Comparative evaluation of transfer learning models and Grad-CAM interpretability for brain tumor detection from MRI

Md. Firoz Hasan[1], Md. Awal Hadi[1], Sumaiya Nasrin[1], Md. Raisul Islam[1], Md. Atik Shahriar[1], Md. Hasan Moon[1], Tanvir Ahmed Momin[1], Dewan Mamun Raza[1]

Department of Computer Science and Engineering, Faculty of Science and Information Technology, Daffodil International University, Dhaka, Bangladesh

Article Info

Article history:

Received Jan 12, 2025
Revised Jun 9, 2026
Accepted Jul 9, 2026

Keywords:

Brain tumor classification
Explainable artificial intelligence
Healthcare artificial intelligence
Medical imaging
Transfer learning

ABSTRACT

Brain tumor classification plays an important role in early diagnosis and treatment planning. The current study aims to evaluate and compare the performance of five pre-trained convolutional neural network (CNN) models, namely VGG16, VGG19, MobileNet, Xception, and InceptionV3 using magnetic resonance imaging (MRI) images categorized into glioma, meningioma, pituitary tumor, and no tumor classes. To enhance model performance and address class imbalance, transfer learning and data augmentation techniques were employed. To boost model interpretability, heatmaps of important areas in tumor classification were produced through gradient-weighted class activation mapping (Grad-CAM). MobileNet was the most accurate with 97% and was more precise and more sensitive. The Grad-CAM visualizations showed the models were attending to clinically relevant features, which increased the interpretability. This comparative study demonstrates the effectiveness of the integration of explainable artificial intelligence (XAI) in deep learning pipelines for reliable brain tumor diagnosis.

Corresponding Author:

Md. Firoz Hasan
Department of Computer Science and Engineering, Faculty of Science and Information Technology
Daffodil International University
Daffodil Smart City, Birulia, Savar, Dhaka, Bangladesh
Email: firoz.cse@diu.edu.bd

1. INTRODUCTION

Brain tumor classification is an important task in the area of medical imaging research because of its significant role in early diagnosis and treatment planning [1]. Traditional methods relying on manual interpretation are often subjective and inconsistent [2]. This has prompted the development of artificial intelligence (AI), especially deep learning, to improve the accuracy and reliability of diagnosis [3].

Convolutional neural networks (CNNs) have shown excellent performance for the analysis of complex magnetic resonance imaging (MRI) data. Recently, pre-trained models such as VGG16, VGG19, MobileNet, Xception, and InceptionV3 have been applied with high accuracy for the tasks of tumor classification [4], [5]. These models have been further improved through transfer learning, which allows efficient feature extraction from smaller medical datasets [6], [7]. However, challenges remain to ensure high generalizability across different tumor types [8]. The present study aims to overcome such limitations by comparing multiple fine-tuned CNN models for the classification of glioma, meningioma, pituitary tumors, and no tumor cases [9] using data augmentation and oversampling to handle class imbalance [10].

The study also includes explainable artificial intelligence (XAI) in the form of gradient-weighted class activation mapping (Grad-CAM) heatmaps to highlight the key areas that influence predictions, thus increasing interpretability for clinicians. MobileNet achieved the best performance with 97% accuracy and it also excelled in precision and recall. The Grad-CAM outputs confirmed the reliability of the model by indicating the important regions of the images that contributed to its decisions.

The reviewing literature highlights rapid progress and increasing impact of deep learning techniques in medical image classification, specifically for detection of brain tumors. Bouhafra and El Bahi [11], reviewed 60 studies (2020-2024) and highlighted the effectiveness of methods such as transfer learning, autoencoders, transformers, and attention mechanisms in enhancing diagnostic precision and computational efficiency, despite challenges such as data heterogeneity and limited interpretability. Bhagyalaxmi *et al.* [12] emphasized the usefulness of deep learning in multi-grade tumor classification. Agarwal *et al.* [13] reported 99.54% accuracy for brain tumor detection using ResNet50. Neamah *et al.* [10] similarly identified some of the key challenges of CNN-based approaches, such as the need for large annotated datasets and high model complexity. Due to their ability to automatically learn complex features [14], [15], CNNs have shown great success in distinguishing the types of tumors as compared to traditional machine learning algorithms, such as glioma, meningioma, and pituitary. By using transfer learning, Improved CNN models such as GoogLeNet and ResNet50 achieved accuracies of 99.67% and 99.91%, respectively with better performance and convergence [16], [17].

Some studies also looked into the grading of gliomas, which shows the versatility of deep learning in classifying brain tumors. The standard classification pipeline consists of preprocessing of MRI images, feature extraction and classification algorithms [18]. Methods like discrete wavelet transform (DWT) and grey level co-occurrence matrix (GLCM) are useful for preprocessing and texture analysis, and principal component analysis (PCA) is used to reduce the dimensionality of features without loss of accuracy. Support vector machines (SVMs) have been shown to be highly effective for both binary and multi-class classification tasks with the right kernels [19]–[21]. Besides brain tumor classification, medical image classification has been extended to other modalities such as MRI, computed tomography (CT), X-ray, and fundus imaging using models like SVMs, Bayesian networks, neural networks, k-nearest neighbors (KNN), and fuzzy logic [22]. Standard classification pipelines are composed of data acquisition, preprocessing, feature selection, and model training [23]. The usual techniques are texture analysis, neural networks, and data mining, and the evaluation is usually based on accuracy, sensitivity, specificity, and runtime [24], [25]. Collectively, these studies highlight the potential of deep learning for improving the accuracy, efficiency, and clinical utility of brain tumor diagnosis, while acknowledging the ongoing challenges associated with data variability and computational requirements.

The study shows the transformative potential of AI-driven solutions for medical diagnostics. It helps to build strong, accurate, and interpretable brain tumor classification systems that can support radiologists and relieve diagnostic workloads in clinical settings by combining sophisticated deep learning models with explainability techniques. Future work will focus on expanding the dataset, exploring novel architectures, and improving interpretability methods to further improve classification accuracy and clinical applicability.

2. METHOD

2.1. Dataset description

The brain tumor classification dataset includes MRI images of four classes: glioma, meningioma, no tumor, and pituitary tumor. This classification facilitates supervised learning by associating each image with its label and thus allowing the models to recognize patterns specific to each type of tumor. The dataset contains 22,848 images. The number of images for each tumor is 6,380 images for no tumor, 5,828 images for pituitary, 5,356 images for meningioma, and 5,284 images for glioma.

2.2. Dataset preprocessing

2.2.1. Resizing and rescaling

Resizing and rescaling normalize the input images and help in computational efficiency. Inconsistent input shapes due to different image sizes can make model training more difficult. Rescaling pixel intensity values to a common scale improves numerical stability and speeds up convergence during training of the model. For deep learning models such as VGG19, Inception, and MobileNet, all the images are resized to 224×224 pixels. This guarantees a uniform input dimension for the dataset, which is necessary for batch processing and efficient memory usage. Images are rescaled such that pixel values range from 0 to 1 to normalize them. The intensity of each pixel is scaled to [0, 1] by dividing by 255 (the maximum pixel intensity value in an 8-bit image). This prevents large input values, which can hinder the learning of the neural network.

2.2.2. Label encoding

The model requires label encoding to convert the categorical labels (glioma, meningioma, and no tumor pituitary) to numeric values to process. Machine learning models expect numeric inputs, so we will convert the labels to integer values that the model can work with. Each category is assigned a separate integer using 'LabelEncoder' from 'Sklearn' library. For example, glioma can be encoded as 0, meningioma as 1, no tumor as 2, and pituitary as 3. This mapping from a categorical label to its encoded integer value is expressed in (1).

$$Y_{encoded} = f(y_{category}) \tag{1}$$

Where $y_{category}$ represents the categorical label (e.g., 'glioma') and f is the function that maps it to an integer value $Y_{encoded}$.

2.2.3. Data augmentation

Data augmentation can increase the diversity of the training set without the need to collect more data. It also helps the model to generalize better (by training on different transformations of the same image). Augmentation techniques such as shifting, zooming, and flipping create new image variations which reduce overfitting and improve robustness.

The image is translated by a ratio of the width in horizontal shifting. For example, if the width shift range is 0.2, the images can be shifted left or right by up to 20% of the image width. This augmentation strategy follows established practice for improving deep learning generalization on limited datasets [26]. This horizontal shifting operation is expressed in (2).

$$x' = x + \alpha \times width \tag{2}$$

Here x is the original horizontal position, and α is a random factor in the range of shift. Vertical shifting translates the image up or down by a fraction of the image height. The range is 0.2, meaning that images can be shifted up or down by up to 20% of the image height. This vertical shifting operation is expressed in (3).

$$y' = y + \beta \times height \tag{3}$$

Where y is the original vertical position and β is a random factor in the range of the shift. Zooming in or out alters the perceived distance of the image. The zoom ratio is 0.2, which means that the image can be zoomed in or out by up to 20%. This zooming transformation is expressed in (4).

$$I_{zoomed} = resize\ (I, (1 \pm Y) \times original\ size) \tag{4}$$

Where Y is a random factor in the zoom range. This method reverses the image horizontally, creating a mirror image. Horizontal flipping assists the model in learning invariance to orientation. This horizontal flipping operation is expressed in (5).

$$I_{flipped} = flip(I, axis = 1) \tag{5}$$

2.2.4. Class balancing (oversampling)

Class balancing is a technique to handle the imbalances in the sample sizes among models to avoid biased prediction. Oversampling duplicates minority instances to ensure that all classes are equally represented. This study use the "RandomOverSampler" from the "Imblearn" library to duplicate samples from under-represented classes until all classes have equal numbers. It helps to increase the size of the dataset while keeping the class balance, thus reducing the possibility of the model favoring any one class. This oversampling strategy is consistent with standard approaches for handling class imbalance in machine learning pipelines [27]. This oversampling process, which equalizes the number of samples across all classes, is expressed in (6).

$$n_i = N_{max} \tag{6}$$

Where n is the new number of samples for class i, equalised among all classes. These preprocessing techniques are combined to effectively prepare the dataset, enabling better model performance and more robust results.

2.3. Xception model with attention mechanism

Xception is a deep CNN [28], which is efficient in extracting complex patterns. It replaces the standard Inception modules with depthwise separable convolutions, which decouple spatial and channel-wise feature learning to capture visual patterns more efficiently. After the feature extraction, the outputs of Xception are reshaped to feed into a multi-head attention layer. This attention mechanism is useful to focus on specific regions in the brain MRI scans that are more likely to contain information about the tumor. The multi-head attention mechanism used here follows the formulation introduced for transformer-based architectures [29], as expressed in (7).

$$Attention\ Output = softmax\ (QK^T/\sqrt{d_k})V \tag{7}$$

Where Q, K, and V are the vectors of the queries, keys, and values, respectively, and d_k is the dimension of the key vectors. The CNN model uses a global average pooling layer to reduce the spatial dimensions while retaining feature information. It also uses dropout layers for regularisation, a softmax output layer for classification into four classes and a multi-head attention layer for enhanced feature extraction. It utilizes dropout layers, Gaussian noise and batch normalization and rectified linear unit (ReLU) activations to prevent overfitting and to retain the feature information. Sparse categorical cross-entropy loss and Adam optimizer are used for model training. The output obtained by combining multiple attention heads is expressed in (8).

$$Attention\ Output = MultiHeadAttention\ (Q, K, V) \tag{8}$$

Where Q, K, and V mechanisms represent the queries, keys and values matrices used in attention models. The Adam optimizer and categorical cross-entropy are used to minimize classification errors. This combination helps to have stable and effective gradient updates. The Adam optimizer is widely adopted for its adaptive learning rate and fast convergence properties [30]. The sparse categorical cross-entropy loss function used to train the model is expressed in (9).

$$Loss\ =\ -\ \sum_{i=1}^{N} y_i.\ log(\hat{y}_i) \tag{9}$$

Where $\hat{y}_i$ is the predicted probability for class i and y_i is the true label. The training is monitored and early stopping is used to stop training if there is no improvement in validation loss over a set number of epochs. This prevents overfitting by returning to the weights that were seen best during training.

2.4. Grad-CAM heatmap visualization

Grad-CAM [31] is a visual explanation method for model predictions that highlights the most important areas of the image for the predicted class, providing insights into which image regions the model deems important. The Grad-CAM localization map, which highlights the region's most relevant to the predicted class, is expressed in (10).

$$L^c_{Grad-CAM} = \Delta I(\sum_k a^c_k A_k) \tag{10}$$

Where A_k are the feature maps of the last convolutional layer and αc are the weights which represent the importance of each feature map. This method is a holistic approach for brain tumor classification, which includes feature extraction, attention mechanisms, and regularisation to improve the model's accuracy, interpretability, and robustness. Each model is trained for the brain tumor classification task by adding multi-head attention mechanisms and regularisation layers on pre-trained ImageNet weights. The models are trained to classify four types of brain tumors: glioma, meningioma, no tumor, and pituitary. Grad-CAM outputs visual explanations of model decisions by visualizing parts of the image most relevant to the predicted class. The corresponding heatmap obtained from these weighted feature maps is expressed in (11).

$$Heatmap = ReLU\ (\sum a_k A_k) \tag{11}$$

Where a_k are the weights obtained by averaging the gradients and A_k are the feature maps of the last convolutional layer. This methodology provides a complete deep learning method for brain tumor classification. Each phase is focused on improving the accuracy, avoiding overfitting, and understanding the decision-making process of the model.

2.5. VGG16 model

VGG16 and VGG19 [32] are pre-trained models that have stacked convolutional layers, which makes them more efficient in capturing local and detailed features in images. They are refined with a multi-head attention layer, regularization by batch normalization and Gaussian noise to improve model robustness. Dense layers with 512 units (ReLU activation) allow learning complex patterns and dropout layers prevent overfitting. The sparse categorical cross-entropy loss function is minimized using the Adam optimizer, which measures the difference between the true labels and the predicted probabilities. In this work, the VGG16 model with the pretrained weights on ImageNet is used as a feature extractor, the layers are frozen during training. VGG19 model like VGG16 has more layers to increase the depth and ability to extract features. Batch normalization and dropout are well-established regularization techniques for stabilizing training and reducing overfitting in deep networks [33], [34]. The output computed by the dense layer within this architecture is expressed in (12).

$$Dense\ Output = ReLU(Wx + b) \tag{12}$$

The pre-trained layers are frozen during training, so that the model can focus on the new tasks and it applies global average pooling, dense layers, Gaussian noise, and dropout to a final softmax classification layer.

2.6. MobileNet model

This study use MobileNet [35] with pre-trained weights and non-trainable features to build the model. Then, the output is passed to a multi-head attention layer, followed by global average pooling, dense layer, batch normalization, Gaussian noise, dropout, and softmax layer to output the probabilities. The loss function used to train the MobileNet-based model is expressed in (13).

$$Loss = -\sum_{i=1}^{N} y_i . log(\hat{y}_i) \tag{13}$$

This study uses the Adam optimizer for stable learning and attention mechanisms to focus on tumor-relevant regions of the image. Batch normalization reduces internal covariate shift, which stabilizes the learning process. The Gaussian noise layers add random noise to the data which helps the network to generalize. Regularization is a technique of randomly picking neurons during the training for overfitting reduction. The combination of batch normalization, Gaussian noise, and dropout improves generalization and prevents overfitting. This ensures robustness and prevents overfitting by ignoring neurons during training The model performance for different tumor classes is evaluated using important classification metrics such as precision, recall, F1-score, and accuracy. These metrics are calculated based on the analysis of true positives, false positives, true negatives, and false negatives for each class. Precision is the proportion of correctly predicted positive cases, recall is the effectiveness of the model in identifying positive cases, F1-score is a balanced assessment, and accuracy is the overall correctness of the model.

3. RESULTS AND DISCUSSION

Figure 1 was created to visualize the class distribution, showing the imbalance in the dataset. This visualization revealed the need of balancing techniques as the classes were imbalanced. As shown in Figure 1, the no-tumor class is the largest group (6,380 images), followed by pituitary (5,828), meningioma (5,356), and glioma (5,284), a difference of about 1,100 images between the largest and smallest classes that motivated the oversampling strategy described in subsection 2.2.5.

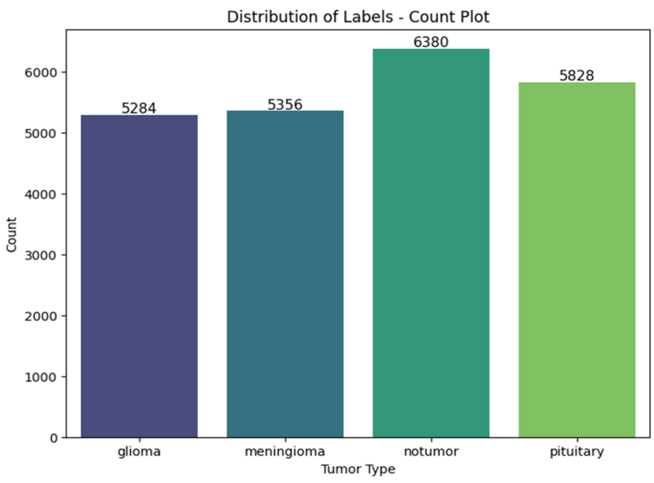


Figure 1. Distribution of class labels

The dataset in Figure 2 was structured into a DataFrame with each row as an image path and label. Labels were encoded numerically to be compatible with model training. Random oversampling was used to tackle the class imbalance problem. After oversampling, each class had the amount of 6,380 images. The study has examined CNN for classifying brain MRI images into 4 categories: glioma, meningioma, no tumor, and pituitary. Transfer learning was applied by using the pre-trained CNN models, which have already learned the patterns from the large-scale datasets. The models VGG16, VGG19, MobileNet, Xception, and InceptionV3 were tested. The models were optimized with Adam optimizer and a sparse categorical cross-entropy loss function for multi-class classification.

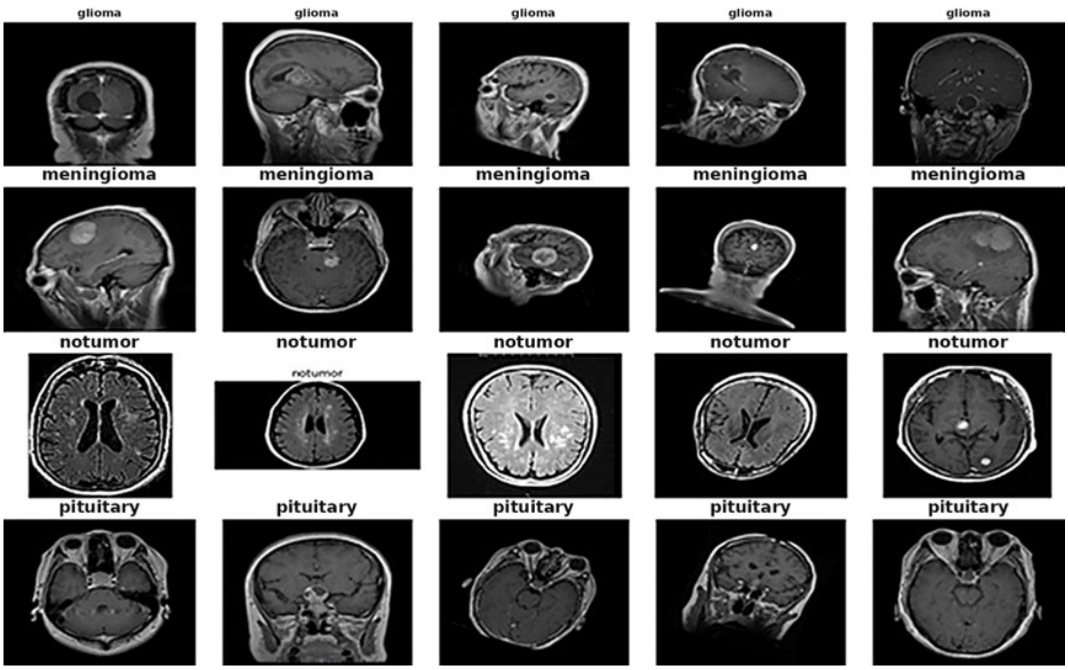


Figure 2. Visual representation of images of different sizes

The objective is to evaluate the performance of various CNN architectures in the classification of brain tumors. Transfer learning was used to take advantage of pre-trained CNN models to achieve the best performance with a small dataset. Each model was resized to the input layers and fine-tuned with the input size (224, 224, 3) pixels. To enhance the generalization of the model, data augmentation operations like horizontal flipping, zooming, and width/height shifting were performed. All the models were trained for 5-10 epochs with early stopping to prevent overfitting.

Table 1 shows the brain tumor classification with multiple metrics. Accuracy, precision, recall, and F1-score are the metrics of this study. The best model, MobileNet, attained 97% accuracy and F1-score which is acceptable for real-time applications. Following closely, VGG19 achieved an accuracy of 92%, showing its ability to capture complex MRI patterns. A simpler version VGG16 had a slightly lower performance with accuracy 91%. Xception was as accurate as VGG16, but slightly less precise in differentiating some tumor types. The lowest performance was achieved by InceptionV3 with 85% accuracy, possibly due to the complexity and the requirement of larger datasets. InceptionV3 showed visible misclassification among similar tumor types like glioma and meningioma, probably due to the visual similarity in Figure 3 of these tumors in MRI images. MobileNet had the highest accuracy of 97% that means that almost all of the tumor images were classified correctly. VGG19 was also quite good with an accuracy of 92%. InceptionV3 was a bit behind with an accuracy of 85% which indicates that it might not have generalized well from the dataset.

Table 1. Classification result for applied model

Model	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
VGG16	0.91	0.91	0.91	91
VGG19	0.92	0.92	0.92	92
MobileNet	0.97	0.97	0.97	97
Xception	0.91	0.91	0.90	91
InceptionV3	0.87	0.85	0.84	85

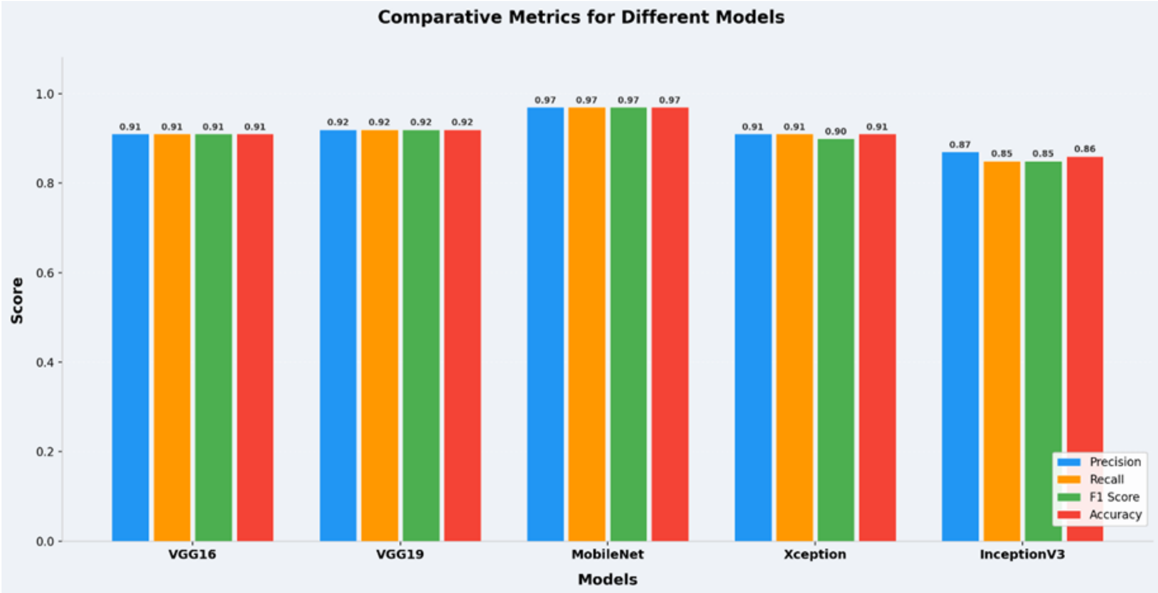
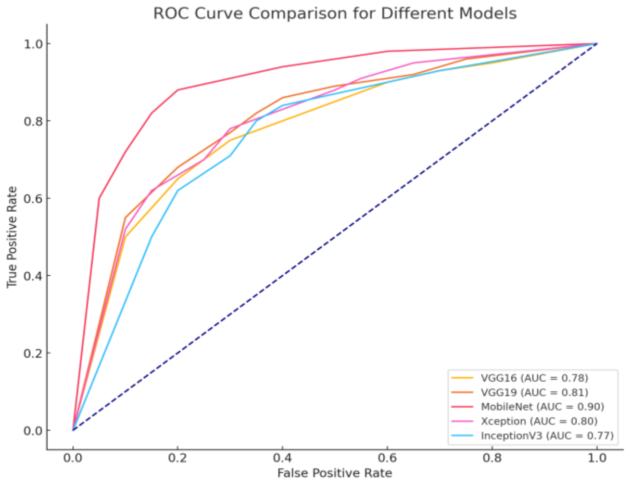


Figure 3. Visualization on precision, recall, F1-score, and accuracy

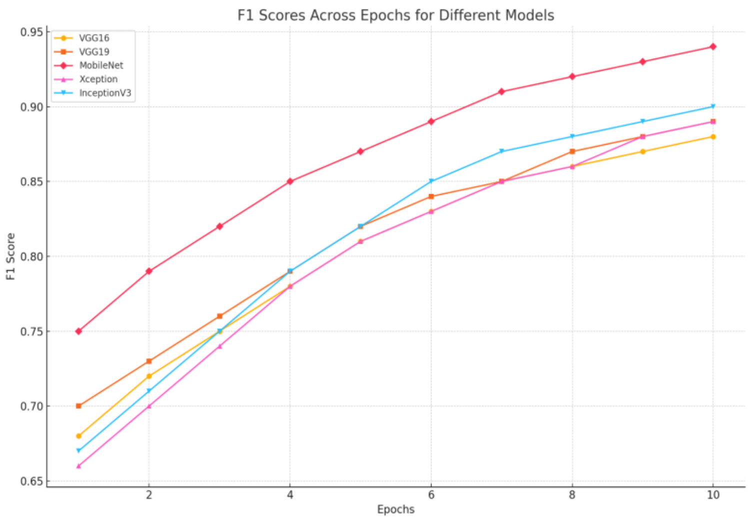
The MobileNet and VGG19 models achieved high precision and recall in predicting the types of tumors, thereby reducing the risk of misclassification and ensuring reliable identification of tumors. The MobileNet and VGG19 models had high recall values, showing that they were able to correctly identify most cases of each type of tumor. But InceptionV3 had recall issues, particularly in discriminating meningioma because it is sensitive to complex patterns and small datasets. F1-score: a combination of precision and recall to know how well a model can avoid false positives and false negatives. MobileNet has the highest F1-score of 0.97 which is a good balance for capturing relevant cases while also avoiding misclassification. VGG19 came next with a high F1-score of 0.92, confirming its reliability for practical applications. InceptionV3 did not perform consistently well in terms of precision and recall, especially in distinguishing visually similar tumor classes like glioma and meningioma, with a relatively lower F1-score. The confusion matrices offered more insight into the misclassification patterns of the models.

Figure 4 shows the model performance evaluation for VGG16, VGG19, MobileNet, Xception, and InceptionV3; where Figure 4(a) shows the receiver operating characteristic (ROC) curve comparison and Figure 4(b) shows the scores across epochs for different models. The ROC curves are used to evaluate the performance of the models in discriminating between classes. The ROC curves of VGG16 and VGG19 are moderate increasing, indicating a good sensitivity and specificity with a solid model performance. MobileNet has a steep curve near the origin, achieving a high true positive rate with small increments of false positive rate and has the highest area under the curve (AUC) of all the models. This shows that MobileNet is good at discriminating between classes and also has high recall and precision with minimal trade-off on false positives. Xception's ROC curve is moderately steep with a steady increase but a little lower AUC than MobileNet. InceptionV3's curve also increases steadily but less steeply at the beginning, indicating a moderate result, where it gets a lot of true positives but possibly with a higher false positive rate early on. The ROC curve and AUC of MobileNet show that it is the most capable model for this classification task. VGG16, VGG19, Xception, and InceptionV3 also perform well with slight differences in the balance of sensitivity and specificity. AUC is a good measure to compare the performance of models in classification problems. This gives a balanced view of the model's ability to discriminate, which is useful if you want to choose a model that does well at all thresholds, and thus be robust and reliable in the real world. If class separation is needed, a model with a better AUC can be often the best option, such as MobileNet. ROC and AUC remain standard tools for evaluating and comparing classifier performance across thresholds [35].

Figure 5 is a line chart that shows the training and validation loss in each model (VGG16, VGG19, MobileNet, Xception, and InceptionV3) over epochs. MobileNet is a suitable model for tasks needing high classification accuracy and a balance between precision and recall. It also has the least training time indicating efficiency in computational resources. VGG16 and VGG19 are consistently improving the precision, recall and F1-score, with moderate AUC values and low tendency of overfitting.



(a)



(b)

Figure 4. Model performance evaluation of (a) ROC curve comparison and (b) scores across epochs for different models

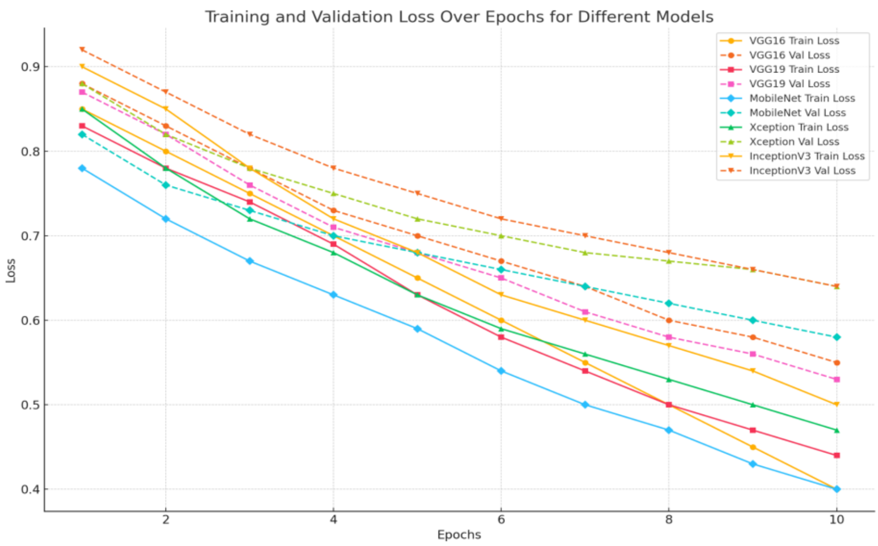


Figure 5. Training and validation loss over epochs for different models

Xception does well but seems a bit overfitted. InceptionV3 obtains reasonable scores for all metrics in Figure 6 but it has the lowest precision and recall among models, indicating that it can be further improved by tuning. MobileNet is chosen because of its efficiency and high performance. Misclassification happens when the model predicts the tumor category of an image wrongly as glioma and meningioma.

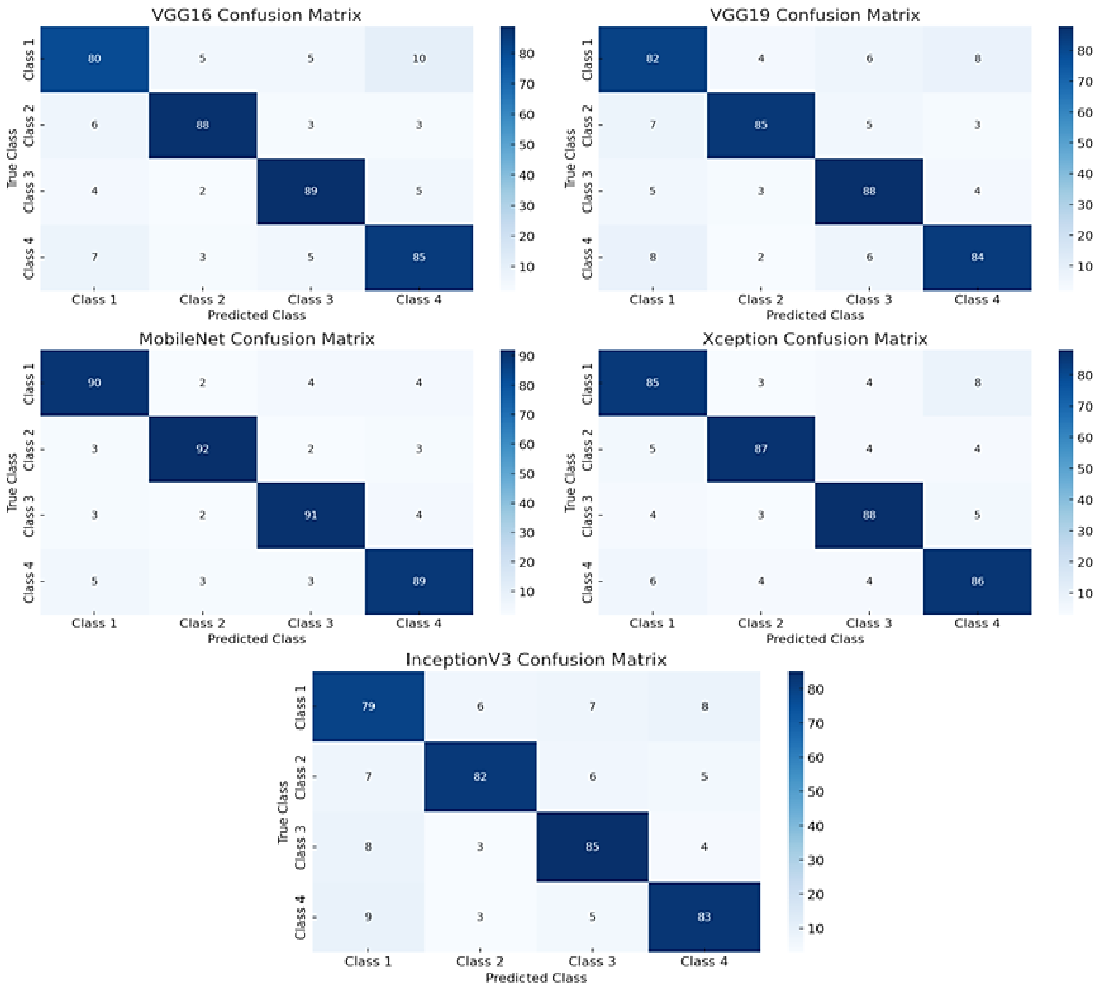


Figure 6. Confusion matrix for different models

In Table 2, cases where the models failed to correctly classify are listed. For each entry, the image ID, the true tumor type, the incorrect prediction, the model used, and the possible reason for the error are provided. All these reasons are visual judgements. Typical problems are that the tumor shapes or sizes are similar and that the image has artefacts that may mislead the model. To better understand these misclassifications, let’s take a look at some of these images. Here is an example of a few selected misclassified images with true and predicted labels shown under each image to help visualize the areas of confusion for each model.

Table 2. Misclassified image summary table

Image ID	True label	Predicted label	Model
3000.jpg	Glioma	Meningioma	InceptionV3
1979.jpg	Meningioma	Glioma	VGG16
4421.jpg	No tumor	Pituitary	MobileNet
5254.jpg	Pituitary	Glioma	VGG19
2599.jpg	Glioma	Meningioma	MobileNet

Figure 7 presents these misclassified MRI images without Grad-CAM overlays, giving a baseline view of the raw inputs before the model's attention regions are visualized in Figure 8. The true and predicted labels for each misclassified image are shown, with the specific regions where the model’s focus might have led to an incorrect prediction. The main source of misclassification in this study is the visual similarity of tumor types and ambiguous image features. Overlapping textures and irregular boundaries of glioma and meningioma often confused them. Frequent false predictions as glioma were made for meningioma images that had blur or indistinctness. In some cases, non-tumorous images with faint artefacts were incorrectly identified as pituitary tumors, suggesting the need for better feature discrimination. Small and localized pituitary tumors were also misclassified as glioma, illustrating limitations in spatial recognition. The results indicate that further preprocessing, feature localization, and model tuning are required to enhance the classification accuracy.

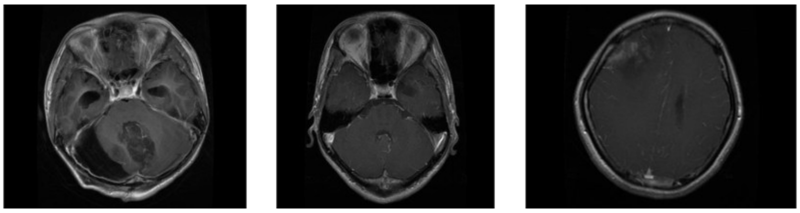


Figure 7. Image without Grad-CAM

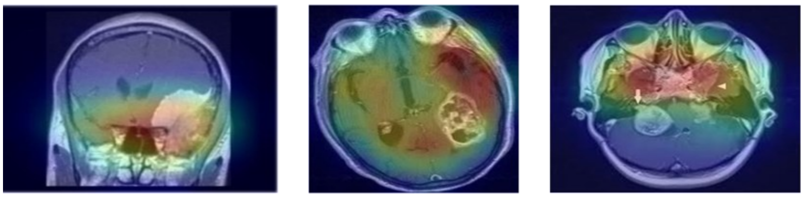


Figure 8. Image after applying Grad-CAM

Misclassified images depict limitations of a model. For example, MRI images of glioma and meningioma can have similar visual traits, which caused InceptionV3 and VGG16 to misclassify glioma as meningioma. These errors may be due to convolutional layers treating shared textures and edges as similar features. The MobileNet architecture was not able to perform well in the presence of slight artefacts and should be adjusted for the sensitivity to texture and shape, perhaps by adjusting the training procedure or the attention mechanisms. Grad-CAM is a method for providing visual explanations of decisions made by deep learning models by utilizing the regions of an image that are most relevant to the model’s classification. An experiment on brain tumor classification with Grad-CAM visualizations provides a heatmap overlay on each image, indicating regions relevant to classifying the tumor types. This interpretability is important for medical applications, as it guarantees that the model attends to the relevant anatomical structures and does not make misclassifications. Grad-CAM heatmaps can also uncover patterns in the model’s decision process that could suggest changes in the model’s sensitivity or preprocessing to better detect true tumor features. This interpretative tool allows developers to identify and fix specific attention errors, thereby increasing the reliability of the model. With the aid of XAI methods such as Grad-CAM in Figure 8, the transparency and trustworthiness of deep learning models can be improved, especially in healthcare. Grad-CAM provides insight into the decision-making process of the model and helps in validating the clinical accuracy of the model and identifying errors that may arise due to inherent limitations or challenges of the dataset. Such interpretability reduces risks in medical diagnosis, where misclassification of tumors may lead to inappropriate treatments. This visual overlap between glioma and meningioma cases has been similarly reported in other multi-class brain tumor classification studies [36], while systematic reviews of XAI in medical imaging confirm that Grad-CAM-based interpretability is among the most widely adopted approaches for building clinician trust in deep learning diagnostic tools [37].

Grad-CAM visualizations highlight the differences in attention, giving actionable suggestions to enhance the model's accuracy and reliability. Now, with the Grad-CAM visualizations integrated into the model evaluation process, developers can see the model’s spatial attention directly and an iterative feedback loop arises. In conclusion, Grad-CAM and XAI techniques are important to enhance deep learning models for medical applications.

4. CONCLUSION

The research finds that deep learning models, especially MobileNet, with 97% accuracy, significantly improve the classification of brain tumors in medical diagnostics. The study emphasizes the significance of model interpretability via Grad-CAM heatmaps, which provide visual insights into AI decision-making and build clinician trust. The study shows the potential to improve diagnostic accuracy and efficiency by integrating these models, which would ultimately benefit radiologists and reduce their workload. The study concludes that future work should focus on expanding datasets, exploring advanced architectures, and making them interpretable to improve the therapeutic usefulness of AI-driven diagnostic tools.

ACKNOWLEDGMENTS

The authors wish to extend our sincere thanks to the Binary Intelligence Lab at Daffodil International University for their invaluable support in the preparation of datasets and conducting laboratory work.

FUNDING INFORMATION

This research received no external funding.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Md. Firoz Hasan	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	
Md. Awal Hadi		✓				✓		✓	✓	✓	✓	✓		✓
Sumaiya Nasrin	✓		✓	✓			✓			✓	✓		✓	✓
Md. Raisul Islam		✓			✓	✓		✓	✓					✓
Md. Atik Shahriar			✓		✓		✓		✓	✓		✓		✓
Md. Hasan Moon	✓			✓		✓				✓	✓			✓
Tanvir Ahmed Momin		✓		✓	✓				✓				✓	
Dewan Mamun Raza	✓	✓	✓		✓	✓		✓		✓				✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

Informed consent was not applicable because this study used publicly available, anonymized data obtained from the online repository and did not involve direct interaction with human participants.

ETHICAL APPROVAL

Ethical approval was not required for this study because it used publicly available datasets obtained from the online repository. The study did not involve human participants, animals, clinical data collection, or any intervention requiring institutional ethical review.

DATA AVAILABILITY

The datasets used in this study are publicly available. It can be accessed through their respective public repositories. The data supporting the findings of this study are available from these public repositories and can be obtained without restriction.

REFERENCES

[1] M. W. Nadeem *et al.*, "Brain tumor analysis empowered with deep learning: a review, taxonomy, and future challenges," *Brain Sciences*, vol. 10, no. 2, 2020, doi: 10.3390/brainsci10020118.
[2] F. J. Dorfner, J. B. Patel, J. K.-Cramer, E. R. Gerstner, and C. P. Bridge, "A review of deep learning for brain tumor analysis in MRI," *npj Precision Oncology*, vol. 9, no. 1, 2025, doi: 10.1038/s41698-024-00789-2.
[3] T. Berghout, "The neural frontier of future medical imaging: a review of deep learning for brain tumor detection," *Journal of Imaging*, vol. 11, no. 1, 2025, doi: 10.3390/jimaging11010002.
[4] M. A. G.-Guzmán *et al.*, "Classifying brain tumors on magnetic resonance imaging by using convolutional neural networks," *Electronics*, vol. 12, no. 4, 2023, doi: 10.3390/electronics12040955.
[5] S. Deepak and P. M. Ameer, "Brain tumor classification using deep CNN features via transfer learning," *Computers in Biology and Medicine*, vol. 111, 2019, doi: 10.1016/j.compbiomed.2019.103345.
[6] O. Özkaraca *et al.*, "Multiple brain tumor classification with dense CNN architecture using brain MRI images," *Life*, vol. 13, no. 2, 2023, doi: 10.3390/life13020349.
[7] R. R. Ali *et al.*, "Learning architecture for brain tumor classification based on deep convolutional neural network: classic and ResNet50," *Diagnostics*, vol. 15, no. 5, 2025, doi: 10.3390/diagnostics15050624.
[8] D. R. Sarvamangala and R. V. Kulkarni, "Convolutional neural networks in medical image understanding: a survey," *Evolutionary Intelligence*, vol. 15, no. 1, pp. 1–22, 2022, doi: 10.1007/s12065-020-00540-3.
[9] S. Pendhari, N. Pendhari, and S. Shroff, "Benchmarking deep learning models for automated MRI-based brain tumor detection: in-depth analysis of CNN, VGG16, VGG19, ResNet-50, MobileNet, and InceptionV3," *International Journal of Computer Applications*, vol. 186, no. 50, pp. 31–35, 2024, doi: 10.5120/ijca2024924233.
[10] K. Neamah *et al.*, "Brain tumor classification and detection based DL models: a systematic review," *IEEE Access*, vol. 12, pp. 2517–2542, 2024, doi: 10.1109/ACCESS.2023.3347545.
[11] S. Bouhafra and H. El Bahi, "Deep learning approaches for brain tumor detection and classification using MRI images (2020 to 2024): a systematic review," *Journal of Imaging Informatics in Medicine*, vol. 38, no. 3, pp. 1403–1433, 2025, doi: 10.1007/s10278-024-01283-8.
[12] K. Bhagyalaxmi, B. Dwarakanath, and P. V. P. Reddy, "Deep learning for multi-grade brain tumor detection and classification: a prospective survey," *Multimedia Tools and Applications*, vol. 83, no. 25, pp. 65889–65911, 2024, doi: 10.1007/s11042-024-18129-8.
[13] S. Agarwal, D. Jain, S. Gupta, S. Sawhney, and Y. Mittal, "Brain tumor detection and classification using deep learning," in *IEEE 2023 5th International Conference on Advances in Computing, Communication Control and Networking, ICAC3N 2023*, 2023, pp. 635–640, doi: 10.1109/ICAC3N60023.2023.10541584.
[14] V. Anand, A. Khajuria, R. K. Pachauri, and V. Gupta, "Multi-class classification of brain tumors using optimized CNN and transfer learning techniques," *Scientific Reports*, vol. 16, no. 1, 2026, doi: 10.1038/s41598-025-34806-6.
[15] M. Z. Khaliki and M. S. Başarslan, "Brain tumor detection from images and comparison with transfer learning methods and 3-layer CNN," *Scientific Reports*, vol. 14, no. 1, pp. 1–12, 2024, doi: 10.1038/s41598-024-52823-9.
[16] A. Raza *et al.*, "A hybrid deep learning-based approach for brain tumor classification," *Electronics*, vol. 11, no. 7, 2022, doi: 10.3390/electronics11071146.
[17] S. K. Mathivanan, S. Sonaimuthu, S. Murugesan, H. Rajadurai, B. D. Shivahare, and M. A. Shah, "Employing deep learning and transfer learning for accurate brain tumor detection," *Scientific Reports*, vol. 14, no. 1, pp. 1–12, 2024, doi: 10.1038/s41598-024-57970-7.
[18] A. Kumar, S. K. Pandey, N. Varshney, K. U. Singh, T. Singh, and M. A. Shah, "Distinctive approach in brain tumor detection and feature extraction using biologically inspired DWT method and SVM," *Scientific Reports*, vol. 13, no. 1, 2023, doi: 10.1038/s41598-023-50073-9.
[19] L. Hussain, S. Saeed, I. A. Awan, A. Idris, M. S. A. Nadeem, and Q.-A. Chaudhry, "Detecting brain tumor using machines learning techniques based on different features extracting strategies," *Current Medical Imaging Formerly Current Medical Imaging Reviews*, vol. 15, no. 6, pp. 595–606, 2019, doi: 10.2174/1573405614666180718123533.
[20] G. Dheepak, A. C. J, and D. Vaishali, "Brain tumor classification: a novel approach integrating GLCM, LBP and composite features," *Frontiers in Oncology*, vol. 13, 2023, doi: 10.3389/fonc.2023.1248452.
[21] F. Khan, Y. Gulzar, S. Ayoub, M. Majid, M. S. Mir, and A. B. Soomro, "Least square-support vector machine based brain tumor classification system with multi-model texture features," *Frontiers in Applied Mathematics and Statistics*, vol. 9, 2023, doi: 10.3389/fams.2023.1324054.
[22] A. Aljuaid and M. Anwar, "Survey of supervised learning for medical image processing," *SN Computer Science*, vol. 3, no. 4, May 2022, doi: 10.1007/s42979-022-01166-1.
[23] E. Zarenia, A. A. Far, and K. Rezaee, "Automated multi-class MRI brain tumor classification and segmentation using deformable attention and saliency mapping," *Scientific Reports*, vol. 15, no. 1, 2025, doi: 10.1038/s41598-025-92776-1.
[24] M. Barstuğan, "An effective flowchart for multimodal brain tumor binary classification with ranked 3D texture features," *Scientific Reports*, vol. 15, no. 1, 2025, doi: 10.1038/s41598-025-11240-2.
[25] E. I. Zacharaki *et al.*, "Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme," *Magnetic Resonance in Medicine*, vol. 62, no. 6, pp. 1609–1618, 2009, doi: 10.1002/mrm.22147.
[26] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0197-0.
[27] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017.
[28] F. Chollet, "Xception: deep learning with depthwise separable convolutions," in *30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017, vol. 2017, pp. 1800–1807, doi: 10.1109/CVPR.2017.195.
[29] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.
[30] D. P. Kingma and J. L. Ba, "Adam: a method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.
[31] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Sep. 2014, *arXiv: 1409.1556*.
[33] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
[34] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *32nd International Conference on Machine Learning, ICML 2015*, 2015, vol. 1, pp. 448–456.
[35] A. G. Howard *et al.*, "MobileNets: efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.
[36] A. Gorenshtein, T. Liba, and A. Goren, "Lightweight transfer learning models for multi-class brain tumor classification: glioma, meningioma, pituitary tumors, and no tumor MRI screening," *Journal of Imaging Informatics in Medicine*, vol. 39, no. 3, pp. 2309–2318, 2026, doi: 10.1007/s10278-025-01686-1.
[37] F. Ahmed, N. S. Naz, S. Khan, A. U. Rehman, W. M. Ismael, and M. A. Khan, "Explainable artificial intelligence (XAI) in medical imaging: a systematic review of techniques, applications, and challenges," *BMC Medical Imaging*, vol. 26, no. 1, 2026, doi: 10.1186/s12880-025-02118-w.

BIOGRAPHIES OF AUTHORS

Md. Firoz Hasan obtained his B.Sc. degree in Computer Science and Engineering from Chittagong University of Engineering and Technology. He is currently pursuing a master in Information Technology at Jahangirnagar University. He serves as a senior lecturer in the Department of Computer Science and Engineering at Daffodil International University. His research interests include natural language processing, machine learning, deep learning, digital image processing, and computer vision. He can be contacted at email: firoz.cse@diu.edu.bd.

Md. Awal Hadi is a Computer Science and Engineering student at Daffodil International University, admitted in Spring 2022. He serves as a campus ambassador for the CDC Club and is interning in Computer Networking and Cybersecurity at Goinnovior Limited. He is also involved in medical science research and works as a student advisor at Skill Jobs. He can be contacted at email: hadi15-5345@diu.edu.bd.

Sumaiya Nasrin received her B.Sc. degree in Computer Science and Engineering from Daffodil International University in 2021. She is currently serving as an IT officer at Uttara Science and Technology Institute, where she began her career as a junior IT officer in 2022 and was promoted in 2024. Her interests include IT systems, technology, and research, with a strong aspiration to contribute to future innovations. She can be contacted at email: sumaiyasristy2017@gmail.com.

Md. Raisul Islam received his B.Sc. degree in Computer Science and Engineering from Daffodil International University, Bangladesh. He is currently pursuing an M.Sc. in Information and Communication Technology at Rajshahi University of Engineering and Technology. His research interests include artificial intelligence, machine learning, cybersecurity, data science, software engineering, and human–computer interaction. He can be contacted at email: raisul15-3813@diu.edu.bd.

Md. Atik Shahriar received his B.Sc. degree in Computer Science and Engineering from Daffodil International University, Bangladesh. He is currently working as an AI specialist at Augmedix Bangladesh and has previous experience as a software engineer and web developer. He also serves as a UN Volunteer. His research interests include artificial intelligence, cybersecurity, data science, machine learning, and image and signal processing. He can be contacted at email: atik15-9990@diu.edu.bd.

Md. Hasan Moon received his B.Sc. degree in Computer Science and Engineering from Daffodil International University in 2021. He is currently serving as a research and administrative assistant and has previous experience as a lab associate, web developer, IT in-charge, programming counselor, and front-end developer. His interests include software development, research, and educational support. He can be contacted at email: hasan15-9768@diu.edu.bd.

Tanvir Ahmed Momin received holds a B.Sc. in Computer Science and Engineering from Daffodil International University, Bangladesh. He is passionate about AI, cybersecurity, software engineering, and embedded systems, with hands-on experience in Python, TensorFlow, Node.js, React, and MongoDB. His interests focus on developing secure and intelligent systems for sectors like healthcare, e-commerce, and smart infrastructure. He can be contacted at email: tanvirdhaka4@gmail.com.

Dewan Mamun Raza is pursuing his Ph.D. degree at the Department of Computer Science and Engineering in Jahangirnagar University. He obtained his B.Sc. and M.Sc. degree in Computer Science from Daffodil International University. Currently, he is serving as an assistant professor at the Department of Computer Science and Engineering in Daffodil International University. His research interest is in natural language processing, especially artificial neural networks, machine learning, and cyber security. He can be contacted at email: dewan.cse@diu.edu.bd.