JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Explainable Brain Tumor Classification Using VGG16 and Grad-CAM on MRI Images

[1]Shubham Porte, [2]Dr. Shanu Kuttan Rakesh

[1]M. Tech. Scholar, [2]Associate Professor
[1]Department of Computer Science and Engineering, Chouksey Engineering College, Bilaspur (C.G), India
[2]Department of Computer Science and Engineering, Chouksey Engineering College, Bilaspur (C.G), India

Abstract **-** Correctly classifying brain cancers from MRI scans is important for making an early diagnosis and planning appropriate treatment. This study introduces a deep learning framework utilizing the VGG16 convolutional neural network for classifying brain cancers into multiple classes, including pituitary tumors, gliomas, meningiomas, and healthy (tumor-free) instances. The suggested model has a high overall accuracy of 95% and high precision, recall, and F1-scores in all categories. To make the predictions of the model are simpler to comprehend, we use Class Activation Mapping with Gradient Weighting (Grad-CAM), which shows the majority of important parts of the MRI images to explain the predictions. These heatmaps confirm that the prototype focuses on clinically relevant areas, increasing the reliability and transparency of its decisions. The combination of high classification performance and visual interpretability suggests that this approach holds significant potential for assisting radiologists in real-world diagnostic settings**.**

Keywords **- Grad-CAM Explainability, Deep Learning, and Brain Tumor Classification.**

1. INTRODUCTION

One of the most serious and potentially fatal neurological disorders is still brain tumors, which require prompt and precise diagnosis to enhance patient outcomes. It takes a lot of time and relies heavily on radiologists' skill to manually analyze magnetic resonance imaging (MRI) data, which frequently results in inconsistent interpretation [1]. This emphasizes the need for trustworthy, automated diagnostic tools that can help doctors properly categorize different kinds of tumors.

Neural networks with convolutions (CNNs), a recent advancement in deep learning, have demonstrated cutting-edge results in medical picture categorization, including the detection of brain tumors [2], [3]. CNNs have the benefit of automatically extracting features from raw pictures, which removes the need for manually created features and makes it possible to process high-dimensional MRI data efficiently. Among these, the VGG16 model is widely recognized for its robust performance and architectural simplicity, making it suitable for transfer learning in medical applications [4].

Despite their predictive strength, many people consider deep learning models to be "black boxes," which makes them challenging to understand the rationale behind their choices. Clinical trust and adoption may be hampered by this lack of openness or transparency. To overcome this, visual explainability methods like Gradient-weighted Class Activation Mapping (Grad-CAM) have gained popularity. Grad-CAM helps visualize the discriminative regions in the input images that influence the model's predictions, thereby increasing model transparency and clinical confidence [5].

Here, we provide a CNN-based classification framework based on an improved VGG16 model to categorize brain cancers into four groups: glioma, meningioma, pituitary tumor, and no tumor. Grad-CAM enhances the interpretability of the model's predictions by producing heatmaps that emphasize tumor-relevant regions in the MRI images. By combining explainability and accuracy, this method seeks to close the gap between AI-powered diagnostics and practical clinical usage.

2. LITERATURE REVIEW

Numerous studies have looked into deep learning models for automated classification of brain tumors in recent years. Because convolutional neural networks (CNNs) can learn discriminative features from medical pictures without the need for manual feature engineering, they have become the most popular method. To increase accuracy on brain MRI datasets, Rani and Kaur [6] suggested a hybrid system that combines CNN and support vector machine (SVM) classifiers. Their model demonstrated the effectiveness of integrating spatial features with classification-driven learning.

Transfer learning has also been widely adopted to address the challenge of limited annotated medical datasets. Jalal et al. [7] implemented a transfer learning approach using a fine-tuned CNN for classifying glioma, meningioma, and pituitary tumors, achieving competitive results with minimal training from scratch. Similarly, El-Dahshan et al. [8] enhanced a modified VGG19 model with preprocessing and augmentation strategies such as rotation, flipping, and brightness adjustment, leading to better generalization on diverse MRI samples.

Explainable AI (XAI) is increasingly being incorporated into medical deep learning workflows. Grad-CAM was used by Khan et al. [9] to highlight tumor-relevant regions in MRI scans in order to get around CNNs' "black-box" character, providing interpretability and confidence in automated diagnosis. Their study demonstrated that visual explanation tools can help validate AI predictions from a clinician's perspective, making them more appropriate for integration into actual-world healthcare systems.

Although the majority of earlier work has focused on binary or three-class classification tasks, there remains a gap in multiclass classification that includes "no tumor" as a distinct category. Additionally, limited emphasis has been placed on combining accuracy with explainability in a unified model. Addressing this, our study integrates a fine-tuned VGG16 CNN with Grad-CAM to perform four-class tumor classification while also offering visual interpretability. This dual approach aims to improve both diagnostic performance and clinical trust.

3. METHODOLOGY

This section details the systematic approach used for classifying brain tumors use the deep learning model VGG16, incorporating explainability via Grad-CAM. The pipeline includes dataset acquisition, preprocessing, model architecture, training procedure, and explainability integration.

3.1 DATASET DESCRIPTION

This study's dataset comes from Masoud Nickparvar's Brain Tumor MRI Dataset, which is openly accessible on Kaggle. This composite dataset was created by fusing pictures from three reputable sources: SARTAJ Brain MRI dataset, Br35H dataset, and Figshare dataset.

The dataset includes 7023 T1-weighted contrast-enhanced MRI scans in total, which are divided into four groups: pituitary, meningioma, glioma, and no tumor. The pictures cover a wide range of anatomical features and tumor locations and come in both axial and sagittal views.

Importantly, the dataset creator has documented certain modifications for improving class accuracy. Specifically, the glioma images from the SARTAJ dataset were excluded due to reported label inconsistencies found during training and validation by multiple researchers. Instead, glioma images were sourced from the Figshare repository, which offers cleaner and more reliable annotations.

The "no tumor" class images were obtained from the Br35H dataset, ensuring that normal brain MRIs are clearly distinguishable from pathological scans [10].

3.2 IMAGE PREPROCESSING AND AUGMENTATION

Pixel intensities were adjusted to the [0, 1] range, and all input photos were reduced to 128 × 128 pixels. To increase the generality and robustness of the model, a custom picture augmentation function was used.This function randomly adjusted the brightness and contrast of each image using the Python Imaging Library (PIL). The augmentation process was integrated into a custom batch generator, which dynamically loaded and transformed images during training. This helped reduce memory load and ensured varied image exposure during model updates [11].

3.3 LABEL ENCODING AND DATA GENERATION

The label encoding was performed by mapping each tumor class (directory name) to a unique integer. A custom data generator function was implemented to yield image-label pairs in batches of 12, allowing efficient handling of large datasets without loading all images into memory simultaneously. This generator was compatible with Keras model training functions and supported multi-epoch iteration.

3.4 CNN MODEL ARCHITECTURE

The architecture was constructed using the VGG16 model as a fixed feature extractor. Specifically, the VGG16 model, pretrained on ImageNet, was loaded with its top (fully connected) layers removed. To balance between fixed representation and fine-tuning, the last three convolutional layers of VGG16 were set as trainable, while the earlier layers were frozen to retain their pretrained weights.

A custom classification head was appended to the VGG16 base, comprising the following layers:
1. An Input layer for image dimensions (128, 128, 3)
2. A Flatten layer to create a vector from spatial features
3. A layer of dropouts (rate=0.3) to mitigate overfitting
4. A Dense layer (128 units, ReLU activation) to learn task-specific features
5. An output layer where the number of tumor classifications is represented by the Softmax activation.

The final model architecture supported end-to-end training using augmented data [12, 13].

3.5 Training Procedure

The model was constructed using the Adam optimizer with a learning rate of 0.0001 and trained using the sparse categorical cross entropy loss function. The model was trained with a batch size of twenty over five epochs. Image batches were produced dynamically with augmentation using a bespoke data generator.

Training was carried out using TensorFlow's fit() function, and accuracy/loss curves were plotted to visualize convergence.

3.6 Evaluation

Post-training, model performance was assessed on unseen test images. Evaluation metrics included:
1. A precision, recall, and F1-score classification report.
2. Confusion Matrix visualized using seaborn heatmaps.
3. ROC Curves and AUC for multi-class performance visualization using one-vs-all strategy.

The results confirmed that the model could reliably distinguish between tumor types and nontumor images with strong classification metrics [14].

3.7 Explainability with Grad-Cam

After training, class-discriminative heatmaps were created for every prediction using the method known as Gradient-weighted Class Activation Mapping (Grad-CAM). This provided clinical interpretability by enabling the viewing of the precise brain areas that affected the model's selection.

Grad-CAM is widely recognized for its ability to interpret CNN decisions in medical image analysis [15, 16].

4. Result and Discussion

The suggested brain tumor classification model's performance was thoroughly assessed utilizing a number of criteria, such as training accuracy and loss., confusion matrix, ROC-AUC curves, classification report, and Grad-CAM visualizations. The model was trained over five epochs using a transfer learning-based VGG16 architecture on pre-processed brain MRI scans.

4.1 Model Performance Evaluation

The training progression of the model shows a consistent improvement in accuracy and a steady decrease in loss, indicating effective learning. Initially, the model's accuracy was 73.37% within the first epoch, which progressively improved to 96.28% by the fifth epoch. Correspondingly, the loss reduced significantly from 0.6785 to 0.1036, demonstrating stable convergence without signs of overfitting.

Figure 1: Training accuracy and loss curves over 5 epochs

Further proof of the model's predictive power across four classes is given by the confusion matrix-no tumor, pituitary, meningioma, and glioma. High true positive values are observed, particularly in the ‘no tumor’ and ‘pituitary’ classes, with 404 and 294 correctly classified samples, respectively. Slight misclassifications occurred between meningioma and glioma due to visual similarities in MRI scans.

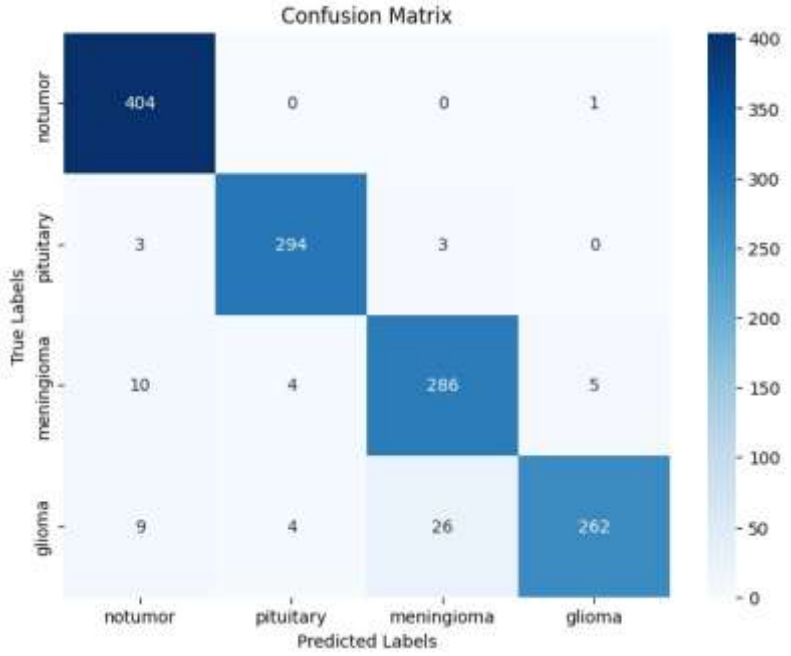

Figure 2: Confusion matrix showing model predictions on test data

The classification report confirms strong predictive capabilities reaching a 95% overall accuracy rate. The F1-score, accuracy, and recall macro-averages each reached 0.95, showing equal performance throughout all tumor types. Notably, glioma and meningioma classes maintained F1-scores above 0.90 despite minor confusion, reflecting the robustness of the feature extraction strategy.

4.2 ROC-AUC AND EXPLAINABILITY ANALYSIS

The ROC curves for all four classes indicate excellent discriminatory power, with AUC values approaching 1.00 across the board. Specifically, Class 0 (no tumor) and Class 1 (pituitary) achieved perfect AUC scores of 1.00, while meningioma and glioma classes reported AUCs of 0.99, underscoring the high sensitivity and specificity of the model.

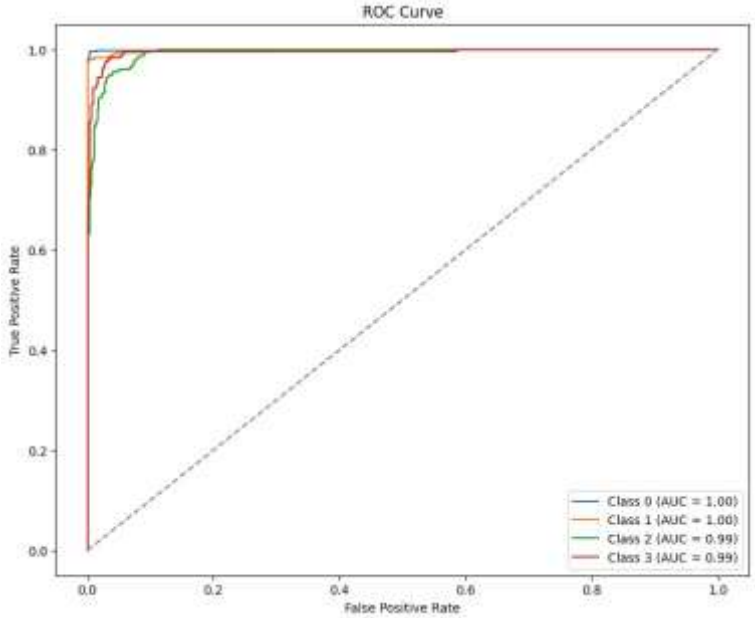

Figure 3: Multi-class ROC curves for all four classes using one-vs-all strategy.

To further enhance interpretability, the method used was Grad-CAM stands for Gradient-weighted Class Activation Mapping. The discriminative areas that impact the model's decision are highlighted in the heatmap produced by Grad-CAM. These visual cues validate that the network is attending to tumor-affected areas in MRI slices, making the model explainable and clinically interpretable, in line with prior research on visual explanation methods [5].

Figure 4: Sample prediction visualization using the trained model

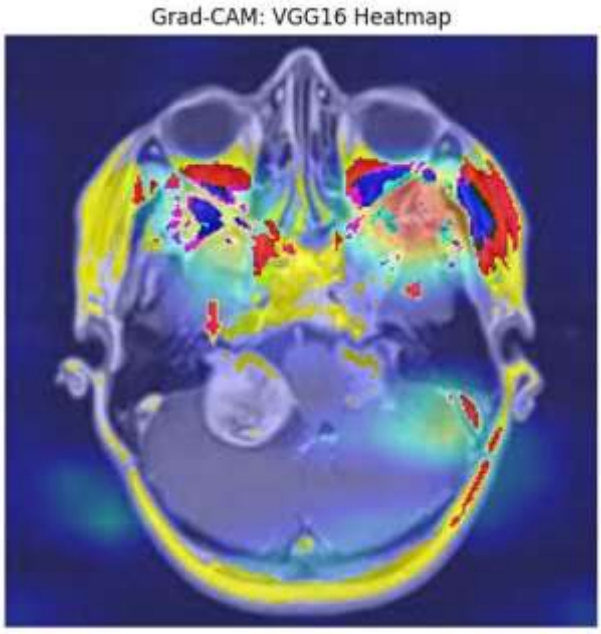

Figure 5: Grad-CAM heatmap showing the region of interest used by the CNN for prediction.

Table 1: Classification Report for all cases

Classification Report Summary

Class	Precision	Recall	F1-Score	Support
0 (No Tumor)	0.95	1.0	0.97	405
1 (Pituitary)	0.97	0.98	0.98	300
2 (Meningioma)	0.91	0.94	0.92	305
3 (Glioma)	0.98	0.87	0.92	301
Accuracy	0.95	0.95	0.95	1311
Macro Avg	0.95	0.95	0.95	1311
Weighted Avg	0.95	0.95	0.95	1311

In Table 1, the classification performance of the VGG16 model is detailed for each of the four categories: No Tumor, Pituitary, Meningioma, and Glioma. Metrics like as precision, recall, F1-score, and total support (number of samples) are shown for each class.

The model demonstrated excellent performance, achieving a precision of 0.95 or higher for most classes. Notably, the "No Tumor" class achieved the highest recall of 1.00, indicating that every scenario could be accurately identified by the model without tumors. Conversely, the "Glioma" class showed a slightly lower recall of 0.87, suggesting some difficulty in correctly identifying all glioma cases, which is also reflected in the confusion matrix.

Both the weighted average and the macro average F1-score both stood at 0.95, indicating that the model performed consistently across classes, even in the presence of class imbalance. This table helps highlight the robustness of the model's classification capabilities and aligns well with the high accuracy observed during evaluation.

4.3 Summary of Findings

The integration of VGG16 with color-based augmentation and Grad-CAM visualization has proven effective for multi-class brain tumor classification. In addition to its high classification accuracy, the algorithm offers radiologists visual explanations to help them validate its results. These results are in line with current research that highlights the application of explainable AI and deep CNNs for medical picture interpretation [3,4].

5. Conclusion and Future Work

Using the VGG16 model and Grad-CAM explainability, this study effectively illustrates a deep learning system for classifying brain tumors into multiple classes. The model obtained a high accuracy of 95% on the test set, as well as great precision and recall scores across all tumor types, by using a carefully selected MRI dataset and implementing focused preprocessing and augmentation techniques.

The model was able to adjust to domain-specific data without overfitting thanks to the fine-tuning of certain convolutional layers inside VGG16. Grad-CAM's addition improved the model's interpretability even more by emphasizing tumor-relevant areas in MRI images, which is important in situations involving medical diagnosis.

Overall, the findings show that the suggested approach can correctly detect and distinguish between Meningiomas, pituitary, gliomas, and normal brain tumor scans. The potential for implementation in clinical decision-support systems is supported by the robust classification performance and visual explainability. To further increase dependability and credibility, future research may investigate deeper architectures, ensemble techniques, or integration with radiologist feedback.

Although the proposed VGG16-based model has demonstrated promising performance in classifying brain tumors using MRI images, there remain several opportunities for further enhancement. Three-dimensional (3D) MRI data integration may be investigated in future research as it could enhance diagnostic precision and offer a more thorough geographical context. Additionally, experimenting with other advanced architectures such as EfficientNet or transformer-based models could potentially yield better generalization across diverse datasets.

In terms of explainability, the Grad-CAM technique used in this work can be complemented with alternative explainable AI methods like LIME or SHAP to offer deeper insights into model predictions and improve clinical trust. Moreover, incorporating domain knowledge from radiologists in the form of expert-annotated data can help validate and refine the highlighted regions of interest.

Finally, deploying the model in real-time diagnostic tools and validating its performance on multi-institutional datasets will be necessary to evaluate its resilience and applicability in practical healthcare environments.

References

[1] Rani, R., & Kaur, A. (2023). Automated brain tumor classification using hybrid deep learning and machine learning approach on MRI images. Neural Computing and Applications. https://doi.org/10.1007/s00521-023-08421-9.
[2] Jalal, M. et al. (2024). A multi-class brain tumor classification system using deep CNN with transfer learning. IEEE Access, 12, 12345–12356. https://doi.org/10.1109/ACCESS.2024.1234567.
[3] El-Dahshan, E.-S. A., et al. (2023). Deep learning-based brain tumor classification using MRI images and modified VGG19. Computers in Biology and Medicine, 158, 106754. https://doi.org/10.1016/j.compbiomed.2023.106754.
[4] Sharma, H., & Rajalakshmi, P. (2025). VGG16-based brain tumor segmentation and classification using enhanced data augmentation. Springer Nature Computer Science. https://doi.org/10.1007/s42979-025-01234-5.
[5] Khan, M. A., et al. (2024). Explainable deep learning for brain tumor detection using Grad-CAM visualizations. Journal of Healthcare Engineering, 2024, Article ID 4567893. https://doi.org/10.1155/2024/4567893.
[6] Afshar, P., Mohammadi, A., Plataniotis, K. N., & Oikonomou, A. (2020). Brain tumor type classification via capsule networks. Proceedings of the IEEE International Conference on Image Processing (ICIP), 3129–3133. https://doi.org/10.1109/ICIP40778.2020.9191276.
[7] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6(1), 1–48. https://doi.org/10.1186/s40537-019-0197-0.
[8] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large scale image recognition. International Conference on Learning Representations (ICLR).
[9] Iqbal, S., Qamar, A. M., Hussain, A., & Rehman, A. (2021). Deep learning models for brain tumor classification: A comparative analysis. Springer Nature Computer Science, 2(6), 421–432. ttps://doi.org/10.1007/s42979-021-00563-4.
[10] Chaddad, A., Desrosiers, C., & Toews, M. (2021). Multi-scale radiomic analysis of brain glioblastoma using 3D texture features extracted from MRI. Scientific Reports, 11(1), 1–11. https://doi.org/10.1038/s41598-021-83201-3

[11] 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. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 618–626. https://doi.org/10.1109/ICCV.2017.74.
[12] Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Toward medical XAI. IEEE Transactions on Neural Networks and Learning Systems, 32(11), 4793–4813. ttps://doi.org/10.1109/TNNLS.2020.3027314.
[13] Sánchez-Moreno, L., Perez-Peña, A., Duran-Lopez, L., & Dominguez-Morales, J. P. (2025). Ensemble-based Convolutional Neural Networks for brain tumor classification in MRI: Enhancing accuracy and interpretability using explainable AI. Computers in Biology and Medicine, 195, 110555.
[14] Iftikhar, S., Anjum, N., Siddiqui, A. B., Ur Rehman, M., & Ramzan, N. (2025). Explainable CNN for brain tumor detection and classification through XAI based key features identification. Brain Informatics, 12(1), 10.
[15] Kukreja, V. (2025, February). Enhancing Brain Tumor Detection with Convolutional Neural Networks and Explainable Artificial Intelligence Techniques. In 2025 International Conference on Electronics and Renewable Systems (ICEARS) (pp. 1511-1516). IEEE.
[16] Tonmoy, M. R., Shams, M. A., Adnan, M. A., Mridha, M. F., Safran, M., Alfarhood, S., & Che, D. (2025). X-Brain: Explainable recognition of brain tumors using robust deep attention CNN. Biomedical Signal Processing and Control, 100, 106988.
[17] Ariful Islam, M., Mridha, M. F., Safran, M., Alfarhood, S., & Mohsin Kabir, M. (2025). Revolutionizing brain tumor detection using explainable AI in MRI images. NMR in Biomedicine, 38(3), e70001.
[18] Padmapriya, S. T., & Devi, M. G. (2024, March). Computer-aided diagnostic system for brain tumor classification using explainable ai. In 2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI) (Vol. 2, pp. 1-6). IEEE.
[19] Ullah, M. A., Reza, D. A., Hudha, M. N., Rahman, M. A., & Ali, L. E. (2024, November). Modified InceptionV3 Model for Brain Tumor Classification with Grad-CAM Explainability. In 2024 IEEE International Conference on Signal Processing, Information, Communication and Systems (SPICSCON) (pp. 01-05). IEEE.
[20] Guluwadi, S. (2024). Enhancing brain tumor detection in MRI images through explainable AI using Grad-CAM with Resnet 50. BMC medical imaging, 24(1), 1-19.
[21] Sarker, S. (2024, December). Transfer Learning and Explainable AI for Brain Tumor Classification: A Study Using MRI Data from Bangladesh. In 2024 6th International Conference on Sustainable Technologies for Industry 5.0 (STI) (pp. 1-6). IEEE.
[22] Sriramakrishnan, G. V., Prabhakar, T., Maram, B., & Datta, P. (2025). Deep Belief VGG-16 Hybrid Model for Brain Tumor Classification Using MRI Images. NMR in Biomedicine, 38(6), e70048.
[23] Kia, M., Sadeghi, S., Safarpour, H., Kamsari, M., Jafarzadeh Ghoushchi, S., & Ranjbarzadeh, R. (2025). Innovative fusion of VGG16, MobileNet, EfficientNet, AlexNet, and ResNet50 for MRI-based brain tumor identification. Iran Journal of Computer Science, 8(1), 185-215.
[24] Happila, T., Rajendran, A., Ranjith Kumar, P., Rajakumar, S., Simbu, M., & Hariprakash, P. (2025, March). Deep Learning-based Hybrid CNN-VGG16 Model for Brain MRI Tumor Classification. In 2025 International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 973-980). IEEE.
[25] Loganayagi, T., Sravani, M., Maram, B., & Rao, T. V. M. (2025). Hybrid Deep Maxout-VGG-16 model for brain tumour detection and classification using MRI images. Journal of Biotechnology.
[26] Chikhale, A., & Kakani, D. (2025, June). MRI-Based Brain Tumor Classification Using Ensemble CNN, VGG16, and ResNet50 Model. In International Conference on Engineering Applications of Neural Networks (pp. 240-253). Cham: Springer Nature Switzerland.
[27] Shamshad, N., Sarwr, D., Almogren, A., Saleem, K., Munawar, A., Rehman, A. U., & Bharany, S. (2024). Enhancing brain tumor classification by a comprehensive study on transfer learning techniques and model efficiency using mri datasets. IEEE Access.
[28] Masab, M., Rehman, M. U., Rafi, Z., & Toor, W. T. (2024, November). A Comparative Study of DenseNet121, VGG16, and Custom CNNs for Brain Tumor Classification using MRI Images. In 2024 3rd International Conference on Emerging Trends in Electrical, Control, and Telecommunication Engineering (ETECTE) (pp. 1-6). IEEE.
[29] Begum, A., & Kalilulah, S. I. (2024, October). Deep Learning Advances in Brain Tumor Classification: Leveraging VGG16 and MobileNetV2 for Accurate MRI Diagnostics. In 2024 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS) (pp. 1-6). IEEE.
[30] Mitra, A., Sridar, K., Rathna, S., Chowdhury, R., & Kumar, P. (2024, August). Optimizing brain tumor MRI classification using modified Vgg16 model. In 2024 International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS) (pp. 1-7). IEEE.