Journal of
Computer & Electrical and Electronics Engineering Sciences
Original Article
DOI: 10.51271/JCEEES-0040

Explainable brain tumor classification using an attention-augmented CNN and Grad-CAM

Maral Anwer Mustafa*[1,2], Osman Ayhan Erdem[3], Esra Söğüt[3]
[1]Department of Computer Engineering, Graduate School of Natural and Applied Sciences, Gazi University, Ankara, Turkiye
[2]Department of Production Mechanics Techniques, Kirkuk Technical Institute, Northern Technical University, Kirkuk, Iraq
[3]Department of Computer Engineering, Faculty of Technology, Gazi University, Ankara, Turkiye

Cite this article as: Mustafa, M. A., Erdem, O. A., & Söğüt, E. (2026). Explainable brain tumor classification using an attention-augmented CNN and Grad-CAM. *J Comp Electr Electron Eng Sci, 4*(2), 77-82. https://doi.org/10.51271/JCEEES-0040

Received: 12.01.2026 ◆ **Accepted**: 21.05.2026 ◆ **Published**: 13.10.2026

ABSTRACT

Aims: This paper explores clinical explainability of deep neural network architectures for classifying brain neoplasms using magnetic resonance imaging (MRI) scans.

Methods: We evaluated multiple architectures, including a baseline convolutional neural network (CNN), a novel attention-augmented CNN, and pre-trained transfer learning models (MobileNetV2, InceptionV3, and Xception). Models were trained using the Adam optimizer on a quaternary-class MRI dataset comprising glioma, meningioma, pituitary adenoma, and healthy cases. Performance was assessed using quantitative metrics (accuracy, precision, sensitivity, and F1-score), while interpretability was evaluated via gradient-weighted class activation mapping (Grad-CAM) visualizations.

Results: MobileNetV2 demonstrated superior performance, achieving a classification accuracy of 91.7%, with highly consistent precision, sensitivity, and F1-scores. The attention-augmented CNN achieved competitive efficacy, while Grad-CAM heatmaps consistently localized pathologically significant regions, thereby confirming the diagnostic reliability and transparency of the models' decision-making processes.

Conclusion: Computationally efficient pre-trained architectures like MobileNetV2 are highly effective for MRI-based neoplasm classification. Furthermore, integrating attention mechanisms with Grad-CAM visualization significantly enhances both diagnostic accuracy and model transparency, highlighting their strong potential for adoption in real-world clinical decision support systems.

Keywords: MRI, deep learning, brain tumor, transfer learning, explainable AI

INTRODUCTION

Deep neural network (DNN) models have proven highly effective for brain tumor classification using magnetic resonance imaging (MRI) and have substantially advanced diagnostic imaging research (Ebiele et al., 2020). Despite these improvements, their clinical adoption remains limited because of persistent concerns regarding model precision, reliability, and transparency. Deep learning systems are often criticized for their inherent opacity, which is commonly described as a black-box problem. This lack of transparency reduces clinicians' confidence in algorithmic predictions and underscores the need for frameworks that offer not only strong predictive performance but also interpretable diagnostic explanations (Brima et al., 2021).

Explainable Artificial Intelligence (XAI) methods are model-agnostic interpretability frameworks designed to provide localized explanations for the predictions of complex machine learning models (Atagün et al., 2026). Recent studies have further emphasized that XAI has become a critical requirement for the reliable adoption of deep learning in medical imaging, particularly in clinically sensitive tasks. Current literature highlights that transparency in medical image analysis is necessary not only to improve clinician trust, but also to support accountability, interpretability, and safer decision-making. Studies have also shown that explainability in medical imaging is not limited to saliency-based visualizations, but may also include textual justifications, and comparative evaluation of explanation quality across methods. In addition, different XAI techniques may vary substantially with some methods providing stronger confidence increase and more clinically meaningful localization of abnormal regions than others. These findings indicate that explainability should be considered a central design objective in AI-based medical imaging systems rather than an optional post hoc add-on (Chaddad et al., 2025; Houssein et al., 2025; Haupt et al., 2025).

Recent studies on brain tumor MRI analysis have increasingly emphasized the joint use of high-performing deep learning architectures and XAI techniques to improve both predictive accuracy and clinical transparency. Recent multiclass classification studies have integrated hybrid deep learning frameworks with Grad-CAM, LIME, and SHAP to

*Corresponding Author: Maral Anwer Mustafa, 24833304401@gazi.edu.tr

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

generate interpretable explanations while achieving strong performance on public brain MRI datasets. Other works have proposed unified diagnostic pipelines combining classification, preprocessing enhancement, and XAI-based visual interpretation, showing that explainability can support the clinical usability of accurate MRI-based tumor detection systems. In addition, explainable deep learning models have also been extended to tumor localization and segmentation tasks, where Grad-CAM and backpropagation-based visualizations were used to improve transparency and support expert interpretation. Collectively, these studies indicate that current brain tumor MRI research increasingly prioritizes not only classification success, but also interpretable and clinically meaningful model behavior (Hasan et al., 2026; Lamrani et al., 2026; Kumar et al., 2026).

Different pre-trained deep learning models have been tested for brain tumor MRI classification using gradient-based attribution methods such as gradient-weighted class activation mapping++ (Grad-CAM++) (Chattopadhyay et al., 2018), Grad-CAM (Selvaraju et al., 2017), and Integrated Gradients (Sundararajan et al., 2017). The Vanilla Gradient method, which relies on gradient propagation for attribution visualization, is susceptible to signal noise and gradient vanishing (Zeineldin et al., 2022). Comparative analyses in NeuroXAI-related studies have shown that Vanilla Gradient often produces noisy and less discriminative saliency maps than alternative techniques (Martí-Juan et al., 2022; Philbrick et al., 2018; Zeineldin et al., 2022). In vascular segmentation, the TorchEsegeta evaluation framework reported that Vanilla Gradient and DeconvNET produced more understandable attributions than DeepLIFT and Grad-CAM++ (Chatterjee et al., 2022). Guided Backpropagation variants were later introduced to reduce signal noise, although sometimes at the cost of oversimplifying feature representations (Wood et al., 2022). layer-wise relevance propagation (LRP) addresses gradient saturation by propagating relevance scores instead of raw gradients, thereby offering an alternative explanation strategy (Saleem et al., 2021). LRP has also shown value in generating class-discriminative activations in computed tomography-based aneurysm identification studies (Montavon et al., 2019).

Recent studies have also explored hybrid and inherently interpretable architectures for brain MRI analysis in order to improve both predictive performance and clinical usability. For example, hybrid frameworks have combined deep feature extraction with machine learning and XAI tools such as Grad-CAM, Grad-CAM++, and LIME to support multiclass brain tumor classification with more transparent decision-making. Other recent work has proposed interpretable MRI pipelines that integrate attention mechanisms, transformer-based components, and retrieval-oriented designs. They improve segmentation, classification, and clinical decision support under heterogeneous imaging conditions. In addition, recent efficient classification architectures have shown that high diagnostic accuracy can be pursued together with built-in interpretability. And these studies suggesting that transparency is increasingly being treated as a core design objective rather than a secondary visualization step (Pacal et al., 2026; Jenifer et al., 2026; Hussain et al., 2026).

Motivated by recent developments, this study investigates the performance of transfer learning architectures and an attention-augmented CNN for multiclass brain tumor classification. Grad-CAM is employed to examine whether model predictions rely on diagnostically meaningful regions in MRI scans. The proposed approach focuses on both classification performance and interpretability.

The key novelty of this study is the integrated evaluation of predictive performance and explainability within a unified framework. Unlike prior studies that mainly focus on accuracy or provide limited qualitative interpretation, this work combines an attention-enhanced CNN with Grad-CAM visualization and evaluates explanation reliability using quantitative metrics. In addition, a comprehensive comparison is conducted across a baseline CNN, an attention-augmented CNN, and state-of-the-art pre-trained models, enabling a more holistic assessment of both performance and interpretability.

METHODS

Ethics

Ethical approval was not required for this study as it does not involve human participants, animal subjects, or identifiable personal data. All procedures were carried out in accordance with the ethical rules and principles.

Study Design

The study provides a detailed procedure on how to classify cerebral neoplasm using MRI sequences with several DNN structures. The reviewed frameworks include CNN architecture with reference, various pre-trained transfer models and an attention-augmented implementation with diagnostic transparency enabled with Grad-CAM interpretability analysis.

The publicly available brain tumor MRI Dataset edited by Msoud Nickparvar (2021) was used as experimental procedures. It contains T1 weighted magnetic resonance sequences stratified in four categories of diagnosis- glioma, meningioma, pituitary adenoma and healthy tissue specimens. The entire data consisted of 5,617 MRI scans that were obtained on 5,617 different individuals. To avoid subject-level data leakage, patient-level independence was rigorously enforced during dataset partitioning. The dataset was divided into training (4,493 sequences from 4,493 individuals), validation (562 sequences from 562 individuals), and testing (562 sequences from 562 individuals), while preserving approximately balanced class distributions across all partitions (Table 1).

Table 1. Class distributions per split

Split	Class	Count	Percentage (%)
Training	Pituitary	1,165	25.93
	No tumor	1,200	26.71
	Meningioma	1,071	23.84
	Glioma	1,057	23.53
Validation	Pituitary	146	25.98
	No tumor	150	26.69
	Meningioma	134	23.84
	Glioma	132	23.49
Test	Pituitary	146	25.98
	No tumor	150	26.69
	Meningioma	134	23.84
	Glioma	132	23.49

Statistical check: χ^2=0.0015, p=1.0000: No significant difference across splits, indicating distributional homogeneity

The Chi-square statistical test was used to identify the lack of significant distributional variance in diagnostic categories between the training, validation and testing cohorts (χ^2=0.0015, p=1.0000), hence achieving statistical equivalence and balanced representation. The independence at the level of patients was strictly maintained with each individual identifier being represented only in one data partition.

Reference convolutional neural network architecture was developed to be used as a baseline. In this structure, three convolutional blocks were stacked in a series with 64, 128, and 256 kernels respectively followed by activation functions of Rectified Linear Units and maximum pooling respectively. A highly connected network of 512 neural units, regularized by a dropout probability of 0.5, dealt with the problem of overfitting, and a final softmax layer produced probabilistic observations in four categories of neoplasms. The process of training was done over 10 epochs using Adam as the optimization algorithm (learning rate=10, -3) and using a validation set at the same time.

The domain-specific adaptation of three pre-trained architectural models (MobileNetV2, InceptionV3, and Xception) to identify neoplasms was performed to increase the ability to extract features and the ability of the model to generalize. Training modules based on original classification on ImageNet were replaced by a global average pooling mechanism, dropout regularization, and a quaternary-class output layer.

The inputs of MobileNetV2 were normalized to the range of -1, +1 as a pre-training requirement. Early training held frozen base layers during 10 epochs then followed by biased unfreezing and refinement of the top 30 layers with a reduction learning rate (10^{-4}), Inception V3 and Xception designs used a bifurcated design: initial training of classifier heads was followed by progressive unfreezing of additional convolutional layers, and extensive fine-tuning was done over the last 20 epochs. This hierarchical model compromised between maintenance of previously learned low-level and intermediate-level representations and domain-specific adaptation to the properties of MRI imaging.

Moreover, the convolutional architecture was enhanced using a convolutional block attention module (CBAM)-inspired attention mechanism to improve both predictive performance and interpretability. Unlike a standalone attention block, the proposed attention modules are directly integrated into the convolutional pipeline to refine intermediate feature representations.

Specifically, the attention mechanism is embedded after convolutional blocks, allowing the network to adaptively recalibrate feature maps during feature extraction. The attention process is applied sequentially, consisting of channel attention followed by spatial attention. In the channel attention stage, global average pooling is used to aggregate channel-wise statistics, which are then passed through fully connected layers to model inter-channel dependencies. This enables the network to emphasize the most informative feature channels.

Subsequently, the spatial attention module focuses on identifying important spatial regions within the feature maps, allowing the model to localize diagnostically relevant tumor structures more effectively. This sequential channel-spatial design improves feature discrimination while introducing minimal computational overhead.

For training, the attention-enhanced CNN was optimized using the Adam optimizer with a learning rate of 0.001, selected due to its stable convergence and adaptive learning capability observed during preliminary experiments. The batch size was set to 32 to balance computational efficiency and generalization performance under hardware constraints. All input images are resized to 128×128 and normalized to the (0,1) range. The model was trained for 15 epochs, which was empirically determined to achieve convergence while avoiding overfitting.

Extensive comparisons of models across all five models (MobileNetV2, InceptionV3, Xception, reference CNN, and attention-augmented CNN) used F1-score, classification accuracy, precision values and sensitivity values. Confusion matrices provided visual representation of category-specific prediction patterns. MobileNetV2 demonstrated superior accuracy, substantiating the efficacy of computationally efficient architectures in medical imaging applications. The attention-enhanced CNN achieved competitive secondary performance, validating the contribution of attention mechanisms toward refined spatial feature localization.

Grad-CAM was used to visualize and interpret model predictions to produce heatmaps, which indicate areas that were most significant to the classification choices. The heatmaps overlaid on MRI images with a JET colormap showed that the most successful models, in this case the attention CNN, paid attention to the tumor-affected regions, which is in accordance with clinical principles and increases the confidence in the work of automated diagnostic systems.

Figure 1 represents the general workflow, comprising of the MRI preprocessing, CNN-based classification, decision evaluation, Grad-CAM visualization, and interpretability analysis.

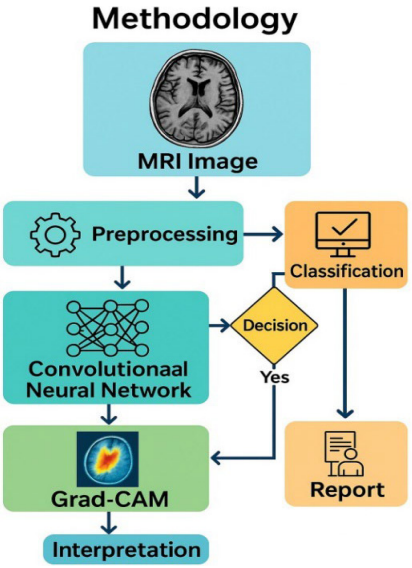


Figure 1. Flow diagram illustrating the deep learning methodology of categorizing tumors of the brain based on MRI images

Grad-CAM: Gradient-weighted class activation mapping, MRI: Magnetic resonance imaging

RESULTS

Table 2 delineates the predictive efficacy of five evaluated architectures on the reserved testing cohort: reference CNN, attention-augmented CNN, MobileNetV2, InceptionV3, and

Xception. All reported metrics represent aggregate values with corresponding standard deviations derived from three independent training iterations, accompanied by 95% confidence intervals (CIs). Across all examined frameworks, MobileNetV2 exhibited superior and most consistent performance characteristics (accuracy=0.917±0.007, F1=0.914±0.008, PR-AUC=0.972±0.002), demonstrating robust discriminative capacity and reliable generalization across experimental replicates. Such high-quality performance can be linked to its highly efficient depthwise separable convolutional computations and learned hierarchies of features that are able to capture the neoplasm-specifically morphological features of MRI series. Attention-enhanced CNN architecture obtained second place ranking (accuracy=0.876±0.018), which is around 6.8 percentage point higher than the reference CNN, therefore vindicating the fact that the channel-spatial attention mechanisms increase the localization of diagnostically relevant tumor structures. The same can be said of InceptionV3 (accuracy=0.874±0.001) and Xception (accuracy=0.909±0.011), which also showed high and consistent performance, demonstrating that the modern factorized and computationally efficient backbone architectures could be transferred to the field of medical imaging. The accuracy of the reference CNN (0.808±0.099) was worse with high variability, which indicates the low optimization stability as compared to pre-trained frameworks. Paired t-tests and Wilcoxon signed-rank tests were statistically validated that MobileNetV2 is of great superiority to InceptionV3 and attention-augmented CNN in terms of accuracy and F1-score measures ($p<0.05$). Further Xception enjoyed expected calibration error (ECE) of 0.032±0.012 which represents the best probabilistic prediction reliability. All of these results support the conclusion that paradigms of transfer learning and attention-mediated mechanisms significantly increase model stability, predictive accuracy, and calibration fidelity in cerebral neoplasm MRI category tasks.

The quantitative analysis of model transparency showed that the Grad-CAM attribution visualizations were highly reliant on the parameters of trained networks, thus establishing them as model-dependent visualization systems of explaining. The validation of integrity provided an average correlation coefficient of -0.0815±0.1137 between the original and parameter randomization activation maps, which implied that saliency visualizations indeed changed greatly during the course of training rather than being a manifestation of inherent structural prejudices. Causal perturbation tests also supported the explanatory fidelity, where deletion area over the perturbation curve (AOPC) scored 94.32 and insertion AOPC scored 87.34, which is consistent with expected loss of confidence with deletion of salient regions and recovery of the confidence in the case of improvement of the diagnostically relevant regions. All these cumulative results justify the soundness and accuracy of Grad-CAM-based explanations, which can be combined in the future with clinical interpretation systems and diagnostic reliability standards.

DISCUSSION

The results show that deep learning models, particularly pre-trained architectures, are effective for brain tumor classification from MRI, with MobileNetV2 achieving the highest accuracy of 91.7%. The attention-augmented CNN also improved interpretability by emphasizing anatomically relevant regions during classification.

Nevertheless, the clinical applicability of these models depends on their robustness to MRI noise, data variability, and potential domain shifts across different scanner settings. Although the integrated attention mechanism and pre-trained backbones may reduce some variability by focusing on diagnostically relevant regions, further evaluation on external datasets and through cross-validation would provide stronger evidence of generalizability.

Table 2. Performance of multi-run tests

Architecture	Precision	F1-score	Classification accuracy	Sensitivity	ECE	PR-AUC
Xception	91.1±1.4% (87.7, 94.6)	90.6±1.3% (87.3, 93.8)	90.9±1.1% (88.2, 93.7)	90.6±1.2% (87.8, 93.5)	3.2±1.2%	96.8±0.2% (96.3, 97.3)
MobileNetV2	91.5±0.8% (89.6, 93.4)	91.4±0.8% (89.4, 93.3)	91.7±0.7% (89.9, 93.5)	91.4±0.7% (89.6, 93.2)	3.6±0.5%	97.2±0.2% (96.6, 97.8)
Attention-augmented CNN	88.2±2.6% (81.6, 94.7)	87.3±2.0% (82.4, 92.2)	87.6±1.8% (83.2, 92.0)	87.2±1.9% (82.6, 91.8)	4.0±2.2%	93.8±2.0% (88.8, 98.7)
InceptionV3	87.7±0.7% (86.0, 89.3)	87.1±0.2% (86.5, 87.7)	87.4±0.1% (87.2, 87.7)	87.1±0.2% (86.7, 87.5)	4.3±1.6%	94.7±0.4% (93.7, 95.7)
Reference CNN	86.6±3.6% (77.6, 95.6)	80.7±10.1% (55.7, 105.7)	80.8±9.9% (56.3, 105.3)	80.6±9.9% (56.0, 105.1)	6.6±0.3%	90.3±2.8% (83.3, 97.2)

ECE: Expected calibration error, PR-AUC: Precision-recall area under the curve, CNN: Convolutional neural network

Figure 2 indicates the visualization of a meningioma MRI by Grad-Cam as the regions of a tumor are drawn where the model is sensitive to the diagnostically significant information, and, therefore, learning behavior can be interpreted and relied on.

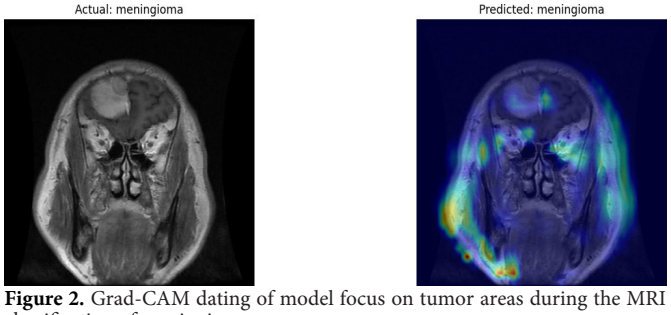


Figure 2. Grad-CAM dating of model focus on tumor areas during the MRI classification of meningioma

Grad-CAM: Gradient-weighted class activation mapping, MRI: Magnetic resonance imaging

In addition, the reliability of the Grad-CAM explanations was assessed beyond visual inspection using parameter randomization and causal perturbation tests, supporting the clinical relevance of the generated heatmaps. Despite these strengths, the study remains limited by the use of a single public dataset, the lack of external validation, and the restricted spatial resolution of Grad-CAM explanations.

Even so, the proposed framework offers practical advantages for clinical deployment, including strong predictive performance, computational efficiency, and improved transparency for MRI-based brain tumor assessment.

CONCLUSION

It was in this study that predictive performance evaluation was combined with transparency analysis to determine the efficacy

of different DNN designs to categorize cerebral neoplasms using MRI sequences. One of the most notable contributions is the serious comparative evaluation of the reference CNN, attention-enhanced CNN, and various pre-trained models in a single experimental framework that has provided the way that computationally efficient models of MobileNetV2 can be used to attain higher classification accuracy at low computational costs. The study also demonstrates that in addition to determining optimal predictive performance, combination of Grad-CAM visual attribution techniques explains the spatial concentration of algorithmic attention on tumor-specific anatomical areas, hence creating meaningful networks between computational predictions and diagnostic utility. These results are indicative of the novelty of combining the comparative efficacy analysis with interpretability mechanisms in a unified framework which confirms the potential of clear deep learning systems as a clinically feasible radiology tool.

Three strategic directions for future research are as follows: developing interpretable clinical decision support interfaces that combine calibrated uncertainty quantification with visual attribution overlays to boost radiologist confidence and enable practical use in real clinical settings; implementing ensemble techniques that learn to combine heterogeneous architectures to enhance predictive stability and probability calibration; and adding multimodal MRI contrasts to improve their ability to distinguish tumor subtypes.

ETHICAL DECLARATIONS

Ethics Committee Approval

Ethical approval was not required for this study as it does not involve human participants, animal subjects, or identifiable personal data.

Peer Review Process

This manuscript was subject to external peer review.

Conflict of Interest

The authors declare no conflicts of interest related to this study. Esra Söğüt serves as the Associate Editor-in-Chief of the journal and was not involved in the editorial evaluation or decision-making process for this article.

Financial Disclosure

The authors received no financial support for the conduct or publication of this research.

Author Contributions

Concept: MAM, OAE, ES; Design: MAM, OAE, ES; Control: MAM, OAE, ES; Resources: MAM, OAE, ES; Materials: MAM, OAE, ES; Data Collection and/or Processing: MAM, OAE, ES; Analysis and/or Interpretation: MAM, OAE, ES; Literature Review: MAM, OAE, ES; Writing the Article: MAM, OAE, ES; Critical Review: MAM, OAE, ES.

Editorial Note

One of the authors of this article serves as the Associate Editor-in-Chief of the journal. To ensure editorial independence, transparency, and the integrity of the peer-review process, the author was excluded from all stages of the editorial evaluation and decision-making process concerning this manuscript. The manuscript was handled independently by an editor with no relevant conflicts of interest and underwent the journal's standard external peer-review procedures. The final decision regarding publication was made independently, based solely on the scientific merit, methodological quality, and relevance of the study, in accordance with the journal's editorial policies and internationally recognized publication ethics principles.

REFERENCES

Atagün, E., Temür, G., & Biroğul, S. (2026). Graph-based interpretable dialogue sentiment analysis: a HybridBERT-LSTM framework with semantic interaction explainer. *Computer Standards & Interfaces, 97*, 104086. https://doi.org/10.1016/j.csi.2025.104086

Brima, Y., Atemkeng, M., Tankio Djiokap, S., Ebiele, J., & Tchakounté, F. (2021). Transfer learning for the detection and diagnosis of types of pneumonia including pneumonia induced by COVID-19 from chest X-Ray images. *Diagnostics, 11*(8), 1480. https://doi.org/10.3390/diagnostics11081480

Chaddad, A., Hu, Y., Wu, Y., Wen, B., & Kateb, R. (2025). Generalizable and explainable deep learning for medical image computing: an overview. *Current Opinion in Biomedical Engineering, 33*, 100567. https://doi.org/10.1016/j.cobme.2024.100567

Chatterjee, S., Das, A., Mandal, C., Mukhopadhyay, B., Vipinraj, M., Shukla, A., ..., & Nürnberger, A. (2022). TorchEsegeta: framework for interpretability and explainability of image-based deep learning models. *Applied Sciences, 12*(4), 1834. https://doi.org/10.3390/app12041834

Chattopadhyay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 839-847). IEEE. https://doi.org/10.1109/WACV.2018.00097

Ebiele, J., Ansah-Narh, T., Djiokap, S., Proven-Adzri, E., & Atemkeng, M. (2020). Conventional machine learning based on feature engineering for detecting pneumonia from chest X-Rays. In Proceedings of the 2020 Conference of the South African Institute of Computer Scientists and Information Technologists (pp. 149-155). Association for Computing Machinery. https://doi.org/10.1145/3410886.3410898

Hasan, M. R., Rudra, S. M., Karmoker, N., Yousuf, M. A., Akhter, J., Al-Moisheer, A. S., ..., & Moni, M. A. (2026). 3D MRI reconstruction and brain tumor diagnosis using deep learning with explainable AI. *Expert Systems with Applications, 315*, 131513. https://doi.org/10.1016/j.eswa.2026.131513

Haupt, M., Maurer, M. H., & Thomas, R. P. (2025). Explainable Artificial Intelligence in radiological cardiovascular imaging-a systematic review. *Diagnostics, 15*(11), 1399. https://doi.org/10.3390/diagnostics15111399

Houssein, E. H., Gamal, A. M., Younis, E. M. G., & Mohamed, E. (2025). Explainable Artificial Intelligence for medical imaging systems using deep learning: a comprehensive review. *Cluster Computing, 28*(7), 469. https://doi.org/10.1007/s10586-025-05281-5

Hussain, M. I., Chowdhury, S. H., Hossain, M. M., & Mamun, M. (2026). NeuroBlend-3: hybrid deep and machine learning framework with explainable AI for multi-class brain tumor detection using MRI scans. *Medinformatics, 3*(1), 56-66. https://doi.org/10.47852/bonviewMEDIN52026540

Jenifer, G. G., & Rajakumar, G. (2026). HyBiUnet-TransMask: a clinical decision system with segmentation and classification for interpretable brain tumor diagnosis in multimodal MRI. *Biomedical Signal Processing and Control, 115*, 109393. https://doi.org/10.1016/j.bspc.2025.109393

Kumar, V. D. R., & Singh, P. (2026). HCAC: trustworthy hybrid attentive framework tackling brain tumor identification with XAI. *IEEE Access, 14*, 16158-16175. https://doi.org/10.1109/ACCESS.2026.3656212

Lamrani, D., Mahjoubi, M. A., Shawki, S., Moutaouakil, W., Hamida, S., ..., & Bahatti, L. (2026). Enhanced multiclass brain tumor segmentation using MRI images and explainable AI techniques. *Multimedia Tools and Applications, 85*(1), 28. https://doi.org/10.1007/s11042-026-21180-2

Martí-Juan, G., Frías, M., Garcia-Vidal, A., Vidal-Jordana, A., Alberich, M., Calderon, W., ..., & Pareto, D. (2022). Detection of lesions in the optic nerve with magnetic resonance imaging using a 3D convolutional neural network. *NeuroImage: Clinical, 36*, 103187. https://doi.org/10.1016/j.nicl.2022.103187

Montavon, G., Binder, A., Lapuschkin, S., Samek, W., & Müller, K. R. (2019). Layer-wise relevance propagation: an overview. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & Müller, K. R. (Eds.), Explainable AI: Interpreting, explaining and visualizing deep learning (Lecture Notes in Computer Science, Vol. 11700, pp. 193-209). Springer. https://doi.org/10.1007/978-3-030-28954-6_10

Nickparvar, M. (2021). Brain tumor MRI dataset. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/2645886

Pacal, I., & Banerjee, T. (2026). Towards accurate and interpretable brain tumor diagnosis: T-FSPANNet with tri-attribute and pyramidal attention-based feature fusion. *Biomedical Signal Processing and Control, 113*, 108852. https://doi.org/10.1016/j.bspc.2025.108852

Philbrick, K. A., Yoshida, K., Inoue, D., Akkus, Z., Kline, T. L., Weston, A. D., ..., & Erickson, B. J. (2018). What does deep learning see? Insights from a classifier trained to predict contrast enhancement phase from CT images. *American Journal of Roentgenology, 211*(6), 1184-1193. https://doi.org/10.2214/AJR.18.20331

Saleem, H., Shahid, A. R., & Raza, B. (2021). Visual interpretability in 3D brain tumor segmentation network. *Computers in Biology and Medicine, 133*, 104410. https://doi.org/10.1016/j.compbiomed.2021.104410

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: visual explanations from deep networks via gradient-based localization. In 2017 IEEE International Conference on Computer Vision (ICCV) (pp. 618-626). IEEE. https://doi.org/10.1109/ICCV.2017.74

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In D. Precup & Y. W. Teh (Eds.), Proceedings of the 34th International Conference on Machine Learning (Vol. 70, pp. 3319-3328). Proceedings of Machine Learning Research. https://proceedings.mlr.press/v70/sundararajan17a.html

Wood, D. A., Kafiabadi, S., Al Busaidi, A., Guilhem, E., Montvila, A., Lynch, J., ..., & Booth, T. C. (2022). Deep learning models for triaging hospital head MRI examinations. *Medical Image Analysis, 78*, 102391. https://doi.org/10.1016/j.media.2022.102391

Zeineldin, R. A., Karar, M. E., Elshaer, Z., Coburger, J., Wirtz, C. R., ..., & Mathis-Ullrich, F. (2022). Explainability of deep neural networks for MRI analysis of brain tumors. *International Journal of Computer Assisted Radiology and Surgery, 17*(9), 1673-1683. https://doi.org/10.1007/s11548-022-02619-x