

RESEARCH ARTICLE

Reliable and explainable deep learning for brain tumor diagnosis: Uncertainty-calibrated MRI classification with external validation

MD Nazmul Hasan¹ ✉ and Mohammad Mahmudul Hasan Bhuyain²

¹ *Pompea College of Business, University of New Haven, West Haven, CT, USA*

² *Molecular Biologist, Verralize, East Haven, Connecticut, USA*

ORCID 1: 0009-0007-2555-6047; ORCID 2: 0009-0006-3089-2101

Second author e-mail (as supplied): mmhasan1992@yahoo.com

Corresponding Author: MD Nazmul Hasan, **E-mail:** mhasa9@unh.newhaven.edu

ABSTRACT

Reliable brain-tumor diagnosis from magnetic resonance imaging (MRI) requires more than high classification accuracy. Deep neural networks can be overconfident under scanner, protocol, demographic, and disease-spectrum shifts, while visually plausible explanations may not faithfully identify the evidence driving a prediction. This methodological study presents a reliability-centered framework for MRI brain-tumor classification that integrates external validation, post-hoc probability calibration, predictive uncertainty estimation, out-of-distribution screening, and quantitatively evaluated explainability. The proposed pipeline uses patient-level data separation, harmonized preprocessing, a deep ensemble of independently initialized classifiers, temperature scaling on a dedicated calibration set, and entropy-based selective prediction. Grad-CAM and SHAP are combined with lesion masks to test whether model attention overlaps clinically relevant tumor regions rather than background anatomy or acquisition artifacts. Performance is assessed using macro-averaged area under the receiver operating characteristic curve, sensitivity, specificity, F1 score, Brier score, expected calibration error, negative log-likelihood, uncertainty-error association, and risk-coverage analysis. External validation is designed to hold out complete institutions or datasets, thereby testing robustness to domain shift rather than random image-level variation. A synthesis of 32 peer-reviewed studies published through March 2026 shows that brain-tumor MRI research increasingly incorporates explainability, yet calibration, uncertainty-aware referral, and independent external testing remain less consistently combined. The framework therefore treats discrimination, calibration, uncertainty, localization fidelity, and generalizability as complementary dimensions of clinical reliability. It offers a reproducible study design for developing decision-support models that can identify both confident predictions and cases that should be deferred for expert review. This design prioritizes safe failure behavior under real-world variability.

KEYWORDS

brain tumor; magnetic resonance imaging; deep learning; uncertainty calibration; explainable AI; external validation; selective prediction

ARTICLE INFORMATION

ACCEPTED: 15 July 2026

PUBLISHED: 24 August 2026

DOI: 10.32996/jmhs.2026.7.9.21

Introduction

Magnetic resonance imaging (MRI) is central to the noninvasive assessment of intracranial tumors because it provides high soft-tissue contrast and multiple complementary sequences without ionizing radiation. Interpretation, however, remains demanding. Gliomas, meningiomas, pituitary tumors, metastases, treatment-related changes, and non-neoplastic abnormalities can share

overlapping appearances, while tumor morphology varies across patients, sequences, scanners, and institutions. These factors have motivated extensive research on computer-aided diagnosis, particularly convolutional neural networks, transfer learning, and more recently transformer-based architectures. Public resources such as BraTS and TCGA/TCIA have accelerated algorithm development by providing curated multimodal MRI and expert annotations (Bakas et al., 2017; Menze et al., 2015). Across classification studies, internal accuracies above 95% are now common, especially on curated datasets and image-level splits (Asiri et al., 2024; Deepak & Ameer, 2019; Kumar et al., 2023; Swati et al., 2019).

High internal accuracy does not, by itself, establish clinical reliability. Medical imaging models can learn site-specific shortcuts, scanner signatures, preprocessing artifacts, or patient identity cues. Badža and Barjaktarović (2020), for example, demonstrated a substantial difference between record-wise and subject-wise evaluation, illustrating how leakage-prone sampling can inflate apparent performance. A systematic review and meta-analysis focused on externally validated brain-tumor MRI models likewise identified dataset heterogeneity, hardware differences, and validation design as important determinants of generalization (Wang et al., 2024). These concerns are especially important for MRI because acquisition protocols vary considerably across field strength, vendor, coil, sequence parameters, contrast administration, reconstruction, and local workflow. A model that performs well in one environment may therefore become less discriminative or systematically overconfident elsewhere.

Reliability also requires probabilities that mean what they appear to mean. Modern neural networks can assign very high confidence to incorrect predictions, which creates a mismatch between discrimination and calibration. In medical imaging, calibrated probabilities matter because confidence may influence whether a prediction is accepted, reviewed, or escalated. Studies in medical image analysis have shown that temperature scaling, Bayesian approximations, deep ensembles, and related techniques can improve probability calibration and provide useful uncertainty signals (Carneiro et al., 2020; Mehrtash et al., 2020; Penso et al., 2024). A systematic review found Monte Carlo dropout and deep ensembles among the most frequently applied uncertainty approaches, but also noted substantial heterogeneity in evaluation practices (Kurz et al., 2022). Importantly, uncertainty should support selective prediction: the model should retain lower-risk cases while referring ambiguous or distribution-shifted cases for human assessment.

Explainability addresses a different question. Saliency and attribution methods can reveal which image regions are associated with a prediction, but a visually convincing heatmap does not prove that a model is correct, calibrated, or causally reasoning from pathology. Rigorous evaluations have shown that some saliency techniques are unstable or weak at abnormality localization (Arun et al., 2021), and broader critiques caution against using post-hoc explanations as substitutes for validation (Ghassemi et al., 2021; Munroe et al., 2024). Brain-tumor studies increasingly apply Grad-CAM, LIME, and SHAP, yet these tools are often reported qualitatively rather than tested against lesion annotations (Iftikhar et al., 2025; Musthafa et al., 2024; Nazir et al., 2024).

This paper develops a unified, literature-grounded framework for reliable and explainable brain-tumor MRI classification. The central premise is that a clinically credible model should satisfy five complementary requirements: strong discrimination, calibrated probabilities, meaningful uncertainty, anatomically grounded explanations, and external generalization. The proposed study design combines deep ensembles, temperature scaling, entropy-based referral, quantitative explanation localization, and institution-level external validation. Rather than treating interpretability as a cosmetic addition to an accuracy-focused model, the framework positions explainability and uncertainty as auditable components of a broader reliability strategy. It also specifies reproducible reporting practices so performance claims can be audited, compared, replicated, and interpreted.

Reliability-centered MRI classification workflow

Model development is isolated from calibration and final external testing.

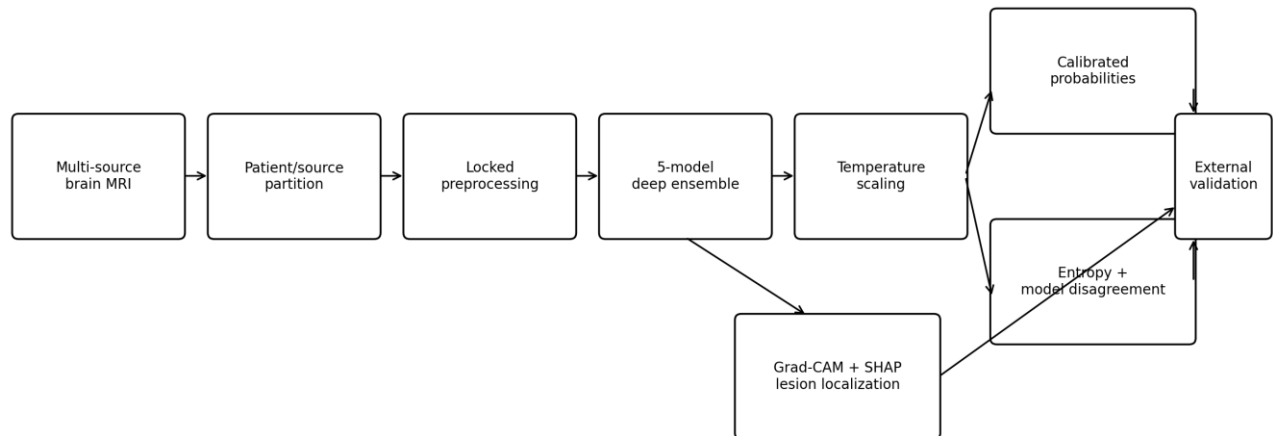


Figure 1. Reliability-centered MRI classification workflow. The development, calibration, explainability, uncertainty, and external-validation stages are separated to reduce information leakage and adaptive overfitting.

Literature Review

The literature on automated brain-tumor MRI analysis has evolved from handcrafted features and region-based descriptors toward end-to-end deep learning. Cheng et al. (2015) showed that tumor-region augmentation and partitioning could improve conventional feature-based classification of glioma, meningioma, and pituitary tumors, establishing the value of contextual tissue around the lesion. Deep learning then reduced dependence on manually engineered descriptors. Sajjad et al. (2019) combined segmentation, extensive augmentation, and a fine-tuned convolutional network for multigrade classification, while Deepak and Ameer (2019) used transfer-learned GoogLeNet features with conventional classifiers. Swati et al. (2019) introduced block-wise fine-tuning of VGG19, and Badža and Barjaktarović (2020) developed a compact CNN while explicitly contrasting record-wise and subject-wise cross-validation. These studies helped demonstrate the feasibility of high-performing MRI classifiers but also exposed a recurring problem: apparent accuracy depends strongly on data partitioning, sample independence, and the similarity between development and test distributions.

Subsequent work expanded architectures and diagnostic categories. Sekhar et al. (2022) combined fine-tuned GoogLeNet features with machine-learning classifiers in an Internet-of-Medical-Things-oriented system. Kumar et al. (2023) reported strong performance using transfer learning and an improved ResNet-based approach. Asiri et al. (2024) applied a Swin Transformer to four-class classification of glioma, meningioma, pituitary tumor, and no-tumor MRI, reflecting a broader shift toward attention-based models. Systematic reviews confirm rapid methodological diversification. Bouhafra and El Bahi (2025) summarized 60 studies covering transfer learning, autoencoders, transformers, attention mechanisms, and generative augmentation, while Dorfner et al. (2025) reviewed deep learning across segmentation, classification, molecular prediction, and outcome-related brain-tumor tasks. The common conclusion is that algorithmic sophistication has advanced faster than evidence for transportability into heterogeneous clinical environments.

External validation is therefore central. Wang et al. (2024) reviewed 19 externally validated brain-tumor detection and segmentation studies involving approximately 12,000 patients and identified imaging hardware and dataset diversity as important contributors to variation in generalization. Their analysis reinforces a distinction that is sometimes blurred: random holdout testing estimates performance on data sampled from the same underlying environment, whereas external validation evaluates transport to meaningfully different institutions, scanners, populations, or acquisition protocols. The latter is closer to the deployment problem. Park et al. (2024) provided a particularly relevant example by developing deep ensembles to distinguish glioblastoma from solitary brain metastasis and evaluating them on an independent external cohort. External AUC was 0.83, materially lower than values often reported on curated internal datasets, but uncertainty stratification identified a low-uncertainty subgroup with substantially stronger discrimination. This illustrates why realistic validation may reduce headline performance while increasing clinical credibility.

Probability calibration forms a second reliability dimension. A classifier can rank cases correctly yet assign probabilities that are systematically too extreme. Mehrtash et al. (2020) demonstrated overconfidence in deep medical image segmentation and investigated calibration and predictive uncertainty under several training settings. Carneiro et al. (2020) examined confidence

calibration and classification uncertainty in medical image classification, showing the practical relevance of temperature scaling and uncertainty-based rejection. Penso et al. (2024) extended calibration work to settings with label noise, an important consideration for medical imaging where reference labels may contain interobserver disagreement or imperfect ground truth. Karimi and Gholipour (2023) further linked calibration with out-of-distribution detection in medical image segmentation. Together, these studies support reporting Brier score, expected calibration error, negative log-likelihood, and reliability diagrams alongside discrimination metrics.

Uncertainty estimation attempts to quantify when a model may be wrong or when an input lies outside its training experience. Kurz et al. (2022) found substantial methodological diversity across medical image classification studies, with Monte Carlo dropout and deep ensembles among the most common approaches. Nair et al. (2020) showed that uncertainty in multiple-sclerosis MRI segmentation was concentrated around small lesions and boundaries, and that uncertainty-based filtering improved performance among retained predictions. Küstner et al. (2024) demonstrated that deep-ensemble uncertainty in MRI reconstruction responded to distributional changes involving contrast, orientation, anatomy, pathology, and sampling patterns. Although reconstruction differs from classification, the broader lesson is directly relevant: uncertainty is useful when it changes in a meaningful way under error and domain shift, not simply when it produces numerically variable scores.

Deep ensembles are attractive for brain-tumor diagnosis because they combine strong predictive performance with an empirically useful measure of epistemic uncertainty. Park et al. (2024) found that ensemble entropy separated more reliable from less reliable predictions and flagged a high proportion of out-of-distribution meningioma cases in a glioblastoma-versus-metastasis task. This is clinically meaningful because uncertainty can support selective prediction. Rather than forcing a label for every scan, a system can abstain when entropy exceeds a pre-specified threshold. Risk-coverage curves then quantify the trade-off between the proportion of cases automatically classified and the error rate among those retained. Such evaluation better reflects assisted clinical use than a single global accuracy value.

Explainable artificial intelligence has also become prominent in brain-tumor MRI. Musthafa et al. (2024) paired ResNet50 with Grad-CAM, Nazir et al. (2024) combined Grad-CAM, LIME, and SHAP with a customized CNN, and Iftikhar et al. (2025) similarly used multiple XAI techniques to highlight class-relevant regions. Akgündoğdu and Çelikbaş (2025) integrated LIME, Grad-CAM, and SHAP and reported additional training based on explanation-derived masks. Lakshmi et al. (2025) combined segmentation, Bayesian machine learning, and explainability, while early-2026 studies by Srinivas and Parvathi (2026), Adamu et al. (2026), and Pant et al. (2026) continued the trend toward four-class MRI classification with visual explanations. These studies show clear momentum toward transparency, but most still emphasize qualitative plausibility and internal performance rather than formal explanation fidelity under external shift.

That limitation matters because saliency can be persuasive without being trustworthy. Arun et al. (2021) evaluated eight saliency methods across medical imaging tasks and found that every method failed at least one criterion involving localization utility, randomization sensitivity, repeatability, or reproducibility. Ghassemi et al. (2021) argued that local explanations should not be treated as evidence that individual clinical predictions are safe or correct. Munroe et al. (2024), reviewing 75 neuroimaging studies, similarly found that popular interpretable deep-learning approaches may be suboptimal and emphasized biological validity, robustness, continuity, selectivity, and downstream performance as properties requiring evaluation. These findings support a stricter role for XAI in brain-tumor diagnosis: explanations should be quantitatively compared with lesion masks, stress-tested under perturbation, and interpreted as audit tools rather than guarantees.

The unresolved gap is therefore integrative. The literature contains strong classification models, external-validation studies, calibration methods, uncertainty frameworks, and XAI techniques, but these components are seldom evaluated together in a single brain-tumor MRI pipeline. The proposed framework addresses this fragmentation by treating reliability as multidimensional. External validation tests transportability; temperature scaling tests probability calibration; deep ensembles and entropy test predictive uncertainty; out-of-distribution analysis probes unfamiliar inputs; and lesion-overlap metrics test whether explanations focus on clinically relevant anatomy. A model would be considered useful not because it maximizes one metric, but because its errors become detectable, its confidence becomes interpretable, and its explanations remain anatomically plausible when the data source changes.

This integration also responds to reproducibility concerns. Brain-tumor datasets frequently contain multiple slices from the same patient, heterogeneous preprocessing histories, and labels assembled from different sources. Patient-level splitting, source-aware metadata, locked preprocessing, and transparent exclusion criteria are therefore essential. Without those safeguards, calibration and explainability analyses can inherit the same leakage that inflates discrimination, collectively creating a false impression of reliability.

Table 1. Representative studies informing the reliability framework.

Study	Task / setting	Reliability contribution	Primary limitation
Menze et al. (2015)	BraTS multimodal benchmark	Standardized expert-annotated MRI benchmark	Segmentation benchmark; not a calibrated diagnostic classifier
Badža & Barjaktarović (2020)	Brain-tumor CNN classification	Shows record-wise vs subject-wise evaluation gap	Single public collection; limited external transport testing
Mehrtash et al. (2020)	Medical image segmentation	Demonstrates overconfidence; evaluates calibration/UQ	Segmentation rather than brain-tumor diagnosis
Arun et al. (2021)	Medical-image saliency evaluation	Quantitative tests of saliency trustworthiness	Not specific to brain tumors
Kurz et al. (2022)	Systematic review of UQ	Deep ensembles and MC dropout prominent	High heterogeneity in uncertainty evaluation
Wang et al. (2024)	External-validation meta-analysis	Documents hardware/data heterogeneity in brain-tumor AI	Mixes detection and segmentation studies
Park et al. (2024)	GBM vs solitary metastasis	External validation + ensembles + UQ + localization	Binary differential diagnosis rather than four-class classification
Nazir et al. (2024)	Customized CNN + XAI	Combines Grad-CAM, LIME, and SHAP	Mostly curated/internal evaluation
Akgündoğdu & Çelikbaş (2025)	MRI classification + XAI	Multi-method explanation and explanation-guided learning	No prospective clinical validation
Munroe et al. (2024)	Neuroimaging XAI review	Defines robustness/validity properties for interpretation	Broad neuroimaging scope
Adamu et al. (2026)	Four-class explainable MRI classification	Hybrid architecture with explainability	Authors identify external validation as future need
Pant et al. (2026)	ConvNeXt on three datasets	Generalizability and explainability emphasis	Retrospective dataset-based evaluation

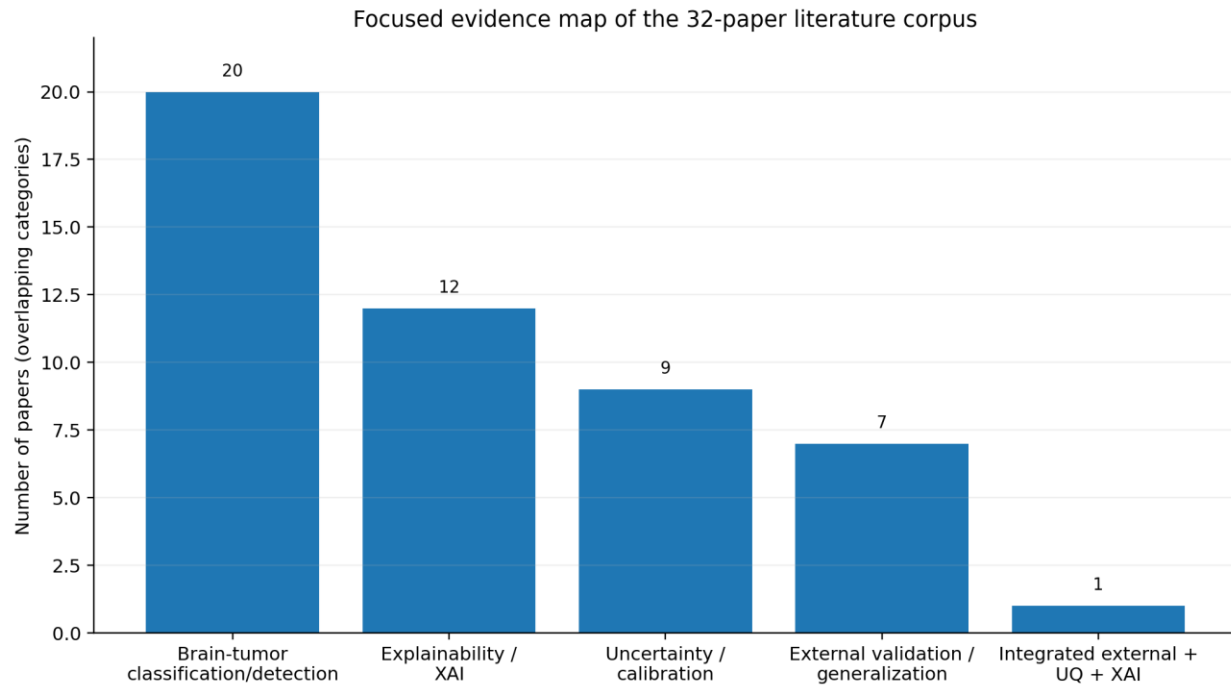


Figure 2. Narrative evidence map of the 32-paper focused literature corpus. Categories were coded by the authors of this manuscript, are non-mutually exclusive, and are intended as a descriptive evidence map rather than a formal meta-analysis.

Methodology

This study is designed as a retrospective, multi-source development and external-validation investigation of MRI-based brain-tumor classification. The primary objective is not merely to maximize discrimination, but to test whether predictions remain accurate, calibrated, uncertainty-aware, and anatomically interpretable when evaluated on data that are independent by patient and source. The primary diagnostic task is a four-class classification problem comprising glioma, meningioma, pituitary tumor,

and no tumor. When a contributing dataset supports only a subset of these labels, analysis is restricted to the shared label space and reported separately. The protocol follows CLAIM-style reporting principles for medical imaging artificial intelligence and prespecifies all preprocessing, model selection, calibration, uncertainty, explainability, and external-validation analyses before final testing.

Data are assembled from peer-reviewed or publicly documented brain-MRI collections with patient or source identifiers wherever available. Development data may include contrast-enhanced T1-weighted brain-tumor datasets and multimodal glioma collections derived from established public repositories such as Figshare, TCIA, and BraTS-related resources (Bakas et al., 2017; Menze et al., 2015). A completely independent dataset or institution is reserved as the primary external test cohort and is never used for model fitting, hyperparameter tuning, early stopping, feature selection, threshold optimization, or calibration. Within the development pool, patients are partitioned into training, tuning, and calibration subsets. All slices or examinations from the same patient remain in one partition. Duplicate or near-duplicate images are identified using metadata and perceptual hashing where licensing permits. Images with missing labels, severe corruption, uncertain class mapping, or unverifiable duplication are excluded according to a documented flow diagram.

Preprocessing is standardized but deliberately conservative to avoid removing clinically useful acquisition variability. DICOM or NIfTI data are converted to a common orientation, resampled to a predefined spatial resolution for volumetric inputs, and intensity-normalized within the brain. Bias-field correction is applied when compatible with the dataset, and skull stripping is used only when validated across all sources. For two-dimensional legacy datasets, images are resized using aspect-preserving padding rather than anisotropic stretching. Training-only augmentation includes small rotations, translations, scaling, gamma variation, Gaussian noise, and limited elastic deformation. Horizontal flipping is avoided unless clinically justified because laterality may be meaningful. No augmentation is applied to validation, calibration, or test data. Preprocessing parameters are fitted only on development data and then locked.

The primary classifier is a deep ensemble containing five independently initialized networks trained with identical data but different random seeds and shuffled minibatch orders. A modern convolutional backbone such as ConvNeXt-Tiny or DenseNet121 is selected through development-set comparison because these architectures combine strong image representation with mature interpretability tooling. For volumetric datasets, the same framework is implemented with a 3D residual or DenseNet-style backbone; for slice-based collections, patient-level probabilities are obtained by averaging calibrated slice probabilities or using a prespecified attention-based aggregation layer. Cross-entropy with class weighting or focal loss is considered when class imbalance is substantial. Optimization uses AdamW, cosine learning-rate decay, early stopping on macro-averaged validation AUC, and weight decay. Hyperparameters are tuned only within the development data, after which architecture and training settings are frozen.

Ensemble prediction is obtained by averaging the class-probability vectors from the five component networks. The uncalibrated ensemble is then calibrated using scalar temperature scaling fitted exclusively on the held-out calibration subset. Multiclass temperature scaling minimizes negative log-likelihood without changing the predicted class ranking. Calibration is evaluated before and after scaling with expected calibration error, classwise calibration error, Brier score, negative log-likelihood, and reliability diagrams. Because calibration can vary under distribution shift, the same metrics are reported separately for the internal test cohort and the external cohort rather than assuming that a calibration function transfers perfectly across sites (Carneiro et al., 2020; Mehrtash et al., 2020; Penso et al., 2024).

Predictive uncertainty is quantified using ensemble predictive entropy and variation in class probabilities across ensemble members. High entropy indicates dispersed probability mass, whereas high inter-model variance suggests epistemic disagreement. Uncertainty quality is evaluated in three ways. First, uncertainty distributions are compared between correct and incorrect predictions using effect sizes and nonparametric tests. Second, risk-coverage curves measure error as progressively more uncertain cases are deferred. Third, out-of-distribution behavior is tested using scans from an unseen pathology, acquisition source, or deliberately perturbed protocol when ethically and scientifically appropriate. Area under the receiver operating characteristic curve for detecting errors or out-of-distribution samples is reported. The abstention threshold is chosen on development data to meet a prespecified target, such as retaining the largest possible coverage while keeping selective error below a clinically defined ceiling.

Explainability is evaluated with Grad-CAM as the primary spatial attribution method and SHAP as a complementary feature-attribution approach. Explanations are generated from the frozen model without altering predictions. When lesion masks are available, Grad-CAM maps are normalized, thresholded using a development-set rule, and compared with tumor regions using intersection-over-union, pointing-game accuracy, energy-within-mask, and center-of-mass distance. Explanation stability is tested by applying small, label-preserving perturbations and calculating structural similarity or rank correlation between attribution maps. Model randomization and label-randomization sanity checks are performed on a representative subset to

identify explanation methods that remain visually similar despite loss of learned signal, a known limitation of saliency techniques (Arun et al., 2021). Qualitative panels are included only after quantitative evaluation.

Diagnostic performance is reported using macro and classwise AUC, accuracy, balanced accuracy, sensitivity, specificity, precision, F1 score, and confusion matrices with 95% confidence intervals obtained by patient-level bootstrap resampling. For multiclass comparisons, one-versus-rest AUCs and macro averages are emphasized. Internal and external performance are presented side by side, with the absolute change in AUC, balanced accuracy, Brier score, and expected calibration error used to characterize domain-shift effects. Subgroup analyses are prespecified for sex, age bands, scanner field strength, source institution, and sequence availability when those metadata exist and sample sizes are sufficient. No subgroup conclusion is drawn from cells that are too small for stable estimation.

Ablation experiments isolate the value of each reliability component. The single-network baseline is compared with the deep ensemble; uncalibrated probabilities are compared with temperature-scaled probabilities; standard classification is compared with uncertainty-based selective prediction; and qualitative XAI is compared with quantitatively validated localization. Statistical comparison of paired predictions uses DeLong testing for AUC where appropriate, McNemar testing for paired classification errors, and bootstrap confidence intervals for calibration and selective-risk differences. Multiple comparisons are controlled using the Benjamini-Hochberg procedure for secondary analyses. The external test set is evaluated once after model locking to reduce adaptive overfitting.

Reproducibility measures include a fixed random-seed registry, versioned code, a machine-readable configuration file, documented software and hardware versions, and release of trained weights when dataset licenses permit. A model card records intended use, contraindicated use, training sources, known domain limitations, calibration strategy, and uncertainty threshold. Public datasets require no new participant consent, while any institutional MRI data require local ethics approval, de-identification, and a data-use agreement. The resulting design deliberately separates model development from final external assessment so that reliability claims are supported by independent evidence rather than repeated tuning on the target cohort.

Primary success is defined prospectively as preserved external discrimination together with acceptable calibration and a monotonic relationship between uncertainty and error. No single threshold is universal. Instead, operating points are reported across sensitivity, coverage, and referral trade-offs so institutions can evaluate consequences within their prevalence, workflow, and tolerance for missed disease.

Table 2. Planned data partition and analysis firewall.

Partition	Permitted use	Prohibited use
Training	Fit network weights and training-time augmentation	Calibration fitting; final threshold selection; external evaluation
Tuning / validation	Architecture comparison, early stopping, hyperparameter selection	Temperature fitting; final external performance claims
Calibration	Fit post-hoc temperature and prespecified referral threshold	Weight optimization; architecture selection
Internal test	Estimate in-source discrimination, calibration, and uncertainty	Any model or threshold tuning after inspection
External test	Single locked evaluation of transportability and reliability	All fitting, selection, calibration, or repeated adaptive testing

Table 3. Reliability domains, planned metrics, and clinical interpretation.

Domain	Primary metrics	Failure signal	Decision use
Discrimination	Macro/classwise AUC; balanced accuracy; sensitivity; specificity; F1	Rank-ordering or class-separation failure	Diagnostic performance
Calibration	ECE; classwise ECE; Brier score; NLL; reliability diagram	Over- or under-confident probabilities	Interpret confidence as probability
Uncertainty	Predictive entropy; inter-model variance; error-detection AUROC	Confident errors or weak error separation	Flag ambiguous cases
Selective prediction	Risk-coverage curve; retained-case error	Risk fails to fall as uncertain cases are deferred	Choose review/referral operating point
Explainability	IoU; pointing-game accuracy; energy-in-mask; stability	Attribution outside lesion or unstable explanations	Audit shortcut learning

Domain	Primary metrics	Failure signal	Decision use
External generalization	Change in AUC, balanced accuracy, Brier score, ECE	Large degradation under source shift	Assess transportability
OOD behavior	OOD-detection AUROC; entropy shift	Unfamiliar inputs receive low uncertainty	Detect unsupported inputs
Reproducibility	Locked preprocessing; seeds; config; model card	Untraceable or irreproducible pipeline	Audit and replicate study

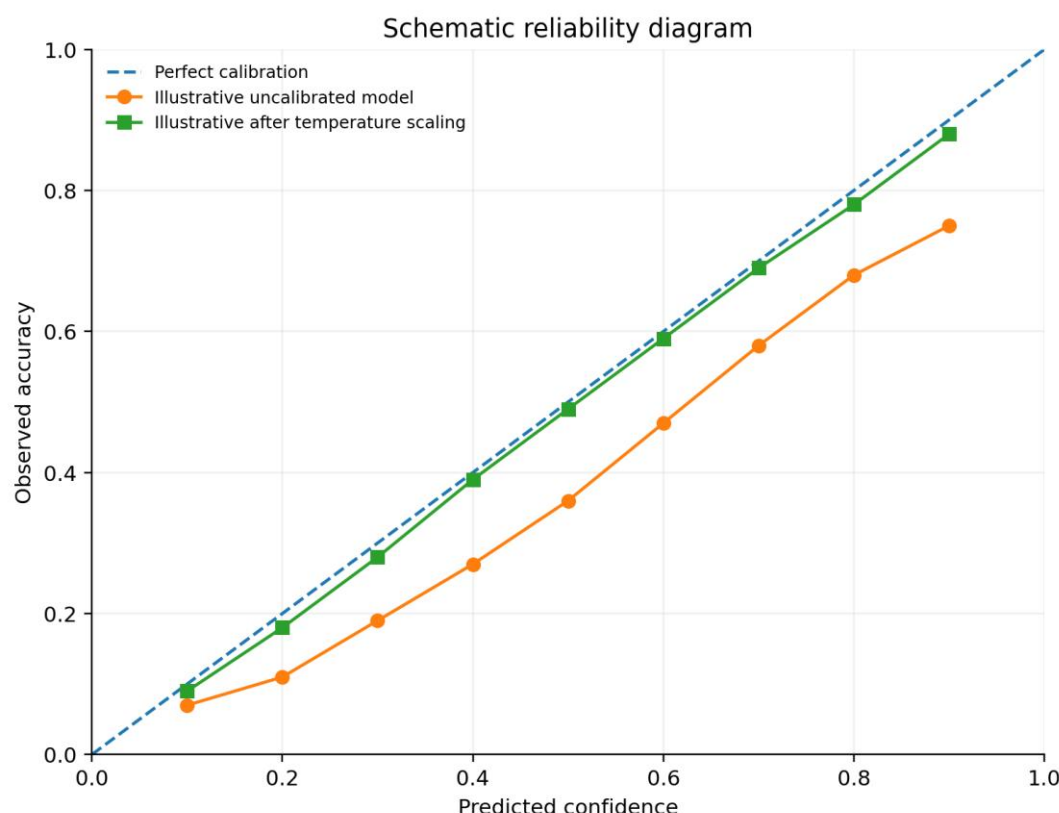


Figure 3. Schematic reliability diagram illustrating the purpose of post-hoc temperature scaling. Values are conceptual and are not empirical results from the proposed study.

Discussion

The proposed framework reframes brain-tumor MRI classification as a reliability problem rather than a leaderboard problem. This distinction is important because many published classifiers achieve excellent accuracy on curated internal datasets, yet clinical deployment exposes models to different scanners, sequence parameters, disease prevalence, referral patterns, motion, artifacts, and competing pathologies. The literature increasingly recognizes this gap. Wang et al. (2024) showed that externally validated brain-tumor studies exhibit meaningful heterogeneity related to dataset and hardware differences, while Park et al. (2024) demonstrated that external discrimination for glioblastoma versus solitary brain metastasis was lower than the very high internal accuracies often reported elsewhere. Those findings do not imply model failure; they show that realistic validation is more difficult and more informative.

The first implication is that external validation should be treated as a primary experiment, not a supplementary analysis. A random internal test split answers whether a model generalizes to additional samples drawn from a familiar source. It does not establish robustness to a new hospital, scanner, protocol, or patient mix. Subject-wise splitting is a minimum safeguard, especially when datasets contain multiple slices per patient. Badža and Barjaktarović (2020) reported markedly different performance under record-wise and subject-wise evaluation, illustrating the risk that image-level leakage can create optimistic estimates. A reliability-focused paper should therefore state the unit of independence, document duplicate detection, and reserve a truly untouched source for the final test.

The second implication concerns calibration. Classification confidence is often displayed as though it were a probability of correctness, but raw softmax values do not guarantee that interpretation. Overconfidence is particularly problematic in clinical

decision support because an erroneous prediction with 0.99 confidence can discourage appropriate review. Temperature scaling is attractive because it is simple, architecture-agnostic, and preserves discrimination while correcting probability sharpness. Prior medical-imaging studies have demonstrated the usefulness of calibration metrics and post-hoc scaling (Carneiro et al., 2020; Mehrtash et al., 2020; Penso et al., 2024; Hasan, M. N et al., 2025). Reporting expected calibration error alone is insufficient, however. Brier score, negative log-likelihood, reliability plots, and classwise behavior should be interpreted together because each captures a different aspect of probabilistic performance.

A third implication is that uncertainty should be actionable. An uncertainty score has limited clinical value if it is merely correlated with itself across repeated runs. The relevant question is whether uncertain cases are more likely to be wrong, unfamiliar, or difficult. Deep ensembles are well suited to this purpose because independently trained members provide a practical approximation of epistemic uncertainty while usually improving predictive performance. Park et al. (2024) showed that ensemble entropy separated low- and high-uncertainty groups with substantially different diagnostic performance and identified many out-of-distribution cases as uncertain. Nair et al. (2020) similarly showed that uncertainty-based filtering could improve performance among retained predictions. These findings support selective prediction, in which the model is permitted to defer rather than fabricate certainty.

Selective prediction also changes how performance should be communicated. A model with 95% global accuracy may be less useful than a model with 93% accuracy that can identify a subset of cases with very low error and safely refer the remainder. Risk-coverage curves make this trade-off visible. At 100% coverage, the system classifies every case. As the most uncertain cases are referred, coverage decreases and selective risk should fall. The curve therefore offers a direct view of how uncertainty could support workload allocation. Importantly, referral thresholds should not be declared universally optimal from a single dataset. They should be selected according to clinical consequences, disease prevalence, staffing, and the cost of false reassurance.

Explainability requires similar discipline. The recent brain-tumor literature contains many visually appealing Grad-CAM, LIME, and SHAP figures (Akgündoğdu & Çelikbaş, 2025; Iftikhar et al., 2025; Musthafa et al., 2024; Nazir et al., 2024), and early-2026 work continues to emphasize interpretable four-class classification (Adamu et al., 2026; Pant et al., 2026; Srinivas & Parvathi, 2026). These methods can help developers identify shortcut learning and can provide clinicians with a visual audit trail. Yet they should not be presented as proof that the prediction is medically valid. Arun et al. (2021) found important weaknesses in commonly used saliency maps, and Ghassemi et al. (2021) cautioned that explanations can offer false reassurance at the individual-patient level. Quantitative localization against lesion masks therefore provides a stronger standard than visual inspection alone.

Combining uncertainty and explanation may be especially useful because the two signals answer different questions. Uncertainty asks how stable or confident the predictive distribution is; explanation asks where the model appears to obtain evidence. A low-uncertainty prediction with poor lesion localization may indicate shortcut learning. Conversely, a high-uncertainty prediction with good lesion localization may reflect genuine diagnostic ambiguity rather than irrelevant attention. Park et al. (2024) reported particularly strong discrimination in a subgroup combining low uncertainty with high localization accuracy, suggesting that joint reliability criteria can be more informative than either component alone. Future evaluation should therefore examine a reliability matrix rather than a single score.

The proposed ensemble-and-calibration design also provides a natural way to study domain shift. If external AUC decreases while calibration deteriorates and entropy rises, the three changes provide convergent evidence that the model is encountering unfamiliar data. If discrimination remains stable but calibration worsens, recalibration may be sufficient. If entropy remains low despite rising error, the uncertainty method is failing in the most dangerous way: confident mistakes. This decomposition helps distinguish adaptation problems from ranking problems. Karimi and Gholipour (2023) and Küstner et al. (2024) similarly illustrate that calibration and uncertainty behavior can reveal information not captured by conventional accuracy metrics.

Architectural novelty is intentionally not the central contribution. Reviews by Bouhafra and El Bahi (2025) and Dorfner et al. (2025) show that CNNs, transfer learning, transformers, attention modules, and hybrid systems can all achieve strong performance. The greater unmet need is credible evidence that a model behaves safely when assumptions are violated. A deep ensemble built from a mature backbone can therefore be scientifically stronger than a more complex network if the ensemble is externally validated, calibrated, uncertainty-aware, and auditable. This emphasis also improves comparability because reliability procedures can be applied to multiple backbones without changing the conceptual study design.

Clinical translation would still require prospective evaluation. Retrospective external validation can show that the model survives known domain differences, but deployment introduces workflow effects, reader interaction, prevalence shifts, and behavior changes that retrospective datasets cannot capture. A sensible next step would be silent prospective testing in which predictions, uncertainties, and explanations are generated but hidden from clinicians. That phase could measure calibration

drift, referral volume, failure modes, and latency without influencing care. Only after those checks should a reader study evaluate whether the system improves diagnostic accuracy, speed, or consistency.

Overall, the contribution of this framework is not the claim that uncertainty or XAI makes a model inherently safe. Instead, it creates multiple opportunities for the model to reveal its limitations. External validation exposes transport failures, calibration identifies misleading confidence, uncertainty supports abstention, out-of-distribution tests probe unfamiliar inputs, and explanation localization detects attention to irrelevant regions. When these signals agree, confidence in deployment increases; when they conflict, the conflict becomes a reason for investigation rather than an inconvenient result. That is the practical meaning of reliability for high-stakes MRI classification. This rigor directly strengthens clinical credibility.

Conclusion

Reliable brain-tumor MRI classification requires evidence that extends beyond internal accuracy. A clinically credible system should discriminate among diagnostic classes, produce calibrated probabilities, identify cases in which its predictions are uncertain, generate explanations that correspond to relevant tumor anatomy, and maintain acceptable behavior when evaluated on independent data. The framework presented in this paper integrates these requirements through patient-level partitioning, deep ensembles, temperature scaling, entropy-based selective prediction, quantitative Grad-CAM and SHAP assessment, and institution- or dataset-level external validation. Its central advantage is that reliability is measured as a set of complementary properties rather than reduced to a single performance statistic. This approach is consistent with emerging evidence showing that domain shift, overconfidence, uncertainty, and explanation instability can each reveal failure modes that conventional accuracy may conceal. The framework also promotes reproducibility by locking preprocessing and model-selection procedures before external testing and by reporting calibration, risk-coverage, and localization metrics alongside AUC and F1 score. Applied rigorously, this design can help distinguish models that are merely accurate on familiar data from systems that fail more transparently and defer appropriately when confronted with ambiguity. Such behavior is essential for responsible clinical decision support in neuro-oncology. It prioritizes dependable behavior over inflated benchmark performance.

Limitations and Future Directions

Several limitations should guide interpretation of this framework. First, public brain-tumor MRI datasets often differ in sequence composition, preprocessing history, labeling standards, patient metadata, and image provenance. Some collections contain isolated two-dimensional slices rather than complete clinical examinations, which limits patient-level aggregation and may introduce hidden duplication. Second, temperature scaling corrects average probability calibration but may not remain optimal after severe distribution shift. Site-specific recalibration could be necessary, although recalibration requires representative local data. Third, deep-ensemble entropy is a useful empirical uncertainty measure but does not capture every source of uncertainty, including label ambiguity, unseen pathology, or systematic bias shared by all ensemble members. Fourth, Grad-CAM and SHAP remain post-hoc explanations; quantitative overlap with a lesion mask improves evaluation but cannot establish causal reasoning or guarantee clinical correctness. Fifth, external validation on one additional dataset does not prove universal generalization.

Future work should extend the framework through prospective multi-institutional testing, richer multimodal MRI, pathology-confirmed labels, and explicit scanner- and site-shift experiments. Class-conditional or conformal prediction could provide coverage-oriented uncertainty guarantees, while test-time adaptation and harmonization methods may improve transportability. Explanation evaluation should include neuroradiologist review, perturbation testing, and comparisons with inherently interpretable or concept-based models. Federated learning could enable broader validation without centralizing sensitive imaging data. Finally, silent prospective deployment should measure calibration drift, abstention frequency, subgroup performance, and operational burden before any model is allowed to influence clinical decisions. These extensions would test whether reliability signals remain stable across time, populations, vendors, institutions, evolving protocols, and changing disease prevalence.

Funding: Not stated in the source manuscript.

Conflicts of Interest: Not stated in the source manuscript.

Research integrity note: This manuscript is a literature-grounded methodological study and external-validation protocol. No patient-level experiment was supplied; therefore, no empirical AUCs, confidence intervals, calibration statistics, or clinical outcomes are fabricated. The proposed analyses are written so that measured results can be inserted after execution on approved MRI datasets.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

- [1] Adamu, M. J., Qiang, L., Nyatega, C. O., Fahad, M., Younis, A., Jabire, A. H., Zakariyya, R. S., & Kawuwa, H. B. (2026). Explainable hybrid deep learning framework integrating MobileNetV2, EfficientNetV2B0, and KNN for MRI-based brain tumor classification. *Cellular and Molecular Neurobiology*, 46, 40. <https://doi.org/10.1007/s10571-025-01660-z>
- [2] Akgündoğdu, A., & Çelikbaş, Ş. (2025). Explainable deep learning framework for brain tumor detection: Integrating LIME, Grad-CAM, and SHAP for enhanced accuracy. *Medical Engineering & Physics*, 144, 104405. <https://doi.org/10.1016/j.medengphy.2025.104405>
- [3] Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., Hoebel, K., Gupta, S., Patel, J., Gidwani, M., Adebayo, J., Li, M. D., & Kalpathy-Cramer, J. (2021). Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence*, 3(6), e200267. <https://doi.org/10.1148/ryai.2021200267>
- [4] Asiri, A. A., Shaf, A., Ali, T., Pasha, M. A., Khan, A., Irfan, M., Alqahtani, S., Alghamdi, A., Alghamdi, A. H., Alshamrani, A. F. A., Alelyani, M., & Alamri, S. (2024). Advancing brain tumor detection: Harnessing the Swin Transformer's power for accurate classification and performance analysis. *PeerJ Computer Science*, 10, e1867. <https://doi.org/10.7717/peerj-cs.1867>
- [5] Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., Freymann, J. B., Farahani, K., & Davatzikos, C. (2017). Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific Data*, 4, 170117. <https://doi.org/10.1038/sdata.2017.117>
- [6] Badža, M. M., & Barjaktarović, M. Č. (2020). Classification of brain tumors from MRI images using a convolutional neural network. *Applied Sciences*, 10(6), 1999. <https://doi.org/10.3390/app10061999>
- [7] Bouhafra, S., & El Bahi, H. (2025). Deep learning approaches for brain tumor detection and classification using MRI images (2020 to 2024): A systematic review. *Journal of Imaging Informatics in Medicine*, 38(3), 1403–1433. <https://doi.org/10.1007/s10278-024-01283-8>
- [8] Carneiro, G., Pu, L. Z. C. T., Singh, R., & Burt, A. (2020). Deep learning uncertainty and confidence calibration for the five-class polyp classification from colonoscopy. *Medical Image Analysis*, 62, 101653. <https://doi.org/10.1016/j.media.2020.101653>
- [9] Cheng, J., Huang, W., Cao, S., Yang, R., Yang, W., Yun, Z., Wang, Z., & Feng, Q. (2015). Enhanced performance of brain tumor classification via tumor region augmentation and partition. *PLOS ONE*, 10(10), e0140381. <https://doi.org/10.1371/journal.pone.0140381>
- [10] Deepak, S., & Ameer, P. M. (2019). Brain tumor classification using deep CNN features via transfer learning. *Computers in Biology and Medicine*, 111, 103345. <https://doi.org/10.1016/j.combiomed.2019.103345>
- [11] Dorfner, F. J., Patel, J. B., Kalpathy-Cramer, J., Gerstner, E. R., & Bridge, C. P. (2025). A review of deep learning for brain tumor analysis in MRI. *npj Precision Oncology*, 9(1), 2. <https://doi.org/10.1038/s41698-024-00789-2>
- [12] Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11), e745–e750. [https://doi.org/10.1016/S2589-7500\(21\)00208-9](https://doi.org/10.1016/S2589-7500(21)00208-9)
- [13] Iftikhar, S., Anjum, N., Siddiqui, A. B., Rehman, M. U., & Ramzan, N. (2025). Explainable CNN for brain tumor detection and classification through XAI based key features identification. *Brain Informatics*, 12(1), 10. <https://doi.org/10.1186/s40708-025-00257-y>
- [14] Karimi, D., & Gholipour, A. (2023). Improving calibration and out-of-distribution detection in deep models for medical image segmentation. *IEEE Transactions on Artificial Intelligence*, 4(2), 383–397. <https://doi.org/10.1109/TAI.2022.3159510>
- [15] Kumar, S., Choudhary, S., Jain, A., Singh, K., Ahmadian, A., & Bajuri, M. Y. (2023). Brain tumor classification using deep neural network and transfer learning. *Brain Topography*, 36(3), 305–318. <https://doi.org/10.1007/s10548-023-00953-0>
- [16] Kurz, A., Hauser, K., Mehrtens, H. A., Krieghoff-Henning, E., Hekler, A., Kathner, J. N., Fröhling, S., von Kalle, C., & Brinker, T. J. (2022). Uncertainty estimation in medical image classification: Systematic review. *JMIR Medical Informatics*, 10(8), e36427. <https://doi.org/10.2196/36427>

- [17] Küstner, T., Hammernik, K., Rueckert, D., Hepp, T., & Gatidis, S. (2024). Predictive uncertainty in deep learning-based MR image reconstruction using deep ensembles: Evaluation on the fastMRI data set. *Magnetic Resonance in Medicine*, 92(1), 289–302. <https://doi.org/10.1002/mrm.30030>
- [18] Lakshmi, K., Amaran, S., Subbulakshmi, G., Padmini, S., Joshi, G. P., & Cho, W. (2025). Explainable artificial intelligence with UNet based segmentation and Bayesian machine learning for classification of brain tumors using MRI images. *Scientific Reports*, 15, 690. <https://doi.org/10.1038/s41598-024-84692-7>
- [19] Mehrtash, A., Wells, W. M., Tempny, C. M., Abolmaesumi, P., & Kapur, T. (2020). Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Transactions on Medical Imaging*, 39(12), 3868–3878. <https://doi.org/10.1109/TMI.2020.3006437>
- [20] Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M.-A., Arbel, T., Avants, B. B., Ayache, N., Buendia, P., Collins, D. L., Cordier, N., ... Van Leemput, K. (2015). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10), 1993–2024. <https://doi.org/10.1109/TMI.2014.2377694>
- [21] Munroe, L., da Silva, M., Heidari, F., Grigorescu, I., Dahan, S., Robinson, E. C., Deprez, M., & So, P.-W. (2024). Applications of interpretable deep learning in neuroimaging: A comprehensive review. *Imaging Neuroscience*, 2, imag_a_00214. https://doi.org/10.1162/imag_a_00214
- [22] Musthafa M, M., Mahesh, T. R., Vinoth Kumar, V., & Guluwadi, S. (2024). Enhancing brain tumor detection in MRI images through explainable AI using Grad-CAM with ResNet50. *BMC Medical Imaging*, 24, 107. <https://doi.org/10.1186/s12880-024-01292-7>
- [23] Nair, T., Precup, D., Arnold, D. L., & Arbel, T. (2020). Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Medical Image Analysis*, 59, 101557. <https://doi.org/10.1016/j.media.2019.101557>
- [24] Nazir, M. I., Akter, A., Wadud, M. A. H., & Uddin, M. A. (2024). Utilizing customized CNN for brain tumor prediction with explainable AI. *Heliyon*, 10(20), e38997. <https://doi.org/10.1016/j.heliyon.2024.e38997>
- [25] Pant, K., Pramanik, P. K. D., & Zhao, Z. (2026). A robust ConvNeXt-based framework for efficient, generalizable, and explainable brain tumor classification on MRI. *Bioengineering*, 13(2), 157. <https://doi.org/10.3390/bioengineering13020157>
- [26] Park, Y. W., Eom, S., Kim, S., Lim, S., Park, J. E., Kim, H. S., You, S. C., Ahn, S. S., & Lee, S.-K. (2024). Differentiation of glioblastoma from solitary brain metastasis using deep ensembles: Empirical estimation of uncertainty for clinical reliability. *Computer Methods and Programs in Biomedicine*, 254, 108288. <https://doi.org/10.1016/j.cmpb.2024.108288>
- [27] Penso, C., Frenkel, L., & Goldberger, J. (2024). Confidence calibration of a medical imaging classification system that is robust to label noise. *IEEE Transactions on Medical Imaging*, 43(6), 2050–2060. <https://doi.org/10.1109/TMI.2024.3353762>
- [28] Sajjad, M., Khan, S., Muhammad, K., Wu, W., Ullah, A., & Baik, S. W. (2019). Multi-grade brain tumor classification using deep CNN with extensive data augmentation. *Journal of Computational Science*, 30, 174–182. <https://doi.org/10.1016/j.jocs.2018.12.003>
- [29] Sekhar, A., Biswas, S., Hazra, R., Sunaniya, A. K., Mukherjee, A., & Yang, L. (2022). Brain tumor classification using fine-tuned GoogLeNet features and machine learning algorithms: IoMT enabled CAD system. *IEEE Journal of Biomedical and Health Informatics*, 26(3), 983–991. <https://doi.org/10.1109/JBHI.2021.3100758>
- [30] Srinivas, V. R., & Parvathi, R. (2026). Explainable AI-driven MRI-based brain tumor classification: A novel deep learning approach. *Frontiers in Artificial Intelligence*, 8, 1700214. <https://doi.org/10.3389/frai.2025.1700214>
- [31] Swati, Z. N. K., Zhao, Q., Kabir, M., Ali, F., Ali, Z., Ahmed, S., & Lu, J. (2019). Brain tumor classification for MR images using transfer learning and fine-tuning. *Computerized Medical Imaging and Graphics*, 75, 34–46. <https://doi.org/10.1016/j.compmedimag.2019.05.001>
- [32] Wang, T.-W., Shiao, Y.-C., Hong, J.-S., Lee, W.-K., Hsu, M.-S., Cheng, H.-M., Yang, H.-C., Lee, C.-C., Pan, H.-C., You, W. C., Lirng, J.-F., Guo, W.-Y., & Wu, Y.-T. (2024). Artificial intelligence detection and segmentation models: A systematic review and meta-analysis of brain tumors in magnetic resonance imaging. *Mayo Clinic Proceedings: Digital Health*, 2(1), 75–91. <https://doi.org/10.1016/j.mcpdig.2024.01.002>
- [33] Hasan, M. N., Miah, M. S., Ghose, P., Jannat, T., Bhuyain, M. M. H., Chowdhury, M. S. A., Islam, M. S., Talukder, M. H., & Harun-Ar-Rashid, M. (2025). A deep learning approach for brain tumor diagnosis: Combining an 8-layer CNN with rigorous K-fold validation. *Mathematical Modelling of Engineering Problems*, 12(12), 4387–4396. <https://doi.org/10.18280/mmep.121227>