NeuroScanXNet: An Explainable CNN Framework for Brain Tumor Classification with Grad-CAM Consistency Analysis and Mini-RAG Diagnostic Reporting

Rasika Yogesh Talwar[1], Dr. Devang Thakar[2]

[1]Data Science and Artificial Intelligence, Chauhan Institute of Science and Amrutben Jivanlal College of Commerce and Economics (Empowered Autonomous), Mumbai

[2]Assistant Professor, Department of Computer Science, Mithibai College of Arts, Chauhan Institute of Science and Amrutben Jivanlal College of Commerce and Economics (Empowered Autonomous), Mumbai

***Abstract*—Deep learning has been widely used in medical imaging, significantly improving the accuracy of brain tumor classification. However, many existing models focus primarily on prediction accuracy without explaining how decisions are made, making their deployment in real clinical settings challenging. Convolutional Neural Networks (CNNs), though effective, are often treated as black-box models, which makes it difficult to trust their outputs. This study proposes a framework — NeuroScanXNet — that addresses both classification accuracy and result interpretability. A CNN-based model is used to classify MRI images into various tumour categories. Grad-CAM is applied to highlight the important regions influencing the model's predictions. In addition, a quantitative consistency analysis using masked cosine similarity is introduced to evaluate whether the model focuses on similar regions across different inputs. A Mini-RAG module based on TF-IDF retrieves relevant medical knowledge to generate a structured diagnostic report. The proposed system achieves a test accuracy of 94.66% while improving transparency and usability for real-world clinical decision support.**

***Keywords*—Brain Tumor Classification; Convolutional Neural Networks; Grad-CAM; Explainable AI; Mini-RAG; Masked Cosine Similarity; MRI; Medical Imaging**

I. Introduction

Over the past few years, the use of Artificial Intelligence in healthcare has risen significantly, especially in medical image analysis. Brain tumor detection and classification represents one of the most impactful application areas, as early and accurate diagnosis directly affects treatment outcomes and patient survival. Traditionally, MRI scans are analysed manually by radiologists, which is time-consuming and subject to inter-observer variability. With the advancement of deep learning, Convolutional Neural Networks have become the dominant approach for image-based classification tasks, demonstrating strong performance in distinguishing tumour types such as meningioma, glioma, and pituitary tumors.

The fundamental challenge with CNN models is their black-box nature: it is difficult to understand the decision-making process, which remains a critical barrier to clinical adoption. Clinicians require not only accuracy but also explanations they can trust. Techniques such as Grad-CAM have been introduced to provide visual explanations; however, these are typically evaluated only qualitatively. There is no established quantitative method to verify whether explanations are consistent or stable across samples.

Furthermore, existing systems often treat classification, explanation, and reporting as separate components, limiting their utility in real-world applications.

This research proposes a unified system — NeuroScanXNet — that integrates CNN-based classification, Grad-CAM explainability, masked cosine similarity-based consistency analysis, and a Mini-RAG module for evidence-grounded diagnostic report generation. The major aim is not only to build an accurate system but also one that is interpretable, reliable, and practical for brain tumor diagnosis.

II. Literature Review

A. CNN-Based Brain Tumor Classification

CNN-based models have been widely applied to classify different types of brain tumors such as glioma, meningioma, and pituitary tumors. These models follow a common process that includes image preprocessing, feature extraction using convolutional layers, and final classification through fully connected layers. Multimodal feature fusion has been found to improve both the accuracy and reliability of tumor classification [13]. Although these methods perform well in terms of prediction, they do not clearly explain how the model arrives at a particular result.

B. Explainable AI in Medical Imaging

To make CNN models more understandable, Explainable AI techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM) have been introduced. Grad-CAM produces heatmaps that highlight the important regions of the image which influence the model's prediction. Many recent studies have used Grad-CAM to visualise tumor-related areas in MRI scans. However, these explanations are usually evaluated only visually, with very little work done to quantitatively measure whether explanations are consistent or reliable across different samples.

C. Advanced Approaches: Segmentation and Grading

Some research has focused on tumor segmentation and grading beyond classification. Models such as U-Net and hybrid CNN-based architectures are used to identify the exact location of tumor regions. These approaches provide more detailed spatial information but often require expert knowledge to interpret and do not focus on generating simple, structured outputs easily understood by clinical users.

D. Retrieval-Augmented Systems in Medical AI

More recently, researchers have explored Retrieval-Augmented Generation (RAG) techniques for medical report generation. RAG combines model predictions with external medical knowledge to produce more informative outputs. Most existing RAG-based methods rely on complex architectures involving large language models and vector databases, making them computationally expensive and difficult to implement in smaller systems.

E. Research Gap

Despite notable advances, several critical gaps remain. First, although Grad-CAM is widely used, most approaches rely only on qualitative heatmap visualisation, lacking quantitative methods to evaluate explanation consistency and reliability. Second, current work primarily focuses on classification accuracy, treating prediction, explanation, and reporting as separate components with very few studies integrating these into a unified pipeline. Third, existing RAG-based approaches are computationally intensive and difficult to implement in resource-constrained research environments. These gaps motivate the proposed NeuroScanXNet framework.

III. Methodology

A. System Architecture

The proposed system is designed as a single unified workflow integrating image classification, explainability, quantitative reliability evaluation, and evidence-based reporting. The pipeline consists of the following components: MRI image input → CNN classification → Grad-CAM explanation → Consistency analysis → Structured output → Mini-RAG retrieval → Diagnostic report. The input MRI image is first processed by the CNN model to predict the tumour class. The predictions are then explained using Grad-CAM, which highlights the important image regions. A consistency analysis then evaluates the reliability of

these explanations. Finally, the Mini-RAG module retrieves relevant medical knowledge to generate a structured diagnostic report.

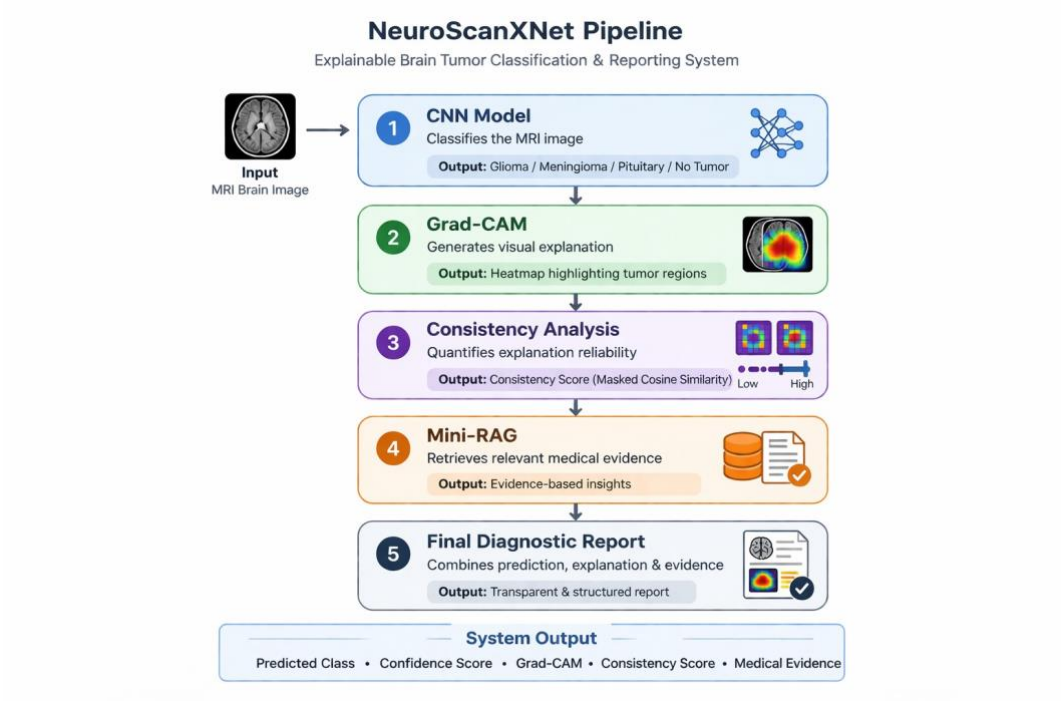

Fig. 1. *Proposed NeuroScanXNet pipeline: MRI → CNN → Grad-CAM → Consistency → Retrieval → Report.*

B. Dataset and Preprocessing

The dataset consists of brain MRI images categorised into four classes: Glioma, Meningioma, Pituitary tumour, and No tumour. The dataset is divided into training and testing subsets organised in class-specific directories. All images are resized to 224×224 pixels and pixel values are normalised to the range [0, 1]. Data is loaded using image generators for efficient batch processing.

C. CNN-Based Classification Model

A Convolutional Neural Network is used as the baseline model for tumour classification. The CNN learns hierarchical spatial features from MRI images through convolutional and pooling layers. Given an input image X, the probability of class y is computed as $P(y \mid X) = \text{Softmax}(f(X))$, and the predicted class is $\hat{y} = \text{argmax } P(y \mid X)$.

D. Grad-CAM for Explainability

Grad-CAM is used to visualise the regions of the image that influence the model's prediction. Let A^k denote the feature maps of the last convolutional layer. The importance weights are computed from the gradients of the predicted class score with respect to each feature map. The Grad-CAM heatmap is then computed as $L_{\text{Grad-CAM}} = \text{ReLU}(\Sigma\ \alpha_k\ A^k)$, where the ReLU function retains only the positive influences. This produces a heatmap highlighting the regions most relevant to the model's decision.

E. Quantitative Grad-CAM Consistency Analysis

To evaluate the reliability of Grad-CAM explanations, a quantitative consistency analysis is performed using masked cosine similarity. For each class, multiple Grad-CAM heatmaps are generated, the top activated regions are extracted using a binary mask M, and pairwise similarity is computed. Let two Grad-CAM heatmaps be represented as vectors A and B. The masked vectors are $A' = A \odot M$ and $B' = B \odot M$. The masked cosine similarity is computed as $\text{Similarity}(A, B) = (A' \cdot B') / (\|A'\| \cdot \|B'\|)$. Higher values indicate consistent attention patterns across samples, while lower values indicate unstable explanations.

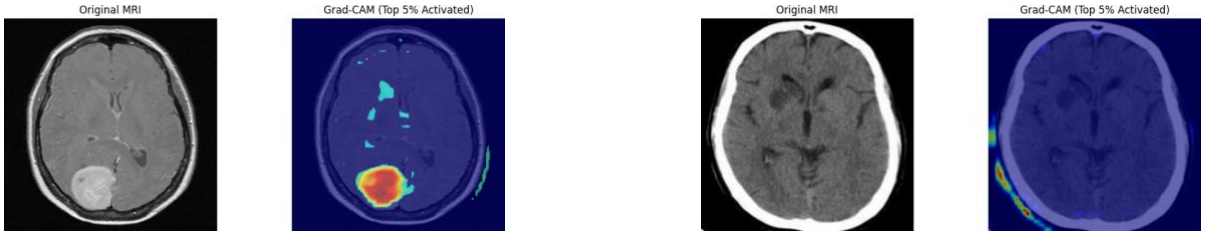

Fig. 2. *Illustration of Grad-CAM consistency analysis. Two heatmaps are generated, top activated regions are selected using a mask, and cosine similarity is computed to measure explanation consistency.*

IV. Results

A. Classification Performance

The CNN-based classification model was trained on the preprocessed MRI dataset and evaluated on the test set, achieving a test accuracy of 94.66%. This demonstrates the model's ability to effectively distinguish between glioma, meningioma, pituitary tumour, and no-tumour classes. However, accuracy alone is insufficient to fully understand the model's behaviour, as it does not explain how predictions are made.

B. Grad-CAM Explainability Results

Grad-CAM was applied to visualise the regions in MRI images that influenced the model's predictions. In many cases, the model highlights regions corresponding to abnormal or tumour-like areas. However, it was also observed that in some samples the highlighted regions did not perfectly align with the tumour region, indicating some level of inconsistency. To improve interpretability, only the top activated regions of the heatmap were visualised, making attention areas clearer and more focused.

(a) Glioma (b) Meningioma

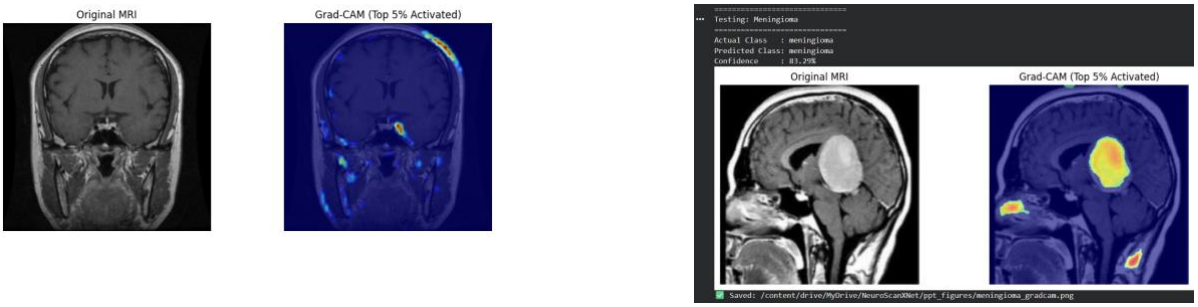

Fig. 3. *Grad-CAM explainability results across tumour classes: (a) Glioma, (b) Meningioma, (c) Pituitary, (d) No Tumour.*

C. Observations on Explainability

Although Grad-CAM provides useful visual insights, several observations were made: in some cases, heatmaps did not perfectly align with tumour regions; a few activations appeared in background or edge regions; and the model sometimes focused on global image patterns rather than localised tumour areas. These observations indicate that the model may rely on spurious correlations or shortcut features, highlighting limitations of visual explainability alone.

D. Quantitative Grad-CAM Consistency Analysis

To evaluate the reliability of Grad-CAM explanations, a quantitative analysis was performed using masked cosine similarity. For each tumour class, 10 Grad-CAM heatmaps were generated, the top 20% of

activated regions were extracted, and pairwise similarity was calculated. The results are presented in Table I.

Table I. Grad-CAM Consistency Scores Across Tumour Classes

Class	Avg Similarity	Std Deviation
Glioma	0.27	0.14
Meningioma	0.17	0.10
No Tumour	0.20	0.12
Pituitary	0.16	0.08

E. Interpretation of Consistency Scores

Glioma shows the highest average similarity (0.27), indicating relatively stable attention patterns. Other classes show moderate stability due to variability in MRI appearances. The standard deviation values reveal variation in explanation patterns across samples. These results confirm that Grad-CAM explanations are not always consistent and that quantitative evaluation is necessary. The consistency scores are visualised in Figure 4.

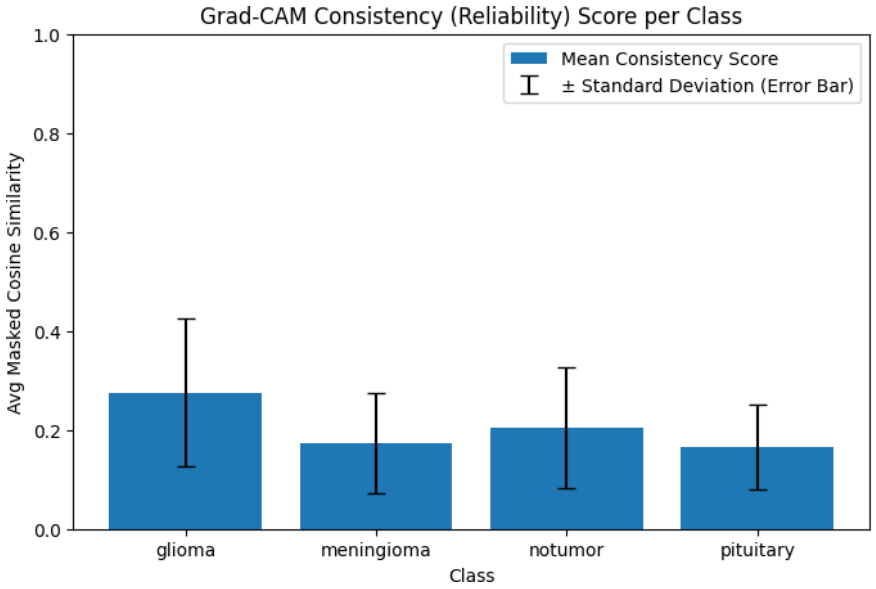

Fig. 4. *Grad-CAM consistency scores across tumour classes (bar chart). Height represents average similarity; error bars indicate variation across samples.*

F. Structured Pipeline Output and Mini-RAG Diagnostic Report

The system generates a structured prediction output summarising the predicted tumour class, actual class label, confidence score, and Grad-CAM visualisation. This is followed by the Mini-RAG diagnostic report, which uses a TF-IDF-based retrieval system to obtain relevant medical knowledge from a predefined knowledge base. The final report includes the predicted class, confidence score, Grad-CAM visualisation, an explanation summary, retrieved supporting medical evidence with similarity scores, and non-prescriptive clinical guidance. This converts the system from a simple classifier into an evidence-assisted diagnostic support framework.

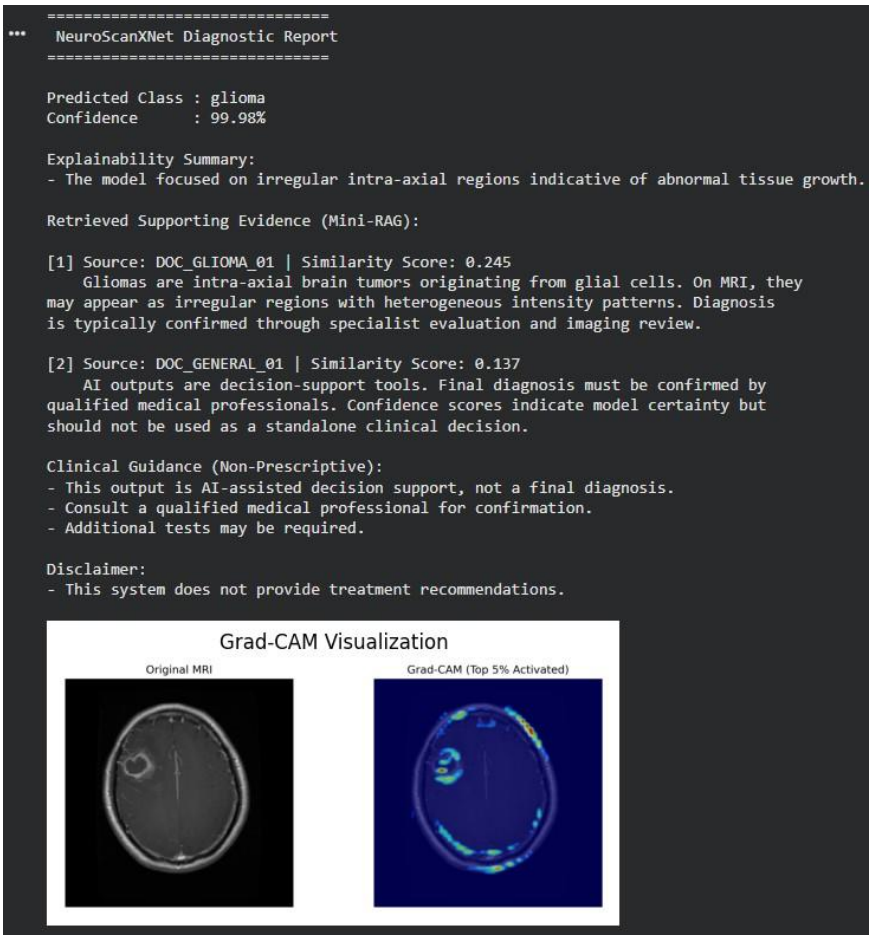

Fig. 5. *Generated diagnostic report combining prediction, Grad-CAM visualisation, and retrieved medical evidence.*

V. Conclusion and Future Scope

A. Conclusion

In this work, NeuroScanXNet — an explainable, system-focused framework for brain tumour classification — was developed using MRI images. A CNN model classifies MRI images into four categories (glioma, meningioma, pituitary tumour, no tumour) with a test accuracy of 94.66%. Grad-CAM provides visual explanations of the model's predictions. A masked cosine similarity consistency analysis quantifies the stability of these explanations. A structured output and a Mini-RAG module using TF-IDF generate an evidence-assisted diagnostic report. Overall, the system combines prediction, explanation, reliability analysis, and report generation into a single framework, improving transparency and practical usability.

B. Limitations

Several limitations were observed. Grad-CAM visualisations did not always align perfectly with actual tumour areas; the model sometimes focused on irrelevant surroundings or background regions. The consistency analysis showed that Grad-CAM outputs were not always stable across similar cases. The Mini-RAG module uses a small, manually created knowledge base, limiting diversity and detail of retrieved evidence. The system does not perform detailed tumour localisation or segmentation, which is often required in clinical settings. Finally, the model was tested on a relatively limited dataset, and performance on real-world clinical data under varying image quality, scanning conditions, and patient variation remains uncertain.

C. Future Scope

Several directions can improve this system: (1) integration of advanced deep learning models such as Vision Transformers for better classification; (2) incorporation of tumour segmentation models such as U-Net to identify tumour boundaries; (3) use of additional explainability techniques such as SHAP and Integrated Gradients alongside Grad-CAM; (4) enhancement of the retrieval mechanism through

embedding-based models and vector databases; (5) multimodal data integration incorporating clinical reports and patient histories; (6) clinical validation and deployment on real clinical datasets; and (7) development of a user-friendly interface for practical medical use. This research demonstrates that high accuracy alone is insufficient for real-world medical AI — transparency, consistency, and evidence-grounding are equally essential.

VI. References

[1] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," Proc. IEEE ICCV, 2017.

[2] G. Litjens, T. Kooi, B. E. Bejnordi, et al., "A Survey on Deep Learning in Medical Image Analysis," Medical Image Analysis, vol. 42, pp. 60–88, 2017.

[3] J. Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries," Proc. First Instructional Conference on Machine Learning, 2003.

[4] P. Lewis, E. Perez, A. Piktus, et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," Advances in Neural Information Processing Systems (NeurIPS), 2020.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," MICCAI, 2015.

[6] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization," Proc. IEEE CVPR, 2016.

[7] Shaheema, S., & Berlin, S. (2025). Brain Tumor Segmentation and Grade Classification using Deep Learning Models and Explainable AI. National Institute of Technology Silchar. http://hdl.handle.net/10603/654733

[8] Sandhiya, B. (2025). An Adapted Generative Adversarial Network and Enhanced Particle Swarm Optimization Based Deep Framework for Brain Tumor Classification using MRI Images. Anna University. http://hdl.handle.net/10603/618264

[9] Guo, Z., Ren, X., Xu, L., Zhang, J., & Huang, C. (2025). RAG-Anything: All-in-One RAG Framework. arXiv:2510.12323. https://arxiv.org/abs/2510.12323

[10] Mandal, S., Chakraborty, S., Tariq, M., et al. (2024). Artificial Intelligence and Deep Learning in Revolutionizing Brain Tumor Diagnosis and Treatment: A Narrative Review. Cureus, 16(8), e66157. https://doi.org/10.7759/cureus.66157

[11] Panboonyuen, T. (2026). Seeing Isn't Always Believing: Analysis of Grad-CAM Faithfulness and Localization Reliability in Lung Cancer CT Classification. arXiv:2601.12826. https://doi.org/10.48550/arXiv.2601.12826

[12] Arabboev, M. (2025). Brain Tumor Classification Using Transfer Learning with MobileNetV2. Techscience.uz, 3(5), 51–63. http://ilmiykutubxona.com/id/eprint/790

[13] Salman, B., et al. (2025). Deep Learning-Based Fusion of Multimodal MRI Features for Brain Tumor Classification. Applied Sciences, 15(24), 13155. https://doi.org/10.3390/app152413155

[14] Ferro, P., Vemanaboina, H., & Prakash, C. (Eds.). (2026). Computational Techniques and Smart Manufacturing. CRC Press. https://doi.org/10.1201/9781003679622

[15] Xu, J. (2025). Research on the Application of Deep Learning in Image Diagnosis in the Field of Healthcare. Applied and Computational Engineering, 178, 88–94. https://ace.ewapub.com/article/view/25819

[16] Khan, M. A., et al. (2025). Transfer Learning for Accurate Brain Tumor Classification in MRI Images. International Journal of Information Technology. https://doi.org/10.1007/s12672-025-02671-4

[17] Dargar, S. K., Birla, S., Dargar, A., Singh, A., & Ganeshaperumal, D. (Eds.). (2025). Sustainable Materials and Technologies in VLSI and Information Processing. CRC Press. https://doi.org/10.1201/9781003641551

[18] Zhang, Y. (2025). A Comparative Analysis Between CNNs and ViTs for MRI-based Brain Tumor Classification. Highlights in Science, Engineering and Technology, 124, 30–37. https://doi.org/10.54097/s64djm51