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

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.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-boxnaturemakesthereasoningbehindeachpredictionhard 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-wiseRelevancePropagation(LRP), andSHapleyAdditiveexPlanations(SHAP)—whichrespectively expose salient regions, per-pixel attributions, and feature-level contributions. We argue that fusing these views into a unifiedXAIpipelineyieldslayered,moretrustworthyrationalesthanany single method, and is a practical path toward auditable clinical decision support.

I. INTRODUCTION

The intersection of Artificial Intelligence (AI) and Deep Learning(DL)hasreshapedhowradiologicaldata—including MRI,CT,andX-raystudies—isexaminedandtriaged.Within thislandscape,MRIisespeciallyvaluedforcerebralpathology becauseofitshighsoft-tissuecontrast,whichmakesanomalous masses easier to delineate.

CNNbasedarchitectures,inparticular,havebecometheworkhorseofautomatedtumourscreening:theyingestrawvoxeldataandprogressivelylearndiscriminativefeatures,typicallyreachingcompetitiveaccuracyonbenchmarkdatasets.

Theflipsideofthisexpressivenessisopacity.Becausethe layered transformations are not directly inspectable, the model's logicishiddenfromendusers—apropertythatispoorly 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 ablack-box model partially transparent. Among the techniques that have gained traction in the medical imaging literature are Grad-CAM,LRP,andSHAP,eachoperatingonadifferentmathematicalprinciple.

Acleartrendinrecentworkistostepbeyondsingle-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.

ContributionsofthisSurvey

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;
- comparesthedominantXAItechniques(Grad-CAM,LRP, SHAP) on common axes;
- contrasts local with global explanation paradigms in the medical context;
- enumeratestheopenproblemsthatlimittoday'sXAI-augmented diagnostic systems; and
- arguesforunifiedexplainabilitystacksasaroutetohigher clinical trust and faster adoption.

Thebroaderobjectiveistoconnectstrongpredictiveperfor-mance with the interpretability requirements of bedside use, while pointing to research directions that could mature the field.

II. BACKGROUND AND MOTIVATION

A. *NeedforIntelligentMedicalImagingSystems*

Inmodernoncology,imagingiscentraltodiagnosingpatholo-gies such as cerebral tumours, and timely, accurate readings haveadirectimpactontherapyselectionandpatientprognosis. Because MRI and CT studies generate a continually growing pool of high-resolution data, fully manual interpretation by radiologistsbecomesabottleneckandissusceptibletofatigue-induced mistakes. The result is a strong demand for computer-assisted screening tools that can sift through scans rapidly and consistently.

CNNshavebecomethedefactobuildingblockforsuchtools, sincetheycanextracttask-relevantrepresentationsdirectlyfrom 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 basisfor each prediction.

B. LimitationsofConventionalDeepLearningModels

While CNNs and related architectures perform well on imaging benchmarks, the "opaque reasoning" problem per-sists. Without an explicit explanation, clinicians cannot easily distinguishapredictiongroundedingenuinepathologyfrom onedrivenbyspuriouscuessuchasscannerartefacts.This uncertaintyfuels reluctancetointegrateAI intocarepathways. Practitionersneedconfirmationthatthenetwork'sfocus alignswithanatomicallymeaningfulregionsbeforetheyaccept itsoutputasasecondreader.Aresearchliteraturethatoptimises chiefly for accuracy — without auditing decision pathways — thereforeofferslimitedoperationalvalueinahospitalsetting.

C. RoleofExplainableAIinMedicalDiagnosis

XAI offers a constructive response by providing tools that translate hidden network computations into human-readable cues.Theoutputmaytaketheformofsaliencymaps,relevance heat-overlays, or per-feature attribution scores.

Forbraintumourstudies,methodssuchasGrad-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 intoa CNN-baseddiagnostic pipelineyields asystem that is both performant and inspectable.

Thisrealisationjustifiesthemovetowardunifiedframeworks solutions that intentionally combine multiple XAI views so thatcliniciansreceivearicher,multi-resolutionjustificationfor every output.

III. LITERATURE SURVEY

A. BrainTumorDetectionusingTransferLearning

AuthorandYear:M.Z.KhalikiandM.S.Bas¸arslan(2024) Methodology: The work investigates how transfer learning canbeexploitedfortumouridentificationonMRI.Established backbones — most notably VGG19 and InceptionV3 — serve as feature extractors, with the upper layers re-trained on themedicaldataset.Byreusingweightslearnedonlarge-scale natural-imagecorpora,theauthorsachievecompetitiveaccuracywithoutrequiringvastamountsofmedicaldata,whichisgenerallyscarceinthisdomain.

Limitation: Performance is tightly coupled to how wellthe source backbone transfers to the target task. Pretrainingon natural photographs can leave domain-specific MRI cues poorlyrepresented,andgeneralisationdeterioratesifthetarget dataset lacks demographic or modality diversity.

B. *Explainable AI using Grad-CAM with ResNet50* Author and Year: M. M. M. et al. (2024)

Methodology:TheauthorspairaResNet50backbone 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; lesionborderscannotberesolvedprecisely.Theheatmapsmay also activate over background structures, producing visually plausible but clinically unhelpful explanations. Quality is boundedbythediscriminativepoweroftheunderlyingfeature maps.

C. *FLAIR MRI for Brain Tumor Segmentation* AuthorandYear:A.Al-Fakihetal.(2024)

Methodology:Thiscontributiontargetssegmentationrather 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 mechanismthatstrengthensthevisibilityofpathologicaltissue. 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. XAITechniquesforBrainTumorAnalysis

Author and Year: P. Narayankar and V. P. Baligar (2024) **Methodology:**ThepaperbenchmarksseveralXAItools appliedtoacommonCNNclassifieronMRIdata.

By placingeachtechniqueonthesamemodelandinputs,the authors compare what each tool reveals about region salience, feature relevance, and decision boundaries, suggesting that the techniquessurfacedifferent—andpartiallycomplementary—facetsofnetworkbehaviour.
Limitation:Althoughthecomparisonisinformative,itstops shortoffusingthetechniquesintoasinglecoherentexplanation pipeline.Readingseveralexplanationssidebysidealsoimposes a cognitive cost and typically requires familiarity with each method's assumptions.

E. CNNwithSHAPforTumorClassification

Author and Year: S. Ahmed, S. N. Nobel, and O. Ullah (2023)
Methodology: A standard CNN is trained for multi-class tumourcategorisation,andSHAPislayeredontoptoattribute importancevaluestofeatures.SHAPreturnsbothper-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 inheritsany errors in the underlying classifier — a confidently wrong prediction will receive a confidently wrong rationale.

F. Grad-CAMBasedExplainableCNN

AuthorandYear:T.HussainandH.Shouno(2023)
Methodology: This study targets multi-class tumour clas-sification with a CNN and uses Grad-CAM to localise the discriminative regions per class. The resulting heatmaps are blendedwiththesourceMRIsothatradiologistscanvisu-ally validate whether the network's attention coincides with anatomically relevant areas.
Limitation:Localisationremainscoarse;smallorirregularly shapedlesionsareoftenonlypartiallyhighlighted.Furthermore, Grad-CAMissensitivetowhichconvolutionallayerisusedto compute gradients, so explanations vary with implementation choices.

G. LRPforModelInterpretability

AuthorandYear:F.Šefčík andW.Benesˇova´(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. DeepLearningforBrainTumorDetection

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 standardpreprocessing—normalisation,resizing,andaugmen-tation—theauthorsreportperformanceusingcommonmetrics anddiscusshowdepth,kerneldesign,andregularisationaffect outcomes.
Limitation: The pipeline depends heavily on dataset scale andclassbalance,andoverfittingcanemergeoncethetraining 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:Theproposalhybridisesrepresentationlearn-ing with classical pattern recognition: a CNN computes high-level descriptors from MRI slices, after which a SupportVector Machine performs the final decision. The combination capitalises on the SVM's effectiveness in low-sample, high-dimensional feature spaces.
Limitation:Two-stagepipelinesarehardertomaintainand 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 AuthorandYear:R.Najjar(2023)
Methodology:Thisisaliteraturereviewratherthana newmethod.Itsurveyshowmachinelearning,andCNNs 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 recommenda-tions 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 tonavigate,TableIcompilestheirmethods,contributions,and weaknesses side by side.

A. DeepLearningApproachesforTumorDetection

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.
BeyondpureDL,hybridpipelineshavebeenproposed.Senanetal.[9]combineCNNderivedembeddingswithadownstreamclassicalclassifier,reportinggainsovereithercomponentused alone. Such designs aim to merge the representational depthofneuralnetworkswiththesample-efficiencyofmargin-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.Thesefactorscontinuetoslowreal-worlddeployment.

B. ExplainableAITechniquesinMedicalImaging

Tomitigateopacity,multipleXAIfamilieshavebeenfolded intoMRIpipelines.Ahmedetal.[5]integrateSHAPwith a CNN to yield local and global feature-importance signals, whileHussainandShouno[6]relyonGrad-CAMtoprojectthe model's attention back onto the input image. LRP, as adapted byŠefčíkandBenesˇova´[7],goesastepfurtherbyattributing relevance at the level of individual pixels.
Narayankar and Baligar [4] explicitly compare several XAI toolsandconcludethatnosinglemethodcaptureseveryaspect ofamodel'sreasoning—differentmethodsilluminatedifferent facets,suggestingthatcombinationsaremoreinformative.
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 thenatureoftheproducedexplanation,sotheselectionshould be matched to the deployment scenario.

C. MRI-BasedTumorAnalysisandImagingTechniques

At the modality level, MRI remains the workhorse for cerebral imaging. Al-Fakih et al. [3] argue that FLAIR sequences,inparticular,sharpenlesioncontrast,whichdirectly benefits downstream segmentation accuracy.
Najjar [10] takes a wider lens, surveying how AI assistants areenteringradiologyworkflows,andflagsgovernance,trans-parency, and validation as the key non-technical bottlenecks for adoption.
Inaggregate,theliteratureshowsexcellentrawperformance from DL models but limited, fragmented progress on interpretability.EachXAItechniquecapturespartofthepicture;

TABLEI
COMPARATIVEANALYSISOFEXISTINGRESEARCHINBRAINTUMORDETECTIONANDXAI

Author	Method	KeyContribution	Limitation
KhalikiandBas¸arslan [1]	TransferLearning(CNN)	Reusedpretrainedbackbonestolift classificationaccuracywithlimitedmedical data.	Pretrainingdomainmismatchmay suppressmedicallyrelevantcues.
M.M.M.etal.[2]	Grad-CAM+ResNet50	Producedoverlayheatmapsthathigh-lighttumourregionsusedbythecla	Spatialresolutioniscoarse;spurious activations occur.

		ssifier.	
Al-Fakihetal.[3]	FLAIRMRIAnalysis	LeveragedFLAIRcontrastandatten-tiontorefinelesionboundaries.	Sensitivetoacquisitionqualityandis computationallyheavy.
NarayankarandBaligar[4]	MultipleXAITechniques	BenchmarkedseveralXAItoolsside bysideonCNNclassifiers.	Nofusionofmethodsintoasingle explanationpipeline.
Ahmedetal.[5]	CNN+SHAP	Deliveredbothlocalandglobal feature-attributionanalyses.	Shapleycomputationscalespoorly withimagesize.
HussainandShouno[6]	Grad-CAMCNN	Class-specificheatmapssupport multi-classlesionlocalisation.	Granularityinsufficientforfine boundaries.
S˘efc˘´ıkandBenes˘ova´[7]	LRP	Pixel-levelattributionviarelevanceback-propagation.	Outputscanbenoisyandrule-dependent.
Mahmudetal.[8]	DeepLearningCNN Mod-els	ComparativeCNNstudyforauto-mateddetection.	Data-hungryandoffersnointer-pretability.
Senanetal.[9]	HybridCNN+ML	CombinedCNNfeatureswithclassi-calclassifiersforhigheraccuracy.	Increasedcomplexity;explanation ab-sent.
Najjar[10]	AIMedicalImagingRe-view	SynthesisesAIuseacrossradiology workflows.	Nonewexperimentsorvalidation.

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

V. TAXONOMY OF EXPLAINABLE AITECHNIQUES

XAImethodscanbegroupedaccordingtothemathematical 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-BasedMethods

This family inspects how the prediction changes when the input—oranintermediateactivation—isperturbed,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 naturallyalignedwiththeinput,itcanbeoverlaidonthesource MRI for direct inspection.

These methods are attractive thanks to their low computa-tional 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-BasedMethods

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 rele-vance through each layer using rules that respect the model's structure,eventuallyproducingaper-pixelimportancemap[7].

The result is much finer than a Grad-CAM heatmap, which makesLRPattractiveforradiologistslookingforpreciselesion edges.

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

C. PerturbationandGame-Theory-BasedMethods

Athirdgroupderivesimportancescoresbymodifyinginput features and observing the resulting change in output. SHAP (SHapleyAdditiveexPlanations)isthemostprominentmember, drawing on cooperative game theory.

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

Theprincipalcostiscomputational:exactShapleyvaluesare exponential,andevensampledapproximationsaredemanding for high-resolution medical images, which can rule SHAP out of real-time applications.

D. ComparativePerspectiveofXAICategories

The three families produce explanations at different granu-larities:

- gradient-basedtechniquesdeliverregion-levelcues;
- backpropagation-basedtechniquesdeliverpixel-levelcues;
- perturbation-basedtechniquesdeliverfeature-levelcues.

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. PerformanceEvaluation

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 intothe misclassified samples — for instance, scans with motion artefacts or partially visible tumours — typically reveals data-quality issues that limit reliability [3], [10].

F. ComparativeAnalysisofXAITechniques

BeyondrunningeachXAItoolindependently,itisimportant 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 networkislooking,butitstopsatcoarselocalisation.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].

Forthesereasons,theframeworkreviewedinthenextsection 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 convergingonaremedy:layermultipleXAItoolsontop of the same predictor so that explanations operate at several resolutions simultaneously.

A unified framework processes MRI inputs through a CNN toobtainthepredictionandthendispatchesthesameactivations 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 diag-nostic signal.

Comparative studies report that such fusion improves the perceivedreliabilityofexplanationsandreducestheambiguity of single-method outputs [4]. Outstanding research questions includetheruntimecostofrunningseveralexplainersin 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 appeal-ing,severalpracticalandscientificobstaclesmustbeaddressed 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. StandardisedEvaluationofExplanationQuality

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. LightweightandReal-TimeExplainers

SeveralpowerfulXAImethods,particularlythosebased 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 deploymentfriction.Developingapproximationstrategies,GPU-friendly implementations, and selective explanation policies (where only ambiguous predictions trigger the full XAI stack) isapromisingdirection.Knowledge-distillationandsurrogate-model techniques may also bring high-fidelity explanations within an acceptable latency budget.

C. Multi-ModalandLongitudinalIntegration

Braintumourmanagementisrarelyasingle-imagedecision. Clinicians cross-reference multiple MRI sequences (T1, T2, FLAIR, contrast-enhanced T1), prior scans, lab results, and patienthistory.FutureXAIframeworksshouldthereforeexplain 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-CentredExplanationDesign

MostcurrentXAIoutputsaredesignedbyAIresearchersfor 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 aligningthelanguageofexplanationswithestablishedradiology reporting templates would substantially raise the practical value of these tools.

E. Robustness,Uncertainty,andTrustCalibration

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

F. Regulatory,Ethical,andPrivacyConsiderations

Deploying interpretable AI in healthcare touches on regu-latory frameworks such as the FDA's Software-as-a-Medical-DeviceguidanceandtheEUAIAct,aswellasdata-protection regimeslikeHIPAAandGDPR.XAIplaysadualrolehere: 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. FoundationModelsandSelf-SupervisedPretraining

A complementary trend is the rise of large, self-supervised foundationmodelstrainedondiversemedicalimagingcorpora. Adapting such models to brain-tumour analysis could relaxthe 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-basedmedicalfoundationmodels—asopposedto beingboltedonafterthefact—isanattractiveopenfrontier.

VIII. CONCLUSION

This survey has consolidated the recent state of the art in deeplearningandExplainableArtificialIntelligenceasapplied 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.

ArangeofXAItechniques—Grad-CAM,LRP,andSHAP beingthemostprominent—havebeenproposedtobridgethis gap.Theyprovidecomplementaryperspectives,fromregional saliency through pixel-level attribution to feature contribution analysis. Each, however, has its own granularity, computational overhead, and reliability trade-offs.

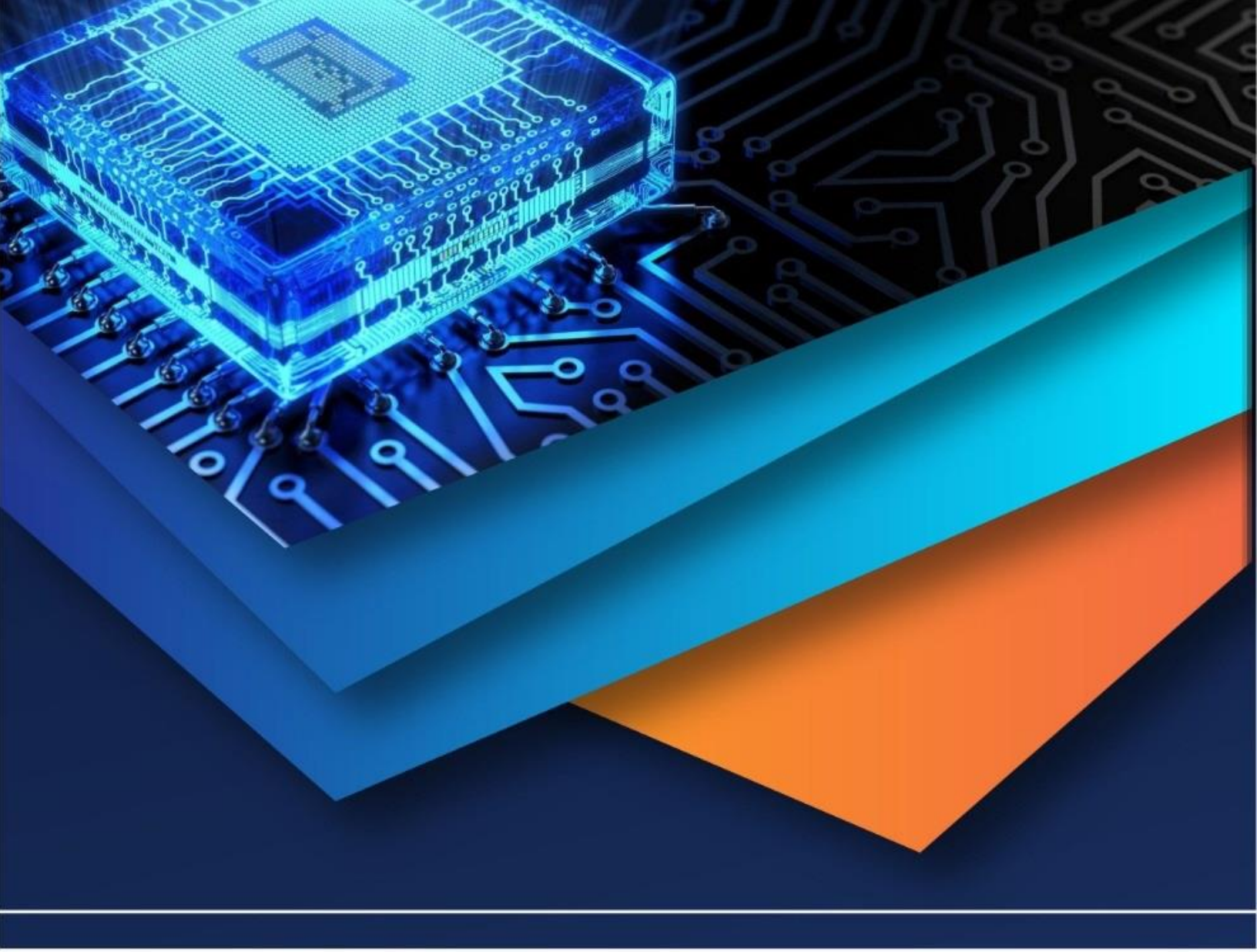
Our principal observation is that no single explanation toolissufficientonitsown.Combiningseveralintoaunifiedframe-work yields multi-resolution rationales that are demonstrably more informative than any single output. This combination is especially valuable in healthcare, where transparency directly affects clinician trust and patient safety.

The survey also flags several open challenges: there is still noconsensusonhowtoevaluateexplanationquality,several techniquesremaintoocostlyforreal-timeuse,andatranslation gappersistsbetweentheartefactsthatXAItoolsproduce and the language clinicians prefer. Tackling these issues — alongsidethefuturedirectionsoutlinedinthepreviussection 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,andclinicallyacceptable—threeprop-erties that, together, are required to make AI a trusted partner in medical decision-making. By coupling rigorous evaluation, lightweight implementations, multi-modal integration, human-centreddesign,robustnessanalysis,andsoundgovernance,the community can move steadily from promising prototypes to dependable clinical assistants.

REFERENCES

[1] M.Z.KhalikiandM.S.Bas¸arslan,“Braintumordetectionfromimagesandcomparisonwithtransferlearningmethods,”ScientificReports,vol. 14, no. 1, 2024.
[2] M. M. M. et al., “Enhancing brain tumor detection in MRI imagesthroughexplainableAIusingGrad-CAMwithResNet50,”BMCMedicalImaging, vol. 24, no. 1, 2024.
[3] A. Al-Fakih et al., “FLAIR MRI sequence synthesis using squeezeattention 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 classifi-cation based on region,” in Proc. Int. Conf. Emerging Technologies inComputer Science (ICETCS), 2024, pp. 1–6.
[5] S.Ahmed,S.N.Nobel,andO.Ullah,“AneffectivedeepCNNmodel for multiclass brain tumor detection using MRI images andSHAP explainability,” in Proc. Int. Conf. Electrical, Computer andCommunication Engineering (ECCE), 2023, pp. 1–6.
[6] T.HussainandH.Shouno,“Explainabledeeplearningapproachformulti-classbrainMRItumorclassificationandlocalizationusingGrad-CAM,”Information, vol. 14, no. 12, 2023.
[7] F. S˘efc˘´ık and W. Benes˘ova´, “Improving a neural network model byexplanation-guidedtrainingforgliomaclassificationbasedonMRIdata,”Int. J. Information Technology, vol. 15, no. 5, pp. 2593–2601, 2023.
[8] M. I. Mahmud, M. Mamun, and A. Abdelgawad, “A deep analysis ofbrain 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 usinghybrid techniques between deep and machine learning,” Computationaland Mathematical Methods in Medicine, 2022.
[10] R. Najjar, “Redefining radiology: A review of artificial intelligenceintegration 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