CPSM | Conference Proceedings in Science and Management

Proceedings of the 2nd International Engineering Data Analytics and Management Conference (EAMCON 2026), 8 to 10 July 2026, Shinawatra University, Thailand

Published: October 2026
DOI: https://doi.org/10.68337/cpsm.v1.i2.2026-024

Vol. 1, Issue 2 | Proceeding Paper

Explainable CNN for Brain Tumor Detection and Classification

Byreddy Sudhakara Reddy[1,*] and J. Balaraju[1]

[1]*Department of Computer Science and Engineering, School of Engineering, Anurag University, Hyderabad, Telangana, India*

[*]Corresponding author: byreddysudh@gmail.com

Abstract. Image analysis is an important part of medical decision-making, where it helps clinicians to reach a diagnosis and plan treatment. Artificial intelligence (AI) and machine learning (ML) methods for medical image analysis have grown rapidly. Explainable artificial intelligence (XAI) methods aim to make AI applications more usable and trustworthy by showing how a model reaches its decisions. This paper reviews recent XAI methods and their use in medical image analysis, including rule-based systems, attention mechanisms, and feature visualization. A convolutional neural network (CNN) was used to classify brain magnetic resonance imaging (MRI) scans by tumor type, and an XAI model explains how the network uses the images to reach each prediction. Of the 15,030 images, 13,563 were used for training and 1,467 for validation and testing. The model reached an accuracy of 80.55%. Local Interpretable Model-Agnostic Explanations (LIME) is used as the example XAI model.

Keywords. AI; LIME; machine learning; XAI

I. Introduction

Explainable artificial intelligence (XAI) has become a main approach to making complex AI models, such as convolutional neural networks (CNNs), understandable to people. XAI and deep learning (DL)-based image processing are important for the healthcare sector. Every diagnostic and treatment decision affects the patient's quality of life, so the interpretability and transparency of AI-driven systems matter in this setting [1], [2].

The aim of this work is to use the core capabilities of XAI to improve the accuracy and value of DL-driven brain tumor classification. The larger goal is better patient care and outcomes, through healthcare decisions that are more informed, transparent, and data-driven [3], [4].

Tumors in different parts of the cerebrum produce different symptoms, depending on the functions of the affected area (Fig. 1). Tumors in the frontal lobes, which govern thought and movement, can cause problems with balance, difficulty in walking, and changes in personality, and relatives often notice that the person seems changed or less engaged. The parietal lobes, located centrally in the brain, process sensory information, and tumors there may cause sensory problems, including vision and hearing deficits. The occipital lobes at the rear of the brain are mainly responsible for vision, and tumors there can impair it [5].

A. Convolutional Neural Networks

In machine learning, CNNs perform well on tasks such as image classification and segmentation. They recognize complex patterns in images while preserving spatial relationships, which makes them useful wherever spatial structure matters. Their hierarchical feature extraction lets them recognize increasingly complex features. Weight sharing reduces the number of trainable parameters, which limits overfitting and improves generalization. Because they scale well and handle large datasets efficiently, CNNs have driven much of the recent progress in AI and computer vision [1].

B. SHAP

SHapley Additive exPlanations (SHAP) gives a detailed account of feature importance and its effect on model predictions by linking specific image characteristics to the outputs of CNN models. By measuring the average contribution of each feature through Shapley values, SHAP identifies the image regions that most influence the classification result. It clarifies the relationship between image features and predictions, which helps to identify the key structures in an image and makes CNN models easier to interpret and more dependable. Its systematic, flexible design allows it to be added to existing CNN models without loss of performance [6].

C. Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) was developed for CNN-based models to show which image regions are relevant to a classification. It identifies the regions of an image that most influence the model's prediction. A multiscale hierarchical visualization, which captures information at several resolutions, shows where the model attends at different anatomical levels and allows image features to be examined at different scales. When subject experts are involved, Grad-CAM visualizations can be related to accepted clinical indicators, which makes the tool more dependable and informative. A user-friendly interface that places Grad-CAM maps beside the original images makes

Cite as: B. Sudhakara Reddy and J. Balaraju, "Explainable CNN for brain tumor detection and classification," Conf. Proc. Sci. Manag., vol. 1, no. 2, pp. 5-8, 2026. https://doi.org/10.68337/cpsm.v1.i2.2026-024.

Copyright © 2026 The Authors. Published by SHORAM Editorial Services, India, under the Creative Commons Attribution 4.0 License (CC BY 4.0).

the two easy to compare. The aim is to make Grad-CAM clinically meaningful, interpretable, and easy to use in practice [7], [8].

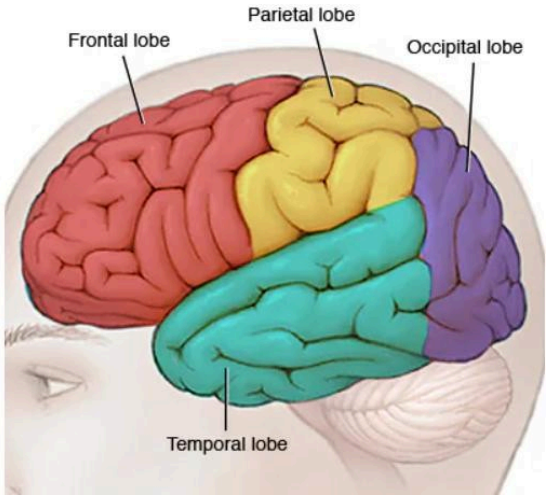


Fig. 1. Lobes of the brain, used to relate tumor symptoms to tumor location.

D. Integrated Gradients

The Integrated Gradients (IG) technique improves the interpretability of CNNs in image classification. Unlike some other interpretability methods, it evaluates the role of neural network activations across all layers. It tracks each neuron's contribution to the final prediction, which is particularly useful for a task as complex as brain tumor classification. Layer-by-layer IG analysis identifies the neurons and features that matter most in each CNN layer and so gives a detailed view of how the network reaches its decisions [9].

E. LIME

Local Interpretable Model-Agnostic Explanations (LIME) is an important step in the development of interpretable AI. LIME addresses the opacity of CNNs by perturbing individual images and generating a set of related but different samples. The local explanations it produces improve understanding of, and confidence in, the model's predictions and show which variables affect each prediction [10].

II. Literature Survey

Yan and colleagues [11] built a framework that analyzes MRI images end to end to improve brain tumor detection. It includes an explainable network made up of segmentation and classification models and classifies low-grade and high-grade gliomas from MRI sequences, which reduces confusion and effort for users. The classification model is based on a VGG-like network with re-parameterization, which gives faster performance and a simpler structure than earlier models such as ResNet and Inception. Besides the segmentation and classification outputs, the framework produces explainability outputs that make the model's operation easier to analyze. It reached a segmentation F1 score of 95% and a classification accuracy of 98.5%.

Pitroda and colleagues developed a framework that detects lung disease in chest X-ray images and separates normal lungs from lung opacity caused by several illnesses, including COVID-19. The dataset contained 3,616 X-rays from patients with COVID-19, 10,192 normal images, and 6,012 images showing lung opacity. A dedicated CNN architecture based on U-Net was used for accurate and efficient image segmentation. In comparisons with DenseNet-121, ResNet-50, and InceptionV3, DenseNet-201 showed the strongest predictive ability. Layer-wise relevance propagation (LRP) produced similar results but also highlighted some irrelevant bone structures in the heat maps. Bone suppression during image processing could improve the model's performance and interpretability.

Esmaeili and colleagues [12] built a system that classifies tumors accurately and also explains its decisions with XAI. Their dataset, obtained from The Cancer Imaging Archive, contains images of glioblastomas (WHO grade IV) and lower-grade gliomas from 354 subjects, resized to 256 × 256 × 128 pixels. The training set contained 19,200 slices containing tumor lesions and 14,800 slices without them. The mean cross-validated prediction accuracies were 92.1% for DenseNet-121, 87.3% for GoogLeNet, and 88.9% for MobileNet.

Naz and colleagues [13] developed an XAI system for the rapid detection of lung-related respiratory disorders. Their study presents a transfer learning method with a CNN that classifies conditions such as tuberculosis, edema, and pneumonia from chest X-rays. An explainability module that extracts activation maps adds to the model's explanatory ability.

Vardhan and colleagues standardized all 780 breast ultrasound images in their dataset to 500 × 500 pixels and labeled them as normal, benign, or malignant. Class imbalance and poor image quality were among the problems of the dataset. The model identified round dark spots well but had difficulty with irregular shapes. It reached a validation accuracy of 99% and a training accuracy of 95% despite the small dataset and doubts about real-world applicability, which shows the promise of combining AI and thermal imaging for breast cancer detection.

Eder and colleagues [14] used visual explanations to interpret network models that predict the survival of glioma patients. Following the winners of the BraTS2020 challenge, they evaluated three network variants and found that the 2D full-size model performed best, with accuracies from 31% to 55.2%. The results were interpreted with SHAP, which showed cases in which good training performance does not guarantee high accuracy. Because few other datasets were available, the authors gave priority to explainability over network optimization. With the required preprocessing, SHAP analysis helps to assess networks trained on limited data. Their future work aims to improve the models and overcome the dataset limits by drawing on several data sources.

Meske, Bunde, Schneider, and Gersch [15] discussed the objectives, stakeholders, and future research directions of XAI. Their study stresses the need for explainability and addresses the risks of opaque AI systems. It reviews earlier research on XAI in information systems and traces the origins of the term, the general goals of XAI, and the stakeholder groups involved. It ends with directions for future XAI research after examining quality criteria for personalized explanations.

Singh, Sengupta, and Lakshminarayanan [16] reviewed explainable DL models for medical image analysis and how

they are currently used across imaging tasks. Besides discussing several approaches, the paper describes the difficulties met in clinical deployment and identifies areas that need more investigation. It takes the viewpoint of a DL researcher developing systems for clinical end users and gives practical advice on applying explainable DL to medical imaging. The study aims to improve current explainability methods so that they cope better with the complexity of medical images. As an example, the Expressive Gradients (EG) technique extends attribution to continuous attributes such as texture and shape and improves the identification of retinal lesions. It builds on the widely used IG approach. Because such extensions can be tailored without creating entirely new techniques, they have advanced work on explainability.

Ghnemat, Alodibat, and Abu Al-Haija [17] used DL models that put explainability and interpretability first. Their work also addresses the difficulty of assessing XAI explanations objectively and acknowledges that explanations can be subjective, particularly with LIME. It stresses the need for human interpretation and expert validation and calls for more research on the effectiveness of XAI models.

Zeineldin and colleagues [18] presented NeuroXAI, an explainability framework that aims to improve understanding of DL networks through advanced visualization methods such as attention maps. NeuroXAI works post hoc on different pretrained deep neural models and gives insight into how they work.

III. Proposed System

The architecture of the proposed system is shown in Fig. 2. MRI images are preprocessed and split into training and test images, a CNN is trained to classify the disease, and an XAI model explains the classifications before evaluation.

A. Dataset

The system uses a Kaggle dataset that combines three sources: figshare, SARTAJ, and Br35H. During preprocessing, each image is labeled with its brain tumor category: glioma, pituitary adenoma, meningioma, or no tumor. Sample images of the four classes are shown in Fig. 3. The model treats the pixel values of the images as features, and the labels are converted to a categorical format so that the model handles the categorical nature of the target variable during training. A CNN is then built with convolution layers (Conv2D), downsampling layers (MaxPooling), and fully connected layers (Dense). After training, the model classifies brain MRI images by tumor type. Several XAI techniques, namely Grad-CAM, SHAP, LIME, and IG, are used to make the model more interpretable and transparent. The explanations produced by SHAP, LIME, IG, and Grad-CAM are assessed systematically to determine the importance of the different factors behind the CNN's decisions.

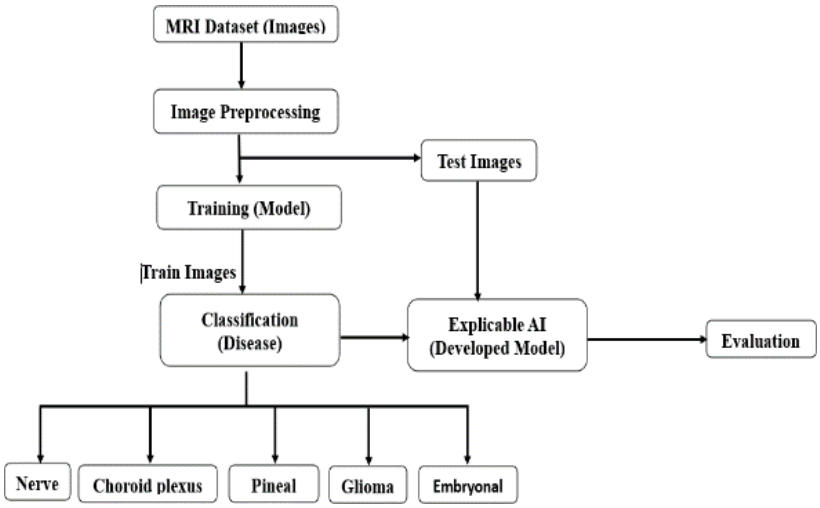


Fig. 2. Proposed system architecture.

B. Model Training

The proposed classifier begins with two convolutional layers, each with 32 filters and a rectified linear unit (ReLU) activation that introduces nonlinearity. Downsampling is done by a max pooling layer with a 2 × 2 pool and a stride of 2. The input shape is 224 × 224 × 3. After ten epochs of training, the model's accuracy was 80.55%. The LIME image explainer applies LIME to explain the predicted brain tumor classes in terms of individual pixels. SHAP assigns each feature a contribution, which shows how the model makes its decisions. Grad-CAM makes the images easier to understand by weighting the areas that matter most for the model's judgments.

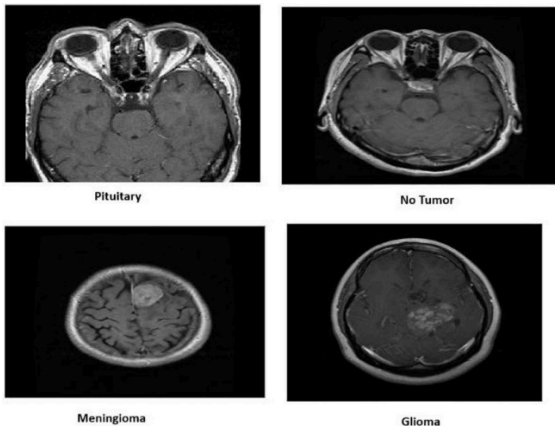


Fig. 3. Sample MRI images of the four classes: pituitary tumor, no tumor, meningioma, and glioma.

IV. Results Analysis

The proposed CNN model classifies brain MRI images into four classes: glioma, meningioma, pituitary tumor, and no tumor. Grad-CAM, LIME, SHAP, and IG are used together with the CNN to improve interpretability, so that the system both classifies the scans and highlights the areas where tumors may be present. On the dataset, the model reached an accuracy of 80.55% over ten epochs.

SHAP gives further insight into the model's predictions through the SHAP values of different pixel regions. The pixel-wise values in the color-coded areas show how each pixel contributed to the classification; a negative SHAP value means that higher intensity in that region lowers the likelihood of the predicted class. Grad-CAM produces heat maps that show the importance of particular areas of the input image. Superimposed on the original image, the heat map shows the regions that had the greatest effect on the network's decision.

V. Conclusion

A CNN system was developed to classify brain tumor MRI images into four groups. To improve the interpretability of its predictions, the model uses XAI methods such as Grad-CAM, LIME, SHAP, and IG in addition to classification. The aim of adding XAI techniques to brain tumor detection systems is to increase diagnostic precision, reduce false positives, and improve patient outcomes. The set of XAI models described here offers a useful toolkit for researchers and experts working to understand complex neural networks in medical imaging and other domains.

Author Contributions

All authors contributed to the conception and preparation of this work and approved the final version of the manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Declaration of Generative AI Use

The authors declare that no generative artificial intelligence tools were used in the preparation of this manuscript. The authors take full responsibility for the content of the publication.

Data Availability

The MRI images used in this study come from a publicly available Kaggle dataset that combines the figshare, SARTAJ, and Br35H collections.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Brij B. Gupta, Akshat Gaurav, and Varsha Arya, "Deep CNN based brain tumor detection in intelligent systems," *Int. J. Intell. Netw.*, vol. 5, pp. 30-37, 2024, doi: 10.1016/j.ijin.2023.12.001.

[2] Mahmoud Khaled Abd-Ellah, Ali Ismail Awad, Ashraf A. M. Khalaf, and Amira Mofreh Ibraheem, "Automatic brain-tumor diagnosis using cascaded deep convolutional neural networks with symmetric U-Net and asymmetric residual-blocks," *Sci. Rep.*, vol. 14, no. 1, Art. no. 9501, 2024, doi: 10.1038/s41598-024-59566-7.

[3] J. Seetha and S. Selvakumar Raja, "Brain tumor classification using convolutional neural networks," *Biomed. Pharmacol. J.*, vol. 11, no. 3, pp. 1457-1461, 2018, doi: 10.13005/bpj/1511.

[4] Wadhah Ayadi, Wajdi Elhamzi, Imen Charfi, and Mohamed Atri, "Deep CNN for brain tumor classification," *Neural Process. Lett.*, vol. 53, no. 1, pp. 671-700, 2021, doi: 10.1007/s11063-020-10398-2.

[5] J. Ricardo McFaline-Figueroa and Eudocia Q. Lee, "Brain tumors," *Am. J. Med.*, vol. 131, no. 8, pp. 874-882, 2018, doi: 10.1016/j.amjmed.2017.12.039.

[6] Radius Tanone, Li-Hua Li, and Shoffan Saifullah, "ViT-CB: Integrating hybrid vision transformer and CatBoost to enhanced brain tumor detection with SHAP," *Biomed. Signal Process. Control*, vol. 100, Art. no. 107027, 2025, doi: 10.1016/j.bspc.2024.107027.

[7] Mohamed Musthafa M, Mahesh T. R., Vinoth Kumar V, and Suresh Guluwadi, "Enhancing brain tumor detection in MRI images through explainable AI using Grad-CAM with Resnet 50," *BMC Med. Imaging*, vol. 24, no. 1, Art. no. 107, 2024, doi: 10.1186/s12880-024-01292-7.

[8] Krishan Kumar and Kiran Jyoti, "Recent advancements in Grad-CAM and variants: Enhancing brain tumor detection, segmentation, and classification," Research Square preprint, 2024, doi: 10.21203/rs.3.rs-5485128/v1.

[9] Huong Hoang Luong and Triet Minh Nguyen, "Improving chicken disease classification based on vision transformer and combine with integrated gradients explanation," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 4, 2024, doi: 10.14569/IJACSA.2024.01504124.

[10] K. Usha Rupni and P. Aruna Priya, "Local interpretable model-agnostic explanations guided brain magnetic resonance imaging classification for identifying attention deficit hyperactivity disorder subtypes," *J. Ambient Intell. Humaniz. Comput.*, vol. 16, no. 2-3, pp. 361-374, 2025, doi: 10.1007/s12652-024-04950-4.

[11] Fei Yan, Yunqing Chen, Yiwen Xia, Zhiliang Wang, and Ruoxiu Xiao, "An explainable brain tumor detection framework for MRI analysis," *Appl. Sci.*, vol. 13, no. 6, Art. no. 3438, 2023, doi: 10.3390/app13063438.

[12] Morteza Esmaeili, Riyas Vettukattil, Hasan Banitalebi, Nina R. Krogh, and Jonn Terje Geitung, "Explainable artificial intelligence for human-machine interaction in brain tumor localization," *J. Pers. Med.*, vol. 11, no. 11, Art. no. 1213, 2021, doi: 10.3390/jpm11111213.

[13] Zubaira Naz, Muhammad Usman Ghani Khan, Tanzila Saba, Amjad Rehman, Haitham Nobanee, and Saeed Ali Bahaj, "An explainable AI-enabled framework for interpreting pulmonary diseases from chest radiographs," *Cancers*, vol. 15, no. 1, Art. no. 314, 2023, doi: 10.3390/cancers15010314.

[14] Matthias Eder, Emanuel Moser, Andreas Holzinger, Claire Jean-Quartier, and Fleur Jeanquartier, "Interpretable machine learning with brain image and survival data," *BioMedInformatics*, vol. 2, no. 3, pp. 492-510, 2022, doi: 10.3390/biomedinformatics2030031.

[15] Christian Meske, Enrico Bunde, Johannes Schneider, and Martin Gersch, "Explainable artificial intelligence: Objectives, stakeholders, and future research opportunities," *Inf. Syst. Manag.*, vol. 39, no. 1, pp. 53-63, 2022, doi: 10.1080/10580530.2020.1849465.

[16] Amitojdeep Singh, Sourya Sengupta, and Vasudevan Lakshminarayanan, "Explainable deep learning models in medical image analysis," *J. Imaging*, vol. 6, no. 6, Art. no. 52, 2020, doi: 10.3390/jimaging6060052.

[17] Rawan Ghnemat, Sawsan Alodibat, and Qasem Abu Al-Haija, "Explainable artificial intelligence (XAI) for deep learning based medical imaging classification," *J. Imaging*, vol. 9, no. 9, Art. no. 177, 2023, doi: 10.3390/jimaging9090177.

[18] Ramy A. Zeineldin, Mohamed E. Karar, Ziad Elshaer, Jan Coburger, Christian R. Wirtz, Oliver Burgert, and Franziska Mathis-Ullrich, "Explainability of deep neural networks for MRI analysis of brain tumors," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 17, no. 9, pp. 1673-1683, 2022, doi: 10.1007/s11548-022-02619-x.