



IJRASET

International Journal For Research in
Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: <https://doi.org/10.22214/ijraset.2026.82087>

www.ijraset.com

Call:  08813907089

E-mail ID: ijraset@gmail.com

A Unified XAI Framework for Interpreting Deep Learning Models in Brain Tumor Detection

Kamalakshi Naganna, AnkithV Hullamani, Aaryan Kumar, Tejaswini K, Yashwanth ES

Abstract—Deep neural networks have shown strong potential for analysing MRI scans in brain tumour detection, but their black-box nature makes the reasoning behind each prediction hard to audit — a serious limitation in safety-critical care. This survey reviews recent work that pairs Convolutional Neural Networks (CNNs) with Explainable Artificial Intelligence (XAI) to mitigate this opacity, taking a CNN trained on the BraTS 2021 dataset as the reference setting. We then examine three complementary explanation families — Gradient-weighted Class Activation Mapping (Grad-CAM), Layer-wise Relevance Propagation (LRP), and SHapley Additive Explanations (SHAP)—which respectively expose salient regions, per-pixel attributions, and feature-level contributions. We argue that fusing these views into a unified XAI pipeline yields layered, more trustworthy rationale than any single method, and is a practical path toward auditable clinical decision support.

I. INTRODUCTION

The intersection of Artificial Intelligence (AI) and Deep Learning (DL) has reshaped how radiological data—including MRI, CT, and X-ray studies—is examined and triaged. Within this landscape, MRI is especially valued for cerebral pathology because of its high soft-tissue contrast, which makes anomalous masses easier to delineate.

CNN-based architectures, in particular, have become the workhorse of automated tumour screening: they ingest raw voxel data and progressively learn discriminative features, typically reaching competitive accuracy on benchmark datasets.

The flipside of this expressiveness is opacity. Because the layered transformations are not directly inspectable, the model's logic is hidden from end users—a property that is poorly tolerated in regulated, high-stakes domains such as oncology. This is exactly the gap that Explainable Artificial Intelligence (XAI) attempts to close. By surfacing which inputs, regions, or learned features drove a given output, XAI tools render a black-box model partially transparent. Among the techniques that have gained traction in the medical imaging literature are Grad-CAM, LRP, and SHAP, each operating on a different mathematical principle.

A clear trend in recent work is to step beyond single-method explanations and instead orchestrate several XAI tools at once. Such “unified” setups blend coarse region maps, fine pixel attributions, and feature-level scores so that the resulting narrative is broader and more robust — particularly helpful when only a fragment of a tumour is visible in a slice.

Contributions of this Survey

The present survey delivers a structured account of XAI as applied to brain tumour analysis. Specifically, it:

- organises the major DL strategies currently employed for MRI-based tumour identification;
- compares the dominant XAI techniques (Grad-CAM, LRP, SHAP) on common axes;
- contrasts local with global explanation paradigms in the medical context;
- enumerates the open problems that limit today's XAI-augmented diagnostic systems; and
- argues for unified explainability stacks as a route to higher clinical trust and faster adoption.

The broader objective is to connect strong predictive performance with the interpretability requirements of bedside use, while pointing to research directions that could mature the field.

II. BACKGROUND AND MOTIVATION

A. Need for Intelligent Medical Imaging Systems

In modern oncology, imaging is central to diagnosing pathologies such as cerebral tumours, and timely, accurate readings have a direct impact on therapy selection and patient prognosis. Because MRI and CT studies generate a continually growing pool of high-resolution data, fully manual interpretation by radiologists becomes a bottleneck and is susceptible to fatigue-induced mistakes. The result is a strong demand for computer-assisted screening tools that can sift through scans rapidly and consistently.

CNNs have become the de facto building block for such tools, since they can extract task-relevant representations directly from pixel data and detect lesions at clinically useful accuracy.

Yet, deploying such systems responsibly demands more than raw scoring metrics — the workflow must also expose the basis for each prediction.

B. Limitations of Conventional Deep Learning Models

While CNNs and related architectures perform well on imaging benchmarks, the “opaque reasoning” problem persists. Without an explicit explanation, clinicians cannot easily distinguish a prediction grounded in genuine pathology from one driven by spurious cues such as scanner artefacts. This uncertainty fuels reluctance to integrate AI into care pathways. Practitioners need confirmation that the network’s focus aligns with anatomically meaningful regions before they accept its output as a second reader. A research literature that optimises chiefly for accuracy — without auditing decision pathways — therefore offers limited operational value in a hospital setting.

C. Role of Explainable AI in Medical Diagnosis

XAI offers a constructive response by providing tools that translate hidden network computations into human-readable cues. The output may take the form of saliency maps, relevance heat-overlays, or per-feature attribution scores.

For brain tumour studies, methods such as Grad-CAM, LRP, and SHAP each address a different layer of the explanation problem: from broad regions of interest down to per-pixel contribution and individual feature weight. Plugging these methods into a CNN-based diagnostic pipeline yields a system that is both performant and inspectable.

This realisation justifies the move toward unified frameworks solutions that intentionally combine multiple XAI views so that clinicians receive richer, multi-resolution justification for every output.

III. LITERATURE SURVEY

A. Brain Tumor Detection using Transfer Learning

Author and Year: M.Z. Khaliki and M.S. Bas, arslan (2024) **Methodology:** The work investigates how transfer learning can be exploited for tumour identification on MRI. Established backbones — most notably VGG19 and InceptionV3 — serve as feature extractors, with the upper layers re-trained on the medical dataset. By reusing weights learned on large-scale natural-image corpora, the authors achieve competitive accuracy without requiring vast amounts of medical data, which is generally scarce in this domain. **Limitation:** Performance is tightly coupled to how well the source backbone transfers to the target task. Pretraining on natural photographs can leave domain-specific MRI cues poorly represented, and generalisation deteriorates if the target dataset lacks demographic or modality diversity.

B. Explainable AI using Grad-CAM with ResNet50

 Author and Year: M. M. M. et al. (2024)

Methodology: The authors pair a ResNet50 backbone with Grad-CAM to deliver post-hoc visual rationales. After classifying MRI inputs, gradient information from the deepest convolutional block is converted into a class-discriminative heatmap. Overlaying this map on the original image lets practitioners check whether the network actually attends to suspected tumour regions.

Limitation: Grad-CAM operates at a coarse spatial scale; lesion borders cannot be resolved precisely. The heatmaps may also activate over background structures, producing visually plausible but clinically unhelpful explanations. Quality is bounded by the discriminative power of the underlying feature maps.

C. FLAIR MRI for Brain Tumor Segmentation

 Author and Year: A. Al-Fakih et al. (2024)

Methodology: This contribution targets segmentation rather than pure classification. Building on the FLAIR sequence — favoured for its suppression of cerebrospinal-fluid signal — the authors design a generative network with an attention mechanism that strengthens the visibility of pathological tissue. The pipeline produces tighter contour estimates and better captures lesion morphology.

Limitation: The combined generative-plus-attention design is heavyweight, raising training cost and inference latency. The method also degrades on noisy or low-resolution scans, and pure segmentation alone does not deliver an end-to-end diagnostic system without an upstream classifier.

D. XAI Techniques for Brain Tumor Analysis

Author and Year: P. Narayankar and V. P. Baligar (2024) **Methodology:** The paper benchmarks several XAI tools applied to a common CNN classifier on MRI data.

By placing each technique on the same model and inputs, the authors compare what each tool reveals about region salience, feature relevance, and decision boundaries, suggesting that the techniques surfaced different—and partially complementary—facets of network behaviour.

Limitation: Although the comparison is informative, it stops short of fusing the techniques into a single coherent explanation pipeline. Reading several explanations side by side also imposes a cognitive cost and typically requires familiarity with each method's assumptions.

E. CNN with SHAP for Tumor Classification

Author and Year: S. Ahmed, S. N. Nobel, and O. Ullah (2023)

Methodology: A standard CNN is trained for multi-class tumour categorisation, and SHAP is layered on top to attribute importance values to features. SHAP returns both per-instance (local) and aggregate (global) signals; the authors render these as colour-coded overlays in which positively and negatively contributing pixels are visually distinguished.

Limitation: Computing Shapley values over high-dimensional images is expensive. Visualisations are also non-trivial for non-experts to read, and the explanation inherits any errors in the underlying classifier — a confidently wrong prediction will receive a confidently wrong rationale.

F. Grad-CAM Based Explainable CNN

Author and Year: T. Hussain and H. Shouno (2023)

Methodology: This study targets multi-class tumour classification with a CNN and uses Grad-CAM to localise the discriminative regions per class. The resulting heatmaps are blended with the source MRI so that radiologists can visually validate whether the network's attention coincides with anatomically relevant areas.

Limitation: Localisation remains coarse; small or irregularly shaped lesions are often only partially highlighted. Furthermore, Grad-CAM is sensitive to which convolutional layer is used to compute gradients, so explanations vary with implementation choices.

G. LRP for Model Interpretability

Author and Year: F. Šefčík and W. Beněšová (2023)

Methodology: The paper applies Layer-wise Relevance Propagation to a CNN that classifies glioma subtypes. By back-propagating the prediction score through the network using relevance redistribution rules, the method assigns an attribution value to each input pixel, providing finer spatial granularity than gradient-based saliency.

Limitation: LRP outputs can appear noisy, with isolated bright pixels that complicate interpretation. The choice of propagation rule (e.g., ϵ -rule, $\alpha\beta$ -rule) materially affects the result, demanding architectural insight from the user.

H. Deep Learning for Brain Tumor Detection

Author and Year: M. I. Mahmud, M. Mamun, and A. Abdelgawad (2023)

Methodology: Several CNN architectures are benchmarked for binary tumour-versus-normal classification on MRI. After standard preprocessing—normalisation, resizing, and augmentation—the authors report performance using common metrics and discuss how depth, kernel design, and regularisation affect outcomes.

Limitation: The pipeline depends heavily on dataset scale and class balance, and overfitting can emerge once the training set becomes too narrow. Crucially, no explainability layer is added, which restricts clinical applicability.

I. Hybrid CNN + SVM Model for Tumor Detection

Author and Year: E. M. Senan et al. (2022)

Methodology: The proposal hybridises representation learning with classical pattern recognition: a CNN computes high-level descriptors from MRI slices, after which a Support Vector Machine performs the final decision. The combination capitalises on the SVM's effectiveness in low-sample, high-dimensional feature spaces.

Limitation: Two-stage pipelines are harder to maintain and tune than end-to-end networks, and the resulting model offers no native explanation channel — feature importance must be reverse-engineered after the fact.

J. AI in Medical Imaging: A Review Author and Year: R. Najjar (2023)

Methodology: This is a literature review rather than a new method. Its survey shows machine learning, and CNNs in particular, are integrated into radiology workflows for diagnosis, triage, and decision support across MRI, CT, and other modalities. The review also touches on data-governance, regulatory and operational considerations.

Limitation: As a synthesis paper, it does not introduce or empirically validate any new algorithm. Many recommendations are necessarily generic and require adaptation before being acted on in a specific clinical setting.

IV. COMPARATIVE ANALYSIS OF EXISTING RESEARCH

To make the trade-offs between the surveyed studies easier to navigate, Table I compile their methods, contributions, and weaknesses side by side.

A. Deep Learning Approaches for Tumor Detection

A consistent message across recent papers is that CNNs are highly capable of extracting tumour-relevant cues from MRI data. Khaliki and Bas, arslan [1] illustrate this by leveraging pretrained backbones to shorten convergence time without sacrificing accuracy, while Mahmud et al. [8] benchmark several architectures to confirm CNNs’ strength in capturing the relevant spatial regularities.

Beyond pure DL, hybrid pipelines have been proposed. Senan et al. [9] combine CNN derived embeddings with a downstream classical classifier, reporting gains over either component used alone. Such designs aim to merge the representational depth of neural networks with the sample-efficiency of margin-based learners.

That said, these systems still struggle with three recurring issues: appetite for large labelled corpora, sensitivity to acquisition noise, and — most importantly for this survey — opacity. These factors continue to slow real-world deployment.

B. Explainable AI Techniques in Medical Imaging

To mitigate opacity, multiple XAI families have been folded into MRI pipelines. Ahmed et al. [5] integrate SHAP with a CNN to yield local and global feature-importance signals, while Hussain and Shouno [6] rely on Grad-CAM to project the model’s attention back onto the input image. LRP, as adapted by Šešić and Benes̃ova [7], goes a step further by attributing relevance at the level of individual pixels. Narayankar and Baligar [4] explicitly compare several XAI tools and conclude that no single method captures every aspect of a model’s reasoning — different methods illuminated different facets, suggesting that combinations are more informative.

M.M.M. et al. [2] reach a similar conclusion: pairing Grad-CAM with deeper architectures yields more diagnostically useful overlays.

Crucially, these methods differ in granularity, runtime, and the nature of the produced explanation, so these selections should be matched to the deployment scenario.

C. MRI-Based Tumor Analysis and Imaging Techniques

At the modality level, MRI remains the workhorse for cerebral imaging. Al-Fakih et al. [3] argue that FLAIR sequences, in particular, sharpen lesion contrast, which directly benefits downstream segmentation accuracy.

Najjar [10] takes a wider lens, surveying how AI assistants are entering radiology workflows, and flags governance, transparency, and validation as the key non-technical bottlenecks for adoption.

In aggregate, the literature shows excellent raw performance from DL models but limited, fragmented progress on interpretability. Each XAI technique captures part of the picture;

TABLE I
COMPARATIVE ANALYSIS OF EXISTING RESEARCH IN BRAIN TUMOR DETECTION AND XAI

Author	Method	Key Contribution	Limitation
Khaliki and Bas, arslan [1]	Transfer Learning (CNN)	Reused pretrained backbone to lift classification accuracy with limited medical data.	Pretraining domain mismatch may suppress medically relevant cues.
M.M.M. et al. [2]	Grad-CAM + ResNet50	Produced overlay heatmap that highlights tumour regions used by the model.	Spatial resolution is coarse; spurious activations occur.

		Classifier.	
Al-Fakihetal.[3]	FLAIRMRIAnalysis	LeveragedFLAIRcontrastandattentiontorefinelesionsboundaries.	Sensitivetoacquisitionqualityanddisputationallyheavy.
NarayankarandBaligar[4]	MultipleXAITechniques	BenchmarkedseveralXAIToolssideonCNNclassifiers.	No fusionofmethodsintoasingle explanationpipeline.
Ahmedetal.[5]	CNN+SHAP	Deliveredbothlocalandglobal feature-attributionanalyses.	Shapleycomputationscalespoorlywithimagesize.
HussainandShouno[6]	Grad-CAMCNN	Class-specificheatmapssupport multi-classlesionslocalisation.	Granularityinsufficientforfine boundaries.
ŠtefčíkandBenešova[7]	LRP	Pixel-levelattributionviarelevanceback-propagation.	Outputscanbenoisyandrulerule-dependent.
Mahmudetal.[8]	DeepLearningCNN Models	ComparativeCNNstudyforautomateddetection.	Data-hungryandoffersnointerpretability.
Senanetal.[9]	HybridCNN+ML	CombinedCNNfeatureswithclassicalclassifiersforhigheraccuracy.	Increasedcomplexity; explanation absent.
Najjar[10]	AIMedicalImagingReview	SynthesisesAIuseacrossradiology workflows.	Nonewexperimentsorvalidation.

none captures all of it. This observation is the core motivation for the unified framework discussed next.

V. TAXONOMY OF EXPLAINABLE AI TECHNIQUES

XAI methods can be grouped according to the mathematical machinery they use to derive an explanation. In the brain-tumour context, three families dominate, each providing a distinct level of insight. Understanding these families helps practitioners pick the right tool — or the right combination of tools — for a given clinical question.

A. Gradient-Based Methods

This family inspects how the prediction changes when the input—or an intermediate activation—is perturbed, computed via partial derivatives. The flagship example is Grad-CAM, which weights the channels of a target convolutional layer using the class-conditional gradient and produces a heatmap localising the regions that drove the decision [6]. Because the heatmap is naturally aligned with the input, it can be overlaid on the source MRI for direct inspection.

These methods are attractive thanks to their low computational footprint and ease of integration with any CNN. The main downside is their relatively coarse spatial granularity, which is inadequate when sub-millimetre detail matters [2].

B. Backpropagation-Based Methods

A second family redistributes the output score backwards through the network’s connections, assigning each input element a share of the prediction.

Layer-wise Relevance Propagation (LRP) is the canonical instance: starting from the output neuron, it propagates relevance through each layer using rules that respect the model’s structure, eventually producing a per-pixel importance map [7].

The result is much finer than a Grad-CAM heatmap, which makes LRP attractive for radiologists looking for precise lesion edges.

The trade-off is implementation complexity and rule-sensitivity: producing reliable maps requires care in choosing propagation rules and verifying conservation properties.

C. Perturbation and Game-Theory-Based Methods

A third group derives importance scores by modifying input features and observing the resulting change in output. SHAP (SHapley Additive Explanations) is the most prominent member, drawing on cooperative game theory.

SHAP estimates each feature's marginal contribution by aggregating over many feature subsets and produces signed attributions—with positive and negative contributions usually rendered in red and blue, respectively [5]. It supports both per-instance reasoning and corpus-level analysis.

The principal cost is computational: exact Shapley values are exponential, and even sampled approximations are demanding for high-resolution medical images, which can rule SHAP out of real-time applications.

D. Comparative Perspective of XAI Categories

The three families produce explanations at different granularities:

- gradient-based techniques deliver region-level cues;
- backpropagation-based techniques deliver pixel-level cues;
- perturbation-based techniques deliver feature-level cues.

Recent evidence [4] indicates that no single tool is enough on its own; layering several methods produces a richer, less ambiguous picture. This observation directly motivates the unified framework explored later in the paper.

E. Performance Evaluation

To evaluate the underlying classifier objectively, several complementary metrics are typically used. Precision, recall, and the F1 score quantify class-specific performance; the area under the ROC curve (AUC-ROC) captures discrimination across thresholds.

A confusion matrix breaks errors into true/false positives and negatives, exposing where the model fails. Drilling into the misclassified samples — for instance, scans with motion artefacts or partially visible tumours — typically reveals data-quality issues that limit reliability [3], [10].

F. Comparative Analysis of XAI Techniques

Beyond running each XAI tool independently, it is important to consider how they behave relative to each other. Their effectiveness varies with the type of insight required.

Grad-CAM is best suited to communicating “where” the network is looking, but it stops at coarse localisation. LRP, in contrast, returns a per-pixel relevance map that supports detailed border analysis. SHAP focuses on “what” contributes to the decision, providing both per-instance and aggregate views.

Empirical comparisons consistently show that any single technique paints only part of the picture [8]; combining them produces a more rounded interpretation. Pairing Grad-CAM with deep backbones such as ResNet, for instance, has been shown to refine the visualisation of relevant regions [9].

For these reasons, the framework reviewed in the next section deliberately combines several XAI tools to produce multi-resolution explanations of tumour predictions.

VI. UNIFIED XAI FRAMEWORK ANALYSIS

Conventional CNN-based diagnostic pipelines are accurate but opaque, which limits adoption. The literature is now converging on a remedy: layer multiple XAI tools on top of the same predictor so that explanations operate at several resolutions simultaneously.

A unified framework processes MRI inputs through a CNN to obtain the prediction and then dispatches the same activations and gradients to several XAI engines in parallel. Grad-CAM produces a regional saliency map; LRP yields a fine-grained pixel attribution; SHAP estimates feature-level contributions. The clinician receives all three views together, rather than choosing one in isolation.

Because the views are complementary, they reinforce one another in difficult cases. Grad-CAM gives a quick answer to “where is the lesion?”, LRP confirms “which pixels exactly?”, and SHAP responds to “which learned features dominated?”. When the methods agree, confidence in the prediction rises; when they disagree, the inconsistency itself is a useful diagnostic signal.

Comparative studies report that such fusion improves the perceived reliability of explanations and reduces the ambiguity of single-method outputs [4]. Outstanding research questions include the runtime cost of running several explainers in production, and how to reconcile cases where the methods disagree.

The pattern shown in Fig. ?? reinforces the central message of this paper: a single XAI lens is rarely enough, whereas a coordinated multi-lens setup gives the radiologist a far more defensible justification for each automated reading.

VII. FUTURE DIRECTIONS AND OPEN RESEARCH CHALLENGES

Although the unified XAI paradigm is conceptually appealing, several practical and scientific obstacles must be addressed before such systems can be integrated into routine clinical workflows. This section consolidates the main directions in which we expect future research to advance.

A. Standardised Evaluation of Explanation Quality

At present, there is no universally accepted benchmark for grading the quality of an explanation. Studies routinely report classification metrics — accuracy, precision, recall — but rarely quantify whether the produced heatmaps, relevance maps, or feature attributions are clinically meaningful. Future work should establish objective, reproducible measures of faithfulness, stability, and localisation accuracy. Such standards will allow researchers to compare XAI techniques on equal footing and help regulators audit clinical AI systems.

B. Lightweight and Real-Time Explainers

Several powerful XAI methods, particularly those based on Shapley values, are computationally costly. In a real-time radiology setting — where a single MRI study may need to be reviewed within minutes — these costs translate directly into deployment friction. Developing approximation strategies, GPU-friendly implementations, and selective explanation policies (where only ambiguous predictions trigger the full XAI stack) is a promising direction. Knowledge-distillation and surrogate-model techniques may also bring high-fidelity explanations within an acceptable latency budget.

C. Multi-Modal and Longitudinal Integration

Brain tumour management is rarely a single-imaged decision. Clinicians cross-reference multiple MRI sequences (T1, T2, FLAIR, contrast-enhanced T1), prior scans, lab results, and patient history. Future XAI frameworks should therefore explain predictions across modalities and across time, highlighting not only spatial regions but also temporal trends — for example, lesion growth between two visits. This longitudinal view is essential for treatment-response monitoring and recurrence detection.

D. Human-Centred Explanation Design

Most current XAI outputs are designed by AI researchers for AI researchers. Heatmaps and SHAP plots, while informative, are not always intuitive to oncologists, neurosurgeons, or general radiologists. Co-designing explanation interfaces with clinicians, validating their utility through user studies, and aligning the language of explanations with established radiology reporting templates would substantially raise the practical value of these tools.

E. Robustness, Uncertainty, and Trust Calibration

A reliable diagnostic assistant must communicate not just its prediction, but also its confidence and the conditions under which its explanations remain valid. Adversarial perturbations, distribution shift, scanner-vendor variability, and demographic bias all threaten this reliability. Coupling XAI with calibrated uncertainty estimates (e.g., Bayesian neural networks, conformal prediction) and stress-testing the resulting explanations against shifted data are important next steps.

F. Regulatory, Ethical, and Privacy Considerations

Deploying interpretable AI in healthcare touches on regulatory frameworks such as the FDA's Software-as-a-Medical-Device guidance and the EU AI Act, as well as data-protection regimes like HIPAA and GDPR. XAI plays a dual role here: it helps satisfy the "right to explanation" for patients, and provides the audit trail regulators increasingly require. Future research should therefore couple technical XAI advances with governance models, accountability mechanisms, and privacy-preserving training paradigms such as federated learning.

G. Foundation Models and Self-Supervised Pretraining

A complementary trend is the rise of large, self-supervised foundation models trained on diverse medical imaging corpora. Adapting such models to brain-tumour analysis could relax the data-hunger problem of fully supervised CNNs, but it also raises new explainability questions: the relevant features may be distributed across very deep networks and hundreds of attention heads.

Designing XAI tools that operate natively on transformer-based medical foundation models—as opposed to being bolted on after the fact—is an attractive open frontier.

VIII. CONCLUSION

This survey has consolidated the recent state of the art in deep learning and Explainable Artificial Intelligence as applied to MRI-based brain tumour analysis. The reviewed evidence confirms that CNN-style architectures are very effective at extracting predictive features, but their black-box nature is the chief barrier to safe deployment in clinical practice.

Among XAI techniques—Grad-CAM, LRP, and SHAP—being the most prominent—have been proposed to bridge this gap. They provide complementary perspectives, from regional saliency through pixel-level attribution to feature contribution analysis. Each, however, has its own granularity, computational overhead, and reliability trade-offs.

Our principal observation is that no single explanation tool is sufficient on its own. Combining several into a unified framework yields multi-resolution rationales that are demonstrably more informative than any single output. This combination is especially valuable in healthcare, where transparency directly affects clinician trust and patient safety.

The survey also flags several open challenges: there is still no consensus on how to evaluate explanation quality, several techniques remain too costly for real-time use, and a translation gap persists between the artefacts that XAI tools produce and the language clinicians prefer. Tackling these issues—alongside the future directions outlined in the previous section—is essential before such systems can move from research prototypes into routine radiology.

In summary, unified XAI frameworks represent a practical step toward DL systems that are not only accurate but also interpretable, auditable, and clinically acceptable—three properties that, together, are required to make AI a trusted partner in medical decision-making. By coupling rigorous evaluation, lightweight implementations, multi-modal integration, human-centred design, robustness analysis, and sound governance, the community can move steadily from promising prototypes to dependable clinical assistants.

REFERENCES

- [1] M.Z. Khaliki and M.S. Basarslan, “Brain tumor detection from images and comparison with transfer learning methods,” *Scientific Reports*, vol. 14, no. 1, 2024.
- [2] M. M. M. et al., “Enhancing brain tumor detection in MRI images through explainable AI using Grad-CAM with ResNet50,” *BMC Medical Imaging*, vol. 24, no. 1, 2024.
- [3] A. Al-Fakih et al., “FLAIR MRI sequence synthesis using squeeze attention generative model for reliable brain tumor segmentation,” *Alexandria Engineering Journal*, vol. 99, pp. 108–123, 2024.
- [4] P. Narayankar and V. P. Baligar, “Explainability of brain tumor classification based on region,” in *Proc. Int. Conf. Emerging Technologies in Computer Science (ICETCS)*, 2024, pp. 1–6.
- [5] S. Ahmed, S. N. Nobel, and O. Ullah, “An effective deep CNN model for multiclass brain tumor detection using MRI images and SHAP explainability,” in *Proc. Int. Conf. Electrical, Computer and Communication Engineering (ECCE)*, 2023, pp. 1–6.
- [6] T. Hussain and H. Shouno, “Explainable deep learning approach for multi-class brain MRI tumor classification and localization using Grad-CAM,” *Information*, vol. 14, no. 12, 2023.
- [7] F. Şefċık and W. Beneṡova, “Improving a neural network model by explanation-guided training for glioma classification based on MRI data,” *Int. J. Information Technology*, vol. 15, no. 5, pp. 2593–2601, 2023.
- [8] M. I. Mahmud, M. Mamun, and A. Abdelgawad, “A deep analysis of brain tumor detection from MR images using deep learning networks,” *Algorithms*, vol. 16, no. 4, 2023.
- [9] E. M. Senan et al., “Early diagnosis of brain tumour MRI images using hybrid techniques between deep and machine learning,” *Computational and Mathematical Methods in Medicine*, 2022.
- [10] R. Najjar, “Redefining radiology: A review of artificial intelligence integration in medical imaging,” *Diagnostics*, vol. 13, no. 17, 2023.



10.22214/IJRASET



45.98



IMPACT FACTOR:
7.129



IMPACT FACTOR:
7.429



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24*7 Support on Whatsapp)