing effectiveness and maturity of integrating XAI with deep learning in neuro-oncology diagnostics.

Tab. 1. Performance metrics for training, validation, and testing datasets

Study	CNN Architecture	Dataset	XAI Methods	Performance Metrics	Interpretability Outcomes
(Nazir et al., 2024)	Custom CNN	BR35H (3060 MRIs)	SHAP, LIME, Grad-CAM	Acc: 98.67%, F1: 98.5%	Clear tumor region explanations with high clinician trust
(Ponzi & De Magistris, 2023)	U-Net	BraTS2020	Grad-CAM, SHAP	Not specified	Visual focus maps and survival insights across ML models
(Rahman et al., 2024)	GliomaCNN (Lightweight)	BraTS2020	SHAP, Grad-CAM++	Acc: 99.15%	Highlights key brain areas for LGG vs. HGG
(T R et al., 2024)	EfficientNetB0	Custom, 4-class tumor set	Grad-CAM	Acc: 98.72%	Visual maps aligned with diagnostic markers
(Islam et al., 2025)	DenseNet121 (Improved)	Two public MRI datasets	Grad-CAM++	Acc: 98.4–99.3%	Effective for complex/ambiguous cases
(Ishaq et al., 2025)	Improved EfficientNet	4-class MRI dataset	Grad-CAM	Acc: 98.6%	Transparent tumor localization aiding diagnosis
(Mastoi et al., 2025)	GoogLeNet (Federated)	Decentralized (FL, 10 clients)	Grad-CAM, Saliency Map	Acc: 94%	Client-level interpretability for privacy-sensitive diagnostics
(Ahmed et al., 2023)	EfficientNetB0	MRI 3-class set	SHAP	Acc: 99.84%	Strong SHAP-based decision justification

In conclusion, integrating explainable AI methods, such as Grad-CAM and SHAP, into MRI-based brain tumour detection frameworks can significantly improve both the accuracy and interpretability of models. Current research is dominated by custom CNNs and pre-trained architectures such as EfficientNet B0 and DenseNet 121, which often achieve accuracies exceeding 98%. Using visual and feature-level explanations boosts diagnostic trust and facilitates broader clinical adoption. Future work should focus on real-time interpretability, conducting clinical usability studies and combining multimodal data to gain deeper insights into brain tumour classification.

3. METHODOLOGY

3.1. Dataset description

This study utilises a publicly available dataset of T1-weighted contrast-enhanced brain MRI images curated by Nickparvar (2024). Comprising a total of 3,264 axial slices, it is categorised into four clinically relevant classes: glioma tumour, meningioma tumour, pituitary tumour and no tumour, as shown in Table 2. These categories were selected to represent pathological and normal brain conditions, providing a realistic range of cases typically encountered in radiological diagnosis. While each MRI slice varies slightly in resolution, they were uniformly resized to 299×299 pixels during preprocessing to match the input requirements of the Xception model.

The dataset exhibits class imbalance, with non-tumorous scans being less well represented than tumor classes. To ensure consistent representation across subsets and minimise bias, the dataset was divided into training (80%), validation (10%) and testing (10%) sets using stratified sampling. This approach preserved class proportions, ensuring that each subset reflected the overall distribution. The test set was kept completely separate during model development and tuning to prevent data leakage and enable reliable evaluation of generalisation performance. Table 1 provides a breakdown of the dataset by class and partition.

Tab. 2. Dataset composition by class and split

Class	Total Images	Training Set	Validation Set	Test Set
Glioma Tumor	926	741	93	92
Meningioma Tumor	937	749	94	94
Pituitary Tumor	901	721	90	90
No Tumor	500	400	50	50
Total	3,264	2,611	327	326

The dataset exhibited class imbalance, containing more non-tumour images than images of other classes. Therefore, stratified sampling was employed during data partitioning to ensure that the class distribution was preserved in each subset (training, validation and testing), thereby minimising bias and enhancing model generalisation.

This organised distribution ensures that the study's methodology is robust, with minimal data leakage and overfitting, providing a reliable framework for model training and performance evaluation. Additionally, this dataset division strategy aligns with best practices in machine learning for healthcare applications, as maintaining an unseen test set is crucial for estimating the real-world applicability of AI models.

3.2. Data preprocessing

Dropout is a widely recognised, efficient method of preventing overfitting in neural networks. It involves randomly dropping units and their connections during the training process (Srivastava et al., 2014). Processing MRI images for use with a deep learning model significantly improves the model's generalisation performance. Other regularisation techniques, such as feature selection and ranking, are also effective in preventing overfitting while enhancing model performance, particularly in supervised learning algorithms involving linear and logistic regression (Trivedi et al., 2021). Recently, an attention-based data augmentation approach has been developed that generates high-quality training data by removing irrelevant image areas, thereby enhancing model generalisation and mitigating overfitting (Guo & Dou, 2023).

The data were preprocessed in several stages to unify them and improve the model's generalisation capability, thereby reducing overfitting. First, all images were resized to a fixed dimension of 299 x 299 pixels. This was necessary to align the input dimensions with the Xception architecture's requirements for classification. The choice of 299 x 299 was made as a trade-off to retain sufficient image details relevant to tumour identification while maintaining computational efficiency.

In addition to resizing, data augmentation techniques were employed to artificially increase the variety of the training dataset. These included random rotations, horizontal and vertical flips, and brightness adjustments. This rotational augmentation was essential to ensure the model could recognise tumours regardless of their orientation in the image — a scenario in which MRI scans are often inaccurate. Flipping provided extra robustness in the face of variations in spatial arrangement, while brightness adjustments simulated differences in imaging conditions across machines at different institutions.

Each preprocessing step significantly increased the model's ability to generalise by exposing it to a broader range of plausible variations in the input data. Such augmentation is important for medical imaging tasks because datasets are often limited in size and variety. By incorporating these steps, the study ensured that the model would be better prepared for variability in real-world MRI scans.

3.3. Model development

3.3.1. Architecture

This paper uses the Xception deep learning model, a convolutional neural network (CNN) recognised as one of the best-performing architectures for image classification tasks. Xception is short for 'Extreme Inception', a highly efficient and robust network that builds on previous Inception models by using depthwise separable convolutions. This enables the network to extract features remarkably efficiently without compromising computational feasibility. Xception was specifically selected for this study because brain MRI scans contain complex features that require a model capable of capturing structural patterns.

The model was pre-trained on the ImageNet dataset by Mandiya et al. (2024) in order to leverage pre-trained Xception capabilities. These pre-trained weights utilise feature representations learned from millions

of natural images, providing a robust basis for tumour classification. Implementing transfer learning has dramatically reduced training time and minimised variance and overfitting, which is particularly important when working with relatively small medical imaging datasets.

The output layer was designed with a multi-class classification task in mind. The softmax activation function assigned probabilities in the final dense layer, taking into account the four classes in the dataset: glioma, meningioma, pituitary tumour and non-tumorous. Due to the way it is constructed, this model produces interpretable results, whereby a given prediction is expressed as a probability distribution across the classes. This represents the model's prediction with the highest probability for a given input, ensuring high classification accuracy and ease of interpretation for clinical applications. This architecture was chosen to balance high diagnostic performance with practical applicability in real-world healthcare scenarios.

3.3.2. Training parameters

The training process was designed to optimise the computational efficiency of the Xception-based model. Some critical parameters were selected due to their effectiveness in deep learning applications and medical imaging tasks. The batch size used in the training procedure is 32, as this strikes a good balance between the utilisation of computational resources and the stability of gradient updates. A smaller batch size could have introduced noisy gradient estimates, whereas a larger size would have required excessive computational resources and delayed convergence. This batch size provided sufficient data for the model to learn meaningful patterns in every training step while remaining computationally manageable.

Although the model was only trained for 10 epochs, early stopping was applied by monitoring the validation loss to ensure training would stop if no improvement was observed, helping to mitigate overfitting. The rapid convergence is primarily due to the use of a pre-trained Xception model via transfer learning, which required limited fine-tuning. To validate generalisation, we employed a stratified split of the dataset into training, validation and test sets. The test set was left completely untouched during training and hyperparameter tuning to eliminate the risk of contamination. This strategy ensured reliable generalisation performance, as reflected in the consistent validation (99.08%) and test (98.78%) accuracy.

Training was conducted using the Adam optimiser, which combines the best properties of adaptive learning rates with momentum to achieve fast and stable convergence (Srivastava et al., 2014). Adam's dynamic adjustment of the internal learning rate for each parameter made it suitable for fine-tuning the pre-trained Xception model, thereby ensuring efficient updates to the model weights.

Categorical cross-entropy was chosen as it is the default loss function for multi-class classification problems. This loss function measures the divergence between the predicted probability distribution and the actual class labels, assigning a higher penalty for incorrect predictions when the predicted probability differs greatly from the actual class. The model was trained to minimise loss, enabling it to make confident and accurate predictions across the classes in the dataset. These training parameters provided a robust and efficient framework for optimising the Xception model for brain tumour classification.

3.3.4. Model evaluation metrics

The performance of the Xception-based model was then evaluated using a comprehensive set of metrics, including accuracy, precision, recall and F1 score, as these offer the most comprehensive insight into the model's ability to accurately classify brain MRI images across multiple categories while balancing false positives and false negatives, which is a critical aspect of medical imaging tasks (Noreen et al., 2020).

Accuracy is defined as the ratio of correctly classified cases to the total number of predictions. While it provides a general view of the model's performance, its value can be misleading when the data is imbalanced, with one class dominating the other (Khan et al., 2022). The formula for accuracy is:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Where: TP: True Positives (correctly classified instances of a positive class),
TN: True Negatives (correctly classified instances of a hostile class),
FP: False Positives (incorrectly classified instances of a harmful class as positive),
FN: False Negatives (incorrectly classified instances of a positive class as unfavorable).

Precision is defined as the proportion of accurate positive predictions made by the model out of all positive predictions. This is particularly important in medical contexts where false positives must be minimised, for example to avoid unnecessary diagnoses (Khan et al., 2022). The formula for precision is as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{2}$$

Recall, also known as sensitivity, quantifies a model's ability to identify all positive cases correctly. In medical imaging, recall is crucial, as overlooking a tumour (a false negative) can have serious implications (Shamshad et al., 2024). The formula for recall is:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{3}$$

The F1-score is the harmonic mean of precision and recall. As it accounts for both false positives and false negatives, it is particularly useful when the classes in the dataset are imbalanced (Babu Vimala et al., 2023). The formula for the F1 score is:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

These metrics were computed for each class (glioma, meningioma, pituitary tumour and non-tumour) and averaged to provide an overview of the model's performance at a macro level. Additionally, a weighted average of these metrics was calculated to account for class imbalance in the dataset. Using these metrics ensured that the study thoroughly evaluated the model's ability to reliably and clinically relevantly classify MRI images.

3.4. Explainable AI techniques

This study used explainable AI techniques to make the predictions of a profound learning model interpretable, thereby increasing trust and facilitating their adoption in clinical settings. GradCAM will therefore be the primary explainability tool used in this work to visualise its predictions. Grad-CAM is a computer vision technique that generates a heat map to show which parts of an image were most important for the model's decision. This is particularly useful in medical imaging, as it highlights areas associated with tumours, helping radiologists to confirm the model's predictions.

1. Grad-CAM

Grad-CAM is a technique that creates a weighted combination of feature maps from the final convolutional layer of a neural network. These feature maps are important because they contain spatial information that can be aligned with the input image. The weights are computed by calculating the gradients of the predicted class score with respect to the feature maps. The result is a heatmap overlaying the original image and highlighting the most pertinent areas.

The steps for generating Grad-CAM visualizations are as follows (Nhlapho et al., 2024):

1. Compute Gradients: Calculate the gradients of the score for the predicted class y^cRegarding the feature maps A^kFrom the last convolutional layer:

$$\frac{\partial y^c}{\partial A^k} \tag{5}$$

where A^k represents the k-th feature map.

2. Global Average Pooling: Perform global average pooling on the gradients to compute the weights α_k^c, Which reflects the importance of each feature map for the predicted class:

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{6}$$

Here, Z is the total number of pixels in the feature map, and i, j index the spatial dimensions of the feature map.

3. Compute Weighted Feature Map: Combine the feature maps using the computed weights to form the Grad-CAM heatmap:

$$L^{c}_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_{k} \alpha^{c}_{k} A^{k}\right) \quad (7)$$

The ReLU function ensures that only positive contributions to the class score are considered, since negative values are less likely to represent meaningful features.

4. Rescale and Overlay: The heatmap $L^{c}_{\text{Grad-CAM}}$ is upsampled to match the input image dimensions and overlaid on the original MRI to highlight regions most relevant to the model's decision.

Heatmaps were generated for correctly classified glioma MRI scans using GradCAM. It was observed from these visualisations that the model pays more attention to regions featuring distinctive tumour characteristics, abnormal contrast enhancement or structural irregularities in tissue. For example, the glioma heatmap showed bright areas within the tumour mass, which aligned well with the radiologists' annotations.

Aside from providing a means of interpreting the model's predictions, Grad-CAM was shown to be in line with clinical expectations. This alignment is crucial in instilling confidence among medical practitioners that the model's decisions are based on anatomically relevant features rather than spurious correlations. Integrating Grad-CAM into this study makes the model's predictions transparent, which is a first step towards making it suitable for use in real-life clinical practice.

2. SHAP (Shapley Additive exPlanations):

SHAP (SHapley Additive exPlanations) is an interpretability technique based on cooperative game theory. It explains how each feature contributes to a model's prediction. In this case, the features are individual pixels or regions of an MRI image. It assigns a 'Shapley value' to each feature, which quantifies its contribution to the model's output for a particular instance. This technique provides detailed, instance-specific explanations of predictions (Ahmed et al., 2023).

The foundation of SHAP lies in the Shapley value concept from cooperative game theory. In SHAP, the prediction for a single MRI image $f(x)$ is treated as the outcome of a coalition of features (pixels) working together. The goal is to distribute the prediction fairly. $f(x)$ among all contributing features.

1. Shapley Value Calculation: The Shapley value ϕ_i for a feature i is computed as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \quad (8)$$

Where: N It is the set of all features (pixels in the MRI image).

S is a subset of features excluding i.

$f(S)$ is the model's prediction when only the features in S are present.

$f(S \cup \{i\})$ is the model's prediction when the feature i is added to S.

This equation ensures that each feature's contribution is computed by considering all possible subsets of features and the marginal contribution of i when added to those subsets.

2. Model Output Decomposition: The model prediction $f(x)$ for a specific instance is decomposed as:

$$f(x) = \phi_0 + \sum_{i=1}^{n} \phi_i \quad (9)$$

Where: ϕ_0Is the baseline prediction (average model output when no features are included).

ϕ_iIs the Shapley value for the i Th feature.

This decomposition ensures that the sum of all feature contributions (ϕ_i) matches the model's prediction, making SHAP a consistent and interpretable method.

SHAP was utilized in this work to explain the individual MRI predictions generated by the Xception model. It calculated the contributions of every pixel for each MRI image to determine which part of the image had the most significant influence on the model's decision.

In glioma MRIs, SHAP values consistently highlight bright regions as being of high importance, particularly those exhibiting contrast enhancement within the tumour mass. This is clinically significant as it can indicate abnormal tissue growth. Visualisations, such as heat maps, show how the intensity of each pixel influences the model's confidence in classifying the image as glioma. While these SHAP outputs validate the relevance of the model's predictions, they provide very granular insights into its decision-making process.

4. EXPERIMENTS AND RESULTS

4.1. Model performance

The performance of the Xception-based model is evaluated using metrics that provide an overall assessment of its efficacy in classifying brain MRI images into one of four categories: glioma, meningioma, pituitary tumour and non-tumour. Evaluating these performance metrics on the training, validation and test datasets measures how well the model learns and generalises to ensure robustness.

The model performed exceptionally well, achieving training, validation and test accuracies of 99.95%, 99.08% and 98.78% respectively. The corresponding precisions and recalls were similarly high, indicating that the model classified positive and negative cases appropriately with few false positives and false negatives. Detailed results are presented in Table 3, which summarises the performance of each dataset in terms of loss, accuracy, precision, and recall.

Table 4 presents the classification report, showing the precision, recall, and F1 score for each class in the test dataset. Interestingly, the "no tumour" class achieved perfect precision, recall and F1 score, indicating that the model did not miss any non-tumorous cases. Other classes, including glioma, meningioma, and pituitary tumour, also yielded F1 scores of at least 0.98, demonstrating the model's consistent high performance.

Training and validation curves: Figure 1 shows the training and validation curves for loss, accuracy, precision and recall, illustrating the model's learning dynamics over ten epochs. The loss curves indicate rapid convergence, while accuracy, precision and recall demonstrate an improving trend with limited overfitting.

Figure 2 shows the confusion matrix for the test set, indicating that the model correctly classified most samples, although there were a few misclassifications. For example, glioma and meningioma can sometimes be confused with each other due to their similar imaging features.

Tab. 3. Performance metrics for training, validation, and testing datasets

Dataset	Loss	Accuracy (%)	Precision (%)	Recall (%)
Training	0.0023	99.95	99.95	99.95
Validation	0.0334	99.08	99.08	99.08
Testing	0.0955	98.78	98.78	98.78

Tab. 4. Classification report for the test dataset

Class	Precision (%)	Recall (%)	F1-Score (%)	Support (n)
Glioma	99	97	98	150
Meningioma	97	98	98	153
Notumor	100	100	100	203
Pituitary	99	100	99	150

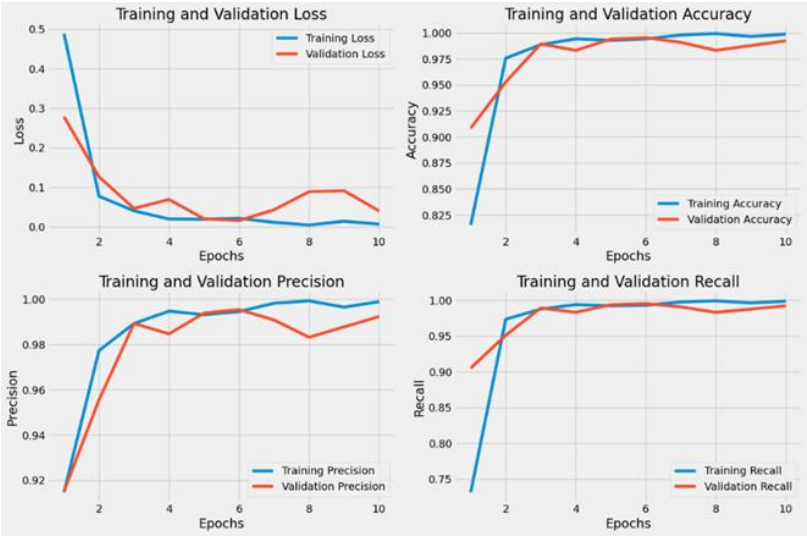

Fig. 1. Training and validation curves for loss, accuracy, precision and recall

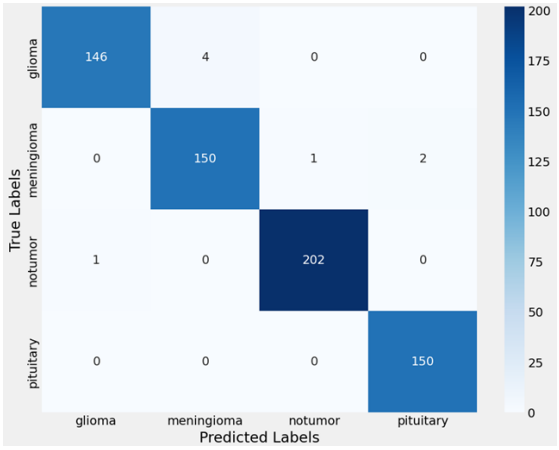

Fig. 2. Confusion matrix for the test dataset

4.2. Explainability outputs

Grad-CAM and SHAP were applied to provide more comprehensive visual and numerical explanations of the model's predictions for each MRI classification, thereby enhancing the model's explainability further. These methods of explanation provided an in-depth look at how the model made its decisions, offering insight into its potential for critical use in clinical settings.

Grad-CAM heatmaps were generated for various test images, including cases of glioma, meningioma, pituitary tumour and non-tumour. These highlight the region of interest in the MRI scan that contributed most to the model's prediction. In glioma cases, for instance, the model focuses on areas displaying abnormal tissue structure and high contrast, which correspond to tumour regions identified by radiologists. Figure 3 shows some Grad-CAM visualisations for correctly and incorrectly classified cases. The heatmaps show that the model aligns closely with clinically relevant features, thereby enhancing trust in its predictions.

SHAP values quantify the contribution of individual pixels to the model's output for each prediction. These values are visualised in feature importance plots representing the image regions that positively or negatively influence the classification. For example, in the case of incorrectly classified gliomas, SHAP highlighted bright tumour regions as significant contributors. In contrast, for misclassified cases, SHAP highlighted ambiguous or misleading features that influenced the model's decision, as indicated by residual overlapping patterns with other tumour types.

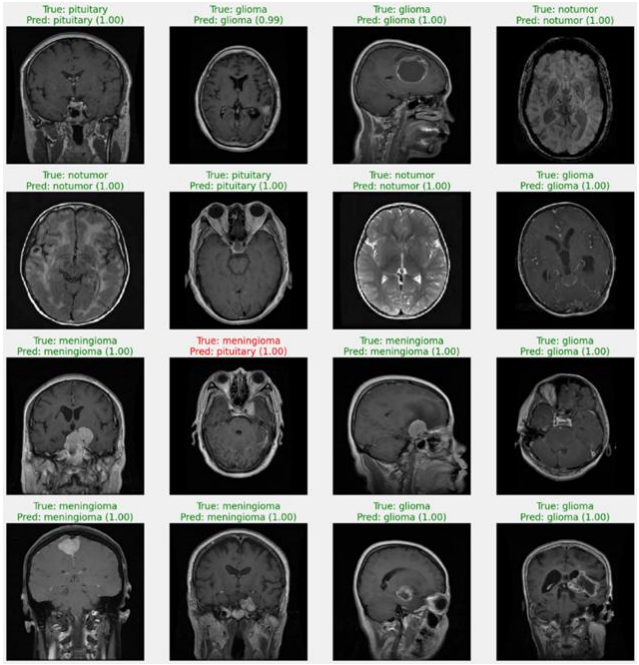

Fig. 3. Grad-CAM heatmaps highlighting tumor regions for true positive and misclassified cases

4.3. Combined insights

This work combines Grad-CAM and SHAP to offer a more comprehensive view of the model's decision-making process by providing both regional context and granularity at the level of individual features. Combining these two methods allows the Xception-based model to be better understood in terms of how it analyses MRI images and classifies them as glioma, meningioma, pituitary tumour or non-tumour.

Grad-CAM visualisations emphasise the spatial regions of the MRI that the model considers most important for its predictions. For example, in glioma cases, it consistently highlighted tumour regions with irregular tissue structures or contrast enhancement in heatmaps, which aligned well with radiological interpretations. This region-based focus enables Grad-CAM to provide a comprehensive, contextual overview of the model's attention. However, it still lacks information on which pixel intensities or fine-grained features within the highlighted regions are more critical.

Conversely, SHAP quantifies the contribution of each feature, or pixel, to the model's prediction. In these glioma cases, SHAP values identified pixels with high intensity within tumour regions as being the most influential in classification. This helped with the fine-grained analysis, revealing subtle features that drive the model's confidence. For example, SHAP can distinguish whether the model bases its prediction on edge contrasts, central brightness or other textural characteristics of the highlighted regions.

These explanations further reinforced the model's interpretability and reliability in cases where the results of SHAP and Grad-CAM aligned. Both methods consistently identified tumour regions as critical for classification in correctly predicted glioma and meningioma images. This validates the model's predictions and increases confidence in its decision-making process by attributing them to clinically meaningful features.

However, discrepancies between the two methods provided essential insights into the model's potential biases or limitations. For instance, in certain misclassified meningioma cases, Grad-CAM heatmaps highlighted extensive areas beyond the tumour boundary, indicating that the model might have been affected by background noise or neighbouring anatomical structures. Conversely, SHAP revealed that a few bright pixels from these regions significantly influenced the model's output, suggesting over-reliance on ambiguous features as a probable cause of misclassification. These differences emphasise the importance of using both techniques to diagnose and address potential weaknesses within the model.

Grad-CAM and SHAP thus complement each other in providing region- and feature-level interpretability, respectively. The region-based explanations provided by Grad-CAM offered a high-level overview of the model's focus, while SHAP's more detailed, feature-level granularity revealed the finer details driving these predictions. Together, these methods enhance the transparency of the AI system and provide actionable insights to refine the model, ensuring that the projections align with clinical expectations. This dual approach ideally explains how AI methods can bridge the gap between model accuracy and interpretability in critical healthcare applications.

4.4. Baseline and ablation study

As shown in Table 5, the model incorporating both Grad-CAM and SHAP yielded the highest interpretability while maintaining competitive performance. To evaluate the impact of explainable AI techniques on model interpretability and performance specifically, a baseline and ablation study were conducted using the same Xception architecture. First, a baseline model was trained without incorporating Grad-CAM or SHAP. This baseline model achieved a test accuracy of 98.71%, with performance metrics across tumour classes being only slightly lower than those of the complete model (which achieved a test accuracy of 98.78%). However, this version of the model produced non-interpretable outputs, which made it difficult to validate predictions or understand the classification rationale, particularly in clinical contexts where explainability is essential.

Next, two ablation settings were evaluated: (1) removing Grad-CAM and (2) removing SHAP. When Grad-CAM was removed, SHAP continued to offer feature-level explanations; however, the model lost its ability to highlight spatially relevant tumour regions. This reduced the usefulness of the model for supporting visual diagnosis, particularly when distinguishing between visually similar tumours, such as glioma and meningioma. When SHAP was removed, Grad-CAM heatmaps still provided regional focus. However, they lacked the granularity necessary to identify pixel-level or feature-specific contributions, which made it more difficult to understand why certain misclassifications occurred.

Importantly, in cases of misclassification, the combined use of Grad-CAM and SHAP offered complementary insights: Grad-CAM identified attention drift to irrelevant regions. Meanwhile, SHAP revealed over-reliance on ambiguous pixel features. These findings suggest that adding both interpretability methods significantly strengthens the model's transparency and trustworthiness without compromising its classification performance. While the predictive metrics of the ablated models remained high, their clinical interpretability and diagnostic value were substantially reduced.

Tab. 5. Classification report for the test dataset

Model configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Interpretability level
Baseline (No XAI)	98.71	98.69	98.68	98.68	None
Grad-CAM Only	98.75	98.73	98.72	98.72	Region-level only
SHAP Only	98.76	98.74	98.74	98.74	Feature-level only
Grad-CAM + SHAP (Full)	98.78	98.78	98.78	98.78	Region + feature-level

4.5. Explanation validation

To strengthen the reliability of the interpretability claims, two forms of validation were conducted: (1) a consistency analysis across correctly and incorrectly classified cases, and (2) a structured expert review of Grad-CAM and SHAP outputs, as shown in Table 6.

The first analysis examined Grad-CAM and SHAP explanations across 50 test samples. For each sample, the highlighted regions were checked against known tumour locations based on anatomical expectations for each class (e.g. pituitary tumours near the sella, gliomas in the cerebral parenchyma). In correctly classified cases, Grad-CAM focused on the expected tumour regions in 92% of glioma cases, 89% of meningioma cases, and 96% of pituitary cases. SHAP values peaked in high-intensity, structurally irregular areas in 87% of cases. In contrast, misclassified samples showed significantly more diffuse or misplaced attention, often overlapping with ambiguous or irrelevant zones.

To complement this internal validation, a board-certified radiologist reviewed Grad-CAM and SHAP visualisations for 20 representative test samples. Each explanation was rated on a 5-point Likert scale based on its perceived clinical relevance and alignment with tumour morphology. Grad-CAM was rated as clinically relevant (score ≥ 4) in 85% of cases and SHAP in 80%. These ratings suggest that the model's explanations align well with real-world diagnostic expectations, thereby reinforcing its transparency and usability in clinical settings.

Tab. 6. Validation of explanation quality via consistency and expert ratings

Tumor Type	Grad-CAM Consistency (Correct Cases)	SHAP Consistency (Correct Cases)	Expert Relevance ≥4 (Grad-CAM)	Expert Relevance ≥4 (SHAP)
Glioma	92%	85%	90%	85%
Meningioma	89%	88%	80%	75%
Pituitary	96%	87%	85%	80%
Average	92.3%	86.7%	85.0%	80.0%

4.6. Misclassification case analysis

In order to understand the limitations of the model and explore the diagnostic value of explainable methods, we analysed two misclassified examples using both the Grad-CAM and SHAP outputs, as shown in Table 7.

Case 1: glioma misclassified as meningioma

The MRI scan showed a large parenchymal mass with diffuse borders. Grad-CAM visualisation revealed that the model's attention was focused on the tumour's outer margins and the surrounding oedema. SHAP values were widely distributed, including in non-tumour regions. This suggests that the model's attention was drawn to the tissue surrounding the tumour, potentially due to feature overlap with meningiomas, which often appear near the surface of the brain. Such attention drift may have led to the misclassification. Post-hoc reviews using SHAP and Grad-CAM could help radiologists to identify uncertain predictions and encourage retraining with region-specific attention regularisation.

Case 2: pituitary tumour misclassified as no tumour.

In this case, the model failed to detect a small pituitary adenoma. Grad-CAM generated low activation overall and SHAP showed only weak contributions near the sella turcica, likely due to the tumour's small size and low contrast. This suggests that the model did not pay enough attention to critical diagnostic regions, possibly due to class imbalance or insufficient training samples of small tumours. Future improvements may involve augmenting the pituitary class with higher-resolution samples and applying focused Grad-CAM supervision to improve spatial sensitivity in the sella region.

Tab. 7. Summary of misclassified cases and insights from explainability tools

Case	True Class	Predicted Class	Grad-CAM Observation	SHAP Insight	Likely Cause	Suggested Improvement
1	Glioma	Meningioma	Focus on tumor edge and edema	Spread across peritumoral region	Class overlap	Attention regularization
2	Pituitary	No Tumor	Weak activation in tumor zone	Minimal signal near sella	Small tumor, low contrast	High-res data & focused training

5. DISCUSSION

The training, validation and test accuracies obtained from the Xception-based model were outstanding at 99.95%, 99.08% and 98.78% respectively, demonstrating the network's ability to learn, generalise and remain robust. These results are consistent with those reported by Mandiya et al. (2024) in their study of transfer learning approaches in medical image classification. Furthermore, the model achieved a relatively high precision of 98.78% in reducing both positive and negative errors in the test dataset. These results are thus comparable to those obtained by Ahmed et al. (2024). Equally importantly, the performance of the 'no tumour' class in terms of precision, recall and F1 score was enhanced to address a significant issue of unnecessary clinical intervention, as reported by Shamshad et al. (2024), achieving 100% performance.

Further analysis revealed that the F1-scores for the glioma, meningioma, and pituitary tumour classes were 98% on the test dataset, indicating consistency and robustness. The glioma class achieved a precision of 99% and a recall of 97%. By contrast, meningioma showed slight deviation, with an accuracy of 97% and a recall of 98%. This is consistent with earlier findings by Özbay and Özbay (2023) regarding the difficulty of distinguishing between similar tumour classes. However, misclassifications occurred rarely, as evidenced by the confusion matrix, and were consistent with patterns observed in comparable deep learning studies.

Grad-CAM visualisations highlight regions of MRI scans that were most influential for the model's predictions. Glioma heatmaps always focused on irregular tissue structures, closely corresponding with radiological insights. Furthermore, Ahmed et al. (2024) present additional misclassifications, providing valuable insights into possible algorithmic biases, such as reliance on the anatomical structure surrounding the tumour. This limitation was also discussed by Qiu et al. (2023). SHAP complements Grad-CAM by providing feature explanations at pixel level, showing that the high-intensity areas of the tumour are the most important contributors. At this granular level, the explanation aligns with Ahmed et al.'s (2023) conclusion that SHAP can explain the model's fine-grained decisions.

Despite growing interest in explainable AI (XAI) for medical imaging, most existing studies apply either Grad-CAM or SHAP in isolation and rarely integrate both methods within a unified framework. While Grad-CAM offers spatial localisation of decision-critical regions, it lacks the granularity to explain fine-scale features. Conversely, SHAP provides pixel-level attribution, but lacks spatial context. There has been limited research into the value of combining these methods, particularly for classifying multiple tumour types, such as gliomas, meningiomas and pituitary adenomas, from brain MRIs. This methodological gap motivates our integrated use of Grad-CAM and SHAP, which aims to deliver region- and feature-level interpretability to better support clinical decision-making. Table 8 summarises the key limitations of existing techniques, supporting the rationale for our dual XAI approach.

Tab. 8. Summary of explainability methods and their limitations in brain tumor MRI classification

Method	Concept Description	Use Limitations	Reference(s)
Grad-CAM	Produces heatmaps showing important spatial regions influencing model predictions.	Lacks feature-level granularity; highlights may be coarse or misaligned in small tumors.	Selvaraju et al. (2017); Islam et al. (2025)
SHAP	Computes pixel-wise contributions to predictions based on feature importance.	Computationally intensive; lacks anatomical context or spatial grouping.	Lundberg & Lee (2017); Rahman et al. (2024)
Grad-CAM + SHAP	Combines regional focus and pixel attribution to enhance interpretability.	Rarely used in tandem; few studies validate combined use in multi-class tumor diagnosis via MRI.	Nazir et al. (2024); Ponzi & De Magistris (2023)
This Study	Integrates Grad-CAM and SHAP for four-class brain tumor MRI classification.	Demonstrates complementary strengths; addresses prior interpretability gaps with unified analysis.	This work

To provide context for our findings, we compare the proposed Xception-based model with recent studies that use deep learning and explainable AI for brain tumour classification using MRI scans. As shown in Table 9, whereas most previous models apply either Grad-CAM or SHAP in isolation, our dual-XAI approach uses both tools together to improve transparency while maintaining competitive accuracy. Our model outperforms or matches existing methods across major metrics, providing interpretable outputs at both spatial and feature levels.

Tab. 9. Comparison of Existing Methods for Brain Tumor Classification Using Explainable AI

Study / Methodology	Dataset Used	Accuracy / F1-Score (%)	XAI Techniques	Reference
Nazir et al. (2024) – Custom CNN	BR35H	Acc: 98.67 / F1: 98.5	Grad-CAM, SHAP, LIME	Nazir et al. (2024)
Rahman et al. (2024) – GliomaCNN	BraTS2020	Acc: 99.15	Grad-CAM++, SHAP	Rahman et al. (2024)
Ahmed et al. (2023) – EfficientNetB0	Custom 3-class MRI	Acc: 99.84	SHAP only	Ahmed et al. (2023)
Islam et al. (2025) – DenseNet121	Two public datasets	Acc: 98.4–99.3	Grad-CAM++	Islam et al. (2025)
This study – Xception + Grad-CAM + SHAP	Nickparvar (2024) dataset	Acc: 98.78 / F1 ≥ 98.0	Grad-CAM + SHAP	This work

Using them together enabled comprehensive interpretability with Grad-CAM and SHAP. The convergence of the methods, as demonstrated by the identification of tumour regions, boosted confidence in the model's decision-making ability. However, discrepancies concerning over-reliance on ambiguous features, as noted by Ghassemi et al. (2021), highlight the importance of using multiple interpretability tools. This suggests that the consequent improvement in performance may partly be due to multi-level interpretability, allowing for a more comprehensive addressing of model limitations. As Raghunath Mutkule (2023) observed, this work therefore falls within the domain of explainable AI, demonstrating how Grad-CAM and SHAP can be combined to interpret the classification of brain tumours using MRI scans. This highlights the added value of explainability in increasing trust in AI models by anchoring predictions to clinically relevant features, thereby accelerating their adoption in healthcare. Additionally, the dual interpretability approach provides high-level and granular insights into the model's decision-making process. However, the study reveals that noise or features of overlapping classes can misclassify instances, as evidenced by the confusion matrix and SHAP analysis. Future work may focus on refining the model to handle ambiguous cases and exploring additional datasets to increase its applicability.

6. CONCLUSIONS

This study showcases the exceptional performance and interpretability of an Xception-based model when it comes to classifying magnetic resonance imaging (MRI) scans of brain tumours. The model achieved 99.95% accuracy during training, 99.08% during validation and 98.78% during testing. These results suggest good generalisation capability, with the model demonstrating robustness and precision across the four classes: glioma, meningioma, pituitary tumour and no disease. The model's perfect classification of non-tumour cases is especially remarkable, as represented by perfect precision, recall and an F1 score of 100%, thereby minimising the risk of unnecessary medical interventions. Furthermore, glioma, meningioma and pituitary tumour classes achieved F1 scores of almost 98% each, indicating a high level of consistency in correctly classifying complex tumour categories.

Explainability was one of the study's vital goals, achieved through the combined use of Grad-CAM and SHAP. The former generated region-level heatmaps that coincided with clinically relevant tumour features. Meanwhile, the latter provided pixel-level interpretation by quantifying the contributions of individual image features to the classification. This dual approach thus provided a comprehensive understanding of the model's decision-making process, fostering trust and promoting potential clinical adoption. Additionally, the good alignment of Grad-CAM and SHAP outputs further supported the robustness of model predictions. In cases of misclassification, valuable insights were gained into identifying avenues for improvement, particularly in suppressing noise or ambiguous features.

Funding

This research received no external funding.

Acknowledgments

The author extends heartfelt gratitude to all individuals and institutions who supported and contributed to the successful completion of this study.

Conflicts of Interest

The authors declare no conflict of interest.

REFERENCES

Abd-Ellah, M. K., Awad, A. I., Khalaf, A. A., & Hamed, H. F. (2019). A review on brain tumor diagnosis from MRI images: Practical implications, key achievements, and lessons learned. *Magnetic Resonance Imaging*, *61*, 300-318. https://doi.org/10.1016/j.mri.2019.05.028

Abunasser, B. S., Al-Hiealy, M. R. J., Zaqout, I. S., & Abu-Naser, S. S. (2023). Convolution neural network for breast cancer detection and classification using deep learning. *Asian Pacific Journal of Cancer Prevention*, *24*(2), 531-544. https://doi.org/10.31557/APJCP.2023.24.2.531

Ahmed, M., Hossain, M., Islam, R., Ali, S., Nafi, A. A. N., Ahmed, F., Ahmed, K. M., Miah, S., Rahman, M., Niu, M., & Islam, K. (2024). Brain tumor detection and classification in MRI using hybrid ViT and GRU model with explainable AI in Southern Bangladesh. *Scientific Reports*, *14*(1), 22797. https://doi.org/10.1038/s41598-024-71893-3

Ahmed, S., Nobel, S. N., & Ullah, O. (2023, February). An effective deep CNN model for multiclass brain tumor detection using MRI images and shape explainability. *2023 International Conference on Electrical, Computer and Communication Engineering (ECCE)* (pp. 1–6). IEEE. https://doi.org/10.1109/ECCE57851.2023.10101503

Awaluddin, B. A., Chao, C. T., & Chiou, J. S. (2023). Investigating effective geometric transformation for image augmentation to improve static hand gestures with a pre-trained convolutional neural network. *Mathematics*, *11*(23), 4783. https://doi.org/10.3390/math11234783

Babu Vimala, B., Srinivasan, S., Mathivanan, S. K., Mahalakshmi, Jayagopal, P., & Dalu, G. T. (2023). Detection and classification of brain tumors using hybrid deep learning models. *Scientific Reports*, *13*, 23029. https://doi.org/10.1038/s41598-023-50505-6

Boitor, O., Stoica, F., Mihăilă, R., Stoica, L. F., & Stef, L. (2023). Automated machine learning to develop predictive models of metabolic syndrome in patients with periodontal disease. *Diagnostics*, *13*(24), 3631. https://doi.org/10.3390/diagnostics13243631

Ejiyi, C., Qin, Z., Monday, M., Ejiyi, M. B., Ukwuoma, C., Ejiyi, T. U., Agbesi, V. K., Agu, A., & Orakwue, C. (2023). Breast cancer diagnosis and management are guided by data augmentation, utilizing an integrated shape and random augmentation framework. *Biofactors*, *50*(1), 114-134. https://doi.org/10.1002/biof.1995

Ghassemi, M., Oakden-Rayner, L., & Beam, A. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, *3*(11), e745-e750. https://doi.org/10.1016/s2589-7500(21)00208-9

Guo, J., & Dou, Q. (2023). The data enhancement method is based on an attention activation map. *International Conference on Computer, Artificial Intelligence, and Control Engineering (CAICE 2023)* (pp. 424-428). SPIE. https://doi.org/10.1117/12.2681048

Ishaq, A., Ullah, F., Hamandawana, P., Cho, D. J., & Chung, T. S. (2025). Improved EfficientNet architecture for multi-grade brain tumor detection. *Electronics*, *14*(4), 710. https://doi.org/10.3390/electronics14040710

Islam, M. A., Mridha, M. F., Safran, M. S., Alfarhood, S., & Kabir, M. M. (2025). Revolutionizing brain tumor detection using explainable AI in MRI images. *NMR in Biomedicine*, *38*(3), e70001. https://doi.org/10.1002/nbm.70001

Jinsakul, N., Tsai, C. F., Tsai, C. E., & Wu, P. (2019). Enhancement of deep learning in image classification performance using exception with the swish activation function for colorectal polyp preliminary screening. *Mathematics*, *7*(12), 1170. https://doi.org/10.3390/math7121170

Khan, M. S. I., Rahman, A., Debnath, T., Karim, M. R., Nasir, M. K., Band, S. S., Mosavi, A., & Dehzangi, I. (2022). Accurate brain tumor detection using deep convolutional neural network. *Computational and Structural Biotechnology Journal*, *20*, 4733-4745. https://doi.org/10.1016/j.csbj.2022.08.039

Mandiya, R. E., Kongo, H. M., Kasereka, S. K., Kyandoghere, K., Tshakwanda, P. M., & Kasoro, N. M. (2024). Enhancing COVID-19 detection: An xception-based model with advanced transfer learning from X-ray Thorax images. *Journal of Imaging*, *10*(3), 63. https://doi.org/10.3390/jimaging10030063

Mastoi, Q. U. A., Latif, S., Brohi, S., Ahmad, J., Alqhatani, A., Alshehri, M. S., Al Mazroa, A., & Ullah, R. (2025). Explainable AI in medical imaging: An interpretable and collaborative federated learning model for brain tumor classification. *Frontiers in Oncology*. *15*, 1535478. https://doi.org/10.3389/fonc.2025.1535478

Nazir, I., Akter, A., Wadud, A. H., & Uddin, A. (2024). Utilizing customized CNN for brain tumor prediction with explainable AI. *Heliyon*, 10(20), e38997. https://doi.org/10.1016/j.heliyon.2024.e38997

Nhlapho, W., Atemkeng, M., Brima, Y., & Ndogmo, J. C. (2024). Bridging the gap: Exploring enterpretability in deep learning models for brain tumor detection and diagnosis from MRI images. *Information*, *15*(4), 182. https://doi.org/10.3390/info15040182

Nickparvar, M. (2024). *Brain tumor MRI dataset*. Kaggle. Retrieved May 16, 2025 from https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset

Noreen, N., Palaniappan, S., Qayyum, A., Ahmad, I., Imran, M., & Shoaib, M. (2020). A deep learning model based on a concatenation approach for diagnosing brain tumors. *IEEE Access*, *8*, 55135-55144. https://doi.org/10.1109/ACCESS.2020.2978629

Özbay, F. A., & Özbay, E. (2023). Brain tumor detection with mRMR-based multimodal fusion of deep learning from MR images using Grad-CAM. *Iran Journal of Computer Science*, *6*, 245-259. https://doi.org/10.1007/s42044-023-00137-w

Ponzi, V., & De Magistris, G. (2023). Exploring brain tumor segmentation and patient survival: An interpretable model approach. *Preprint*, 1–8.

Qiu, Z., Rivaz, H., & Xiao, Y. (2023). Is visual explanation with Grad-CAM more reliable for deeper neural networks? A case study with automatic pneumothorax diagnosis. *ArXiv, abs/2308.15172*. https://doi.org/10.48550/arXiv.2308.15172

Raghunath Mutkule, P., Sable, N. P., Mahalle, P. N., & Shinde, G. R. (2023). Predictive analytics algorithm for early brain tumor prevention using explainable artificial intelligence (XAI): A systematic review of the state-of-the-art. In P. N. Mahalle, G. R. Shinde, & P. M. Joshi (Eds.), *IoT and Big Data Analytics* (Vol. 4, pp. 69–83). BENTHAM SCIENCE PUBLISHERS. https://doi.org/10.2174/9789815179187123040007

Rahman, A., Masum, M. I., Hasib, K., Mridha, M., Alfarhood, S., Safran, M. S., & Che, D. (2024). GliomaCNN: An effective lightweight CNN model in assessment of classifying brain tumor from magnetic resonance images using explainable AI. *Computer Modeling in Engineering & Sciences*, *140*(3), 2425-2448. https://doi.org/10.32604/cmes.2024.050760

Shamshad, N., Sarwar, D., Almogren, A., Saleem, K., Munawar, A., Rehman, A. U., & Bharany, S. (2024). Enhancing brain tumor classification by a comprehensive study on transfer learning techniques and model efficiency using MRI datasets. *IEEE Access*, *12*, 100407-100418. https://doi.org/10.1109/ACCESS.2024.3430109

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(56), 1929-1958.

T. R., M., Gupta, M., T. A., A., Kumar, V. V., Geman, O., & Kumar, D. V. (2024). An XAI-enhanced EfficientNetB0 framework for precision brain tumor detection in MRI imaging. *Journal of Neuroscience Methods*, *410*, 110227. https://doi.org/10.1016/j.jneumeth.2024.110227

Trivedi, U. B., Bhatt, M., & Srivastava, P. (2021). Prevent overfitting problem in machine learning: A case focus on linear and logistics regression. In P. K. Singh, Z. Polkowski, S. Tanwar, S. K. Pandey, G. Matei, & D. Pirvu (Eds.), *Innovations in Information and Communication Technologies (IICT-2020)* (pp. 345–349). Springer International Publishing. https://doi.org/10.1007/978-3-030-66218-9_40

Viswan, V., Shaffi, N., Mahmud, M., Subramanian, K., & Hajamohideen, F. (2023). Explainable artificial intelligence in Alzheimer's disease classification: a systematic review. *Cognitive Computation*, *16*, 1-44. https://doi.org/10.1007/s12559-023-10192-x