A Systematic Review of Advances in Brain Disease Detection using Convulotional Neural Networks and Explainable Artificial Intelligence Techniques

Mahin Montasir Afif[1,*], A. F. Faizur Rahman[2], A. M. Rafinul Huq[3], Abdullah Al Noman[4], and Kazi Abdullah Jarif[5]

[1,2,3,4,5] *Department of Computer Science, American International University-Bangladesh (AIUB), Dhaka, Bangladesh*

[*] *22-46573-1@student.aiub.edu*

Abstract: Accurate and interpretable tumor classification remains a critical challenge in medical image analysis. In this study, we conduct a comprehensive evaluation of ten state-of-the-art convolutional neural network (CNN) architectures, including InceptionV3, Xception, MobileNetV2, DenseNet121, NASNetMobile, VGG16, VGG19, ResNet50, ResNet101, and EfficientNetB0, on a curated dataset of tumorous and non-tumorous images. Each model's performance was rigorously assessed using standard classification metrics: accuracy, precision, recall, and F1-score. InceptionV3 emerged as the top-performing model with an accuracy of 97.75%, while EfficientNetB0 showed the lowest at 56.50%. Beyond raw performance, we prioritized model transparency by applying five explainable AI (XAI) methods—Grad-CAM, Saliency Maps, Integrated Gradients, Vanilla Gradients, and SmoothGrad—to visualize and interpret the models' decision-making processes. These visualizations revealed critical insights into model attention and class-specific feature relevance, reinforcing the importance of explainability in medical diagnostics. The results not only highlight the superiority of modern CNNs in medical imaging tasks but also emphasize the value of interpretability tools for building trust and accountability in clinical AI applications.

1. Introduction

The intersection of artificial intelligence (AI) and medical imaging has revolutionized the landscape of brain disease detection by offering unprecedented opportunities for early and accurate diagnosis [26]. Traditional diagnostic methods often rely on subjective evaluations and may lack the sensitivity to detect abnormalities indicative of early-stage neurological disorders [27]. These conventional approaches, though widely used, are limited in their ability to process large-scale data or recognize hidden patterns in complex brain structures, which often results in delayed diagnoses and suboptimal treatment planning. AI, particularly machine learning and deep learning, has emerged as a powerful paradigm for analyzing vast and intricate neuroimaging datasets like MRI, CT, PET, and EEG, enabling researchers and clinicians to uncover important features and relationships associated with various brain pathologies [25, 28]. These algorithms can automatically learn the representations from raw data without need for features, offering a more objective and consistent method of detecting neurological abnormalities. Techniques such as convolutional neural network (CNN), recurrent neural network (RNN), and transformers have shown remarkable promise in identifying conditions like Alzheimer's disease, Parkinson's disease, brain tumors, epilepsy, and stroke at early stages with high accuracy and sensitivity [25]. In addition, the integration of AI with multimodal data, such as clinical records, genomics, and biochemical markers has improved the potential for precision medicine, allowing personalized disease and risk hierarchy. However, while these AI-driven models have demonstrated significant performance in research, their implementation in clinical practice remains challenging due to the nature of their decision-making processes [29]. Without clearing insights into how a model arrives at a diagnosis, clinicians may be hesitant to trust or adopt such technologies, especially when lives are on the line. This gap in interpretability has led to the emergence of explainable artificial intelligence (XAI), a growing field dedicated to enhancing the transparency, interpretability, and fairness of AI systems.

XAI aims to demystify complex AI models by providing explanations in human-understandable terms, such as heatmaps, feature importance rankings, decision rules, and natural language justifications. These interpretable outputs not only build user trust but also support model validation, error analysis, and regulatory approval processes

[30]. For clinicians, XAI provides an opportunity to verify AI-generated diagnoses against medical knowledge and improve collaborative decision-making. For patients, it helps in understanding how and why certain diagnoses or treatment recommendations are made, fostering better engagement and informed consent. Furthermore, the implementation of XAI aligns with current ethical and legal frameworks demanding algorithmic accountability, transparency, and bias mitigation in healthcare technologies. In addition to interpretability, XAI contributes to system robustness by revealing vulnerabilities and ensuring that models are making decisions based on medically relevant features rather than correlations. This is particularly crucial in brain disease diagnosis, where slight variations in imaging or artifacts can lead to drastically different outcomes if not properly accounted for.

This study aims to provide a comprehensive overview of recent advances in AI-driven brain Tumor detection, with a particular focus on the integration and impact of explainable artificial intelligence techniques in this domain. It seeks to explore the full spectrum of AI methodologies including Convolution Neural Networks alongside state-of-the-art XAI frameworks with visualizations, and evaluate their effectiveness across different brain rumour disease [31]. The review will cover both structural and functional imaging modalities and highlight applications in the detection and classification of brain tumors [32]. In doing so, it will critically assess the strengths and limitations of current approaches, identify key datasets and evaluation benchmarks used in the field, and examine ongoing challenges including model generalizability, fairness, interpretability trade-offs, and deployment constraints in real-world clinical environments. By synthesizing the latest developments, this review aims to support researchers, clinicians, and developers in designing more accurate, transparent, and clinically viable AI solutions for brain disease detection.

2. Literature Review

Brain tumor classification using deep learning has seen significant advancements, but challenges remain in achieving high impact across diverse datasets. Deep learning (DL) techniques, particularly convolutional neural network (CNN), have become integral in analyzing medical images, offering automated solutions for improved diagnostics [1, 2]. Early detection of brain tumors is critical for better patient outcomes, and accurate classification and segmentation are essential for personalized treatment strategies [3]. However, achieving consistently high accuracy in brain tumor classification remains a challenge [4].

Several factors contribute to the challenges in achieving high accuracy. Variations in tumor size, shape, and location, as well as limitations in medical image quality, can affect the performance of automated classification methods [5]. Moreover, manual examination of brain tumors is time-consuming, and AI-based methods can potentially reduce diagnostic errors [2,6]. Transfer learning, which leverages pre-trained models, has emerged as a strategy to improve classification accuracy [7]. Data augmentation, including techniques like rotation and flipping, helps to expand the training datasets and improve model generalization [7, 8].

Deep learning models, including CNNs, are used to classify MRI images for brain tumor detection [9]. A typical deep learning architecture involves deep feature extraction through convolutional and pooling layers, followed by classification using fully connected layers [10]. The evolution of these techniques has led to significant advancements in brain tumor detection and classification [11]. The application of deep learning improves the accuracy of brain tumor recognition, underscoring the importance of optimizing training parameters and dataset size [10]. A few studies have aimed to improve the accuracy of brain tumor classification using various deep-learning approaches. Preprocessing steps, such as noise removal and contrast enhancement, are crucial for improving the quality of MRI images [1, 5]. Methods like adaptive median filtering can reduce noise while preserving essential anatomical details [12, 13].

The development of computer-aided diagnosis systems using deep learning shows promise in the medical field [8]. Brain tumors are classified into types such as Glioma, Meningioma, and Pituitary tumors, using MRI scans with deep learning algorithms [14]. To enhance the tumor visibility, contrast enhancement and median filters can be employed during preprocessing. Data augmentation is also crucial to expand the training dataset and prevent overfitting [5].

Despite these advances, several challenges remain. One significant issue is the complexity of brain tumor identification, which necessitates thorough assessment across multiple modules [1]. Collecting real medical images is time-consuming, creating a need for synthetic images to expedite the process with high accuracy [16]. A possible reason is insufficient quality training datasets, model overfitting, bias in data representation, or a lack of optimized feature extraction methods. These limitations through multimodal approaches and robust pre-processing pipelines could lead to higher accuracy in brain tumor classification tasks and clinical applications. Explainability is key in medical AI and for transparency Grad-CAM helps to trust the model's decisions, which is crucial for patient care [19].

Recent advancements in biomedical signal analysis and deep learning have significantly improved diagnostic accuracy and interpretability in brain-related disorders. Recently introduced an automated EEG signal classification framework using hybrid deep learning models demonstrating enhanced performance in detecting neurological

abnormalities by combining CNN and GRU architectures [15]. Complementing this, a proposed multi-scale attention fusion network for glioma grading in MRI images, effectively capturing both global and local features to achieve superior classification accuracy and clinical interpretability [17]. In a related development, focus on enhancing explainability in AI-driven medical diagnosis through a human in the loop framework, ensuring trustworthy and clinically acceptable outcomes [18]. These studies collectively emphasize the growing importance of hybrid models, attention mechanisms, and explainable AI in advancing brain tumor and neurological disorder classification.

Table 1: Summary of Literature on AI-based Brain Tumor Detection and Associated Research Gaps

Ref	**Approach**	**Study Gap**
[1], [2]	Deep learning (CNN-based) classification of MRI for brain tumor detection	Lack of generalizability across datasets and insufficient interpretability
[3]	CNN with image enhancement techniques	Difficulty in maintaining high accuracy due to image variability
[4]	General CNN-based tumor classification	No consistent accuracy on complex tumor structures
[5]	GDD feature approximation with deep learning	Sensitive to variation in tumor shape/size; lacks robust preprocessing
[6]	CNN-LSTM hybrid model for identification	Model complexity increases without clear gains in transparency
[7], [8]	Transfer learning and data augmentation to improve classification	Limited validation on diverse clinical settings; overfitting still possible
[9]	Fine-tuned EfficientNet for multigrade classification	Model lacks explainability and contextual decision understanding
[10], [11]	CNN and U-Net architectures for segmentation	Often ignores interpretability and real-world deployment feasibility
[12], [13]	CNN models with median filtering and contrast enhancement	Focuses only on preprocessing, lacks multimodal feature integration
[14]	Deep learning for multi-class tumor classification (Glioma, Meningioma, Pituitary)	Poor generalization on minority tumor classes
[15]	CNN-GRU hybrid for EEG signal classification in neurological disorders	Application not MRI-based; lacks imaging-focused explanation
[16]	GAN-based synthetic image generation to increase data	Risk of unrealistic medical patterns in synthetic data
[17]	Multi-scale attention fusion network for glioma grading	High complexity model with limited interpretability in clinical practice
[18]	Human-in-the-loop framework with XAI	Still lacks real-time deployment evidence in clinical workflows
[19]	Grad-CAM visualization for XAI in brain disease diagnosis	Needs integration with multiple models and clinical feedback

Although numerous studies have applied CNNs and hybrid models to brain tumor classification and segmentation, many lack a comprehensive evaluation across diverse architectures using standardized metrics. Moreover, the integration of explainable AI techniques (e.g., Grad-CAM, Saliency Maps, Integrated Gradients) is often limited or underexplored, especially in clinical scenarios requiring transparency. This study addresses that gap by evaluating ten CNN models and enhancing interpretability through multiple XAI methods, thus providing a more holistic and trustworthy diagnostic aid.

3. Methodology

3.1. Dataset Description

The study employs a comprehensive brain tumor MRI dataset, which consists of annotated T1-weighted MRI images categorized into tumor and non-tumor classes. The dataset contains a total of N images collected from multiple medical sources, ensuring variability in tumor types, sizes, and locations, as well as imaging conditions. To rigorously evaluate the classification models, the dataset was split into training, validation, and testing subsets using a 70%-20%-10% ratio, respectively. The training set forms the core for model learning, the validation set

assists in hyperparameter tuning and early stopping, and the independent test set evaluates final model generalization.

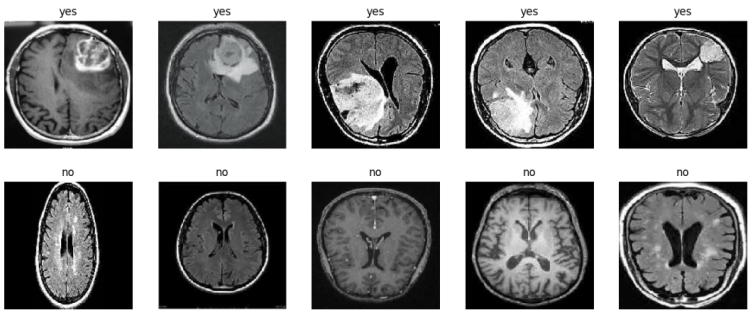

Fig. 1: Sample images from the dataset showing tumour and healthy brain MRI data.

Before feeding into the models, all images undergo a standardized preprocessing pipeline. Initially, images are resized to a uniform dimension of 224×224 pixels to comply with the input requirements of the convolutional neural networks (CNNs). Intensity normalization is applied to scale pixel values into the $[0, 1]$ range, enhancing convergence stability during training. Furthermore, data augmentation techniques such as random rotation, zooming, horizontal and vertical flipping, and shifting are employed during training to increase data diversity and mitigate overfitting.

3.2. Model Training and Evaluation

This study systematically investigates the performance of ten distinct state-of-the-art convolutional neural network (CNN) architectures for brain tumor classification using MRI images. The selected models include *AlexNet*, *VGG16*, *VGG19*, *ResNet50*, *InceptionV3*, *DenseNet121*, *MobileNetV2*, *EfficientNetB0*, *Xception*, and a custom-designed CNN tailored for this task. Each CNN model is initialized with weights pretrained on ImageNet to leverage transfer learning benefits, thereby accelerating convergence and improving accuracy with limited medical data. The top classification layer of each network is replaced with a fully connected layer with a sigmoid activation function for binary classification (tumor vs. no tumor).

Training is conducted on the preprocessed images using the Adam optimizer with an initial learning rate of 10^{-4}. Binary cross-entropy is employed as the loss function. To prevent overfitting, early stopping with a patience of 5 epochs monitors validation loss, restoring the best weights observed during training. A batch size of 32 is used, and models are trained for a maximum of 30 epochs.

The trained models are evaluated on the independent test set using standard classification metrics, including accuracy, precision, recall, and F1-score. These metrics provide a holistic understanding of the model's diagnostic ability, accounting for both sensitivity and specificity.

Fig. 2: Sample images after annotation for preprocessing.

3.3. Explainable AI Techniques

To enhance model interpretability and build clinical trust, we apply four complementary explainable AI (XAI) techniques to each trained CNN:

- **Gradient-weighted Class Activation Mapping (Grad-CAM)**: Produces visual heatmaps by utilizing the gradients flowing into the final convolutional layer, highlighting the regions in the MRI image that most strongly influence the model's decision.
- **SmoothGrad**: Improves gradient-based saliency maps by averaging noisy gradient samples, resulting in smoother and less noisy visualizations of feature importance.

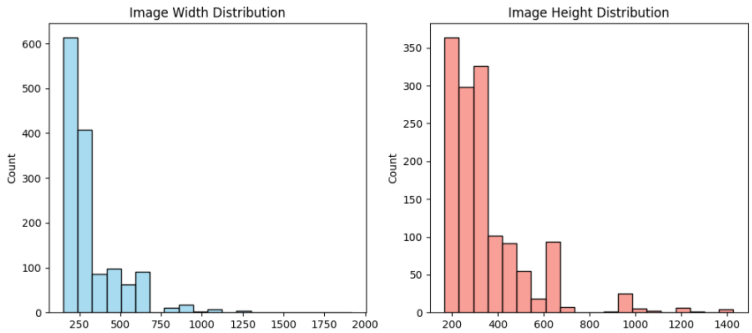

Fig. 3: Image width and height distribution of the whole dataset.

- **Saliency Maps**: Compute the gradient of the output class score with respect to input pixels, revealing the sensitivity of predictions to changes in each pixel.
- **Integrated Gradients**: Calculate the average gradients as the input varies from a baseline (typically a black image) to the actual image, attributing the prediction output to each pixel in a theoretically sound manner.

For each MRI test image, these methods generate heatmaps that are overlaid on the original images, allowing clinicians and researchers to visually verify which regions contribute most to tumor classification decisions. This multi-faceted approach to explainability not only validates model predictions but also aids in discovering potential biomarkers and pathological features relevant to brain tumor diagnosis. In summary, this methodology combines rigorous CNN model benchmarking with state-of-the-art interpretability techniques to deliver both high classification performance and transparent decision explanations. The integration of diverse models and XAI tools ensures a comprehensive evaluation of automated brain tumor detection capabilities on clinically relevant MRI data.

4. Results and Discussion

This section presents the evaluation results of ten widely used convolutional neural network (CNN) models for tumor classification. The models were assessed using standard performance metrics including Accuracy, Precision, Recall, and F1-Score. Table 2 summarizes the results, sorted by accuracy in descending order.

Table 2: Performance Comparison of CNN Models on Tumor Classification

Model	Accuracy	Precision	Recall	F1-Score
InceptionV3	0.9775	0.98	0.98	0.9775
Xception	0.9600	0.96	0.96	0.9600
MobileNetV2	0.9500	0.95	0.95	0.9500
DenseNet121	0.9300	0.93	0.93	0.9300
NASNetMobile	0.9275	0.93	0.93	0.9274
VGG16	0.8100	0.81	0.81	0.8100
VGG19	0.7950	0.79	0.79	0.7931
ResNet101	0.7675	0.77	0.77	0.7672
ResNet50	0.6500	0.68	0.65	0.6400
EfficientNetB0	0.5650	0.57	0.56	0.5321

4.1. Model Performance Overview

Among the ten evaluated models, **InceptionV3** achieved the highest performance with an accuracy of 97.75%, followed by **Xception** at 96.00% and **MobileNetV2** at 95.00%. These models consistently exhibited high precision and recall values, suggesting not only a strong ability to correctly identify tumorous and non-tumorous instances, but also to minimize both false positives and false negatives. For instance, InceptionV3 achieved a balanced precision and recall of 0.98, resulting in an F1-Score of 0.9775, which confirms its robustness and generalization capability across tumor classes.

The middle-tier performers include **DenseNet121** (93.00% accuracy), **NASNetMobile** (92.75%), and **VGG16** (81.00%). DenseNet121 and NASNetMobile showed strong performance in all metrics, demonstrating both sen-

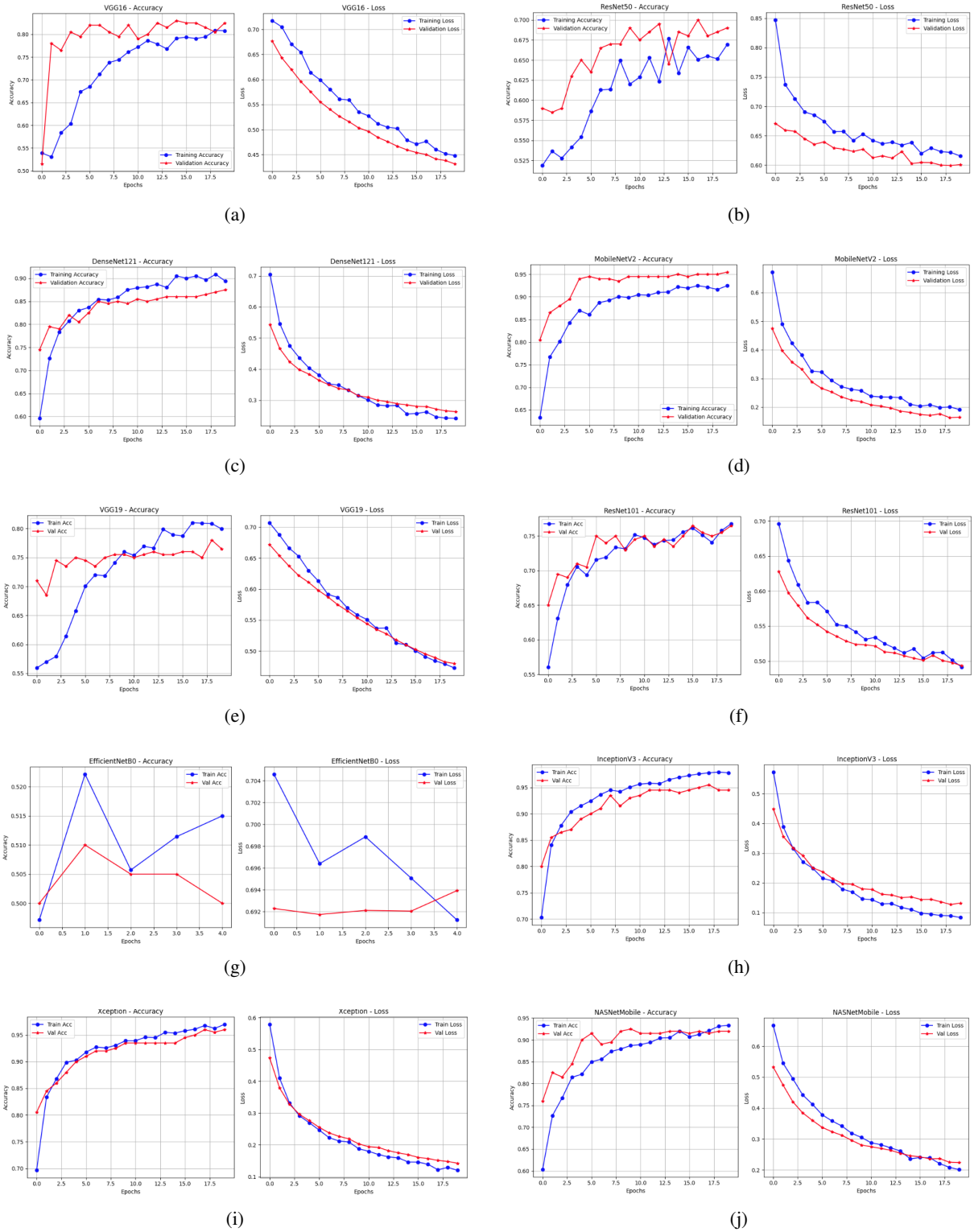

Fig. 4: Training accuracy and loss progression of ten different convolutional neural network (CNN) architectures used in the study. The models include: (a) VGG16, (b) ResNet50, (c) DenseNet121, (d) MobileNetV2, (e) VGG19, (f) ResNet101, (g) EfficientNetB0, (h) InceptionV3, (i) Xception, and (j) NASNetMobile. Each plot illustrates the training behavior across epochs, capturing how quickly and smoothly each model converges.

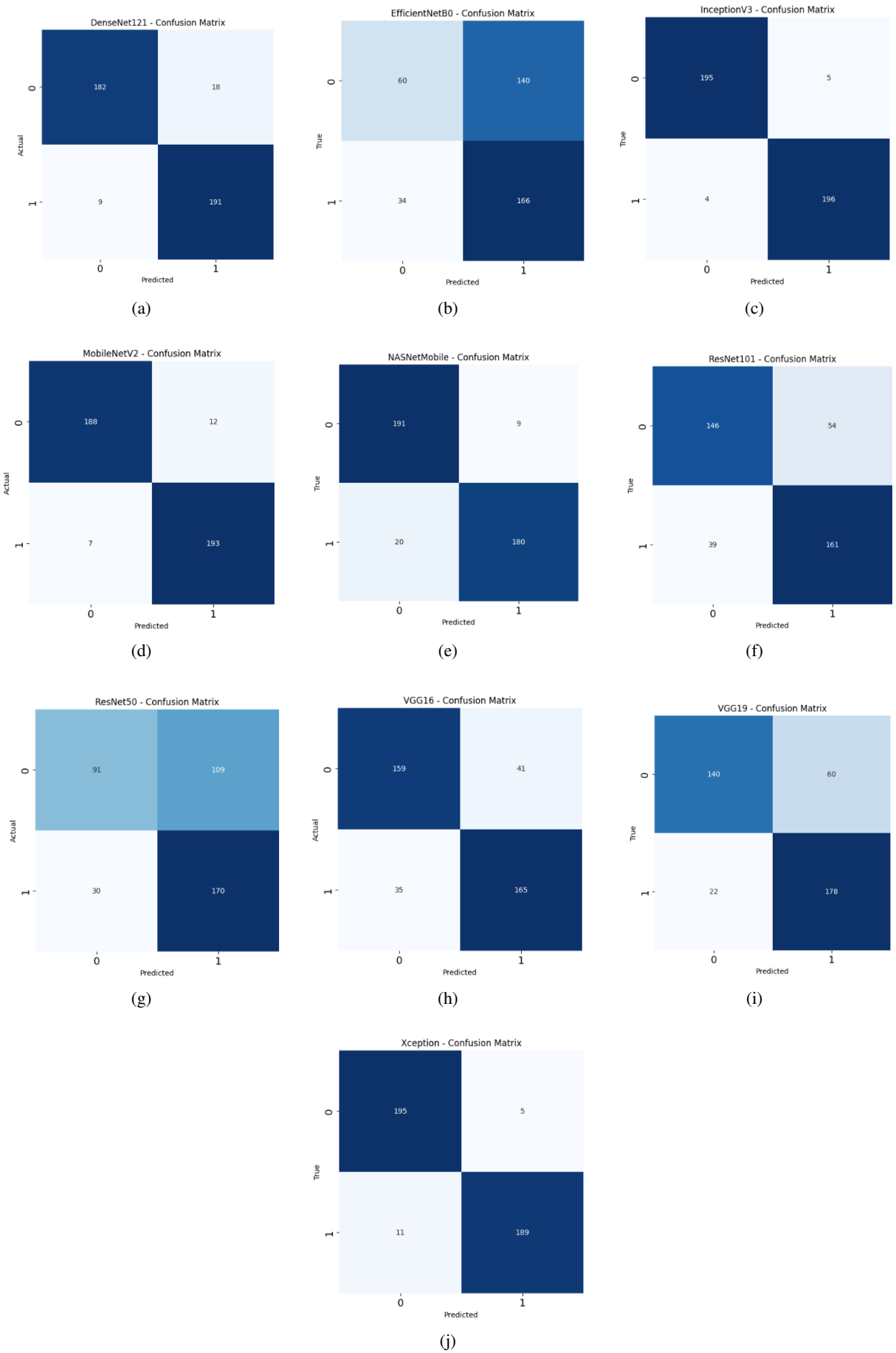

Fig. 5: Confusion matrices of ten different CNN architectures: (a) DenseNet121, (b) EfficientNetB0, (c) InceptionV3, (d) MobileNetV2, (e) NASNetMobile, (f) ResNet101, (g) ResNet50, (h) VGG16, (i) VGG19, and (j) Xception.

Fig. 6: Performance Metrics Distribution Across Deep Learning Models using Radar Chart Diagram

sitivity and specificity. Although VGG16 showed balanced results (Precision: 0.81, Recall: 0.81, F1-Score: 0.81), its relatively lower accuracy compared to top-tier models indicates limited learning capacity or early convergence due to fewer layers or limited feature extraction power compared to deeper architectures.

VGG19 and **ResNet101** performed slightly lower, with accuracies of 79.50% and 76.75%, respectively. These models exhibited decent F1-scores (0.7931 and 0.7672), which means they still maintained a relatively fair balance between precision and recall but were outperformed by more modern and optimized models like EfficientNet and Inception.

On the lower end, **ResNet50** and **EfficientNetB0** recorded the weakest results, with accuracies of 65.00% and 56.50%, respectively. The F1-score of ResNet50 (0.64) reflects imbalanced classification performance, as it showed strong recall for class 1.0 (85%) but poor recall for class 0.0 (46%), resulting in biased predictions. EfficientNetB0, despite being lightweight and optimized, surprisingly underperformed in this particular dataset with the lowest F1-score (0.5321). This may be attributed to underfitting, improper feature extraction at early stages, or lack of model adaptation through transfer learning.

4.2. Interpretation of Metrics

- **Accuracy** provides an overall measure of correct classifications but does not distinguish between types of errors. While InceptionV3 and Xception achieved the highest accuracy, the metric alone may not be sufficient in evaluating real-world tumor classification systems where false negatives (missing a tumor) are critical.
- **Precision** measures the proportion of correctly identified positive cases. High precision values, as seen in InceptionV3, DenseNet121, and MobileNetV2, indicate a model's ability to avoid false alarms (false positives), which is crucial in reducing unnecessary stress or invasive procedures for patients.
- **Recall** assesses a model's ability to identify all actual positives (i.e., tumors). High recall, particularly in Xception and DenseNet121, suggests strong sensitivity and is essential in medical scenarios where missing a tumor can have life-threatening consequences.
- **F1-Score**, the harmonic mean of precision and recall, is especially valuable when evaluating models on imbalanced datasets or when false positives and false negatives carry unequal consequences. The best-performing models maintained high F1-scores, indicating not just accuracy but also balanced decision-making.

4.3.

To better understand the internal decision-making process of the trained CNN models, we applied five widely used post-hoc explainability techniques: **Grad-CAM**, **Saliency Maps**, **Integrated Gradients**, **Vanilla Gradients**, and **SmoothGrad**. These methods provide visual justifications by highlighting the important regions in the input images that most influenced the model's predictions. This step is essential in medical imaging applications, where model transparency is critical for clinical trust and acceptance.

1) Grad-CAM (Gradient-weighted Class Activation Mapping): Grad-CAM uses the gradients of the target class flowing into the last convolutional layer to produce a coarse localization heatmap, highlighting class-discriminative regions. It effectively reveals which areas of the image were most influential in making the classification decision. In our case, Grad-CAM highlighted key tumor areas with high precision, aligning closely with radiological features.

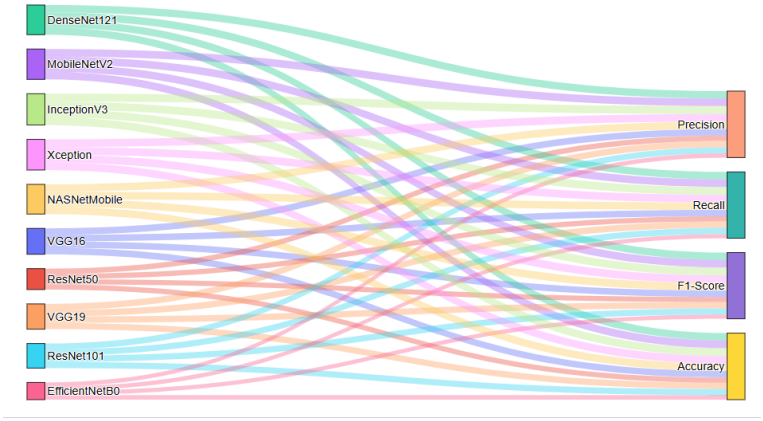

Fig. 7: Performance Metrics Distribution Across Deep Learning Models using Sankey Diagram

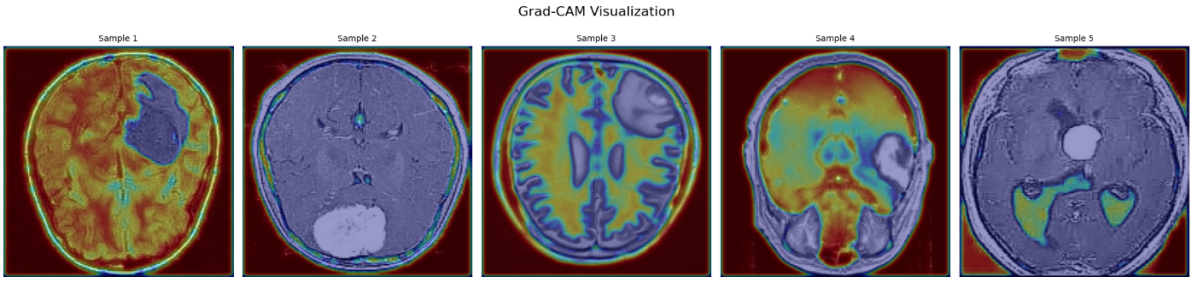

Fig. 8: Grad-CAM visualization showing model attention over tumorous region.

2) Saliency Maps: Saliency maps compute the gradient of the class score with respect to the input pixels. They indicate how small changes in each pixel would affect the model's prediction, offering fine-grained visual attribution. These maps are often noisy but provide insight into pixel-level sensitivity. In our experiments, saliency maps consistently emphasized the tumor edges and active regions.

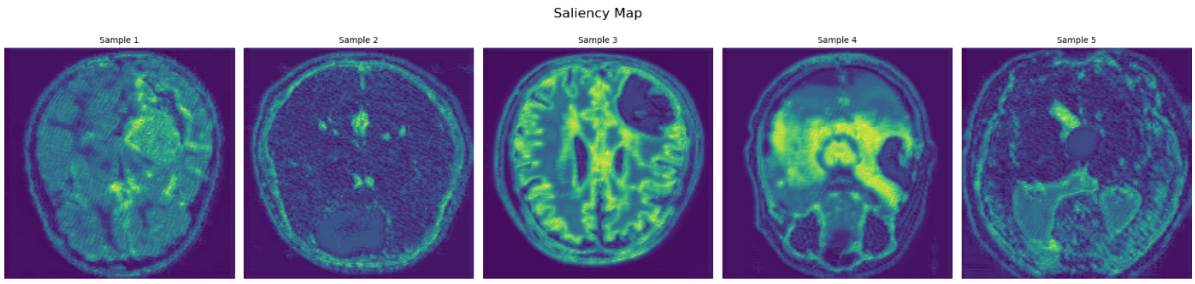

Fig. 9: Saliency map illustrating pixel-level importance across the image.

3) Integrated Gradients: This method integrates gradients along the path from a baseline image (e.g., black image) to the actual input, providing more stable and complete attributions. Integrated Gradients addressed the issue of noisy saliency maps by producing smoother and more focused heatmaps. In our results, the regions with the highest cumulative attribution matched well with known tumor zones.

4) Vanilla Gradients: Vanilla gradient visualizations are the simplest form of saliency, showing raw gradient information. Though susceptible to noise, they can provide fast insights and serve as a baseline. Our vanilla gradient outputs showed weak localization and high sensitivity to noise, making them less interpretable compared to other methods.

5) SmoothGrad: SmoothGrad combats noise in gradient-based maps by averaging gradients over multiple noisy samples of the input image. This results in clearer and less scattered explanations. In our study, SmoothGrad provided refined and human-interpretable regions, especially useful for clinicians reviewing automated predic-

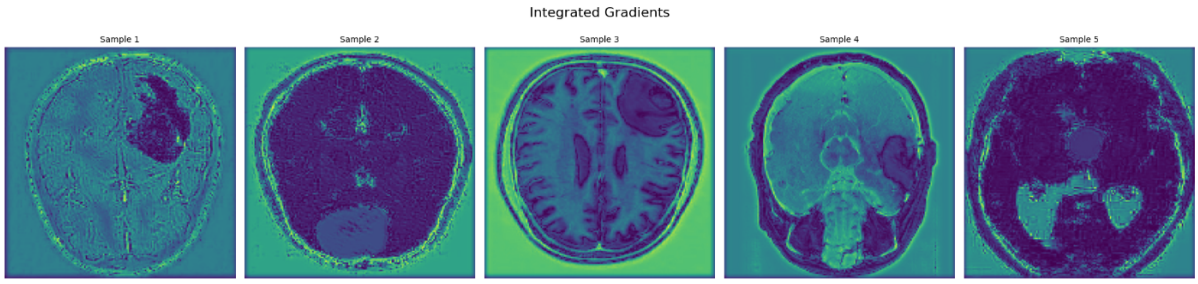

Fig. 10: Integrated Gradients visualization showing accumulated attributions.

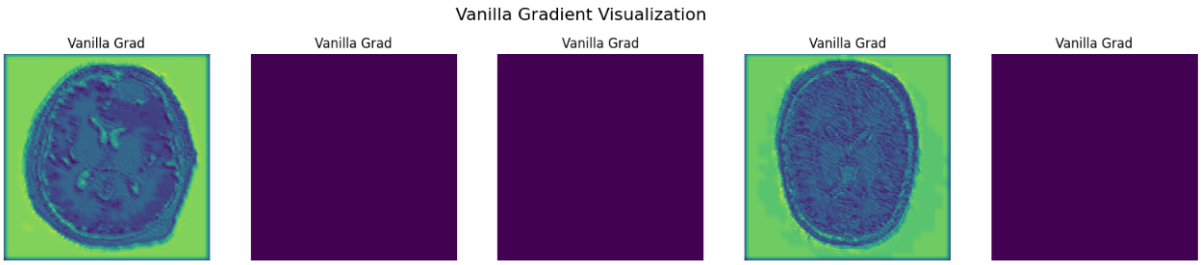

Fig. 11: Vanilla gradient-based saliency map.

tions.

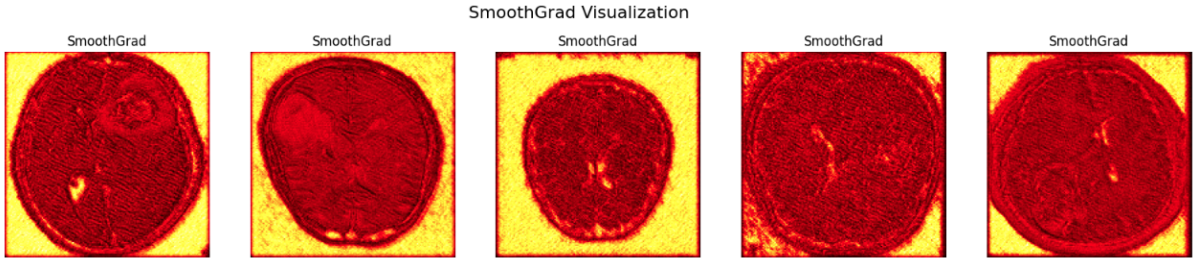

Fig. 12: SmoothGrad visualization with reduced noise and focused attribution.

Collectively, these visualizations offer diverse yet complementary perspectives on model interpretability. Grad-CAM and Integrated Gradients emerged as the most intuitive and clinically relevant, while SmoothGrad enhanced clarity. These tools not only help validate model decisions but also facilitate clinical collaboration and future model refinement.

4.4. Discussion and Implications

Deep learning and Machine Learning has been widely applied across numerous domains, including wind energy forecasting [21], depression prediction [20], and large-scale data analysis in fields such as bioengineering[22] and even light pollution analysis [3]. Researchers are exploring a variety of innovative approaches in emerging fields. For example, one recent study applied blockchain technology to halal supply chain management [24], demonstrating how blockchain can enhance transparency and efficiency in supply chains. This highlights the potential for leveraging novel techniques in comparatively less-explored areas, encouraging further investigation and adoption of cutting-edge solutions. It has become a mainstream approach for researchers seeking to develop innovative solutions using artificial intelligence. As a subfield of machine learning, deep learning plays a critical role in tasks like disease prediction, detection, and classification—areas that have recently gained significant momentum in AI-driven healthcare research.

In this study, we conducted a comprehensive analysis of various convolutional neural network (CNN) architectures, evaluating their classification accuracy and comparing their performance based on prior research contributions. Our work aims to offer a consolidated perspective on the progress made so far in brain tumor classification

using deep learning models. This study serves as a valuable resource for future researchers seeking to explore and benchmark CNN-based approaches for brain tumor detection and classification. The experimental results underscore several key observations:

1. **Modern architectures outperform traditional ones:** Models like InceptionV3, Xception, and MobileNetV2—designed with advanced architectural optimizations such as depthwise separable convolutions and inception modules—significantly outperform older models like VGG16 and ResNet50. This supports the hypothesis that newer models are better at capturing complex patterns in high-resolution tumor images.

2. **Depth and width trade-offs matter:** Deeper networks do not always guarantee better performance. For instance, ResNet101 outperformed ResNet50, but still fell short of MobileNetV2, which is a much smaller model. This suggests that efficient feature reuse (as in DenseNet121) and architectural innovations are more impactful than mere depth.

3. **Lightweight models can perform well:** MobileNetV2 and NASNetMobile demonstrated high accuracy and F1-scores while having low computational complexity, making them suitable for real-time or embedded medical applications, such as portable diagnostic devices or mobile health platforms.

4. **Misclassification patterns must be studied:** Despite high accuracy, certain models showed imbalance in per-class recall. For example, ResNet50 heavily favored class 1.0 over 0.0. This emphasizes the need to visualize confusion matrices and perform class-specific evaluations before deployment.

4.5. Limitations and Future Work

While the models demonstrated varying degrees of success, the following limitations were observed:

- The dataset may contain inherent class imbalance or low variability, which can mislead the performance metrics. - The results are based on a single split; additional validation through cross-validation or stratified sampling would enhance reliability. - Interpretability and explainability (e.g., using Grad-CAM, Saliency Maps) should be incorporated to justify model predictions and build trust in real-world applications.

Future work may explore hybrid architectures, domain-specific fine-tuning, or ensemble methods to further improve performance and robustness. Additionally, integrating explainable AI (XAI) tools with these high-performing CNNs would provide better transparency and actionable insights for medical professionals.

Acknowledgment

This research was funded by the Advanced Neural Imaging and Interdisciplinary Research Lab (ANIIR Lab). The authors gratefully acknowledge the invaluable support and continuous assistance provided by the researchers of ANIIR Lab throughout the entire project.

5. Conclusion

In this study, we conducted a comprehensive evaluation of ten prominent convolutional neural network (CNN) architectures for the classification of brain tumors using magnetic resonance imaging (MRI). These included both traditional and modern deep learning models such as VGG16, VGG19, ResNet50, ResNet101, DenseNet121, MobileNetV2, EfficientNetB0, InceptionV3, Xception, and NASNetMobile. Each model was rigorously trained and validated on a balanced dataset consisting of tumorous and non-tumorous MRI images. We assessed their performance across standard classification metrics — accuracy, precision, recall, and F1-score — and provided comparative insights to highlight their individual strengths and limitations. The results revealed that models like InceptionV3, Xception, and MobileNetV2 outperformed others, demonstrating exceptional classification performance with accuracy and F1-scores exceeding 95%. DenseNet121 and NASNetMobile also exhibited strong capabilities, whereas deeper networks such as ResNet101 and newer models like EfficientNetB0 showed moderate performance, likely influenced by dataset size, architectural complexity, and feature extraction efficiency.

Beyond raw performance metrics, we also emphasized the importance of explainability and transparency in AI models, particularly in critical applications like healthcare. To this end, we employed a suite of Explainable AI (XAI) techniques — including Grad-CAM, Saliency Maps, SmoothGrad, Integrated Gradients, and Vanilla Gradients — to visualize and interpret the decision-making processes of our models. These visualizations provided valuable insights into the regions of the MRI images that were most influential in classification, thereby reinforcing trust in the model outputs for clinical use. The integration of XAI methods not only enriched model interpretability but also ensured that potential biases or overfitting patterns could be identified and addressed early in the deployment pipeline.

The comprehensive analysis presented in this work underscores the transformative potential of deep learning in medical imaging and diagnosis. However, several challenges remain. Variability in MRI image quality, class

imbalance, and limited availability of annotated datasets continue to impact the robustness and generalization of deep learning models. Future research should focus on addressing these challenges through the incorporation of multimodal data, synthetic data generation (e.g., GAN-based augmentation), domain adaptation, and ensemble methods that combine the strengths of multiple models. Additionally, more collaborative datasets involving diverse demographics and imaging conditions are essential to develop universally reliable diagnostic systems.

Moreover, the clinical adoption of deep learning tools must be accompanied by standardized validation procedures, regulatory compliance, and continuous feedback from medical practitioners. While our study lays a foundational benchmark by offering a side-by-side evaluation of widely used CNN models, it also opens avenues for more targeted innovations — such as hybrid models combining CNNs with transformers or attention-based architectures, as well as lightweight mobile-friendly networks suitable for deployment in resource-constrained environments.

In conclusion, this research offers a valuable reference for researchers and practitioners aiming to explore and improve brain tumor classification using deep learning. By unifying performance evaluation with explainability, we bridge the gap between model accuracy and clinical trust. Our findings reinforce the notion that explainable, efficient, and well-validated AI models can significantly contribute to timely and precise brain tumor diagnosis, thereby improving patient care and supporting the future of AI-assisted healthcare systems.

References

[1] M. A. Abid and K. Munir, "A systematic review on deep learning implementation in brain tumor segmentation, classification and prediction," *Multimedia Tools and Applications*, 2025. [Online]. Available: https://doi.org/10.1007/s11042-025-20706-4

[2] H. Sadr, M. Nazari, S. Yousefzadeh-Chabok, H. Emami, R. Rabiei, and A. Ashraf, "Enhancing brain tumor classification in MRI images: A deep learning-based approach for accurate diagnosis," *Image and Vision Computing*, vol. 159, p. 105555, 2025. [Online]. Available: https://doi.org/10.1016/j.imavis.2025.105555

[3] Z. Rasheed, Y.-K. Ma, I. Ullah, Y. Y. Ghadi, M. Z. Khan, M. A. Khan, A. Abdusalomov, F. Alqahtani, and A. M. Shehata, "Brain tumor classification from MRI using image enhancement and convolutional neural network techniques," *Brain Sciences*, vol. 13, no. 9, p. 1320, 2023. [Online]. Available: https://doi.org/10.3390/brainsci13091320

[4] S. Arora and M. Sharma, "Deep learning for brain tumor classification from MRI images," in *Proc. Int. Conf. Image Inf. Process. (ICIIP)*, Shimla, India, 2021, pp. 409–412. doi: 10.1109/ICIIP53038.2021.9702609.

[5] M. Vimala, S. Palanisamy, S. Guizani, and H. Hamam, "Efficient GDD feature approximation based brain tumour classification and survival analysis model using deep learning," *Egyptian Informatics Journal*, vol. 28, p. 100577, 2024. [Online]. Available: https://doi.org/10.1016/j.eij.2024.100577

[6] R. Vankdothu, M. A. Hameed, and H. Fatima, "A brain tumor identification and classification using deep learning based on CNN-LSTM method," *Computers and Electrical Engineering*, vol. 101, p. 107960, 2022. [Online]. Available: https://doi.org/10.1016/j.compeleceng.2022.107960

[7] Z. Rasheed, Y.-K. Ma, I. Ullah, M. Al-Khasawneh, S. S. Almutairi, and M. Abohashrh, "Integrating convolutional neural networks with attention mechanisms for magnetic resonance imaging-based classification of brain tumors," *Bioengineering*, vol. 11, no. 7, p. 701, 2024. [Online]. Available: https://doi.org/10.3390/bioengineering11070701

[8] S. Ahmmed, P. Podder, M. R. H. Mondal, S. M. A. Rahman, S. Kannan, M. J. Hasan, A. Rohan, and A. E. Prosvirin, "Enhancing brain tumor classification with transfer learning across multiple classes: An in-depth analysis," *BioMedInformatics*, vol. 3, no. 4, pp. 1124–1144, 2023. [Online]. Available: https://doi.org/10.3390/biomedinformatics3040068

[9] P. Priyadarshini, P. Kanungo, and T. Kar, "Multigrade brain tumor classification in MRI images using fine-tuned EfficientNet," *e-Prime – Advances in Electrical Engineering, Electronics and Energy*, vol. 8, p. 100498, 2024. [Online]. Available: https://doi.org/10.1016/j.prime.2024.100498

[10] M. H. Al-Jammas, E. A. Al-Sabawi, A. M. Yassin, and A. H. Abdulrazzaq, "Brain tumors recognition based on deep learning," *e-Prime – Advances in Electrical Engineering, Electronics and Energy*, vol. 8, 2024, Art. no. 100500. [Online]. Available: https://doi.org/10.1016/j.prime.2024.100500

[11] A. Akter, N. Nosheen, S. Ahmed, M. Hossain, M. A. Yousuf, M. A. A. Almoyad, K. F. Hasan, and M. A. Moni, "Robust clinical applicable CNN and U-Net based algorithm for MRI classification and segmentation for brain tumor," *Expert Systems with Applications*, vol. 238, 2024, Art. no. 122347. [Online]. Available: https://doi.org/10.1016/j.eswa.2023.122347

[12] Y. Xie, F. Zaccagna, L. Rundo, C. Testa, R. Agati, R. Lodi, D. N. Manners, and C. Tonon, "Convolutional neural network techniques for brain tumor classification (from 2015 to 2022): Review, challenges, and future perspectives," *Diagnostics*, vol. 12, no. 8, p. 1850, 2022. doi: 10.3390/diagnostics12081850.

[13] M. Saradha, V. Agil, M. Danesha, and M. Vignesh, "A literature review on brain tumor classification

using deep learning," *Int. J. Res. Appl. Sci. Eng. Technol. (IJRASET)*, vol. 12, no. 3, pp. 284, Mar. 2024. doi: 10.22214/ijraset.2024.58808

[14] M. Nazir, S. Shakil, and K. Khurshid, "Role of deep learning in brain tumor detection and classification (2015 to 2020): A review," *Computerized Medical Imaging and Graphics*, vol. 91, p. 101940, 2021. doi: 10.1016/j.compmedimag.2021.101940

[15] T. R. Mahesh, M. Gupta, A. T. A. Anupama, V. Kumar, O. Geman, and V. D. Kumar, "An XAI-enhanced EfficientNetB0 framework for precision brain tumor detection in MRI imaging," *J. Neurosci. Methods*, vol. 410, p. 110227, 2024. doi: 10.1016/j.jneumeth.2024.110227

[16] M. M. Afif, A. A. Noman, K. M. Kabir, M. M. Ahmmed, M. M. Rahman, M. Mahmud, and M. A. Babu, "Proportional Sensitivity in Generative Adversarial Network (GAN)-Augmented Brain Tumor Classification Using Convolutional Neural Network," *arXiv preprint* arXiv:2506.17165, 2025. [Online]. Available: https://arxiv.org/abs/2506.17165

[17] J. G. Melekoodappattu, C. K. Puthiyapurayil, A. Vylala, and A. S. Dhas, "Brain cancer classification based on multistage ensemble generative adversarial network and convolutional neural network," *Cell Biochemistry and Function*, vol. 41, no. 8, pp. 1357–1369, 2023. doi: 10.1002/cbf.3870

[18] N. Thenmoezhi, B. Perumal, and A. Lakshmi, "Multi-view image fusion using ensemble deep learning algorithm for MRI and CT images," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 23, no. 3, Art. no. 40, pp. 1–24, Mar. 2024. doi: 10.1145/3640811

[19] S. N. Eity, M. M. Afif, T. Fairooz, M. M. Ahmmed, and M. S. Miah, "DGG-XNet: A hybrid deep learning framework for multi-class brain disease classification with explainable AI," *arXiv preprint* arXiv:2506.14367, 2025. [Online]. Available: https://arxiv.org/abs/2506.14367

[20] M. M. Ahmmed, A. A. Noman, M. M. Afif, K. M. T. Kabir, M. M. Rahman, and M. Mahmud, "A model-mediated stacked ensemble approach for depression prediction among professionals," *arXiv preprint* arXiv:2506.14459, 2025. [Online]. Available: https://arxiv.org/abs/2506.14459

[21] M. M. Afif, K. M. T. Kabir, A. A. Noman, M. E. A. Islam, and M. M. Ahmmed, "Forecasting wind energy potential in Chattogram, Bangladesh: Statistical modeling incorporating Rayleigh and Weibull distributions," in *Proc. Undergraduate Conf. on Intelligent Computing and Systems (UCICS)*, Varendra University, Rajshahi, Bangladesh, Feb. 2025. [Online]. Available: https://www.researchgate.net/publication/389589352

[22] M. E. A. Islam, K. M. T. Kabir, M. M. Afif, A. A. Noman, and M. M. Ahmmed, "Biomedical engineering for sustainable health: A qualitative study on advancing SDG 3 in Bangladesh," in *Proc. Undergraduate Conf. on Intelligent Computing and Systems (UCICS)*, Varendra University, Rajshahi, Bangladesh, Feb. 2025. [Online]. Available: https://www.researchgate.net/publication/389588995

[23] M. M. Afif, A. A. Noman, M. E. A. Islam, and M. M. Ahmmed, "Losing the night: A comprehensive analysis of trends in artificial light pollution patterns in Bangladesh using VIIRS data," in *Proc. Int. Conf. Electronics and Informatics (ICEI)*, 2024. [Online]. Available: https://www.researchgate.net/publication/387488597

[24] A. A. Noman, M. M. Afif, A. M. R. Huq, A. F. Faizur Rahman, and M. E. A. Islam, "Blockchain-driven halal supply chains: Enhancing transparency and efficiency while ensuring Shariah adherence," *International Journal of Innovative Science and Research Technology*, vol. 10, no. 4, pp. [page numbers if available], Apr. 2025, doi: 10.38124/ijisrt/25apr1001. [Online]. Available: https://tinyurl.com/pz7ymvs2

[25] A. Shoeibi, M. Khodatars, M. Jafari, N. Ghassemi, D. Sadeghi, P. Moridian, A. Khadem, R. Alizadehsani, S. Hussain, A. Zare, Z. A. Sani, F. Khozeimeh, S. Nahavandi, U. R. Acharya, and J. M. Gorriz, "Automated detection and forecasting of COVID-19 using deep learning techniques: A review," *Neurocomputing*, vol. 577, p. 127317, 2024, doi: https://doi.org/10.1016/j.neucom.2024.12731710.1016/j.neucom.2024.127317.

[26] T. Hulsen, "Explainable Artificial Intelligence (XAI): Concepts and Challenges in Healthcare," *AI*, vol. 4, no. 3, pp. 652–666, 2023, doi: https://doi.org/10.3390/ai403003410.3390/ai4030034.

[27] M. Shrivastava and L. Ye, "Neuroimaging and artificial intelligence for assessment of chronic painful temporomandibular disorders—a comprehensive review," *Int. J. Oral Sci.*, vol. 15, no. 1, p. 58, Dec. 2023, doi: https://doi.org/10.1038/s41368-023-00254-z10.1038/s41368-023-00254-z.

[28] S. Zolfaghari, S. Suravee, D. Riboni, and K. Yordanova, "Sensor-Based Locomotion Data Mining for Supporting the Diagnosis of Neurodegenerative Disorders: A Survey," *ACM Comput. Surv.*, vol. 56, no. 1, Art. no. 10, pp. 1–36, Aug. 2023, doi: https://doi.org/10.1145/360349510.1145/3603495.

[29] J. Chaki and G. Deshpande, "Brain Disorder Detection and Diagnosis using Machine Learning and Deep Learning – A Bibliometric Analysis," *Current Pharmaceutical Design*, vol. 22, no. 13, pp. 2191–2216, May 2024, doi: 10.2174/1570159X22999240531160344.

[30] S. Lee and K.-S. Lee, "Predictive and Explainable Artificial Intelligence for Neuroimaging Applications," *Diagnostics*, vol. 14, no. 21, p. 2394, 2024, doi: https://doi.org/10.3390/diagnostics1421239410.3390/diagnostics14212394.

[31] R. Gupta, S. Kumari, A. Senapati, R. K. Ambasta, and P. Kumar, "New era of artificial intelligence and

machine learning-based detection, diagnosis, and therapeutics in Parkinson's disease," *Ageing Res. Rev.*, vol. 90, p. 102013, 2023, doi: https://doi.org/10.1016/j.arr.2023.10201310.1016/j.arr.2023.102013.

[32] Y. Maeda *et al*., "Rewiring the primary somatosensory cortex in carpal tunnel syndrome with acupuncture," *Brain*, vol. 140, no. 4, pp. 914–927, Apr. 2017, doi: https://doi.org/10.1093/brain/awx01510.1093/brain/awx015.