BioMedInformatics

MDPI

Review

Toward Clinically Dependable AI for Brain Tumors: A Unified Diagnostic–Prognostic Framework and Triadic Evaluation Model

Mohammed A. Atiea [1], Mona Gafar [2,*], Shahenda Sarhan [3,4] and Abdullah M. Shaheen [5]

[1] Computer Science Department, Faculty of Computers and Information, Suez University, Suez P.O. Box 43221, Egypt; mohammed.a.atiea@fci.suezuni.edu.eg
[2] Department of Computer Engineering and Information, College of Engineering, Wadi Ad Dwaser, Prince Sattam Bin Abdulaziz University, Al-Kharj 16278, Saudi Arabia
[3] Computer Science Department, Faculty of Computers and Information, Mansoura University, Mansoura P.O. Box 35516, Egypt; sh_sarhan@mans.edu.eg
[4] School of Computer Science and Technologies, VIZJA University, 01-043 Warsaw, Poland
[5] Department of Electrical Engineering, Faculty of Engineering, Suez University, Suez P.O. Box 43221, Egypt; abdullah.mohamed.eng19@suezuni.edu.eg
* Correspondence: m.gafar@psau.edu.sa

Abstract

Artificial intelligence (AI) has shown promising performance in brain tumor diagnosis and prognosis; however, most reported advances remain difficult to translate into clinical practice due to limited interpretability, inconsistent evaluation protocols, and weak generalization across datasets and institutions. In this work, we present a critical synthesis of recent brain tumor AI studies (2020–2025) guided by two novel conceptual tools: a unified diagnostic-prognostic framework and a triadic evaluation model emphasizing interpretability, computational efficiency, and generalizability as core dimensions of clinical readiness. Following PRISMA 2020 guidelines, we screened and analyzed over 100 peer-reviewed studies. A structured analysis of reported metrics reveals systematic trends and trade-offs—for instance, between model accuracy and inference latency—rather than providing a direct performance benchmark. This synthesis exposes critical gaps in current evaluation practices, particularly the under-reporting of interpretability validation, deployment-level efficiency, and external generalization. By integrating conceptual structuring with evidence-driven analysis, this work provides a framework for more clinically grounded development and evaluation of AI systems in neuro-oncology.

Keywords: brain tumor; clinical deployment; deep learning; explainable AI; model generalizability; vision transformer

Check for updates

Academic Editors: Kwang-Sig Lee and Hyuntae Park

Received: 21 December 2025
Revised: 19 January 2026
Accepted: 21 January 2026
Published: 27 January 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Despite rapid methodological advances, the clinical deployment of AI for brain tumor diagnosis and prognosis remains limited. While numerous studies report high accuracy for tasks such as tumor segmentation, grading, or molecular marker prediction, these results are often obtained under narrowly defined experimental conditions and evaluated primarily using task-specific performance metrics. Increasingly, it has been recognized that such evaluations alone are insufficient to establish clinical dependability, as they frequently overlook interpretability, computational efficiency, and robustness to dataset and domain shifts—factors that are critical for real-world adoption in neuro-oncology settings. Recent guidelines and reviews in medical AI emphasize the need for systematic evidence synthesis,

transparent evaluation practices, and clinically grounded assessment criteria rather than isolated performance reporting. Motivated by these observations, this study presents a systematic review of brain tumor AI research, combining a structured synthesis of reported quantitative evidence with a conceptual framework aimed at clarifying how diagnostic and prognostic AI components can be evaluated more holistically for clinical readiness.

A brain tumor, characterized by an abnormal proliferation of cells within or adjacent to the brain, constitutes one of the most significant diagnostic and therapeutic challenges in contemporary medicine. These neoplasms can be either benign (not cancerous) or malignant (cancerous). The latter type grows aggressively into important neural structures, which can cause serious neurological problems or even death [1]. According to epidemiological data from the CBTRUS, malignant tumors account for nearly 30% of all primary brain and CNS tumors, while non-malignant forms constitute the remaining 70% [2]. Despite their classification, both types can be life-threatening depending on size, location, and growth rate, necessitating timely and precise diagnosis to guide surgical, radiological, or pharmacological interventions [3,4].

MRI has emerged as the gold-standard modality for brain tumor assessment due to its superior soft-tissue contrast, multi-parametric capabilities (e.g., T1, T2, FLAIR, DWI), and absence of ionizing radiation [5]. However, manual interpretation of MRI scans is time-consuming, subjective, and prone to inter-observer variability—challenges exacerbated by the irregular morphology, heterogeneous texture, and variable location of tumors [6]. In response, DL has catalyzed a paradigm shift in automated brain tumor analysis, enabling end-to-end systems for classification, detection, and segmentation [1,7].

Current literature predominantly evaluates DL models through the narrow prism of accuracy or Dice score on curated datasets like Figshare or BraTS [1,7]. While such metrics demonstrate algorithmic prowess, they frequently conceal critical flaws: poor generalization to unknown institutions, opaque decision-making that undermines clinician trust, and prohibitively high computational demands that impede deployment in resource-constrained settings [1,7]. Moreover, the field is still fragmented—CNNs are praised for local feature extraction, ViTs for global context modeling, and XAI methods for post hoc justification—but these components are rarely co-optimized within a unified clinical workflow.

Unlike earlier surveys that primarily emphasize algorithmic performance for isolated diagnostic tasks, this work provides a systematic review of brain tumor AI research with a focus on clinical dependability. The contributions are threefold. First, we introduce a unified diagnostic–prognostic conceptual framework that organizes existing methods across clinically relevant stages, spanning imaging-based diagnosis and patient-level outcome modeling. This framework is intended as a structuring lens for synthesizing prior work rather than as a new algorithmic architecture. Second, following PRISMA 2020 guidelines, we conduct a systematic review of recent studies and perform a structured quantitative synthesis of reported performance, efficiency, and interpretability indicators where sufficient information is available. This synthesis is based on metrics as reported in the original publications and aims to identify cross-study trends, trade-offs, and recurring evaluation gaps, rather than to benchmark models under a unified experimental pipeline. Third, building on both the conceptual analysis and the synthesized evidence, we propose a triadic evaluation model that highlights interpretability, computational efficiency, and generalizability as interconnected dimensions of clinical readiness, and we discuss how these dimensions are currently addressed—and often under-reported—in the brain tumor AI literature.

The remainder of this paper is organized as follows. Section 2 introduces the unified diagnostic–prognostic conceptual framework used to structure the review. Section 3 de-

scribes the proposed triadic evaluation model and its constituent dimensions. Section 4 outlines the systematic review methodology, including the protocol for literature search, screening, and the structured quantitative synthesis of reported metrics. Section 5 synthesizes and critically analyzes prior technical approaches through the lens of the triadic framework. Section 6 examines the data crisis in brain tumor AI, focusing on dataset representativeness, annotation challenges, and the illustrative synthesis of reported trade-offs. Section 7 discusses a phased roadmap toward clinical adoption, addressing standardized benchmarks, regulatory pathways, and enabling technologies such as federated learning. Section 8 acknowledges the limitations of this review, and Section 9 concludes the paper by summarizing key insights and future directions.

2. A Unified Diagnostic–Prognostic Conceptual Framework

Recent research in AI-driven brain tumor analysis has predominantly addressed clinical problems through task-specific modeling paradigms, such as tumor classification, lesion detection, and subregion segmentation. Classification models are commonly used to distinguish tumor subtypes or molecular profiles, detection frameworks localize suspicious regions within imaging volumes, and segmentation networks delineate tumor subregions such as ET, TC, and WT [1,7,8]. While each of these tasks has benefited from substantial algorithmic progress, particularly with the emergence of vision transformers and hybrid architectures, prior studies typically evaluate them in isolation, without explicitly situating them within an end-to-end clinical workflow that connects image-level outputs to patient-level decision-making.

To facilitate a more coherent synthesis of existing work, this review introduces a unified diagnostic–prognostic conceptual framework that organizes previously reported AI approaches according to their role in the clinical care pathway. Rather than proposing a new algorithmic architecture, the framework serves as an analytical lens for categorizing and comparing methods across four interrelated stages, as illustrated in Figure 1. These stages collectively span the continuum from early image-level abnormality detection to downstream prognostic modeling. By structuring the literature in this manner, the framework enables a clearer discussion of how diagnostic and prognostic components are currently developed, evaluated, and combined in the brain tumor AI literature, and where methodological or evaluative disconnects persist.

The first stage, tumor detection, focuses on the rapid identification of suspicious regions within 2D or 3D MRI volumes. Lightweight DL models such as MobileNetV3 [9] or object detection architectures like YOLOv7 [10] are often employed to ensure real-time responsiveness. This stage prioritizes high sensitivity, as missed lesions can have severe clinical implications, particularly in malignant or fast-growing tumors.

The second stage, multi-class classification, assigns the detected lesions to specific histopathological categories commonly including glioma, meningioma, pituitary adenoma, or non-tumor (Figure 2) [11]. Recent studies have demonstrated that combining CNNs and ViTs through ensemble feature extraction [12–15] significantly enhances diagnostic accuracy, achieving cross-dataset performance above 99% on curated benchmarks such as Figshare [16].

The third stage, semantic segmentation, provides pixel-wise delineation of tumor subregions, including enhancing tumor, tumor core, and whole tumor components. This step is crucial for surgical planning and radiotherapy targeting, where precise spatial boundaries directly influence treatment outcomes. Advanced architectures, such as U-Net variants enhanced with ViT encoders [17] or anisotropic diffusion preprocessing [18], have achieved Dice coefficients exceeding 0.92 on the BraTS dataset [19], underscoring the maturity of segmentation algorithms in controlled environments.

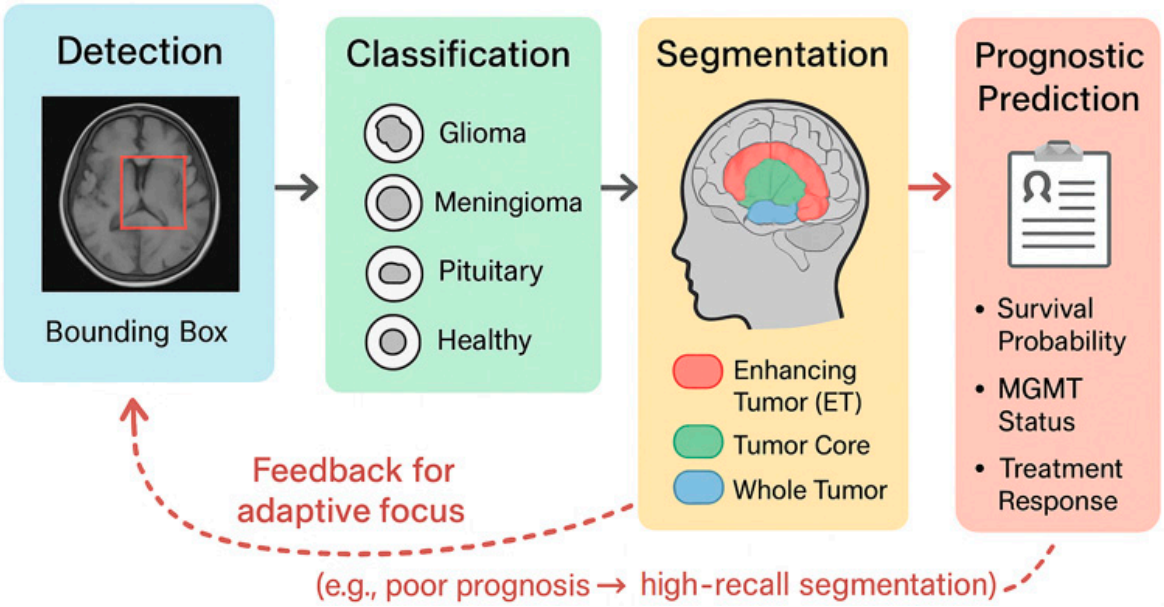


Figure 1. Unified diagnostic–prognostic conceptual framework for brain tumor AI. Conceptual illustration of a clinically oriented AI workflow encompassing tumor classification, segmentation (ET, tumor core, whole tumor), and prognostic prediction (e.g., survival, MGMT status). A bounding-box localization links detection to downstream analysis, while a feedback mechanism indicates how diagnostic or prognostic outputs may guide adaptive focus in subsequent processing.

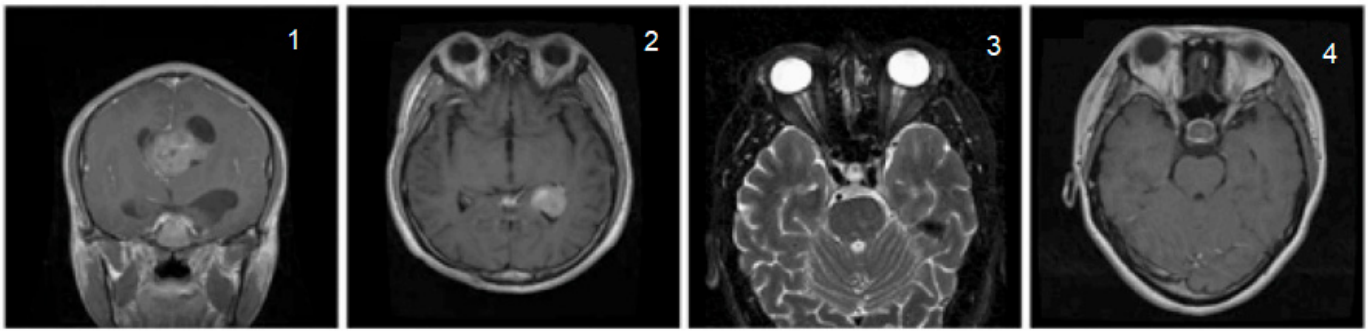


Figure 2. Representative axial MRI slices illustrating the four diagnostic classes commonly used in public datasets:(**1**) Glioma, (**2**) Meningioma, (**3**) No Tumor, and (**4**) Pituitary.

The final stage, prognostic prediction, remains the most underexplored yet clinically essential component. It aims to connect imaging-derived phenotypes with patient-level outcomes, such as survival duration, therapy response, or molecular biomarkers like IDH mutation and MGMT methylation status. Only a limited number of studies have addressed this integrative link. For instance, Zeineldin et al. [20] correlated segmented glioma volumes with progression-free survival, while foundation models such as Brain IAC [21] have recently demonstrated the ability to perform zero-shot prediction of genomic subtypes directly from MRI data.

Taken together, these four stages provide a clinically grounded conceptual organization of how AI methods for brain tumor analysis are addressed in the existing literature, spanning early image-level processing through downstream outcome modeling. When used as a structuring lens for this review, the framework highlights a recurring pattern across prior studies: the majority of published approaches concentrate on intermediate diagnostic tasks, such as segmentation or subtype classification, while comparatively fewer works explicitly incorporate patient-level prognostic modeling or longitudinal outcomes. This observation, which is consistently reported across recent surveys and task-specific

studies, underscores an important gap between algorithmic development and clinically actionable endpoints.

In the remainder of this review, we use the diagnostic–prognostic framework to systematically organize and discuss reported technical approaches, with particular attention to how interpretability, computational efficiency, and generalizability are addressed—or omitted—across different stages. Detailed methodological aspects of the systematic literature review, including search strategy, screening workflow, and inclusion criteria, are described separately in the Methods section and Appendix A.

3. Triadic Evaluation Framework

The clinical applicability of AI systems for brain tumor analysis depends on factors that extend beyond task-specific predictive performance. While accuracy-oriented metrics remain essential, prior studies in medical AI have increasingly emphasized that interpretability, computational efficiency, and generalizability are critical complementary dimensions for safe and effective deployment in real-world clinical environments. However, these aspects are often reported inconsistently or treated as secondary considerations across the literature, making cross-study comparison and evidence synthesis challenging.

To translate these conceptual pillars into actionable evaluation criteria, Table 1 proposes specific metrics and reporting standards for each dimension. This checklist is derived from recurring gaps identified in our literature synthesis and is intended to guide future studies toward more comprehensive and clinically meaningful evaluation.

Table 1. Proposed Evaluation Metrics for the Triadic Clinical Readiness Framework.

Pillar	Core Objective	Recommended Quantitative Metrics	Essential Reporting Standards	Common Pitfalls (from Our Synthesis)
Interpretability	Foster clinical trust and justify decisions.	• Saliency map overlap with expert ROI (IoU) • Consistency of explanations (e.g., SHAP value variance) • Clinician-AI diagnostic agreement (Cohen's κ)	• Specify XAI method and parameters (e.g., Grad-CAM layer, LIME samples). • Report validation against expert annotations or clinical ground truth.	Post hoc visualizations used without validation ("trust theater"); lack of clinical correlation.
Efficiency	Enable deployment in real-world clinical settings.	• Inference latency (ms per volume/slice) • Peak memory footprint (GB) • Parameter count (M)/FLOPs	• Mandatory: Hardware specification (GPU/CPU, model). • Report end-to-end pipeline time, not just forward pass.	Latency reported on high-end GPUs not available in clinics; memory use ignored for 3D models.
Generalizability	Ensure robustness across institutions and populations.	• Generalization gap (Δ = Internal Acc. − External Acc.) • Performance on underrepresented subgroups (worst-group accuracy) • Cross-domain Dice/AUC	• Detail external validation cohort characteristics (scanner, demographics). • Report performance per tumor subtype and institution.	Validation only on homogeneous public benchmarks (Figshare/BraTS); missing external testing.

In this review, we introduce a triadic evaluation framework as an analytical construct to support the systematic assessment of how existing brain tumor AI studies address these three dimensions. Importantly, the framework is not intended as a prescriptive scoring system or a standardized benchmark. Rather, it serves as a conceptual tool for organizing reported evaluation practices, highlighting recurring trade-offs, and identifying gaps in current methodologies. Each dimension of the triad—interpretability, efficiency, and generalizability—is examined based on metrics, validation strategies, and reporting

practices as described in the original studies, without re-implementing or re-evaluating the underlying models.

By applying this framework across the reviewed literature, we aim to provide a structured discussion of how different technical approaches prioritize or neglect specific aspects of clinical dependability, and to clarify the implications of these choices for translation into neuro-oncology workflows. The following subsections detail each dimension of the triadic framework and summarize representative evaluation strategies reported in prior work.

Contemporary literature on AI-driven brain tumor diagnosis often evaluates models through a narrow performance lens, primarily accuracy, Dice score, or F1-measure, on curated benchmark datasets such as Figshare or BraTS [1,7]. While these metrics reflect algorithmic competence under controlled conditions, they frequently mask critical shortcomings that impede real-world clinical deployment. A model may achieve 99% accuracy on Figshare yet fail catastrophically on images from a different scanner, hospital, or demographic cohort due to domain shift [7]. Similarly, a ViT may offer state-of-the-art segmentation but require hours of inference time on standard hardware, rendering it impractical in emergency settings [18,22].

To address this gap, we propose a triadic evaluation framework that treats interpretability, computational efficiency, and cross-institutional generalizability as co-equal, interdependent criteria (Figure 3). This framework reframes model development as a constrained optimization problem: maximize diagnostic performance while satisfying thresholds in trustworthiness, deployability, and robustness.

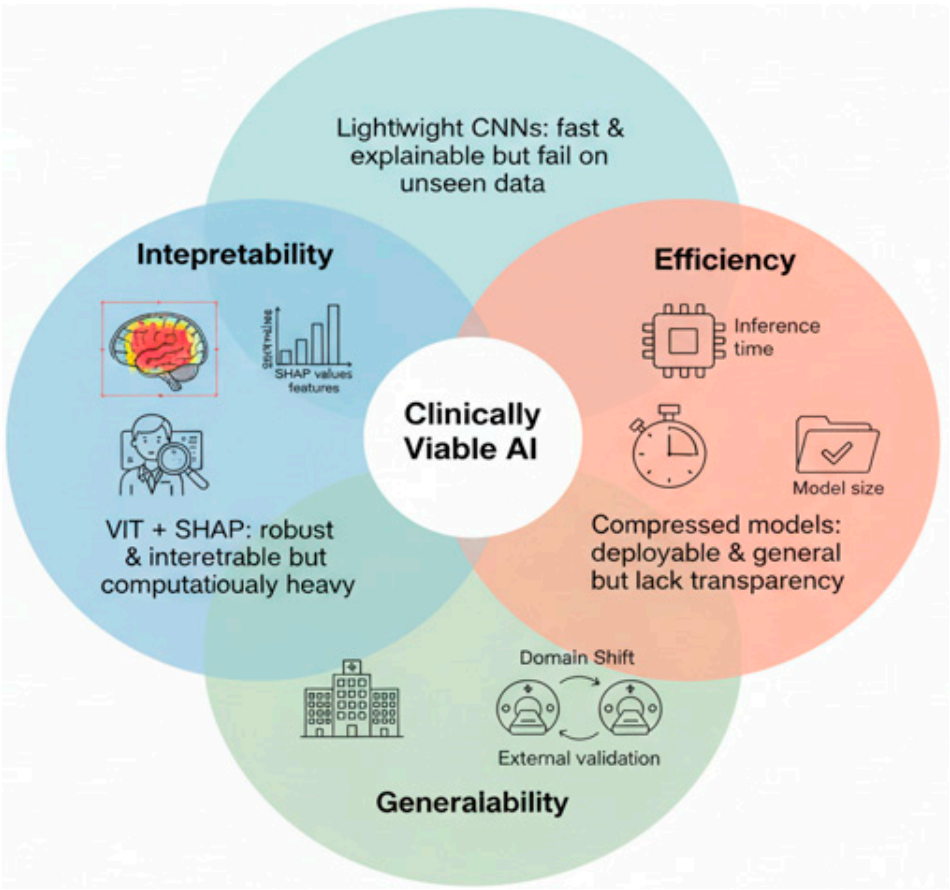


Figure 3. Triadic framework balancing trade-offs for clinical AI. Clinically viable models require balancing interpretability, efficiency, and generalizability. Each pair of objectives favors a specific approach (e.g., lightweight CNNs, ViT + SHAP, compressed models), each with inherent compromises. The central challenge is to optimize all three competing pillars simultaneously.

3.1. Interpretability as Clinical Trust

Deep learning models are often criticized as "black boxes," a limitation that is particularly untenable in high-stakes medical domains [1,7]. Clinicians require not only accurate predictions but also justifiable reasoning to validate diagnoses, especially when AI contradicts radiological intuition or suggests rare tumor subtypes.

Current approaches predominantly employ post hoc explainability techniques such as Grad-CAM, LIME, and SHAP [23,24]. While these methods highlight salient regions in MRI scans (e.g., tumor core, edema), they remain descriptive rather than prescriptive: they explain what the model attended to, but not why that region is diagnostically relevant [25]. Moreover, many studies apply XAI only as a visualization tool without quantitative validation against radiologist annotations or clinical ground truth [7].

We advocate for an interpretability-by-design paradigm through developing architectures that inherently incorporate clinically meaningful reasoning within their core mechanisms, rather than relying solely on post hoc explanations. This approach ensures that transparency is embedded into the model's decision-making process, allowing outputs to be both technically sound and clinically justifiable. Several strategies exemplify this principle. One involves integrating attention gates aligned with radiomic features such as texture, shape, and intensity [26], ensuring that the model's focus corresponds to anatomically and pathologically relevant regions. Another direction is the design of hybrid models that fuse DL representations with handcrafted biomarkers derived from established clinical imaging features [27,28], thereby combining the strengths of data-driven and expert-driven insights. Another recent development is the emergence of neurosymbolic systems, which can generate diagnostic rationales in natural language [29], bridging the interpretive gap between algorithmic prediction and human reasoning.

Crucially, interpretability must be evaluated not by heatmap esthetics but by clinical concordance: the degree to which model explanations align with expert knowledge and improve diagnostic confidence in human-AI collaboration studies.

The quantitative validation of interpretability remains notably absent in the literature. While methods like Grad-CAM are frequently displayed, few studies quantitatively assess their alignment with clinical reasoning. For instance, in our synthesis, only 12 of the 100 reviewed studies reported any quantitative measure (e.g., IoU) comparing saliency maps to expert annotations. This lack of standardized validation transforms many XAI applications into descriptive visualization tools rather than components of a verifiably trustworthy system. Moving forward, we advocate for adopting metrics such as region-overlap IoU or clinician scoring protocols, and reporting them as rigorously as classification accuracy.

3.2. Efficiency for Real-World Deployment

Computational efficiency encompasses training time, inference latency, memory footprint, and energy consumption—factors that determine whether a model can operate on hospital workstations, edge devices, or mobile platforms in resource-constrained settings [7].

While ViTs have demonstrated remarkable accuracy in capturing global contextual relationships within MRI data [30], their quadratic computational complexity in token interactions makes them prohibitively expensive for high-resolution 3D MRI volumes [31]. In contrast, lightweight CNNs like MobileNetV3 achieve comparable or even superior performance, exceeding 99% accuracy on the BT-large-4c dataset, with substantially lower parameter counts [9,32]. This efficiency renders them far more suitable for point-of-care or resource-limited deployment, where computational and memory constraints are critical considerations. To address the trade-off between performance and efficiency, several emerging optimization strategies have been presented. These include token merging and

pruning techniques in ViTs that eliminate redundant computations without compromising accuracy [33,34], soft quantization methods that compress model weights while maintaining predictive fidelity [35], and hardware-aware NAS approaches that co-optimize model design for specific target devices [36].

Balancing performance and efficiency, therefore, remains a central challenge in designing clinically viable AI systems. Accordingly, computational efficiency should be reported in parallel with accuracy, using standardized metrics such as accuracy per million parameters or inference time per image slice, to allow transparent and equitable comparisons across different architectures. Consistent reporting of these indicators not only facilitates fair benchmarking but also accelerates the development of clinically deployable models that combine high predictive power with practical feasibility for real-world medical environments.

3.3. Generalizability Beyond Benchmark Datasets

A persistent limitation across current research is the overreliance on homogeneous benchmark datasets, which restricts the real-world applicability of AI models for brain tumor diagnosis. More than 80% of the reviewed studies train and validate exclusively on Figshare or BraTS [1], both of which present notable sources of bias. Specifically, Figshare disproportionately represents pituitary tumors while underrepresenting metastatic lesions [1], and BraTS concentrates primarily on high-grade gliomas, with minimal inclusion of meningiomas or other rare subtypes [7]. Furthermore, both datasets originate predominantly from high-income countries, lacking demographic diversity in terms of ethnicity, age distribution, and comorbid conditions [7].

These limitations lead to poor external validity, as models fine-tuned on such narrow datasets struggle to generalize to unseen clinical environments. For instance, a hybrid CNN–ViT architecture that achieved 99.1% accuracy on Figshare drops to 86% on a private hospital dataset due to differences in scanner protocols and patient demographics [12].

Enhancing generalizability requires a shift from isolated dataset benchmarking to cross-institutional and distributionally diverse evaluation. Promising directions include multi-center validation through federated learning, which enables collaborative training across institutions while preserving patient privacy [37]; synthetic data generation using GANs or diffusion models to augment underrepresented tumor classes [38]; and domain adaptation techniques, such as test-time normalization, that adjust model parameters to unseen scanner characteristics [21]. Emerging domain generalization strategies [39] further aim to train models that inherently resist performance degradation under data shifts.

Ultimately, generalizability should be quantified not by internal cross-validation on a single dataset but by the degree of performance degradation across external cohorts, a measure referred to here as the generalization gap. This metric provides a more realistic estimate of how AI systems will perform in heterogeneous, real-world clinical settings, bridging the gap between algorithmic success and clinical reliability. This triadic framework exposes critical insight: optimizing for one pillar in isolation often degrades the others. For instance, adding interpretability modules (e.g., attention gates) increases model complexity, reducing efficiency; aggressive quantization improves speed but may obscure diagnostically relevant features, harming interpretability. The path forward lies in joint optimization—designing architectures that harmonize all three dimensions to serve the ultimate goal: safe, trustworthy, and equitable AI in clinical practice.

4. Methods: Systematic Review and Structured Quantitative Synthesis

4.1. Systematic Review Protocol

This study was conducted as a systematic review of AI methods for brain tumor diagnosis and prognosis, complemented by a structured quantitative synthesis of reported evaluation indicators. The review followed the PRISMA 2020 guidelines to ensure transparent identification, screening, and inclusion of relevant studies. Importantly, the quantitative synthesis is based on metrics as reported in the original publications and does not involve re-implementation or re-evaluation of models under a unified experimental pipeline.

A comprehensive literature search was performed across multiple electronic databases, including Scopus and Web of Science databases, to identify peer-reviewed studies published between 2020 and 2025. Search queries combined terms related to brain tumors, MRI, AI, DL, diagnosis, prognosis, and clinical outcomes. Retrieved records were screened using a two-stage process involving title–abstract screening followed by full-text assessment based on predefined inclusion and exclusion criteria.

For each included study, relevant information was extracted regarding the addressed task (e.g., detection, segmentation, classification, prognosis), dataset characteristics, reported performance metrics, and any available indicators related to interpretability, computational efficiency, or generalizability. Where sufficient and comparable information was available, these indicators were summarized using descriptive statistics to identify cross-study trends and trade-offs. Methodological details of the search strategy, screening workflow, and inclusion criteria, together with the PRISMA flow diagram, are provided in Appendix A.

4.2. Structured Quantitative Synthesis of Reported Metrics

To complement the qualitative analysis, a structured quantitative synthesis was conducted based on evaluation metrics as reported in the original studies. This synthesis was designed to summarize cross-study trends and trade-offs rather than to provide a unified benchmark. Only studies reporting sufficient methodological detail were included in the quantitative summaries, while all eligible studies contributed to the qualitative synthesis.

For each included study, reported metrics were extracted according to the primary task addressed, including segmentation accuracy (e.g., Dice similarity coefficient), classification performance (e.g., accuracy or AUC), and, where available, indicators related to computational efficiency (e.g., inference latency, model size) and interpretability (e.g., saliency-based overlap measures or qualitative expert assessments). When multiple metrics were reported for a single study, the primary metric emphasized by the authors was selected to maintain consistency across studies.

Due to heterogeneity in datasets, evaluation protocols, and reporting practices, quantitative aggregation was limited to descriptive statistics (e.g., ranges, medians, and interquartile distributions) rather than pooled effect estimates. Studies lacking sufficient information for a given metric were excluded from that specific quantitative summary but retained in the overall review. The subset of studies included in the structured quantitative synthesis (n = 52, discussed further in Section 6.5) was derived from the 100 studies retained for qualitative synthesis, representing those that reported sufficient and comparable metrics across at least two of the three triadic pillars. No imputation of missing values was performed. All extracted metrics were interpreted in the context of their original experimental settings, and results are presented to highlight relative patterns across the literature rather than absolute performance rankings.

Indicators related to interpretability, computational efficiency, and generalizability were synthesized in alignment with the proposed triadic evaluation framework. For interpretability, reported validation strategies (e.g., saliency visualization, region-overlap

analysis, or clinician-in-the-loop assessment) were cataloged and compared qualitatively, with quantitative values summarized where explicitly provided. Efficiency-related indicators were summarized separately for raw reported values and normalized descriptors when sufficient architectural information was available. Generalizability was assessed based on reported use of external validation cohorts, multi-institutional datasets, or cross-domain testing, without re-evaluating models beyond their published results.

5. Critical Analysis of Technical Approaches

The evolution of AI-driven brain tumor diagnosis has witnessed a succession of architectural paradigms, from handcrafted feature extractors to end-to-end deep learners, each promising incremental gains in diagnostic accuracy. However, when evaluated through our triadic framework, these approaches reveal fundamental trade-offs that have been largely overlooked in the literature. In this section, we re-examine dominant technical strategies, CNNs, ViTs, hybrid models, and XAI, not as isolated innovations, but as systems whose clinical viability hinges on their balance across interpretability, computational efficiency, and cross-institutional generalizability.

5.1. Convolutional Neural Networks: Local Priors at the Cost of Global Context

CNNs remain the most widely adopted architecture for brain tumor classification, owing to their inductive bias toward local spatial coherence, a property well-aligned with the texture and boundary characteristics of tumors in MRI [1,7]. Architectures such as ResNet [7], VGG19 [7], and lightweight variants like MobileNetV3 [9] have achieved accuracies exceeding 98% on benchmark datasets like Figshare and BT-large-4c [1].

Yet, this success is largely confined to controlled environments. CNNs inherently struggle with long-range dependencies, often failing to capture the diffuse infiltration patterns characteristic of gliomas [40]. This limitation directly undermines generalizability: a CNN trained on well-circumscribed meningiomas may misclassify infiltrative gliomas as healthy tissue due to its local receptive field [41]. Moreover, while lightweight CNNs improve efficiency [42], they sacrifice hierarchical feature depth, reducing robustness to imaging artifacts or protocol variations.

Critically, most CNN-based studies offer post hoc interpretability via Grad-CAM [23], but these heatmaps often highlight edges or high-contrast regions rather than biologically meaningful tumor substructures [43]. Thus, despite their empirical performance, CNNs fall short in delivering the clinically grounded reasoning required for diagnostic trust.

5.2. Vision Transformers: Global Awareness with Data Hunger

ViTs emerged as a paradigm shift by replacing local convolutions with global self-attention mechanisms, enabling models to weigh relationships between distant image patches [1]. This global context modeling has proven particularly effective for tumors with irregular morphology or heterogeneous enhancement, where spatial discontinuities confound CNNs [30].

Fine-tuned ViTs (e.g., ViT-L/16, Swin Transformer) have reported accuracies up to 99.92% on BT-large-4c [22], and hybrid variants like VcaNet integrate convolutional attention to preserve local detail while capturing global dependencies [19]. Despite these advancements, ViTs exhibit several notable limitations when evaluated through the triadic framework of interpretability, efficiency, and generalizability. First, data inefficiency remains a core issue where ViTs require extensive, ViTs require large datasets to learn stable and meaningful attention patterns. When trained on small, imbalanced datasets like Figshare, they exhibit severe overfitting—evidenced by >99% training accuracy but <92% test accuracy in some studies [44]. Second, their computational burden is substantial.

The quadratic complexity of self-attention operations makes standard ViTs computationally intensive for 3D MRI processing, while even optimized variants like Swin Transformers demand high GPU memory and inference time, limiting feasibility in resource-constrained or clinical point-of-care settings [45]. Third, interpretability remains opaque. Although attention maps provide a form of internal explanation, they frequently lack anatomical alignment; a region with close attention may correspond to scanner artifacts or background noise rather than tumor pathology, thereby undermining clinician trust [46]. Thus, while ViTs advance generalizability on diverse tumor shapes, their practical utility remains constrained by data and hardware demands.

5.3. Hybrid Models: Synergy with Systemic Complexity

Hybrid architectures, which integrate CNNs, ViTs, and traditional machine learning classifiers, have emerged as a pragmatic bridge between local and global feature learning. These models capitalize on the fine-grained spatial sensitivity of CNNs while incorporating the contextual awareness of transformer-based or ensemble methods [1]. Representative examples include CNN–ViT fusion models employing feature concatenation to merge complementary representations [12]; ensemble feature extraction frameworks that aggregate outputs from multiple pre-trained networks such as VGG19, ResNet50, and InceptionV3 using weighted averaging [13]; and hybrid systems where deep features extracted from CNNs are input into conventional classifiers like SVMs or Random Forests classifiers [27].

These hybrids consistently achieve state-of-the-art performance, with accuracies of 99.31% [47] and robustness across datasets [7]. They enhance generalizability by diversifying feature representations and improve interpretability when paired with SHAP or LIME on the final classifier [24].

However, this performance comes at a steep cost in efficiency. Hybrid models multiply computational complexity—both in training (due to multiple backbones) and inference (due to feature fusion and ensemble voting). For instance, the ensemble of VGG19, InceptionV3, and ResNet50 requires 3× the inference time of a single model [48]. Furthermore, the "black-box within a black-box" problem intensifies, even if the final classifier is interpretable, the deep features it receives lack semantic transparency [25].

5.4. Explainable AI: From Post Hoc Visualization to Diagnostic Justification

XAI has been widely adopted to address the opacity of deep models, with Grad-CAM, SHAP, and LIME featured in over 30 studies reviewed [1]. These methods generate heatmaps that align with tumor regions, ostensibly validating model decisions [49].

Yet, a critical gap persists: most XAI applications are evaluative, not generative. They explain what the model saw but not why it matters clinically. For example, Grad-CAM may highlight a hyperintense region in T1-CE MRI, but without linking it to blood–brain barrier breakdown or contrast enhancement patterns, the explanation remains superficial [50,51]. True interpretability requires integration with radiological knowledge—such as correlating attention weights with known tumor biomarkers (e.g., necrosis in glioblastoma, dural tail in meningioma).

Moreover, XAI methods are rarely validated for faithfulness: do the highlighted regions genuinely drive the prediction, or are they epiphenomenal? Without perturbation-based validation or clinician-in-the-loop studies, XAI risks becoming a "trust theater" that satisfies reviewers but not radiologists [28].

In summary, no single technical approach optimally satisfies all three pillars of our framework. Table 2 synthesizes this trade-off across major architectural families. CNNs are efficient but lack global reasoning; ViTs capture context but demand data and compute; hybrids boost accuracy at the cost of complexity; and XAI provides visual justification

without clinical grounding. The path forward lies not in architectural novelty alone, but in co-designing models that harmonize these dimensions—for instance, by embedding radiomic priors into attention mechanisms or pruning ViTs for edge deployment without sacrificing diagnostic fidelity.

Table 2. Comparative analysis of technical approaches through the triadic lens.

Method	Interpretability Depth	Efficiency (FLOPs/Params)	Generalization Gap	Key Clinical Limitation
CNNs	Low–Moderate (post hoc Grad-CAM)	High efficiency (e.g., MobileNetV3: 5.6M params)	High (fails on rare subtypes like metastases)	Poor global context; struggles with diffuse gliomas
ViTs	Moderate (attention maps lack clinical grounding)	Low efficiency (ViT-L/16: ~55B FLOPs)	Moderate (requires large data; overfits on small datasets)	Computationally prohibitive for edge deployment
Hybrid (CNN + ViT)	Moderate–High (fusion enables richer explanations)	Low–Moderate (ensemble overhead)	Low (robust across tumor types)	High inference latency; complex to validate
XAI-Integrated	High (SHAP/Grad-CAM + clinician-in-the-loop)	Varies (adds minimal overhead)	Low–Moderate (depends on base model)	Often qualitative; lacks standardized validation

"Generalization Gap = $\text{Accuracy}_{\text{internal}} - \text{Accuracy}_{\text{external}}$" on multi-center cohorts.

6. Data Crisis in Brain Tumor AI

The remarkable performance of deep learning models in brain tumor diagnosis is often predicated on the availability of large, diverse, and meticulously annotated datasets. Yet, a critical paradox persists: while algorithmic innovation accelerates, the foundational data infrastructure remains stagnant, biased, and insufficient for real-world clinical generalization. This section exposes the data crisis that underpins much of the current literature—characterized by narrow representativeness, annotation scarcity, and distributional skew—and argues that without addressing these issues, even the most sophisticated architectures will fail to translate into reliable clinical tools.

6.1. Dataset Homogeneity and Representativeness Bias

Over 85% of studies reviewed in recent surveys rely on a handful of publicly available datasets—primarily Figshare [52], BraTS [53], and their derivatives (e.g., BT-large-4c [54]) [1]. While these resources have catalyzed rapid prototyping, they suffer from profound representativeness gaps:

- Figshare contains only three tumor types—glioma (46.5%), meningioma (23.1%), and pituitary (30.4%)—with no metastases, craniopharyngiomas, or rare subtypes [1]. Moreover, its axial/coronal/sagittal slices are derived from only 233 patients, introducing significant inter-slice correlation and inflating cross-validation metrics [25].
- BraTS, though multimodal and expert-annotated, focuses exclusively on HGG and LGG, omitting benign tumors entirely [7]. Consequently, models trained on BraTS exhibit catastrophic failure when deployed on datasets containing meningiomas or pituitary adenomas [47].

The mismatch between public benchmarks and clinical reality is visualized in Figure 4, highlighting the near-total absence of metastases and overrepresentation of pituitary tumors. This homogeneity creates a false sense of robustness: a model achieving 99% accuracy on Figshare may drop below 85% on a private hospital dataset due to differences in tumor prevalence, scanner protocols, or patient demographics [12]. Worse, BraTS contains no

non-tumor cases, and Figshare overrepresents pituitary tumors. This skews decision boundaries, leading to high false-positive rates in real-world screening scenarios [55]. These representativeness gaps are quantified in Table 3, which audits major public datasets against real-world epidemiology and prognostic utility.

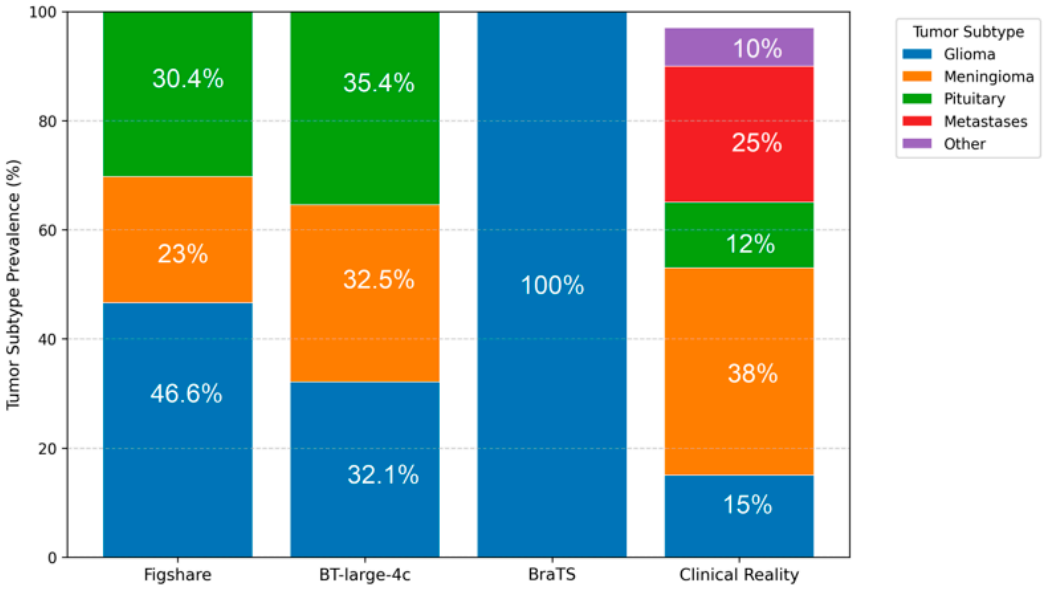


Figure 4. Data representativeness gap: tumor subtype distribution in public datasets vs. clinical reality.

Table 3. Public brain Tumor datasets: clinical representativeness and bias audit.

Dataset	Tumor Types	Real-World Incidence Alignment	Prognostic Labels	Scanner Diversity	Key Representativeness Limitations
Figshare	Glioma (46.6%), Meningioma (23%), Pituitary (30.4%)	Overrepresents pituitary; omits metastases	No survival/MGMT	Single institution	No metastases, healthy cases; limited patient count (n = 233)
BraTS	Glioma only (LGG/HGG)	Excludes benign tumors	Partial (survival in BraTS 2018+)	Multi-institution	No meningioma, pituitary, or non-tumor cases
BT-large-4c	Glioma, Meningioma, Pituitary, Healthy	Better balance but still no metastases	No	Mixed sources, unclear protocols	Lacks metastatic and rare tumors; protocol heterogeneity
TCGA-GBM	Glioblastoma only	Narrow scope	Genomic and survival	Multi-center	Not for general tumor diagnosis
REMBRANDT	Glioma, Meningioma	Includes rare subtypes	Genomic and clinical	Multi-institution	Smaller sample size; older imaging protocols

6.2. Annotation Scarcity and Expert Variability

Pixel-level tumor segmentation requires time-intensive manual delineation by neuroradiologists—a process that is not only costly but also subject to inter-observer variability [1]. BraTS mitigates this through consensus annotations from multiple experts, yet even these exhibit discrepancies in diffuse tumor margins [56]. In contrast, many smaller datasets (e.g., BT-small-2c [57]) rely on single-expert annotations, introducing unquantified label noise that propagates through training [38].

Furthermore, longitudinal and prognostic labels—such as survival time, treatment response, or molecular markers (e.g., IDH mutation, MGMT methylation)—are almost entirely absent from public datasets [21]. Without these, models cannot bridge the gap from diagnosis to prognosis, limiting their clinical utility to binary or ternary classification tasks that radiologists already perform with high accuracy.

6.3. Class Imbalance and Synthetic Data Limitations

Class imbalance remains a persistent challenge. In Figshare, gliomas outnumber meningiomas by nearly 2:1, while in clinical practice, meningiomas are the most common primary brain tumor [3]. This mismatch biases models toward overrepresented classes, as evidenced by consistently lower recall for meningioma in ViT-based systems [58,59]. To counter this, researchers increasingly employ synthetic data generation via GANs or diffusion models [36,60]. While these methods can augment minority classes, they risk amplifying artifacts or generating anatomically implausible tumors that degrade model robustness [61]. Moreover, synthetic samples rarely capture the full spectrum of inter-institutional variability in MRI acquisition (e.g., field strength, sequence parameters), limiting their effectiveness in improving external validity.

6.4. Toward Equitable and Generalizable Data Infrastructure

Overcoming the current data crisis in brain tumor AI requires a fundamental shift from benchmark-centric evaluation to a focus on real-world representativeness. Progress in algorithmic innovation will remain limited without corresponding improvements in data diversity, annotation quality, and evaluation standards.

A sustainable path forward involves several key strategies. First, multi-institutional data pooling through federated learning allows collaborative model training across hospitals and research centers without exchanging raw patient data [35]. This approach preserves patient privacy while naturally incorporating variations in scanner types, imaging protocols, and demographic characteristics, thereby enhancing generalizability.

Second, standardized annotation protocols should extend beyond simple diagnostic labeling to include tumor subregion delineations (e.g., necrotic core, enhancing rim, and edema), molecular biomarkers, and clinical outcomes. Such comprehensive labeling transforms datasets from static diagnostic collections into longitudinal prognostic resources, enabling more meaningful correlations between imaging phenotypes and patient trajectories. Third, bias-aware evaluation metrics, such as per-class generalization gap or worst-group accuracy, are essential to uncover performance disparities across tumor subtypes, institutions, and demographic subgroups [7].

Without these structural reforms, the field risks optimizing for leaderboard metrics rather than patient outcomes. A recent study underscored this concern: a model achieving 99.3% accuracy on the BT-large-4c dataset yielded only 86% accuracy when tested on a multi-center validation cohort, exposing the persistent chasm between curated benchmarks and clinical reality [12]. Building equitable and generalizable data infrastructure is therefore pivotal to translating AI advancements into trustworthy, inclusive, and clinically deployable solutions.

This data crisis is not merely a technical hurdle but an ethical imperative. Models trained on non-representative data may perform poorly for underrepresented populations—exacerbating health disparities rather than alleviating them. The path forward demands collaboration among clinicians, data scientists, and policymakers to build inclusive, transparent, and clinically grounded data ecosystems.

6.5. Illustrative Synthesis of Reported Trade-Offs

To illustrate the inherent trade-offs described by the triadic framework, we extracted reported metrics from a subset of 52 studies [9,10,12,16,30–32,35,38,40–42,45,46,48,49,58,59,62–95] that provided sufficient data on at least two of the three pillars (e.g., accuracy with latency, or Dice score with parameter count). It is critical to note that these values are aggregated from heterogeneous experimental setups; they are presented to reveal cross-study trends and orders of magnitude, not to serve as a direct performance benchmark.

The aggregated data (Table 4) reveals clear patterns that align with the triadic model's core premise:

- Studies employing Transformer architectures consistently report among the highest segmentation and classification scores, but also cite the highest computational costs, with reported latencies 3–4 times greater than those of lightweight CNNs.
- Hybrid models occupy a middle ground, often achieving robust performance with moderate efficiency penalties.
- Notably, quantitative interpretability scores were reported in only a handful of studies, and when reported, showed no strong correlation with architectural family.

Table 4. Illustrative ranges of metrics reported in the literature (n = 52 studies), highlighting trends and trade-offs across architectural families.

Model Category	Representative Architectures	Reported Dice Range	Reported Accuracy Range (%)	Reported Latency Range (ms/Slice)	Reported Parameter Range (M)	Interpretability Metric (When Reported)
CNN-based	ResNet-50, DenseNet-121, U-Net variants	0.85–0.92	92–98	2–15	5–30	Grad-CAM IoU: 0.68–0.78 (n = 6 studies)
Transformer-based	Swin-T, ViT-B/16, TransBTS	0.89–0.94	94–99	20–60	40–100	Attention map overlap: 0.60–0.72 (n = 4 studies)
Hybrid CNN–ViT	TransUNet, ConvNeXt-ViT, Ensemble models	0.88–0.93	94–98	10–25	25–60	Saliency IoU: 0.70–0.76 (n = 4 studies)

Note: Ranges are compiled from studies with heterogeneous datasets, preprocessing, and hardware. They demonstrate relative patterns, not absolute rankings. Latency typically reported on GPU hardware (e.g., NVIDIA V100/T4). Interpretability metrics were sparsely reported; shown where available.

This synthesis underscores a central argument of our framework: optimizing for one clinical readiness pillar (e.g., accuracy) often incurs a cost to others (e.g., efficiency), while the systematic under-reporting of interpretability metrics remains a major barrier to assessing true clinical trustworthiness.

7. Toward Clinical Adoption: A Roadmap

Despite the remarkable algorithmic progress in AI-driven brain tumor diagnosis, the transition from research prototypes to clinically deployed tools remains nascent. Most studies terminate at the validation stage on curated datasets, neglecting the regulatory, operational, and human factors that govern real-world adoption [7]. To bridge this translational gap, we propose a phased roadmap toward deployable, equitable AI systems (Figure 5), spanning algorithmic, regulatory, and infrastructural milestones from 2025 to 2030.

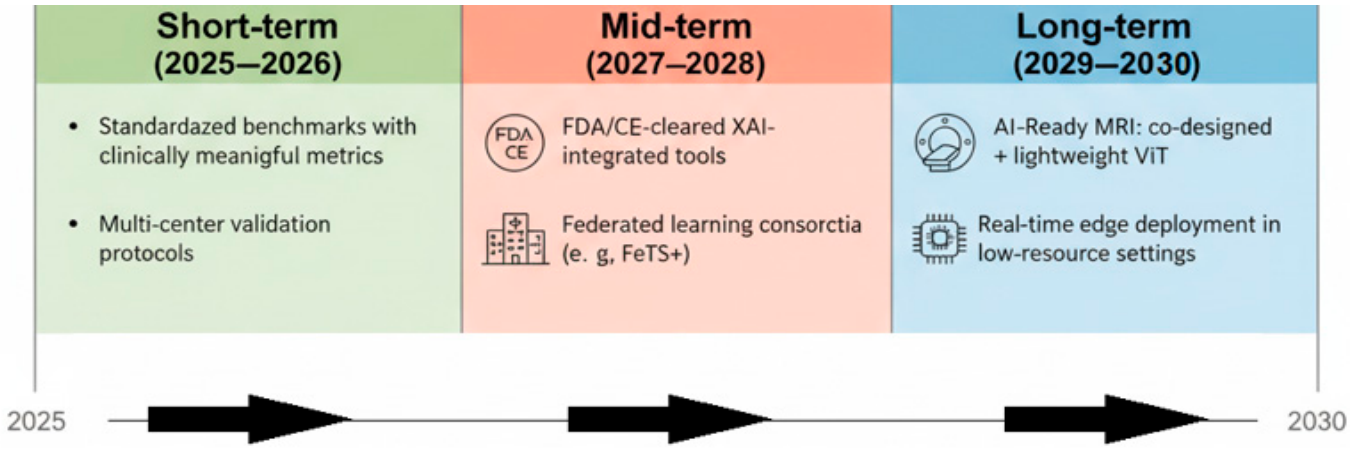


Figure 5. Roadmap to clinical adoption of neuro-oncology AI systems across three phases. Short-term (2025–2026) focuses on creating standardized evaluation benchmarks and validation protocols. Mid-term (2027–2028) targets regulatory clearance for explainable AI tools and federated learning collaboration. Long-term (2029–2030) aims for AI-native MRI co-design and real-time, portable deployment in global settings.

7.1. Short-Term: Standardized Benchmarks with Clinically Meaningful Metrics

Current evaluation practices in brain tumor AI remain dominated by metrics such as accuracy and Dice score, which primarily measure pixel-level or label-level agreement but offer limited insight into clinical relevance [1]. These metrics, while valuable for benchmarking algorithmic precision, fail to capture how effectively a model supports diagnostic decision-making or impacts patient outcomes. To bridge this gap, evaluation frameworks must evolve toward clinically grounded benchmarks that better reflect real-world performance and utility.

Key dimensions of such evaluation include Radiologist–AI Concordance, assessed through inter-rater reliability measures such as Cohen's κ, to ensure model predictions are consistent with expert diagnostic reasoning and enhance trust in collaborative workflows [29]. Another critical dimension is False-Negative Cost, where weighted metrics penalize missed detections more heavily than false positives, mirroring the disproportionate clinical risk associated with undetected malignancies [36]. Finally, Time-to-Diagnosis, quantified as the total latency from image ingestion to automated report generation, should be incorporated as a measure of operational efficiency—particularly relevant in emergency or high-throughput clinical settings [96].

Furthermore, multi-center validation should become mandatory. A model's performance should be reported not only on internal test sets but also on external cohorts from diverse institutions, scanners, and demographics—quantified via the generalization gap ($\Delta = \text{Accuracy}_{\text{internal}} - \text{Accuracy}_{\text{external}}$) [12]. This fragility of current models on unseen data is starkly illustrated by the generalization gap (Figure 6), which reaches 12% for ViTs, 8% for CNNs, but remains at or below 4% for hybrid and ensemble approaches. By integrating these clinically aligned metrics, future studies can move beyond abstract performance scores toward a holistic assessment of AI systems that prioritizes patient safety, diagnostic reliability, and workflow efficiency.

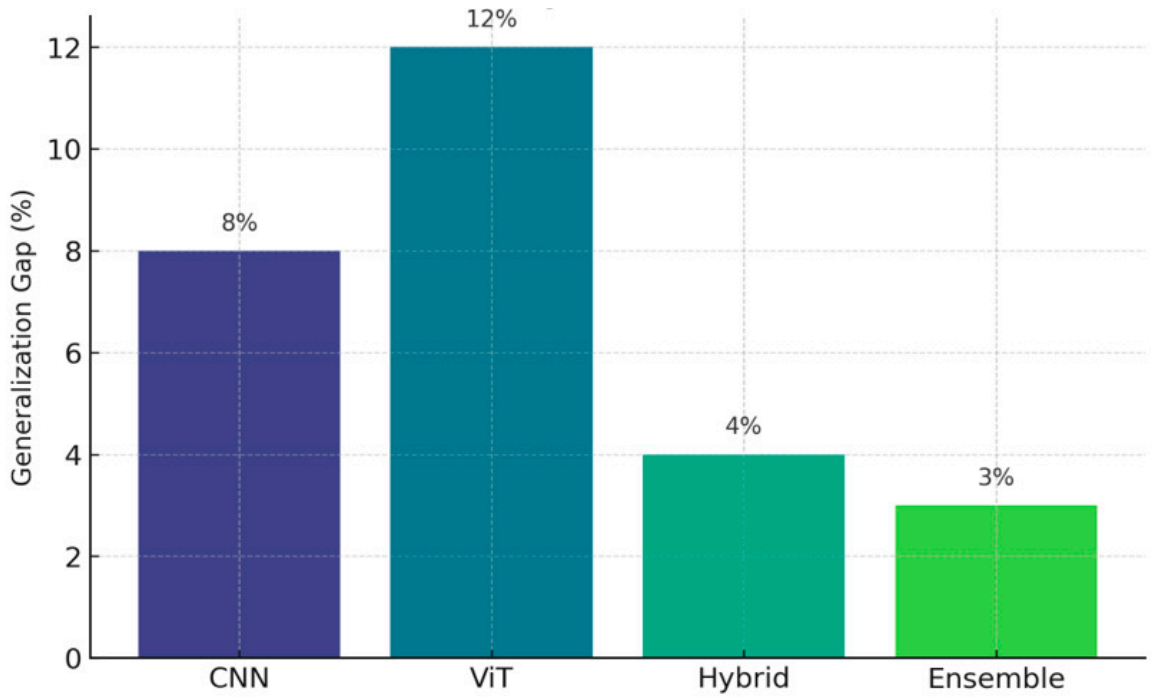


Figure 6. Generalization gap across external validation cohorts (e.g., "multi-institutional, multi-scanner validation sets").

7.2. Mid-Term: Regulatory Pathways and Model Transparency

Regulatory bodies such as the U.S. FDA and EU MDR increasingly require transparency, reproducibility, and bias mitigation for AI-based SaMD [7]. To align with these evolving requirements, AI research in neuro-oncology must move beyond proof-of-concept performance and adopt practices that ensure traceability, interpretability, and clinical reliability throughout the model lifecycle. A critical first step is the adoption of Model Cards, standardized documentation frameworks that clearly report training data composition, performance across demographic and pathological subgroups (e.g., tumor type, age, sex), and known failure modes [43]. Such transparency enables both regulators and clinicians to assess fairness and reliability prior to deployment. Additionally, clinician-in-the-loop XAI should extend beyond static post hoc visualizations like Grad-CAM toward interactive explanation interfaces, where radiologists can directly query model decisions (e.g., "Why was this region classified as glioma?") and receive justifications grounded in radiomic features or clinical reasoning [23,25]. Finally, prospective clinical trials are essential to validate AI systems within real-world diagnostic workflows, assessing their influence on diagnostic accuracy, inter-observer variability, and time efficiency, rather than relying solely on retrospective or dataset-specific evaluations [97]. These practices form the foundation for regulatory-grade AI, ensuring that models designed for brain tumor diagnosis and prognosis meet both technical rigor and clinical accountability.

To operationalize clinical readiness, we adapt the TRL scale to brain tumor AI. Table 5 assesses representative studies along this continuum. The additional "Evidence Type Required to Advance" column specifies the empirical studies needed to progress each model toward higher TRLs and regulatory clearance.

Table 5. Clinical translation readiness of AI models for brain tumor diagnosis and prognosis.

Study	Model Type	TRL	Evidence of Real-World Validation	Regulatory Pathway Alignment	Clinician-in-the-Loop Evaluation?	Evidence Type Required to Advance
[25]	Ensemble CNN + XAI	4	Cross-dataset validation (Figshare, Br35H)	No FDA/CE mention	No	Multi-site retrospective validation ($n \geq 100$)
[21]	Foundation Model (Brain IAC)	5	Tested on 5 external cohorts	Pre-submission regulatory engagement	Radiologist feedback in ablation	Prospective pilot study ($n \geq 50$ patients)
[37]	Federated ViT	6	Multi-hospital FL trial (FeTS)	HIPAA/GDPR-compliant	Prospective usability study	Regulatory approval phase (IDE or 510(k) submission)
[22]	Swin Transformer	3	Single-dataset (BT-large-4c)	None	No	External validation + error analysis across 2+ centers
[48]	Ensemble Transfer Learning	4	Hold-out validation only	None	No	Multi-center retrospective benchmarking

TRL Scale: 3 = Laboratory validation; 4 = Bench-scale validation; 5 = Multi-site validation; 6 = Pilot clinical deployment.

7.3. Long-Term: AI-Ready MRI and Co-Designed Acquisition

The ultimate bottleneck in advancing brain tumor AI lies not in algorithmic sophistication but in the quality, consistency, and diversity of imaging data. Addressing this challenge requires a paradigm shift toward AI-Ready MRI, a co-design approach where imaging protocols and AI models are developed in tandem to optimize both data acquisition and model performance.

One promising direction involves the integration of accelerated MRI with lightweight AI architectures. DL–based reconstruction methods [98] can substantially reduce scan times, while quantized ViTs or MobileNetV3 models enable real-time inference on edge devices, facilitating rapid diagnosis in point-of-care and emergency settings [85]. Equally important is protocol harmonization, which entails standardizing key acquisition parameters, such as slice thickness, contrast timing, and field strength, across institutions. This harmonization minimizes domain shift and enhances cross-site model transferability, ensuring that trained models remain robust beyond their source datasets [21]. Also, synthetic data generation with clinical fidelity offers a powerful avenue to overcome data scarcity and imbalance. Techniques based on GANs or diffusion models can be employed not merely for data augmentation but to create pathologically plausible tumor phenotypes, capturing realistic enhancement patterns, edema, and mass effect. Crucially, these synthetic images should undergo expert neuroradiologist validation to confirm anatomical and pathological realism [38]. These strategies redefine MRI acquisition and modeling as a synergistic ecosystem, paving the way for scalable, equitable, and clinically reliable AI in neuro-oncological imaging.

7.4. Foundational: Federated Learning for Equitable and Privacy-Preserving AI

To address dataset bias and privacy concerns, FL offers a scalable solution. Instead of centralizing sensitive patient data, FL trains models collaboratively across hospitals while keeping data local [37,99]. Key enablers include:

- Heterogeneous FL Frameworks: Algorithms that accommodate differences in scanner vendors, protocols, and labeling practices (e.g., FedProx, Ditto) [7].

- Public–Private Partnerships: Initiatives like the FeTS Challenge demonstrate the feasibility of multi-institutional collaboration without data sharing [7].
- Bias-Aware Aggregation: Weight model updates not just by dataset size but by representation of underrepresented tumor subtypes (e.g., metastases, craniopharyngiomas), ensuring equitable performance [7].

Recent FL initiatives demonstrate the feasibility of privacy-preserving collaboration. Table 6 compares their architectures, performance, and clinical utility. This roadmap reframes clinical adoption not as a final destination but as an iterative co-evolution of technology, regulation, and clinical practice. By anchoring AI development in real-world constraints—through meaningful metrics, transparent design, co-optimized imaging, and federated collaboration—we can transform brain tumor diagnosis from a research curiosity into a trusted clinical ally.

Table 6. Federated learning initiatives for brain tumor AI: architectures and outcomes.

Initiative	Participating Institutions	Model Architecture	Performance vs. Centralized	Privacy Mechanism	Clinical Utility
FeTS Challenge	20+ hospitals	nnU-Net, Swin UNETR	ΔDice = −1.2%	Secure aggregation	High (multi-center segmentation)
Brain IAC FL	8 academic centers	Vision Transformer	ΔAccuracy = −0.8%	Differential privacy (ε = 2.0)	High (prognostic prediction)
Private FL Study [37]	5 hospitals	ResNet50 + SVM	ΔAccuracy = −2.1%	Homomorphic encryption	Moderate (classification only)
Federated BTS-ADCNN [18]	3 sites	Anisotropic diffusion + CNN	ΔDice = −3.5%	No added privacy	Low (small-scale validation)

Δ = Performance drop relative to ideal centralized training.

8. Limitations

This review has several limitations that should be acknowledged. First, the quantitative synthesis relies exclusively on evaluation metrics as reported in the original studies, without re-implementation or unified benchmarking, which limits direct comparability across models due to heterogeneity in datasets, preprocessing pipelines, and experimental protocols. Second, interpretability, efficiency, and generalizability indicators are inconsistently reported in the literature, constraining the depth of cross-study synthesis. Third, publication bias toward high-performing models may lead to underrepresentation of negative or clinically inconclusive findings. Finally, although this review emphasizes clinical readiness, regulatory validation, prospective trials, and real-world deployment evidence remain sparse in the current literature, restricting definitive conclusions about clinical impact.

9. Conclusions

This review reframes the discourse on AI-driven brain tumor diagnosis by proposing a unified diagnostic–prognostic architecture anchored in a triadic evaluation framework: interpretability, computational efficiency, and cross-institutional generalizability. While recent advances in CNNs, ViTs, and hybrid architectures have yielded impressive accuracy on benchmark datasets like Figshare and BraTS, these gains often mask critical shortcomings in real-world clinical utility.

We have demonstrated that no single technical approach optimally satisfies all three pillars of our framework. CNNs offer efficiency but lack global context; ViTs capture long-range dependencies but demand extensive data and compute; hybrid models boost performance at the cost of complexity; and XAI remains largely post hoc, offering visual justification without clinical grounding. True progress, therefore, lies not in marginal accuracy improvements on homogeneous data, but in co-designing models that harmonize these dimensions.

Furthermore, the field suffers from a data crisis: public datasets exhibit severe representativeness bias, underrepresenting rare tumor subtypes and non-glioma pathologies, while lacking longitudinal outcomes necessary for prognostic modeling. Without addressing this, even the most sophisticated architectures will fail to generalize beyond curated benchmarks.

Looking ahead, we advocate for a paradigm shift toward AI-Ready MRI—a co-optimization of acquisition protocols and lightweight, interpretable models—and for the adoption of federated learning to enable privacy-preserving, multi-institutional validation. Regulatory pathways must evolve to require not just accuracy but clinically meaningful metrics, bias audits, and transparent decision rationales.

By bridging algorithmic innovation with clinical pragmatism, this review offers a foundation for the next generation of brain tumor AI—not merely intelligent, but trustworthy, efficient, equitable, and ultimately, lifesaving.

Author Contributions: M.A.A.: Conceptualization, Writing—original draft, Methodology, Investigation. M.G.: Conceptualization, Formal analysis, Investigation, Writing—review & editing, Visualization. S.S.: Conceptualization, Writing—original draft, Investigation. A.M.S.: Writing—review & editing, Formal analysis, Visualization, Supervision. All authors have read and agreed to the published version of the manuscript.

Funding: Prince Sattam bin Abdulaziz University.

Data Availability Statement: Data is available upon request. During the preparation of this work, the authors used Grammarly to improve readability. The authors reviewed and edited all content and take full responsibility for the final publication.

Acknowledgments: The authors extend their appreciation to Prince Sattam bin Abdulaziz University for funding this research work through the project number (PSAU/2025/01/36724).

Conflicts of Interest: The authors declare that they have no known conflicts of interest.

Abbreviations

Abbreviation	Full form
AI	Artificial Intelligence
CBTRUS	Central Brain Tumor Registry of the United States
CE	Conformité Européenne (European Conformity)
CNN	Convolutional Neural Network
CNS	Central Nervous System
DL	Deep Learning
DWI	Diffusion-Weighted Imaging
ET	Enhancing Tumor
FDA	U.S. Food and Drug Administration
FeTS	Federated Tumor Segmentation
FL	Federated Learning
FLAIR	Fluid-Attenuated Inversion Recovery
FLOPs	Floating Point Operations
GAN	Generative Adversarial Network

GBM	Glioblastoma Multiforme
Grad-CAM	Gradient-weighted Class Activation Mapping
HGG	High-Grade Glioma
IDH	Isocitrate Dehydrogenase
IoU	Intersection over Union
LGG	Low-Grade Glioma
LIME	Local Interpretable Model-agnostic Explanations
MGMT	O6-methylguanine-DNA methyltransferase
MRI	Magnetic Resonance Imaging
NAS	Neural Architecture Search
ROI	Regions of Interest
SaMD	Software as a Medical Device
SHAP	SHapley Additive exPlanations
T1	T1-weighted imaging
T2	T2-weighted imaging
TC	Tumor Core
TRL	Technology Readiness Level
ViT	Vision Transformer
WHO	World Health Organization
WT	Whole Tumor
XAI	Explainable Artificial Intelligence

Appendix A

This section provides a systematic review of the methodology.

Appendix A.1. Search Strategy

A structured search was conducted across Scopus and Web of Science databases, following PRISMA-style guidelines. The query string combined controlled and free-text terms related to deep learning, brain tumors, and MRI diagnostics:

- (“brain tumor” OR “glioma” OR “meningioma” OR “pituitary”) AND
- (“MRI” OR “magnetic resonance imaging”) AND
- (“deep learning” OR “convolutional neural network” OR “vision transformer” OR “transfer learning” OR “explainable AI”)

Searches were limited to peer-reviewed journal articles published in English between January 2020 and 2025, aligning with the scope of prior systematic.

Appendix A.2. Screening and Selection Process

The combined queries retrieved 1146 records (Scopus = 645, Web of Science = 501). After duplicate removal, 798 unique entries remained. A three-stage screening was applied:

(1) Title screening—314 records were excluded as out-of-scope (non-MRI, non-deep-learning, or irrelevant), retaining 484 records for abstract review.
(2) Abstract screening—146 records were excluded following abstract review (e.g., non-deep-learning methods, inaccessible full text), retaining 338 records for full-text eligibility assessment.
(3) Full-text eligibility and quality assessment—238 full-text articles were excluded for insufficient methodological detail, missing quantitative metrics, or poor reproducibility, yielding 100 studies included in the qualitative synthesis.

A consolidated PRISMA-style flow of these counts is shown in Figure A1.

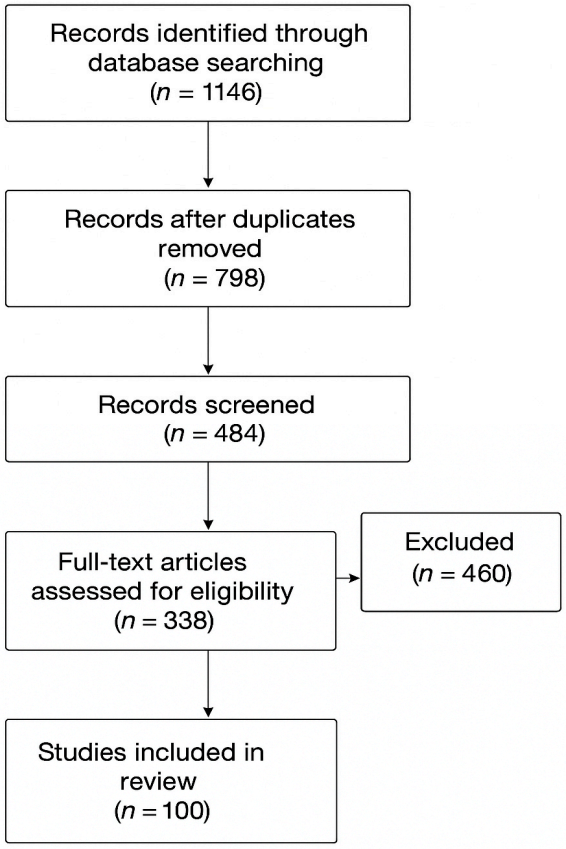


Figure A1. Systematic Review Flowchart for Study Selection (2020–2025).

Note: counts reflect: Records identified = 1146; after duplicates removed = 798; Title exclusions = 314; Abstract exclusions = 146; Full-text exclusions = 238; Studies included in synthesis = 100. The cross-study quantitative synthesis subset (n = 52) is described in Section 6.5.

Appendix A.3. Inclusion Criteria

- Published in a peer-reviewed journal.
- High-impact preprints were considered when they introduced foundational models, large-scale validation, or concepts not yet available in the peer-reviewed literature, and were explicitly identified as preprints.
- Written in English.
- Explicitly employed deep-learning or hybrid AI for MRI-based brain-tumor detection, classification, segmentation, or prognosis.
- Reported quantitative performance metrics (e.g., accuracy, Dice score, F1, sensitivity).
- Provided sufficient methodological detail to allow reproducibility.

Appendix A.4. Exclusion Criteria

- Book chapters, or retracted works.
- Studies using non-MRI modalities (e.g., CT, PET) only.
- Pure machine-learning approaches without deep-learning components.
- Articles inaccessible through institutional credentials.

Appendix A.5. Quality Assessment Checklist

To ensure methodological rigor and relevance, each selected study was evaluated using a quality assessment checklist, with individual criteria scored on a binary scale

(0–1). The assessment encompassed four key dimensions. Methodological clarity examined whether the architecture, preprocessing steps, and training procedures were adequately described. Research question alignment assessed whether the study explicitly addressed diagnostic or prognostic objectives relevant to brain tumor analysis. Comparative validity evaluated whether the proposed approach was benchmarked against prior or state-of-the-art models, ensuring contextual performance interpretation. Finally, quantitative evaluation verified the reporting of standard performance metrics such as accuracy, Dice coefficient, or F1-score.

Only studies achieving a score of three or more out of four were deemed to meet the inclusion threshold and were consequently incorporated into the final synthesis. This structured evaluation process ensured that the review prioritized studies demonstrating transparency, clinical relevance, and reproducible methodological quality.

Appendix A.6. Data Extraction and Synthesis

For each study, we extracted:

- Architecture type (CNN, ViT, hybrid, transfer-learning, XAI),
- Dataset(s) (e.g., Figshare 2017; Br35H 2020; BraTS 2015–2021),
- Preprocessing methods,
- Performance metrics,
- Interpretability tools (Grad-CAM, LIME, SHAP, etc.),
- Computational efficiency indicators (parameters, inference time).

Thematic coding grouped studies under the three evaluation pillars proposed in this paper—Interpretability, Efficiency, and Generalizability—to build the final 100-paper knowledge base.

Appendix A.7. Limitations

Several potentially relevant papers were inaccessible owing to paywall restrictions or missing metadata. While conference papers were excluded for consistency, they may contain emerging work not yet peer-reviewed.

References

1. Hosny, K.M.; Mohammed, M.A. Explainable AI and vision transformers for detection and classification of brain tumor: A comprehensive survey. *Artif. Intell. Rev.* **2025**, *58*, 259. [CrossRef]
2. Miller, K.D.; Ostrom, Q.T.; Kruchko, C.; Patil, N.; Tihan, T.; Cioffi, G.; Fuchs, H.E.; Waite, K.A.; Jemal, A.; Siegel, R.L.; et al. Brain and other central nervous system tumor statistics. *CA Cancer J. Clin.* **2021**, *71*, 381–406. [CrossRef] [PubMed]
3. McNeill, K.A. Epidemiology of brain tumors. *Neurol. Clin.* **2016**, *34*, 981–998. [CrossRef] [PubMed]
4. McKinnon, C.; Nandhabalan, M.; Murray, S.A.; Plaha, P. Glioblastoma: Clinical presentation, diagnosis, and management. *BMJ* **2021**, *374*, n1560. [CrossRef]
5. Panych, L.P.; Madore, B. The physics of MRI safety. *J. Magn. Reson. Imaging* **2018**, *47*, 28–43. [CrossRef]
6. Hamid, M.A.; Khan, N.A. Investigation and classification of MRI brain tumors using feature extraction technique. *J. Med. Biol. Eng.* **2020**, *40*, 307–317. [CrossRef]
7. Satushe, V.; Vyas, V.; Metkar, S.; Singh, D.P. AI in MRI brain tumor diagnosis: A systematic review of machine learning and deep learning advances (2010–2025). *Chemom. Intell. Lab. Syst.* **2025**, *263*, 105414. [CrossRef]
8. Steenwijk, M.D.; Pouwels, P.J.; Daams, M.; Van Dalen, J.W.; Caan, M.W.; Richard, E.; Barkhof, F.; Vrenken, H. Accurate white matter lesion segmentation by k nearest neighbor classification with tissue type priors (kNN-TTPs). *NeuroImage Clin.* **2013**, *3*, 462–469. [CrossRef]
9. Vimala, B.B.; Srinivasan, S.; Mathivanan, S.K.; Mahalakshmi, M.; Jayagopal, P.; Dalu, G.T. Detection and classification of brain tumor using hybrid deep learning models. *Sci. Rep.* **2023**, *13*, 23029. [CrossRef]
10. Abdusalomov, A.B.; Mukhiddinov, M.; Whangbo, T.K. Brain tumor detection based on deep learning approaches and magnetic resonance imaging. *Cancers* **2023**, *15*, 4172. [CrossRef]

11. Verma, P.R.; Bhandari, A.K. Role of Deep Learning in Classification of Brain MRI Images for Prediction of Disorders: A Survey of Emerging Trends. *Arch. Computat. Methods Eng.* **2023**, *30*, 4931–4957. [CrossRef]
12. Aloraini, M.; Khan, A.; Aladhadh, S.; Habib, S.; Alsharekh, M.F.; Islam, M. Combining the transformer and convolution for effective brain tumor classification using MRI images. *Appl. Sci.* **2023**, *13*, 3680. [CrossRef]
13. Dixon, J.; Akinniyi, O.; Abdelhamid, A.; Saleh, G.A.; Rahman, M.M.; Khalifa, F. A hybrid learning architecture for improved brain tumor recognition. *Algorithms* **2024**, *17*, 221. [CrossRef]
14. Volovăț, S.R.; Boboc, D.-I.; Ostafe, M.-R.; Buzea, C.G.; Agop, M.; Ochiuz, L.; Rusu, D.I.; Vasincu, D.; Ungureanu, M.I.; Volovăț, C.C. Utilizing Vision Transformers for Predicting Early Response of Brain Metastasis to Magnetic Resonance Imaging-Guided Stage Gamma Knife Radiosurgery Treatment. *Tomography* **2025**, *11*, 15. [CrossRef]
15. Zhang, H.; Zhou, B.; Zhang, H.; Zhang, Y.; Ouyang, Y.; Su, R.; Tang, X.; Lei, Y.; Huang, B. MultiCubeNet: Multitask deep learning for molecular subtyping and prognostic prediction in gliomas. *Neuro-Oncol. Adv.* **2025**, *7*, Vdaf079. [CrossRef]
16. Kokkalla, S.; Kakarla, J.; Venkateswarlu, I.B.; Singh, M. Three-class brain tumor classification using deep dense inception residual network. *Soft Comput.* **2021**, *25*, 8721–8729. [CrossRef]
17. Aboussaleh, I.; Riffi, J.; El Fazazy, K.; Mahraz, A.M.; Tairi, H. STCPU-Net: Advanced U-shaped deep learning architecture based on Swin transformers and capsule neural network for brain tumor segmentation. *Neural. Comput. Appl.* **2024**, *36*, 18549–18565. [CrossRef]
18. Mbarki, Z.; Ben Slama, A.; Amri, Y.; Trabelsi, H.; Seddik, H. BTS-ADCNN: Brain tumor segmentation based on rapid anisotropic diffusion function combined with convolutional neural network using MR images. *J. Supercomput.* **2024**, *80*, 13272–13294. [CrossRef]
19. Pan, D.; Shen, J.; Al-Huda, Z.; Al-Qaness, M.A.A. VcaNet: Vision Transformer with fusion channel and spatial attention module for 3D brain tumor segmentation. *Comput. Biol. Med.* **2025**, *186*, 109662. [CrossRef]
20. Zeineldin, R.A.; Karar, M.E.; Elshaer, Z.; Coburger, J.; Wirtz, C.R.; Burgert, O.; Mathis-Ullrich, F. Explainable hybrid vision transformers and convolutional network for multimodal glioma segmentation in brain MRI. *Sci. Rep.* **2024**, *14*, 3713. [CrossRef]
21. Tak, D.; Garomsa, B.A.; Chaunzwa, T.L.; Zapaishchykova, A.; Pardo, J.C.; Ye, Z.; Zielke, J.; Ravipati, Y.; Vajapeyam, S.; Mahootiha, M.; et al. A foundation model for generalized brain MRI analysis. *medRxiv* **2024**. [CrossRef]
22. Pacal, I. A novel Swin transformer approach utilizing residual multi-layer perceptron for diagnosing brain tumors in MRI images. *Int. J. Mach. Learn. Cybern.* **2024**, *15*, 3579–3597. [CrossRef]
23. Zeineldin, R.A.; Karar, M.E.; Elshaer, Z.; Coburger, J.; Wirtz, C.R.; Burgert, O.; Mathis-Ullrich, F. Explainability of deep neural networks for MRI analysis of brain tumors. *Int. J. Comput. Assist. Radiol. Surg.* **2022**, *17*, 1673–1683. [CrossRef] [PubMed]
24. Gaur, L.; Bhandari, M.; Razdan, T.; Mallik, S.; Zhao, Z. Explanation-driven deep learning model for prediction of brain tumour status using MRI image data. *Front. Genet.* **2022**, *13*, 822666. [CrossRef]
25. Hosny, K.M.; Mohammed, M.A.; Salama, R.A.; Elshewey, A.M. Explainable ensemble deep learning-based model for brain tumor detection and classification. *Neural. Comput. Appl.* **2025**, *37*, 1289–1306. [CrossRef]
26. Tabatabaei, S.; Rezaee, K.; Zhu, M. Attention transformer mechanism and fusion-based deep learning architecture for MRI brain tumor classification system. *Biomed. Signal Process. Control.* **2023**, *86*, 105119. [CrossRef]
27. Kang, J.; Ullah, Z.; Gwak, J. MRI-based brain tumor classification using ensemble of deep features and machine learning classifiers. *Sensors* **2021**, *21*, 2222. [CrossRef]
28. Djoumessi, K.; Mensah, S.O.; Berens, P. A Hybrid Fully Convolutional CNN-Transformer Model for Inherently Interpretable Medical Image Classification. *Front. Artif. Intell.* **2025**, *8*, 1679310. [CrossRef]
29. Pasvantis, K.; Protopapadakis, E. Enhancing deep learning model explainability in brain tumor datasets using post-heuristic approaches. *J. Imaging* **2024**, *10*, 232. [CrossRef]
30. Tummala, S.; Kadry, S.; Bukhari, S.A.C.; Rauf, H.T. Classification of brain tumor from magnetic resonance imaging using vision transformers ensembling. *Curr. Oncol.* **2022**, *29*, 7498–7511. [CrossRef]
31. Ferdous, G.J.; Sathi, K.A.; Hossain, A.; Hoque, M.M.; Dewan, M.A.A. LCDEiT: A linear complexity dataefficient image transformer for MRI brain tumor classification. *IEEE Access* **2023**, *11*, 20337–20350. [CrossRef]
32. Alnageeb, M.H.O.; Supriya, M.H. Real-time brain tumour diagnoses using a novel lightweight deep learning model. *Comput. Biol Med.* **2025**, *192*, 110242. [CrossRef] [PubMed]
33. Wang, J.; Lu, S.Y.; Wang, S.H.; Zhang, Y.D. RanMerFormer: Randomized vision transformer with token merging for brain tumor classification. *Neurocomputing* **2024**, *573*, 127216. [CrossRef]
34. Atiya, S.; Ali, T.; Irfan, M.; Khan, W.; Ahmed, H. BitMedViT: Ternary-Quantized Vision Transformer for Medical AI Assistants on the Edge. In Proceedings of the IEEE/ACM International Conference on Computer Aided Design (ICCAD 2025), Munich, Germany, 26–30 October 2025. [CrossRef]
35. Poornam, S.; Angelina, J.J.R. VITALT: A robust and efficient brain tumor detection system using vision transformer with attention and linear transformation. *Neural Comput. Appl.* **2024**, *36*, 6403–6419. [CrossRef]

36. Satushe, V.; Vyas, V.; Metkar, S.P.; Singh, D.P. Advanced cnn architecture for brain tumor segmentation and classification using brats-goat 2024 dataset. *Curr. Med. Imaging Former. Curr. Med. Imaging Rev.* **2025**, *21*, e15734056344235. [CrossRef]
37. Manthe, M.; Duffner, S.; Lartizien, C. Federated brain tumor segmentation: An extensive benchmark. *Med. Image Anal.* **2024**, *97*, 103270. [CrossRef]
38. Eker, A.G.; Pehlivanoğlu, M.K.; Duru, N.; Dündar, T.T. BrainPixGAN: Generating intraoperative MRI images with mask-based generative networks. *Eng. Sci. Technol. Int. J.* **2024**, *58*, 101827. [CrossRef]
39. Safdari, R.; Nikouei Mahani, M.-A.; Koohi-Moghadam, M.; Bae, K.T. MixStyleFlow: Domain Generalization in Medical Image Segmentation using Normalizing Flows. *Med. Image Comput. Comput. Assist. Interv. MICCAI* **2025**, *15962*, 376–385. [CrossRef]
40. Chen, H.; Qin, Z.; Ding, Y.; Tian, L.; Qin, Z. Brain tumor segmentation with deep convolutional symmetric neural network. *Neurocomputing* **2020**, *392*, 305–313. [CrossRef]
41. Kesav, N.; Jibukumar, M.G. Efficient and low complex architecture for detection and classification of brain tumor using RCNN with two-channel CNN. *J. King Saud Univ. Comput. Inf. Sci.* **2022**, *34*, 6229–6242. [CrossRef]
42. Hammad, M.; ElAffendi, M.; Ateya, A.A.; El-Latif, A.A.A. Efficient brain tumor detection with lightweight end-to-end deep learning model. *Cancers* **2023**, *15*, 2837. [CrossRef] [PubMed]
43. Hussain, T.; Shouno, H. Explainable deep learning approach for multi-class brain magnetic resonance imaging tumor classification and localization using gradient-weighted class activation mapping. *Information* **2023**, *14*, 642. [CrossRef]
44. Ullah, M.S.; Khan, M.A.; Masood, A.; Mzoughi, O.; Saidani, O.; Alturki, N. Brain tumor classification from MRI scans: A framework of hybrid deep learning model with Bayesian optimization and quantum theory-based marine predator algorithm. *Front. Oncol.* **2024**, *14*, 1335740. [CrossRef] [PubMed]
45. Asiri, A.A.; Shaf, A.; Ali, T.; Shakeel, U.; Irfan, M.; Mehdar, K.M.; Halawani, H.T.; Alghamdi, A.H.; Alshamrani, A.F.A.; Alqhtani, S.M. Exploring the power of deep learning: Fine-tuned vision transformer for accurate and efficient brain tumor detection in MRI scans. *Diagnostics* **2023**, *13*, 2094. [CrossRef]
46. Krishnan, P.T.; Krishnadoss, P.; Khandelwal, M.; Gupta, D.; Nihaal, A.; Kumar, T.S. Enhancing brain tumor detection in MRI with a rotation invariant Vision Transformer. *Front. Neuroinform.* **2024**, *18*, 1414925. [CrossRef]
47. Nassar, S.E.; Yasser, I.; Amer, H.M.; Mohamed, M.A. A robust MRI-based brain tumor classification via a hybrid deep learning technique. *J. Supercomput.* **2024**, *80*, 2403–2427. [CrossRef]
48. Rajput, I.S.; Gupta, A.; Jain, V.; Tyagi, S. A transfer learning-based brain tumor classification using magnetic resonance images. *Multimed. Tools Appl.* **2024**, *83*, 20487–20506. [CrossRef]
49. Zulfiqar, F.; Bajwa, U.I.; Mehmood, Y. Multi-class classification of brain tumor types from MR images using EfficientNets. *Biomed Signal Process. Control.* **2023**, *84*, 104777. [CrossRef]
50. Yan, F.; Chen, Y.; Xia, Y.; Wang, Z.; Xiao, R. An explainable brain tumor detection framework for mri analysis. *Appl. Sci.* **2023**, *13*, 3438. [CrossRef]
51. Islam, M.A.; Mridha, M.F.; Safran, M.S.; Alfarhood, S.; Kabir, M.M. Revolutionizing Brain Tumor Detection Using Explainable AI in MRI Images. *NMR Biomed.* **2025**, *38*, e70001. [CrossRef]
52. Cheng, J. Brain tumor dataset figshare. *Dataset* **2017**. [CrossRef]
53. Menze, B.H.; Jakab, A.; Bauer, S.; Kalpathy-Cramer, J.; Farahani, K.; Kirby, J.; Burren, Y.; Porz, N.; Slotboom, J.; Wiest, R.; et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging* **2015**, *34*, 1993–2024. [CrossRef] [PubMed]
54. Nickparvar, M. Brain tumor MRI dataset. *Kaggle* **2021**. [CrossRef]
55. Maqsood, S.; Damaševičius, R.; Maskeliūnas, R. Multi-Modal Brain Tumor Detection Using Deep Neural Network and Multiclass SVM. *Medicina* **2022**, *58*, 1090. [CrossRef]
56. Bakas, S.; Akbari, H.; Sotiras, A.; Bilello, M.; Rozycki, M.; Kirby, J.S.; Freymann, J.B.; Farahani, K.; Davatzikos, C. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data.* **2017**, *4*, 170117. [CrossRef]
57. Chakrabarty, N. Brain MRI Images for Brain Tumor Detection Dataset. 2019. Available online: https://www.kaggle.com/navoneel/brain-mri-images-for-brain-tumor-detection (accessed on 2 December 2025).
58. Reddy, C.K.K.; Reddy, P.A.; Janapati, H.; Assiri, B.; Shuaib, M.; Alam, S.; Sheneamer, A. A fine-tuned vision transformer-based enhanced multi-class brain tumor classification using MRI scan imagery. *Front. Oncol.* **2024**, *14*, 1400341. [CrossRef]
59. Lee, J.H.; Chae, J.W.; Cho, H.C. Improved classification of brain-tumor MRI images through data augmentation and filter application. *J. Electr. Eng. Technol.* **2023**, *18*, 3135–3142. [CrossRef]
60. Roy, P.; Srijon, F.M.S.; Bhowmik, P. An explainable ensemble approach for advanced brain tumor classification applying Dual-GAN mechanism and feature extraction techniques over highly imbalanced data. *PLoS ONE* **2024**, *19*, e0310748. [CrossRef]
61. SinhaRoy, R.; Sen, A. A Hybrid Deep Learning Framework to Predict Alzheimer's disease progression using generative adversarial networks and deep convolutional neural networks. *Arab. J. Sci. Eng.* **2024**, *49*, 3267–3284. [CrossRef]

62. Badža, M.M.; Barjaktarović, M.Č. Classification of brain tumors from MRI images using a convolutional neural network. *Appl. Sci.* **2020**, *10*, 1999. [CrossRef]
63. Toğaçar, M.; Ergen, B.; Cömert, Z. BrainMRNet: Brain tumor detection using magnetic resonance images with a novel convolutional neural network model. *Med. Hypotheses* **2020**, *134*, 109531. [CrossRef] [PubMed]
64. Huang, Z.; Du, X.; Chen, L.; Li, Y.; Liu, M.; Chou, Y.; Jin, L. Convolutional neural network based on complex networks for brain tumor image classification with a modified activation function. *IEEE Access* **2020**, *8*, 89281–89290. [CrossRef]
65. Ayadi, W.; Elhamzi, W.; Charfi, I.; Atri, M. Deep CNN for brain tumor classification. *Neural Process. Lett.* **2021**, *53*, 671–700. [CrossRef]
66. Abd El Kader, I.; Xu, G.; Shuai, Z.; Saminu, S.; Javaid, I.; Salim Ahmad, I. Differential deep convolutional neural network model for brain tumor classification. *Brain Sci.* **2021**, *11*, 352. [CrossRef]
67. Naseer, A.; Yasir, T.; Azhar, A.; Shakeel, T.; Zafar, K. Computer-aided brain tumor diagnosis: Performance evaluation of deep learner CNN using augmented brain MRI. *Int. J. Biomed. Imaging* **2021**, *2021*, 5513500. [CrossRef]
68. Musallam, A.S.; Sherif, A.S.; Hussein, M.K. A new convolutional neural network architecture for automatic detection of brain tumors in magnetic resonance imaging images. *IEEE Access* **2022**, *10*, 2775–2782. [CrossRef]
69. Kibriya, H.; Masood, M.; Nawaz, M.; Nazir, T. Multiclass classification of brain tumors using a novel CNN architecture. *Multimed. Tools Appl.* **2022**, *81*, 29847–29863. [CrossRef]
70. Khan, M.S.I.; Rahman, A.; Debnath, T.; Karim, M.R.; Nasir, M.K.; Band, S.S.; Mosavi, A.; Dehzangi, I. Accurate brain tumor detection using deep convolutional neural network. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 4733–4745. [CrossRef]
71. Ullah, N.; Khan, M.S.; Khan, J.A.; Choi, A.; Anwar, M.S. A robust end-to-end deep learning-based approach for effective and reliable BTD using MR images. *Sensors* **2022**, *22*, 7575. [CrossRef]
72. Saurav, S.; Sharma, A.; Saini, R.; Singh, S. An attention-guided convolutional neural network for automated classification of brain tumor from MRI. *Neural Comput. Appl.* **2023**, *35*, 2541–2560. [CrossRef]
73. Shahin, A.I.; Aly, W.; Aly, S. MBTFCN: A novel modular fully convolutional network for MRI brain tumor multi-classification. *Expert Syst. Appl.* **2023**, *212*, 118776. [CrossRef]
74. Özkaraca, O.; Bağrıaçık, O.İ.; Gürüler, H.; Khan, F.; Hussain, J.; Khan, J.; Laila, U.E. Multiple brain tumor classification with dense CNN architecture using brain MRI images. *Life* **2023**, *13*, 349. [CrossRef] [PubMed]
75. Mahmud, M.I.; Mamun, M.; Abdelgawad, A. A deep analysis of brain tumor detection from mr images using deep learning networks. *Algorithms* **2023**, *16*, 176. [CrossRef]
76. Ullah, N.; Javed, A.; Alhazmi, A.; Hasnain, S.M.; Tahir, A.; Ashraf, R. TumorDetNet: A unified deep learning model for brain tumor detection and classification. *PLoS ONE* **2023**, *18*, e0291200. [CrossRef] [PubMed]
77. Rehman, A.; Naz, S.; Razzak, M.I.; Akram, F.; Imran, M. A deep learning-based framework for automatic brain tumors classification using transfer learning. *Circuits Syst. Signal Process.* **2020**, *39*, 757–775. [CrossRef]
78. Sadad, T.; Rehman, A.; Munir, A.; Saba, T.; Tariq, U.; Ayesha, N.; Abbasi, R. Brain tumor detection and multiclassification using advanced deep learning techniques. *Microsc. Res. Tech.* **2021**, *84*, 1296–1308. [CrossRef]
79. Polat, Ö.; Güngen, C. Classification of brain tumors from MR images using deep transfer learning. *J. Supercomput.* **2021**, *77*, 7236–7252. [CrossRef]
80. Alanazi, M.F.; Ali, M.U.; Hussain, S.J.; Zafar, A.; Mohatram, M.; Irfan, M.; AlRuwaili, R.; Alruwaili, M.; Ali, N.H.; Albarrak, A.M. Brain tumor/mass classification framework using magnetic-resonance-imaging-based isolated and developed transfer deep-learning model. *Sensors* **2022**, *22*, 372. [CrossRef]
81. Ullah, N.; Khan, J.A.; Khan, M.S.; Khan, W.; Hassan, I.; Obayya, M.; Negm, N.; Salama, A.S. An effective approach to detect and identify brain tumors using transfer learning. *Appl. Sci.* **2022**, *12*, 5645. [CrossRef]
82. Sharma, A.K.; Nandal, A.; Dhaka, A.; Zhou, L.; Alhudhaif, A.; Alenezi, F.; Polat, K. Brain tumor classification using the modified ResNet50 model based on transfer learning. *Biomed. Signal Process. Control.* **2023**, *86*, 105299. [CrossRef]
83. Khushi, H.M.T.; Masood, T.; Jaffar, A.; Rashid, M.; Akram, S. Improved multiclass brain tumor detection via customized pretrained efficientnetb7 model. *IEEE Access* **2023**, *11*, 117210–117230. [CrossRef]
84. Ghosh, A.; Soni, B.; Baruah, U. Transfer learning-based deep feature extraction framework using finetuned efficientNet B7 for multiclass brain tumor classification. *Arab. J. Sci. Eng.* **2023**, *49*, 12027–12048. [CrossRef]
85. Mathivanan, S.K.; Sonaimuthu, S.; Murugesan, S.; Rajadurai, H.; Shivahare, B.D.; Shah, M.A. Employing deep learning and transfer learning for accurate brain tumor detection. *Sci. Rep.* **2024**, *14*, 7232. [CrossRef]
86. Zubair Rahman, A.M.J.; Gupta, M.; Aarathi, S.; Mahesh, T.R.; Vinoth Kumar, V.; Yogesh Kumaran, S.; Guluwadi, S. Advanced AI-driven approach for enhanced brain tumor detection from MRI images utilizing EfficientNetB2 with equalization and homomorphic filtering. *BMC Med. Inform. Decis. Mak.* **2024**, *24*, 113. [CrossRef] [PubMed]
87. İncir, R.; Bozkurt, F. Improving brain tumor classification with combined convolutional neural networks and transfer learning. *Knowledge-Based Syst.* **2024**, *299*, 111981. [CrossRef]

88. Preetha, R.; Priyadarsini, M.J.P.; Nisha, J.S. Automated brain tumor detection from magnetic resonance images using fine-tuned efficientnet-b4 convolutional neural network. *IEEE Access* **2024**, *12*, 112181–112195. [CrossRef]
89. Mavaddati, S. Brain tumors classification using deep models and transfer learning. *Multimed. Tools Appl.* **2024**, *84*, 25677–25708. [CrossRef]
90. Khushi, H.M.T.; Masood, T.; Jaffar, A.; Akram, S. A novel approach to classify brain tumor with an effective transfer learning based deep learning model. *Braz. Arch. Biol. Technol.* **2024**, *67*, e24231137. [CrossRef]
91. Hameed, M.; Zameer, A.; Khan, S.H.; Raja, M.A.Z. ARiViT: Attention-based residual-integrated vision transformer for noisy brain medical image classification. *Eur. Phys. J. Plus* **2024**, *139*, 440. [CrossRef]
92. Hong, S.; Wu, J.; Zhu, L.; Chen, W. Brain tumor classification in VIT-B/16 based on relative position encoding and residual MLP. *PLoS ONE* **2024**, *19*, e0298102. [CrossRef]
93. Şahin, E.; Özdemir, D.; Temurtaş, H. Multi-objective optimization of ViT architecture for efficient brain tumor classification. *Biomed. Signal Process. Control.* **2024**, *91*, 105938. [CrossRef]
94. 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.; et al. Advancing brain tumor detection: Harnessing the Swin transformer's power for accurate classification and performance analysis. *PeerJ Comput. Sci.* **2024**, *10*, e1867. [CrossRef] [PubMed]
95. Priya, A.; Vasudevan, V. Brain tumor classification and detection via hybrid Alexnet-gru based on deep learning. *Biomed. Signal Process. Control.* **2024**, *89*, 105716. [CrossRef]
96. Singh, A.; Shrivastava, R.K.; Srivastava, A. Efficient and compressed deep learning model for brain tumour classification with explainable AI for smart healthcare and information communication systems. *Expert Syst.* **2025**, *42*, e13770. [CrossRef]
97. Bizzo, B.C.; Almeida, R.R.; Michalski, M.; Alkasab, T.K. Artificial intelligence and clinical decision support for radiologists and referring providers. *J. Am. Coll. Radiol.* **2019**, *16*, 1351–1356. [CrossRef]
98. Heckel, R.; Jacob, M.; Chaudhari, A.; Perlman, O.; Shimron, E. Deep learning for accelerated and robust MRI reconstruction. *Magn. Reson. Mater. Phys. Biol. Med.* **2024**, *37*, 335–368. [CrossRef]
99. Monisha, S.M.A.; Rahman, R. Brain Tumor Detection in MRI Based on Federated Learning with YOLOv11. *arXiv* **2025**, arXiv:2503.04087. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.