



VGG19 in MRI Brain Tumor Classification: A Systematic Review of Validation Rigor and Clinical Readiness

Zhandark. Khalife Soltani¹, Mehdi. Sadeghzadeh^{2*}, Abbas. Toloie Eshlaghy³

¹ Department of Information Technology Management, SR.C., Islamic Azad University, Tehran, Iran

² Department of Computer Engineering, SR.C., Islamic Azad University, Tehran, Iran

³ Department of Industrial Management, SR.C., Islamic Azad University, Tehran, Iran

* Corresponding author email address: mehdi.sadeghzadeh@iau.ac.ir

Article Info

ABSTRACT

Article type:

Original Research

Section:

Emerging Technologies in Health

How to cite this article:

Khalife Soltani, Z., Sadeghzadeh, M., & Toloie Eshlaghy, A. (2027). VGG19 in MRI Brain Tumor Classification: A Systematic Review of Validation Rigor and Clinical Readiness. *Health Nexus*, 5(1), 1-22.

<https://doi.org/10.61838/kman.hn.6426>



© 2027 the authors. Published by KMAN Publication Inc. (KMANPUB), Ontario, Canada. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

VGG19 remains a widely used reference backbone for MRI-based brain tumor classification; however, the literature is difficult to interpret because model modifications, dataset provenance, data-splitting strategies, and evaluation practices vary substantially. This systematic review provides a VGG19-centered evidence map linking architectural adaptations to validation rigor and indicators of clinical readiness. The review was updated in accordance with PRISMA 2020 principles. Searches were conducted in PubMed, IEEE Xplore, ScienceDirect, and SpringerLink, supplemented by citation searching, to identify peer-reviewed studies published from January 2015 through December 2025; the search was last updated on July 27, 2026. Because the original database export files were not retained, database-specific retrieval and duplicate counts are not reported. The final evidence base was reconciled at the reference level: 50 unique references were classified, 20 contextual or methodological sources were separated, and 30 empirical studies were included. Data were extracted on classification task, MRI source, class structure, role of VGG19, preprocessing procedures, splitting unit, validation design, performance metrics, interpretability, and reproducibility. Reporting quality was appraised using the CLAIM 2024 domains. Results were not statistically pooled because of substantial outcome heterogeneity; instead, a quantitative descriptive synthesis and structured comparative evidence map were used. Research output increased after 2021. VGG19 was employed as a stand-alone classifier, transfer-learning feature extractor, fine-tuned backbone, hybrid-model component, ensemble member, and component of explainable and federated-learning frameworks. All 30 studies reported accuracy or an accuracy-derived outcome; 25 reported class-specific metrics, 12 employed cross-validation, six included external or distributed validation, seven incorporated explainability methods, and four provided code or other implementation materials. High classification accuracies were frequently reported for Figshare and related contrast-enhanced T1-weighted MRI datasets; however, the strength of the evidence was often limited by image-level data splitting, incomplete reporting of patient provenance, lack of calibration assessment, and limited multicenter validation. The principal contribution of this review is a three-axis synthesis demonstrating that architectural improvements should be interpreted jointly with validation rigor and evidence of clinical readiness. Fine-tuned and hybrid VGG19-based systems represent competitive baselines; however, reliable clinical translation requires patient-level data separation, external validation, class-specific performance reporting, calibration assessment, explainability, and transparent documentation of computational methods and implementation.

Keywords: brain tumor classification; magnetic resonance imaging; VGG19; transfer learning; systematic review; explainable artificial intelligence

1. Introduction

Brain tumors constitute a heterogeneous group of intracranial neoplasms that differ substantially in their cellular origin, anatomical location, biological aggressiveness, imaging appearance, prognosis, and therapeutic requirements. Accurate differentiation among tumor types and grades is therefore a critical component of clinical decision-making, because diagnostic classification directly influences surgical planning, treatment selection, surveillance strategies, and estimation of patient prognosis. Magnetic resonance imaging (MRI) has become one of the principal imaging modalities for brain tumor assessment because of its high soft-tissue contrast, multiplanar capability, absence of ionizing radiation, and ability to characterize complementary anatomical and pathological properties through multiple sequences. T1-weighted, contrast-enhanced T1-weighted, T2-weighted, and fluid-attenuated inversion recovery images provide different information regarding tumor morphology, enhancement, edema, necrosis, infiltration, and surrounding brain tissue. Large-scale initiatives such as the Brain Tumor Segmentation benchmark and associated glioma MRI collections have also supported the development of standardized imaging resources and expert-generated annotations for computational research (1, 2). Nevertheless, manual interpretation of MRI remains demanding because tumor appearance may vary markedly within and between diagnostic categories, while overlapping radiological characteristics can complicate discrimination among clinically distinct lesions. These challenges have stimulated extensive interest in computer-assisted methods capable of extracting quantitative imaging features and supporting automated tumor detection, segmentation, classification, and grading.

The emergence of deep learning has substantially transformed medical image analysis by replacing or complementing manually engineered feature extraction with hierarchical representation learning. Convolutional neural networks (CNNs) can automatically learn increasingly complex spatial representations from imaging data, with early convolutional layers typically identifying local structures such as edges and textures and deeper layers capturing more task-specific morphological configurations. The broader medical-imaging literature has demonstrated

the ability of deep neural networks to support detection, segmentation, classification, registration, and other imaging tasks across numerous specialties (3). Subsequent developments have expanded this paradigm toward deeper architectures, multimodal learning, attention mechanisms, transformers, ensembles, and increasingly sophisticated strategies for improving robustness and generalization (4). In brain tumor analysis specifically, CNN-based frameworks have shown considerable potential for distinguishing tumor types directly from MRI images. Early work demonstrated that tumor-region augmentation and partitioning could improve discrimination by emphasizing diagnostically relevant portions of the image (5), while later CNN-based studies reported strong classification performance across multiple brain tumor categories (6, 7). These developments established deep learning as a major methodological direction in MRI-based neuro-oncology research.

Despite the effectiveness of deep CNNs, medical imaging datasets are generally much smaller than the natural-image datasets used to train modern computer-vision architectures. Transfer learning has consequently become one of the dominant strategies for MRI classification. In transfer learning, a network pretrained on a large source dataset is repurposed for a target clinical task either by using the pretrained convolutional layers as fixed feature extractors or by selectively fine-tuning some or all network weights. The rationale is that low- and intermediate-level representations learned from large image collections may remain useful despite the substantial domain difference between photographs and medical images. Fine-tuning then enables higher-level representations to adapt to disease-specific morphology and texture. Numerous studies have applied this strategy to brain tumor MRI classification, demonstrating that pretrained architectures can provide effective feature representations even when labeled medical datasets are comparatively limited (8-11). However, reported improvements depend not only on the selected backbone but also on the depth of fine-tuning, augmentation strategy, optimizer, regularization, input resolution, class distribution, and data-partitioning protocol. Consequently, high classification accuracy alone cannot establish that one transfer-learning architecture is inherently superior to another.

Among pretrained CNN architectures, VGG19 has remained particularly prominent as a reference backbone for brain MRI classification. Its architecture is characterized by a regular sequence of small 3×3 convolutional filters arranged into progressively deeper blocks, interspersed with spatial downsampling. This relatively uniform organization facilitates feature extraction, selective fine-tuning, architectural modification, and controlled comparison with alternative CNNs. Although newer residual, densely connected, and transformer-based models introduce mechanisms that can improve optimization or long-range feature modeling, VGG19 remains attractive because of its mature pretrained weights, straightforward implementation, well-understood internal structure, and compatibility with multiple downstream classifiers and fusion strategies. Brain tumor studies have therefore used VGG19 not only as a direct classifier but also as a feature extractor, a fine-tuned backbone, a component of hybrid architectures, and a comparator against other transfer-learning models. For example, ROI-oriented preprocessing and synthetic data augmentation have been integrated with deep architectures to improve tumor classification (12), while comparative transfer-learning studies have continued to include VGG19 as an important reference model against more recent architectures (9, 11). This sustained use makes VGG19 particularly suitable for examining the relationship between architectural adaptation and methodological rigor over time.

The literature has increasingly moved beyond direct application of pretrained CNNs toward hybrid and ensemble designs. Hybrid systems combine representations from deep networks with other classifiers, feature streams, segmentation components, or decision mechanisms in an attempt to exploit complementary information. Rasool et al., for example, demonstrated the potential of combining deep features with machine-learning classification for brain tumor recognition (13), while Veeramuthu et al. integrated image-based and feature-based classification to improve MRI tumor categorization (14). Ensemble methods similarly seek to reduce model-specific errors by combining predictions or learned representations from multiple networks. Aurna et al. proposed a two-stage feature-level ensemble of deep CNN models and reported high performance across multiple brain MRI datasets (15), whereas Nassar et al. applied a hybrid deep-learning strategy incorporating multiple predictive

components (16). Such methods demonstrate the expanding flexibility of deep-learning systems, but they also introduce greater parameter counts, more hyperparameters, increased computational requirements, and additional opportunities for overfitting. Consequently, improvements in headline accuracy should be balanced against model complexity, inference cost, reproducibility, and the strength of the validation procedure.

Recent research has also incorporated transformer-based models, multimodal information fusion, and joint segmentation-classification frameworks. Vision Transformers provide an alternative to conventional convolution by modeling global relationships between image regions and have been investigated in ensemble settings for brain tumor MRI classification (17). More recent systems have combined transformer-based representations, transfer learning, ensemble modeling, and explainable AI to improve both predictive performance and model transparency (18). Multimodal approaches attempt to exploit complementary information available across MRI sequences rather than treating all scans as equivalent representations. Rohini et al. investigated a multimodal hybrid CNN framework for brain tumor grade classification (19), while Ullah et al. integrated optimized segmentation, multimodal classification, information fusion, and explainable AI within a unified framework (20). These developments are technically important because they move the field toward richer representations of tumor phenotype; nevertheless, increasing architectural sophistication does not eliminate fundamental concerns regarding sampling, validation, data leakage, reproducibility, and external generalizability.

Interpretability has become another increasingly important dimension of medical artificial intelligence. Deep-learning predictions may be difficult for clinicians to evaluate when the features responsible for a decision are not visible or intuitively understandable. Gradient-based visualization methods such as Grad-CAM were developed to identify image regions that contribute strongly to a network prediction and have become widely used for examining whether model attention corresponds to diagnostically meaningful anatomy (21). Attention mechanisms provide another approach by adaptively weighting informative channels or spatial regions. The convolutional block attention module, for example, sequentially models channel

and spatial attention and illustrates how feature emphasis can be incorporated efficiently into CNN architectures (22). Brain tumor classification studies have increasingly incorporated explanation methods into their systems. NeuroNet19 combined a VGG19-related design with explainable analysis to support interpretation of brain tumor predictions (23), and Alshuhail et al. incorporated explainability into a refined neural-network framework for MRI classification (24). Kothadiya et al. further proposed a gradient-explainer architecture explicitly associated with VGG-based brain tumor detection (25). These studies represent an important progression beyond accuracy-only evaluation; however, saliency and attention maps should be interpreted as evidence of association with model output rather than proof of causal reasoning, diagnostic correctness, or clinical safety.

The validity of reported performance also depends critically on dataset provenance and partitioning. Brain MRI datasets frequently contain multiple slices from the same patient. If images rather than patients are randomly divided into training and test sets, anatomically similar slices originating from the same individual may appear in both partitions. Such leakage can inflate apparent predictive performance because the model is evaluated partly on information highly related to its training observations. The same concern applies when augmentation is conducted before data splitting or when nominally separate public repositories contain duplicate or repackaged versions of the same images. Studies involving ROI extraction, augmentation, preprocessing, or complex feature engineering must therefore ensure that all data-dependent operations are performed within appropriately separated training data and that test data remain independent. Approaches based on segmentation, localization, or deep feature extraction, including those proposed by Nawaz et al. and other MRI classification studies, illustrate the diversity of processing pipelines whose validity ultimately depends on rigorous partitioning and transparent reporting (26). Thus, methodological rigor cannot be inferred from architecture alone.

Generalizability is especially important because MRI data are sensitive to acquisition-related variation. Scanner manufacturer, magnetic field strength, coil configuration, pulse-sequence parameters, reconstruction methods, slice

thickness, image resolution, and institutional preprocessing can all alter image distributions. A model trained on a single curated dataset may therefore perform substantially worse when applied to scans from another hospital. Public datasets are indispensable for benchmarking, but repeated development and testing on the same repositories can encourage optimization to dataset-specific characteristics rather than clinically generalizable tumor features. Cross-dataset testing, site-level separation, multicenter validation, and prospective evaluation are therefore necessary steps toward credible clinical translation. Recent research trends toward multimodal and more diverse datasets are encouraging, but the distinction between internal performance and external clinical generalization remains fundamental. High performance reported on a limited retrospective dataset should be interpreted as evidence of algorithmic feasibility rather than proof of readiness for routine diagnostic use.

This distinction aligns with the broader transition in medical AI from performance-centered evaluation toward trustworthy and deployable systems. Transparent reporting standards have long emphasized that predictive models require clear descriptions of participant selection, predictors, outcomes, validation, and model performance if they are to be critically appraised and reproduced (27). For systematic reviews, PRISMA 2020 provides an updated framework for transparent reporting of information sources, selection processes, eligibility criteria, and synthesis procedures (28). More specifically for artificial intelligence in medical imaging, the updated Checklist for Artificial Intelligence in Medical Imaging emphasizes clear description of datasets, reference standards, partitions, model development, performance assessment, error analysis, and validation (29). These reporting principles are directly relevant to brain tumor classification because numerical performance values are difficult to interpret when patient counts, split units, model-selection procedures, or external testing are incompletely described.

The concept of clinical readiness extends beyond discrimination performance. Trustworthy medical AI should also address robustness, fairness, usability, explainability, traceability, reproducibility, safety, and the capacity to function reliably under real-world distribution shifts. The FUTURE-AI international consensus guideline reflects this

broader orientation toward trustworthy and deployable healthcare AI and emphasizes that technical performance is only one component of responsible clinical implementation (30). In MRI brain tumor classification, this means that a clinically meaningful system should ideally demonstrate patient-level independence, strong class-wise performance, calibrated probability estimates, external or multicenter validation, transparent error analysis, explainability where appropriate, and sufficient implementation detail to support replication. Such evidence remains more limited than the abundance of studies reporting very high internal accuracy. Therefore, the most important question is increasingly not whether VGG19 or a related architecture can achieve a high score on a particular dataset, but whether that score is supported by a validation design capable of estimating performance in unseen patients and settings.

The growing diversity of VGG19 applications makes this issue particularly timely. VGG19 now occupies several distinct methodological roles: it may be used unchanged as a benchmark, employed as a frozen feature extractor, selectively fine-tuned, integrated with handcrafted or deep features, combined within ensembles, modified through attention or pooling mechanisms, or paired with explainability frameworks. Across this continuum, newer studies have reported increasingly sophisticated architectures and high classification performance (10, 11, 16, 23, 25). Yet comparisons between these systems are often complicated by differences in class definitions, MRI sequences, preprocessing, dataset provenance, augmentation, patient counts, validation strategy, and metric reporting. A model evaluated with patient-level cross-validation or independent external testing provides a qualitatively different form of evidence from a model evaluated through a random image-level hold-out split, even when both report similar accuracy. A synthesis centered solely on numerical performance would therefore risk conflating architectural sophistication with evidential strength.

Accordingly, a systematic review of VGG19 in MRI brain tumor classification should evaluate the architecture within the broader methodological ecosystem in which its results are produced. This requires simultaneous consideration of the clinical task, imaging source, dataset independence, role of VGG19, preprocessing and

augmentation procedures, validation unit, cross-validation design, external testing, class-wise metrics, explainability, reproducibility, and indicators of clinical readiness. By integrating architectural taxonomy with validation rigor and reporting quality, it becomes possible to distinguish technically promising systems from those supported by evidence more compatible with clinical translation. Such an approach also provides a more informative basis for future research than simple ranking by maximum reported accuracy because it identifies where methodological improvements—not merely architectural modifications—are most urgently required.

Therefore, the aim of this systematic review was to critically synthesize the use of VGG19 in MRI-based brain tumor classification by mapping its architectural adaptations, datasets and clinical tasks, validation practices, performance reporting, explainability and reproducibility characteristics, and evidence of clinical readiness.

2. Methods and Materials

This study was designed as a systematic review and structured evidence-mapping study of the use of VGG19 in magnetic resonance imaging (MRI)-based brain tumor classification. The review was updated in accordance with the PRISMA 2020 reporting principles to ensure transparent identification, screening, eligibility assessment, and inclusion of relevant studies. Because the available literature differed substantially in terms of clinical task, tumor class definitions, MRI datasets and sequences, sampling units, preprocessing procedures, model architectures, data-partitioning strategies, and validation protocols, statistical pooling of classification accuracy was considered methodologically inappropriate. Accordingly, the study employed a systematic evidence-map approach combined with quantitative descriptive synthesis. The reporting and methodological strength of the included studies was additionally evaluated according to relevant domains of the CLAIM 2024 framework, particularly those concerning data-source description, reference standards, data partitioning, model specification and training, performance evaluation, error analysis, external testing, explainability, and reproducibility. The scope of the review was defined before the updated search and was restricted to studies using human brain MRI for tumor detection, tumor-type

classification, tumor grading, or combined analytical tasks that included a classification or grading outcome. Studies were eligible when they explicitly employed VGG19, a modification or extension of VGG19, or a directly relevant convolutional neural network comparator in a context that contributed to evaluation of VGG19-based approaches. Only peer-reviewed primary research published in English between January 1, 2015, and December 31, 2025, was considered eligible. The updated literature search was completed on July 27, 2026, to allow the identification of publications from the predefined 2025 endpoint that may have been indexed after their initial publication.

The eligible sample consisted of peer-reviewed journal articles and full conference papers based on human brain MRI datasets containing labeled tumor or non-tumor classes and reporting at least one quantitative model-performance measure. Eligible outcomes included accuracy, sensitivity, specificity, precision, recall, F1-score, area under the receiver operating characteristic curve, or other clearly defined quantitative classification or grading metrics. Studies based exclusively on animal, phantom, non-brain, or non-MRI data were excluded. Studies employing only conventional image-processing or machine-learning techniques without a relevant deep-learning comparator were also excluded, as were studies addressing segmentation alone without reporting a tumor-classification or grading outcome. Reviews, editorials, conference abstracts without full papers, theses, dissertations, and preprint-only reports were not considered primary evidence. The final evidence base was reconciled at the reference level rather than reconstructed from unverifiable historical database counts. The revised corpus contained 50 unique references with no duplicate entries remaining in the final bibliography. Of these, 20 were classified as contextual or methodological sources, including reporting guidelines, foundational deep-learning architectures, dataset descriptions, explainability methods, generalization studies, and a historical non-deep-learning comparator. These sources informed the conceptual and methodological framework of the review but were not treated as primary empirical evidence. The remaining 30 publications were full-text empirical studies of MRI-based brain tumor analysis that satisfied the predefined eligibility criteria and constituted the final study sample for the qualitative and quantitative descriptive synthesis. Because

the original database export files had not been retained, database-specific retrieval counts, duplicate-removal counts from the initial searches, and non-auditable exclusion counts were not reconstructed or reported. Instead, methodological transparency was maintained through explicit eligibility criteria, exact database search procedures, a defined publication window and search date, a complete final bibliography, and study-level extraction and coding of all empirical reports included in the synthesis.

Data collection was conducted through a structured literature-search and evidence-extraction procedure. The principal information sources were PubMed, IEEE Xplore, ScienceDirect, and SpringerLink, selected to provide complementary coverage of biomedical research, medical imaging, electrical and computer engineering, artificial intelligence, and computer-science literature. The electronic database search was supplemented by backward and forward citation searching of eligible studies and relevant methodological publications, with candidate references checked at the title and digital object identifier level when available. The search strategy was organized around four conceptual domains: neural-network architecture, brain-tumor condition, MRI modality, and classification or grading task. Search syntax was adapted to the technical requirements of each database while maintaining conceptual equivalence across information sources. In PubMed, the principal search combined terms referring to “VGG19,” “Visual Geometry Group,” or convolutional neural networks with terms for brain tumors, glioma, meningioma, and pituitary tumors, MRI or magnetic resonance imaging, and classification or grading. The IEEE Xplore search similarly combined metadata terms for VGG19 or convolutional neural networks with brain tumor, glioma, or meningioma, MRI, and classification or grading. ScienceDirect was searched using combinations of VGG19 or convolutional neural network terminology with brain tumor, glioma, meningioma, or pituitary tumor terminology, MRI or magnetic resonance imaging, and classification or grading terms. SpringerLink was searched using combinations of VGG19 or CNN with brain tumor, glioma, meningioma, or pituitary tumor, MRI, and classification or grading. Citation searching was performed by reviewing the reference lists of eligible studies and examining publications that cited key empirical and methodological studies. The search strategy

was intentionally broad enough to identify studies in which VGG19 was the principal model, a transfer-learning backbone, a feature extractor, an ensemble component, a hybrid architecture component, or a comparator relevant to the review question.

A standardized data-extraction form served as the primary data-collection instrument for the evidence synthesis. For every included empirical study, information was recorded on bibliographic characteristics, year of publication, clinical or computational task, MRI source, MRI sequence where reported, dataset provenance, sample characteristics, number and definition of diagnostic classes, role of VGG19 within the modeling pipeline, preprocessing procedures, image augmentation, optimization approach, training configuration, validation strategy, splitting unit, evaluation metrics, best reported performance outcome, use of interpretability or explainability methods, external or cross-dataset validation, distributed or multicenter evaluation, and availability of source code or other implementation materials. When an article evaluated several neural-network architectures, the VGG19 configuration or modification most relevant to the review objective was extracted in detail; when appropriate, the best clearly identifiable comparator was also recorded to preserve the within-study comparative context. Performance estimates were not detached from their underlying validation design, because identical numerical accuracy values may reflect substantially different evidential strength depending on whether the test set was created using image-level random partitioning, patient-level separation, cross-validation, external datasets, or multicenter data. Consequently, validation practices were coded descriptively into categories representing hold-out validation, cross-validation, explicit patient-level separation, cross-dataset or external validation, and distributed or multicenter evaluation. This validation-strength coding was used exclusively as an evidence-mapping device and was not interpreted as a formal risk-of-bias scale or as a numerical measure of clinical performance.

A structured reporting-strength appraisal was also incorporated into the data-collection process using domains derived from CLAIM 2024. Rather than calculating a composite numerical quality score, each study was examined for the transparency and completeness of information necessary to interpret its findings. The extracted appraisal

domains included clarity of the MRI data source, reporting of patient and image counts, definition and distribution of diagnostic classes, description of the clinical or annotation reference standard, transparency regarding the unit used for training-validation-test partitioning, specification of the neural-network architecture and modifications, description of training and optimization procedures, reporting of class-specific performance measures, error or failure analysis, use of internal or external validation, incorporation of explainability methods, and provision of source code, trained models, supplementary implementation details, or other reproducibility resources. This domain-based approach was selected because the purpose of the review was not merely to identify the numerically highest-performing VGG19 configuration, but to determine whether reported architectural advances were supported by validation practices and reporting standards compatible with reproducible and clinically meaningful evidence.

Data analysis consisted of qualitative comparative synthesis, quantitative descriptive analysis, and structured evidence mapping. The extracted study characteristics were first organized according to publication year, clinical task, dataset family, class structure, MRI source, VGG19 configuration, preprocessing strategy, validation procedure, evaluation metrics, use of explainability, availability of external validation, and reproducibility resources. Frequencies and proportions were calculated to characterize temporal publication patterns, major dataset families, types of VGG19 implementation, coverage of performance metrics, use of cross-validation, external or distributed validation practices, explainability methods, and availability of implementation materials. VGG19 applications were categorized according to their functional role, including stand-alone classifier, transfer-learning feature extractor, fine-tuned pretrained backbone, component of a hybrid architecture, ensemble member, and component of explainable, federated, or distributed-learning systems. Performance outcomes were interpreted together with the characteristics of the underlying task and validation design rather than treated as directly comparable estimates across studies. In particular, reported accuracy values were retained in their original study context because accuracy is a bounded performance measure whose interpretation depends on class prevalence, number of diagnostic categories, test-set

construction, dataset difficulty, patient independence, and data-partitioning procedures. Accordingly, no pooled mean accuracy, meta-analytic effect size, or aggregated superiority ranking was calculated.

Comparative interpretation focused on three interrelated dimensions: architectural adaptation, validation rigor, and clinical-readiness evidence. Architectural adaptation was assessed according to how VGG19 was incorporated or modified, including fine-tuning, feature extraction, hybridization, ensembling, integration with explainability techniques, and implementation in federated or distributed environments. Validation rigor was assessed from the explicitly reported experimental design, particularly the use of cross-validation, patient-level separation, independent external datasets, cross-dataset testing, or multicenter evaluation. Clinical readiness was evaluated descriptively using indicators such as transparent patient provenance, class-wise performance reporting, calibration assessment, interpretability or explainability, external validation, multicenter or distributed evaluation, error analysis, and availability of sufficient computational documentation to support replication. Direct statements that one model outperformed another were restricted to comparisons conducted within the same study using the same dataset, preprocessing procedure, partition, and evaluation protocol. Numerical results obtained from different studies were therefore described comparatively but were not interpreted as evidence of formal superiority when the underlying experimental conditions differed. The final synthesis integrated these dimensions into a structured evidence map intended to distinguish high reported predictive performance from stronger evidence of generalizability and clinical

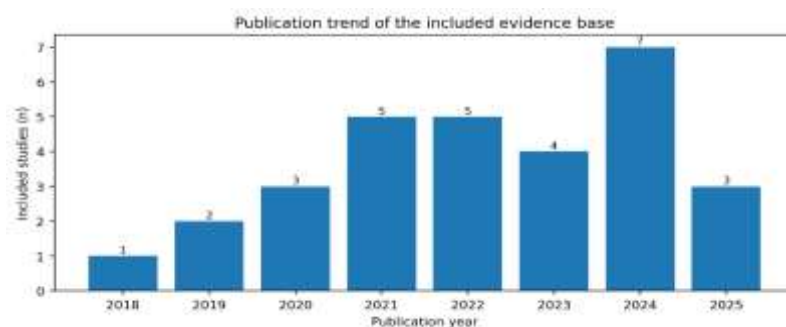
applicability. This analytical strategy enabled the review to identify not only which VGG19-based configurations achieved high reported classification performance, but also whether those results were supported by sufficiently rigorous validation, transparent reporting, and reproducibility practices to justify progression toward clinical evaluation.

3. Findings and Results

The final audited evidence corpus comprised 30 eligible empirical studies addressing deep-learning-based analysis of brain tumors on magnetic resonance imaging (MRI), with particular emphasis on VGG19 as a classifier, transfer-learning backbone, feature extractor, hybrid component, ensemble member, or methodological benchmark. The evidence base showed a clear increase in publication activity after 2021. Earlier studies primarily investigated conventional convolutional neural networks (CNNs), pretrained feature extraction, and comparatively simple transfer-learning approaches, whereas more recent studies increasingly incorporated fine-tuning, multi-model fusion, attention mechanisms, transformer-assisted architectures, explainable artificial intelligence (XAI), and federated learning. This temporal pattern demonstrates increasing methodological complexity and diversification within the field, but it should not be interpreted as equivalent to increased clinical maturity. A substantial proportion of recent studies continued to depend on retrospective public datasets, internal hold-out testing, or incompletely documented patient-level partitioning.

Figure 1

Publication trend of the 30 included empirical studies by year.



The descriptive profile of the evidence base further demonstrated an imbalance between reported predictive performance and the methodological features required to establish generalizability. Accuracy or an accuracy-derived outcome was reported in all 30 studies, whereas only 25 provided class-wise performance metrics, 12 used cross-validation, seven explicitly described patient- or site-level separation, six incorporated external, cross-dataset, or

distributed validation, seven included an explainability component, and only four reported code or sufficiently explicit implementation materials. Calibration was not systematically documented in the extraction matrix. These findings indicate that relatively minor differences in headline accuracy should be interpreted cautiously when they arise from studies with substantially different validation designs.

Table 1

Quantitative profile of the included evidence base

Characteristic	n	Interpretation
Included empirical studies	30	Final audited primary evidence base
Studies using VGG19 as a primary backbone or feature source	11	Direct evidence relevant to the principal review focus
Studies using cross-validation	12	Provides stronger internal repeatability than a single hold-out partition
Studies with external, cross-dataset, or distributed validation	6	Provides direct evidence regarding generalization beyond one development dataset
Studies with explicit patient-level or site-level separation	7	Reduces the risk of slice-level or subject-level information leakage
Studies reporting class-wise metrics	25	Provides more clinically informative evaluation than overall accuracy alone
Studies incorporating explainability	7	Supports plausibility assessment, visualization, and error analysis
Studies reporting code or implementation materials	4	Reproducibility resources remained uncommon; reported links were not independently tested
Studies reporting accuracy or an accuracy-derived outcome	30	Demonstrates strong dependence on accuracy as the dominant evaluation metric
Studies with systematic calibration reporting	Not systematically identified	Calibration remains a major clinical-readiness gap

Three-class classification of glioma, meningioma, and pituitary tumors using contrast-enhanced T1-weighted MRI was the most common task. However, the included studies were not homogeneous with respect to either diagnostic objective or outcome definition. Other investigations examined binary tumor detection, benign-versus-malignant discrimination, four-class classification incorporating normal MRI images, glioma grading, and combined segmentation-classification frameworks. Because task difficulty, class prevalence, class number, and disease spectrum differed markedly across studies, a single ranking based on reported accuracy would be methodologically misleading. An accuracy of 99% in a binary or highly curated three-class dataset is not directly comparable with a similar value obtained in a four-class, multimodal, multicenter, or grading task.

Figshare and derivative contrast-enhanced T1-weighted collections were the most frequently used sources of imaging

data. Their widespread use has facilitated comparison between algorithms and supported rapid experimentation, particularly for glioma, meningioma, and pituitary tumor classification. However, repeated reuse of the same underlying images, or repackaged versions of them across different repositories, creates an important methodological concern. A dataset described as an independent external source may not necessarily be fully independent if duplicate or derivative images are shared with the development dataset. This possibility represents an under-recognized threat to external validation.

BraTS datasets were more commonly used in glioma-focused studies, especially multimodal investigations involving T1-weighted, contrast-enhanced T1-weighted, T2-weighted, and FLAIR sequences and tasks integrating tumor localization, segmentation, or grading. Kaggle-based and mixed public repositories were frequently used for rapid benchmarking. Institutional datasets provided more

clinically realistic heterogeneity but often showed weaker reproducibility because acquisition provenance, scanner characteristics, patient composition, annotation procedures, and preprocessing pipelines were less completely documented.

The imaging sequence itself represented an important source of methodological heterogeneity. T1-weighted images primarily provide anatomical information, whereas contrast-enhanced T1-weighted imaging emphasizes regions associated with blood-brain-barrier disruption and enhancement. T2-weighted and FLAIR imaging emphasize water content, edema, and non-enhancing tumor

components. Consequently, a classifier developed using only contrast-enhanced T1-weighted images addresses a different representation problem from a multimodal system. Scanner vendor, field strength, coil configuration, sequence parameters, slice thickness, reconstruction procedures, and post-processing further influence MRI intensity distributions and may create substantial domain shift between institutions. Strong internal performance therefore cannot be assumed to transfer to external clinical environments without multicenter testing, harmonization, or explicit domain-generalization procedures.

Table 2

Principal task, dataset, and VGG19 adaptation categories identified across the evidence base

Analytical dimension	Major categories identified	Principal interpretation
Classification task	Three-class glioma/meningioma/pituitary classification; binary tumor detection; four-class classification including normal MRI; benign/malignant classification; glioma grading; joint segmentation-classification	Heterogeneity in task definition prevents meaningful global ranking by accuracy
Dominant MRI source	Contrast-enhanced T1-weighted MRI	Most common setting, particularly in Figshare-derived studies
Other MRI settings	Multimodal BraTS MRI; institutional multi-sequence MRI; mixed public repositories	Better representation of clinical variability but greater methodological heterogeneity
Dataset families	Figshare derivatives; BraTS; Kaggle and related public datasets; mixed repositories; institutional datasets	Dataset provenance is central to interpreting generalization
Direct VGG19 classification	VGG19 used as the principal classifier	Provides a conventional transfer-learning baseline
Frozen transfer features	VGG19 convolutional representations extracted without extensive weight updating	Computationally efficient but vulnerable to natural-image-to-MRI domain mismatch
Partial or full fine-tuning	Selected or all pretrained layers adapted to MRI data	Improves task adaptation but increases overfitting risk on small datasets
Hybrid architectures	VGG19 combined with recurrent units, SVMs, segmentation modules, pyramid pooling, transformers, or additional feature streams	Can improve representation but increases architectural complexity
Ensemble models	VGG19 or VGG-family models combined with multiple independent classifiers or backbones	May reduce model-specific variance at the cost of latency and interpretability
Attention/XAI/federated extensions	Attention modules, Grad-CAM/LIME-based interpretation, or federated VGG19	Extends evaluation toward interpretability, governance, and distributed generalization

The included literature supported six major categories of VGG19 adaptation: direct VGG19 classification, frozen transfer-feature extraction, partial or full fine-tuning, hybrid feature or multitask systems, ensemble architectures, and attention-, explainability-, or federated-learning extensions. This taxonomy was based on the functional role of VGG19 within the analytical pipeline rather than simply recording whether VGG19 appeared among a list of comparator architectures.

Stand-alone VGG19 and frozen-feature configurations were among the most computationally straightforward approaches and remain valuable as reproducible reference baselines. Frozen feature extraction is particularly attractive when sample sizes are limited because it reduces the number of trainable parameters and can be paired with support-vector machines, k-nearest-neighbor classifiers, extreme learning machines, or shallow neural networks. Nevertheless, the approach preserves features originally learned from natural-image datasets and therefore may retain

substantial source-to-target domain mismatch. High performance obtained using frozen features must consequently be interpreted together with feature-selection procedures, patient-level partitioning, and the independence of the test set.

Fine-tuning was used to adapt higher-level pretrained features to MRI-specific texture, morphology, and intensity distributions. Selective fine-tuning of later convolutional blocks generally provides a practical compromise between representational adaptation and overfitting. Full-network fine-tuning may be beneficial when datasets are sufficiently large and regularization is appropriately configured, but fair evaluation requires identical partitions and augmentation pipelines across comparator models. Without controlled ablation, reported gains cannot reliably be attributed to architectural modification because differences in preprocessing, augmentation, optimizer settings, or learning-rate schedules may explain part of the improvement.

Hybrid approaches combined VGG19-derived features with recurrent networks, support-vector machines, segmentation modules, pyramid pooling, transformers, alternative CNN feature streams, or other decision mechanisms. These models frequently reported high classification performance because complementary feature spaces can represent different aspects of tumor morphology and texture. Their principal methodological disadvantage is the expansion of the hypothesis space and the corresponding increase in parameters, hyperparameters, optimization complexity, and opportunities for overfitting. Such systems therefore require stronger ablation, preferably nested validation, and explicit reporting of computational cost.

Ensemble models similarly achieved high reported performance by reducing model-specific variance through voting, averaging, or feature-level fusion. However, comparison between a five-model ensemble and a single VGG19 architecture is incomplete unless parameter count, hardware, training time, inference latency, and memory requirements are considered. For clinical deployment, a marginal increase in classification accuracy may not justify substantially greater computational requirements or reduced interpretability.

Attention-based systems offered a more targeted extension by emphasizing diagnostically informative feature

channels or spatial regions. Channel attention is particularly compatible with VGG19 because it reweights learned feature maps without requiring a complete redesign of the network. However, attention weights should not automatically be interpreted as causal explanations. Explanation methods such as Grad-CAM and LIME can show the regions associated with model predictions and are useful for identifying implausible model behavior, but visual saliency does not establish causal reasoning, diagnostic validity, or clinical safety. Ideally, explanation maps should be compared against tumor annotations or assessed by domain experts.

Federated VGG19 represented one of the most clinically relevant recent developments because it shifted part of the research problem from architecture optimization toward data governance and inter-site heterogeneity. Federated learning allows participating sites to train a shared model without centralizing patient images. Its importance therefore lies not only in accuracy but also in testing whether a familiar architecture can function under non-identically distributed data and privacy constraints. Nevertheless, prospective workflow evidence remained absent, and federated simulation using partitioned public datasets should not be equated with true multicenter clinical deployment.

Common preprocessing procedures included image resizing, intensity normalization, contrast enhancement, denoising, region-of-interest extraction, and, in multimodal investigations, registration and skull stripping. Augmentation strategies included rotation, reflection, translation, cropping, zooming, intensity perturbation, and synthetic image generation. These procedures can increase effective training variability and improve robustness, particularly in small datasets, but their implementation has direct implications for validity.

A critical finding across the evidence base was the risk of data leakage associated with slice-level datasets. Many public brain MRI datasets contain several highly correlated slices originating from the same patient. When random image-level partitioning is used, anatomically similar or nearly identical slices from the same subject may appear in both the training and test sets. The resulting test set is therefore not fully independent and can yield overly optimistic performance estimates. The problem becomes more severe when augmentation occurs before data

partitioning because augmented versions of the same source image can be distributed across training and testing subsets.

Patient-level partitioning must therefore occur before augmentation and before any preprocessing procedure whose parameters are estimated from the data. Studies should explicitly identify the split unit, provide patient counts in addition to image counts, describe whether subject identifiers were available, and report whether duplicates or repackaged images were checked across repositories. The relatively limited number of studies explicitly reporting patient- or site-level separation was therefore one of the most important methodological findings of the review.

Accuracy dominated the evaluation literature. Although 25 studies reported class-wise metrics, sensitivity, specificity, precision, recall, F1-score, and area under the receiver operating characteristic curve were not uniformly reported. The predominance of accuracy is problematic in imbalanced datasets because a high aggregate value can conceal poor recognition of a clinically important minority class.

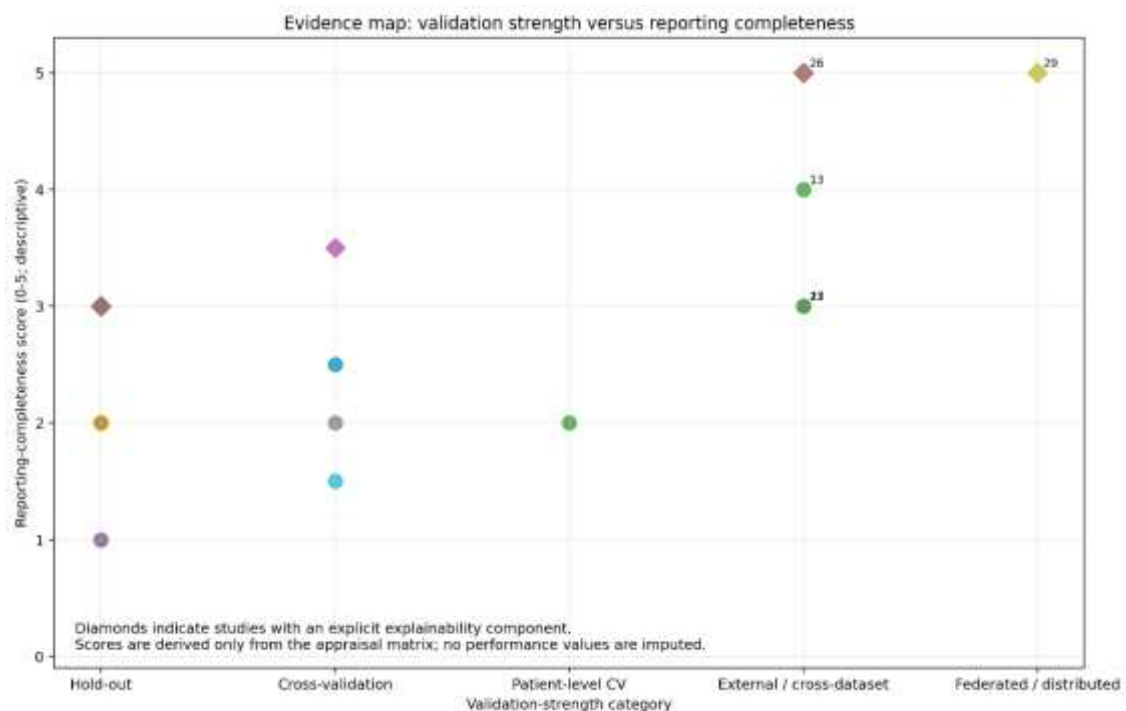
Cross-validation was used in 12 studies and provided more robust assessment of internal repeatability than a single

hold-out split. Nevertheless, cross-validation does not replace independent external testing because folds drawn from the same source population generally share acquisition characteristics, preprocessing history, and disease spectrum. Only six studies incorporated external, cross-dataset, distributed, or otherwise broader validation. Explicit patient- or site-level separation was identified in seven studies. Seven investigations incorporated some form of XAI, whereas only four provided code or implementation resources. Calibration, which is important when model probabilities may inform clinical decisions, was not systematically reported.

The evidence therefore indicated that validation rigor and reporting completeness varied independently. A study could provide extensive architectural details while relying entirely on internal hold-out testing, whereas another could undertake stronger cross-dataset validation without releasing code or complete computational documentation. Consequently, no single quality score adequately represented evidence strength.

Figure 2

Evidence map relating validation strength, reporting completeness, and clinical-readiness indicators across the included studies.



The comparative results demonstrated that high reported accuracy was common but that its interpretation depended strongly on validation context. Afshar et al. (2018) showed that capsule-based representations could improve robustness to spatial variation in tumor-type classification, but the evidence remained based on internal hold-out testing. Deepak and Ameer (2019) reported an accuracy of 98.00% using pretrained transfer features with support-vector machine or k-nearest-neighbor classification and, importantly, described patient-level five-fold evaluation. Ismael et al. (2020) reported a best accuracy of 99.00% using an enhanced residual architecture, although evaluation remained internally based. Gab Allah et al. (2021) combined VGG19 with PGGAN augmentation and CNN, GRU, or BiGRU classification heads and reported 98.54% for the CNN configuration, with synthetic augmentation supporting minority-class representation.

Aurna et al. (2022) reported an overall accuracy of 99.13%, with dataset-specific performance ranging from 98.16% to 99.76%, using a two-stage feature-level ensemble across four public MRI datasets. Islam et al. (2023)

evaluated VGG19, InceptionV3, MobileNet, and DenseNet121 using Figshare, SARTAJ, and Br35H datasets; MobileNet achieved 99.60%, InceptionV3 achieved 98.00%, and VGG19 remained a competitive reference architecture. Nassar et al. (2024) reported a best accuracy of 99.31% using majority voting across multiple deep-learning models, although the five-model architecture increased computational complexity. Haque et al. (2024) introduced NeuroNet19, combining VGG19 with improved pyramid pooling and LIME-based interpretation on 7,023 images. Hossain et al. (2024) used a Vision Transformer ensemble with explainability across multiple MRI sources, extending evaluation toward cross-dataset robustness. Kothadiya et al. (2025) reported a validation accuracy of 98.81% for the VGX framework, although inconsistency between the architecture named in the title and that described in the abstract limited interpretability. Muntaqim and Smrity (2025) evaluated federated VGG19 under non-identically distributed partitions and reported accuracies of 97.18% for FedAvg, 98.24% for FedProx, and 98.45% for SCAFFOLD.

Table 3

Representative comparative evidence interpreted according to validation context

Study	Method	Data	Principal reported outcome	Validation context
Afshar et al. (2018)	Capsule network for MRI tumor-type classification	Figshare CE-MRI	Capsule representation improved robustness to spatial variation	Internal hold-out
Deepak and Ameer (2019)	Transfer features with SVM/KNN	Figshare CE-MRI	98.00% accuracy; patient-level five-fold evaluation	Patient-level cross-validation
Ismael et al. (2020)	Enhanced residual deep network	Benchmark brain MRI	Best reported accuracy of 99.00%	Internal hold-out
Gab Allah et al. (2021)	VGG19 + PGGAN + CNN/GRU/BiGRU	Figshare CE-MRI	CNN head achieved 98.54%; synthetic augmentation supported minority classes	Internal hold-out
Aurna et al. (2022)	Two-stage feature-level ensemble	Four public MRI datasets	Overall 99.13%; dataset-wise 98.16–99.76%	Cross-validation/cross-dataset evidence
Islam et al. (2023)	VGG19, InceptionV3, MobileNet, DenseNet121 comparison	Figshare, SARTAJ, Br35H	MobileNet 99.60%; InceptionV3 98.00%; VGG19 remained a strong reference	Cross-dataset evaluation
Nassar et al. (2024)	Hybrid deep learning with majority voting	Figshare CE-MRI	Best reported accuracy 99.31%; five-model voting increased complexity	Cross-validation
Haque et al. (2024)	NeuroNet19: VGG19 + improved pyramid pooling	7,023 MRI images	High multiclass performance with LIME-based interpretation	Internal evaluation with XAI
Hossain et al. (2024)	Vision Transformer ensemble + explainability	Multi-source MRI	Ensembling and explanation maps improved robustness and transparency	Cross-dataset evaluation
Kothadiya et al. (2025)	VGX VGG-based gradient-explainer framework	Public brain MRI	Validation accuracy 98.81%; architecture-label inconsistency identified	Internal validation
Muntaqim and Smrity (2025)	Federated VGG19 under non-IID distributions	Distributed public MRI partitions	FedAvg 97.18%, FedProx 98.24%, SCAFFOLD 98.45%	Federated/distributed evaluation

A study-level appraisal demonstrated substantial variation in architectural contribution, validation strength, and interpretive limitations. The complete set of 30

empirical investigations is summarized in Table 4 to preserve the breadth of the evidence base while avoiding interpretation based solely on performance percentages.

Table 4

Critical profiles of all included empirical studies

Study	Technical contribution and principal result	Validation evidence and principal limitation
Afshar et al. (2018)	Capsule network evaluated on three-class Figshare CE-MRI; capsule representation improved robustness to spatial variation	Internal hold-out evaluation; no independent external validation; no explicit XAI; patient-level separation unclear; incomplete class-wise reporting
Sultan et al. (2019)	Deep CNN with augmentation on Figshare and an additional MRI dataset; reported 96.13% and 98.70% in two settings	Cross-dataset evidence and stronger separation of evaluation units; no explicit explanation analysis
Deepak and Ameer (2019)	Pretrained transfer features with SVM/KNN; 98.00% accuracy	Patient-level five-fold cross-validation; no independent external validation; no explicit XAI; incomplete class-wise reporting
Ismael et al. (2020)	Enhanced residual deep network; best accuracy 99.00% with precision, recall, F1-score, and balanced accuracy	Internal hold-out; no external validation; no explicit XAI; patient-level split unclear
Badza and Barjaktarovic (2020)	Custom 22-layer CNN on Figshare; best accuracy 96.56%	Cross-validation; no external validation; no XAI; patient-level partition unclear
Khan et al. (2020)	CNN-based multiclass MRI classification; strong baseline performance	Internal hold-out; no external validation; no explicit XAI; patient-level separation unclear
Alhassan and Zainon (2021)	Hard-swish CNN on Figshare; best reported accuracy 98.60%	Internal hold-out; no external validation; no XAI; patient-level partition unclear
Gab Allah et al. (2021)	VGG19 + PGGAN with CNN/GRU/BiGRU heads; CNN achieved 98.54%	Internal hold-out; no external validation; no explicit XAI; patient-level partition unclear
Noreen et al. (2021)	Fine-tuned Inception/Xception ensemble; best single-model result 94.34% and improved ensemble stability	Internal hold-out; no external validation; no XAI; patient-level split unclear
Nawaz et al. (2021)	Improved CornerNet with DenseNet features; best accuracy 98.80%	Internal hold-out; no external validation; no XAI; patient-level partition unclear
Khan et al. (2021)	K-means ROI extraction, fine-tuned VGG19, and synthetic augmentation	Internal hold-out; no external validation; no XAI; patient-level separation unclear
Alanazi et al. (2022)	CNN framework with augmentation on public MRI collections; high reported performance	Internal hold-out; provenance and patient-level splitting remain important interpretive concerns
Aurna et al. (2022)	Two-stage feature-level ensemble across four datasets; overall 99.13%	Cross-validation and cross-dataset evidence; no explicit XAI
Rasool et al. (2022)	GoogleNet features + SVM; hybrid result 98.10% versus 93.10% for fine-tuned network	Internal hold-out; no external validation; no XAI; patient-level split unclear
Veeramuthu et al. (2022)	Combined feature- and image-based classifier; 98.97% accuracy, 98.86% sensitivity, 97.14% specificity	Cross-validation; no external validation; no XAI; patient-level partition unclear
Tummala et al. (2022)	Vision Transformer ensemble for four-class MRI classification	Cross-validation and explainability; no independent external validation; patient-level split unclear
Islam et al. (2023)	VGG19/InceptionV3/MobileNet/DenseNet121 comparison; MobileNet 99.60%, InceptionV3 98.00%	Cross-dataset testing; stronger generalization evidence; no explicit XAI
Fauzi et al. (2023)	VGG19 versus ResNet101 across institutional MRI sequences; VGG19 81% for MBT and ResNet101 83% for MPBT	Cross-validation; no independent external validation; patient-level details limited
Rohini et al. (2023)	Customized CNN with VGG19 and other pretrained models; best accuracy 99.43%	Internal hold-out; no external validation; no XAI; patient-level split unclear
Ben Slama et al. (2023)	Res-Net-VGG19 hybrid on BraTS for joint segmentation and glioma grading	Cross-validation; independent external validation and dedicated XAI not established; stronger ablation and patient-level documentation needed
Nassar et al. (2024)	Deep-learning ensemble with majority voting; best accuracy 99.31%	Cross-validation; no independent external validation; increased computational complexity
Haque et al. (2024)	NeuroNet19 using VGG19 and improved pyramid pooling; LIME interpretation	Explicit XAI; no independent external validation; patient-level split unclear
Rasa et al. (2024)	Fine-tuned transfer-learning models across two public datasets; 99.96% on smaller data and 96.34–98.20% on larger data	Five-fold cross-validation and cross-dataset evaluation; no explicit XAI
Alshuhail et al. (2024)	Sequential CNN with dropout and Grad-CAM; 98.00% test accuracy and approximately 97–98% class-wise metrics	Explicit XAI; no independent external validation; patient-level split unclear
Gupta et al. (2024)	Augmented CNN for binary benign/malignant classification	Internal hold-out; no independent external validation; no explicit XAI; incomplete class-wise reporting

Hossain et al. (2024)	Vision Transformer ensemble with explainability across multi-source MRI	Cross-dataset testing and XAI; uncertainty remains regarding prospective clinical utility and calibration
Ullah et al. (2024)	Multimodal segmentation-classification framework with XAI	Cross-validation and XAI; no independent external validation; patient-level split unclear
Kothadiya et al. (2025)	VGX gradient-explainer framework; reported validation accuracy 98.81%	Architecture-label inconsistency between VGG19 title and VGG16 description; external validation and patient-level separation not established
Muntaqim and Smrity (2025)	Federated VGG19 under non-IID data; SCAFFOLD achieved 98.45%	Federated/distributed validation and XAI; prospective clinical utility and calibration remain uncertain
Rastogi et al. (2025)	Fine-tuned VGG19, Xception, MobileNetV2, and InceptionResNetV2; strong comparative performance	Cross-validation; no external validation; no XAI; patient-level split unclear

The study-level profiles showed that methodological strength and headline performance were not equivalent. Some of the highest reported accuracies originated from internally evaluated public datasets without explicit patient-level separation or independent validation. Conversely, studies incorporating cross-dataset, distributed, or patient-level testing occasionally reported less extreme numerical performance but produced more informative evidence about generalization.

This distinction was particularly important for Figshare-derived contrast-enhanced MRI datasets, which frequently produced accuracies approaching or exceeding 98%. Such results establish technical feasibility under the reported experimental conditions, but they should not be interpreted as estimates of clinical diagnostic performance unless subject independence, duplicate removal, calibration, and external validity have been demonstrated.

The findings therefore support a three-axis interpretation of the VGG19 literature. The first axis concerns architectural adaptation, including transfer learning, fine-tuning, hybridization, attention, ensembling, and federated implementation. The second concerns validation rigor, ranging from image-level hold-out experiments to patient-level cross-validation, cross-dataset testing, and distributed or multicenter evaluation. The third concerns clinical readiness, including transparent data provenance, calibration, class-specific evaluation, error analysis, explainability, computational reporting, and reproducibility. An improvement along one axis does not imply improvement along the others.

The observed performance of VGG19-based models must be interpreted within the clinical characteristics of brain tumors and the physical characteristics of MRI. Gliomas arise from glial lineages and encompass biologically diverse entities with infiltrative growth. Meningiomas generally arise from the meninges and are

commonly extra-axial, whereas pituitary adenomas originate in the sellar region. Their management, surgical approach, prognosis, and treatment pathways differ, making reliable classification clinically meaningful.

The technical findings further showed that preprocessing and model architecture cannot be interpreted independently of MRI acquisition. CNNs learn local spatial patterns through repeated convolution, nonlinear activation, and pooling, with early layers typically encoding edges and textures and deeper layers representing increasingly task-specific configurations. In brain MRI, however, the learned representation is strongly shaped by intensity normalization, segmentation or ROI selection, sequence type, label quality, and sampling strategy.

Transfer learning reduces the amount of labeled medical data required by reusing representations learned from large natural-image datasets. The success of this approach does not imply equivalence between natural images and MRI; rather, it suggests that some low- and intermediate-level visual primitives remain transferable. The findings showed that the degree of fine-tuning, learning rate, augmentation, regularization, and size of the classification head can materially influence whether transferred representations adapt successfully or overfit.

VGG19 itself is characterized by repeated 3×3 convolutional kernels arranged in five convolutional blocks, interspersed with max-pooling and followed by fully connected classification layers in the original architecture. Its principal advantage in the reviewed literature was not architectural novelty but stability and modularity. Intermediate VGG19 representations could be used for classification, fusion, segmentation-related tasks, or input to secondary classifiers. Common modifications replaced the original fully connected layers with global average pooling, smaller dense heads, dropout, weight decay, and selectively fine-tuned layers.

The principal architectural disadvantages were the relatively large parameter count and substantial memory and computational requirements. In comparison with residual or densely connected networks, VGG19 also lacks explicit skip connections that facilitate gradient propagation. Its continuing relevance therefore derives from mature pretrained weights, straightforward layer-wise interpretation, widespread implementation support, and suitability as a benchmark for controlled ablation.

The findings indicated that frozen VGG19 representations remain useful when datasets are small or computational resources are limited. Their modular structure allows pretrained feature vectors to be combined with SVMs, shallow neural classifiers, or related methods. However, performance gains should be interpreted with careful attention to data provenance, feature selection, and test-set independence because a fixed ImageNet representation can preserve substantial domain mismatch.

Fine-tuning later convolutional blocks generally provided a stronger mechanism for adapting VGG19 to MRI than maintaining completely frozen pretrained weights. Nevertheless, the evidence did not support a universal conclusion that deeper fine-tuning necessarily yields better generalization. When sample sizes are small, full fine-tuning introduces a large number of trainable parameters and may amplify overfitting. Controlled within-study comparison is therefore more informative than cross-study comparison of headline values.

Hybrid architectures showed considerable flexibility. VGG19 representations were combined with recurrent units, support-vector machines, segmentation pipelines, pyramid pooling, transformers, and additional CNN backbones. Such combinations can capture complementary spatial and contextual information, but they introduce additional hyperparameters and increase the importance of rigorous ablation. Without testing the contribution of each component under identical partitions, it remains difficult to establish whether gains derive from VGG19 adaptation, feature fusion, preprocessing, augmentation, or optimization.

Ensembles demonstrated similar trade-offs. Majority voting and feature fusion frequently increased performance stability, but stronger prediction alone does not establish superior clinical utility. Parameter count, inference latency, hardware requirements, model size, energy use, and

interpretability should accompany accuracy when evaluating deployability. A marginal improvement in predictive performance may be insufficient to justify a several-fold increase in computational burden.

Attention mechanisms were increasingly incorporated to emphasize diagnostically informative channels or regions. Such mechanisms are compatible with VGG19 feature maps because they can reweight intermediate representations without replacing the entire backbone. Nevertheless, the evidence did not support interpreting attention weights as direct evidence of model reasoning.

Grad-CAM, LIME, and related explanation techniques were used in several recent studies to identify regions associated with model predictions. Their most defensible role in the reviewed literature was as tools for plausibility assessment and failure analysis. Explanations can reveal whether a classifier focuses on the tumor, skull margins, acquisition artifacts, annotations, or unrelated anatomy. However, explanations do not demonstrate causal diagnostic reasoning and cannot independently establish safety. Comparison against lesion masks, neuroradiologist assessment, or predefined anatomical criteria would provide stronger validation.

Federated learning expanded the methodological problem beyond model performance. The federated VGG19 evidence demonstrated that a familiar CNN backbone can be optimized under non-identically distributed data without centralizing the underlying MRI images. This directly addresses data-governance concerns relevant to healthcare. Nevertheless, splitting a public dataset into simulated sites remains different from prospective federation across institutions with distinct scanners, patient populations, clinical workflows, and information systems.

Studies from heart disease prediction, Parkinson disease detection, liver tumor segmentation, liver lesion classification, and explainable or quantum-assisted cardiovascular prediction were retained only as methodological context and were not counted as brain-tumor evidence. Their value lay in illustrating general principles that were directly applicable to the reviewed MRI literature.

Heart disease studies using SVM, k-nearest neighbors, and logistic regression demonstrated that prediction performance should be accompanied by runtime and computational efficiency when systems are intended for

practical use. Parkinson disease studies emphasized the importance of modality-specific analysis and subject-level partitioning. Liver tumor segmentation research showed that task-specific metrics such as Dice coefficient, Hausdorff distance, and volumetric similarity are essential where spatial overlap rather than categorical accuracy is the primary endpoint. Liver lesion classification studies highlighted the importance of macro- and micro-averaged F1-scores and Matthews correlation coefficient under class imbalance. Explainable cardiovascular models further demonstrated that performance and explanation can be reported jointly without treating explanation as a substitute for validation.

These cross-domain observations reinforce the central finding that evaluation metrics should correspond to the intended clinical task. Classification requires class-specific discrimination metrics; segmentation requires overlap and boundary metrics; imbalanced classification requires metrics that are sensitive to minority classes; probabilistic decision support requires calibration; and clinically deployable systems require assessment of computation, reproducibility, and external generalization.

The synthesis provided consistent answers to the review questions. VGG19 has evolved from a relatively fixed

transfer-learning baseline into a flexible component of fine-tuned, hybrid, ensemble, attention-based, explainable, and federated systems. Contrast-enhanced T1-weighted Figshare-derived data remain dominant, whereas BraTS has supported multimodal glioma-related work and institutional datasets have provided greater clinical diversity at the expense of standardized provenance. Accuracy continues to dominate evaluation despite its limitations, and external validation, calibration, patient-level partitioning, reproducibility, and prospective testing remain substantially less common.

The evidence did not support the conclusion that increasingly complex architectures are inherently superior. Fine-tuning, hybridization, and ensembling can improve performance under controlled conditions, but the most defensible evidence arises from within-study comparison using identical partitions, explicit ablation, patient-level separation, and transparent computational reporting. Persistent methodological barriers included possible data leakage, small and imbalanced samples, uncertain patient counts, inconsistent preprocessing, limited external testing, weak calibration reporting, and restricted implementation transparency.

Table 5

Evidence-based answers to the research questions

Research question	Evidence-based answer
RQ1	VGG19 has progressed from a fixed transfer-learning baseline to a flexible backbone used for feature extraction, partial or full fine-tuning, hybrid fusion, ensemble learning, attention, explainability, and federated learning.
RQ2	Contrast-enhanced T1-weighted Figshare-derived datasets dominate three-class tumor classification; BraTS supports multimodal glioma-focused analysis, whereas mixed public and institutional datasets broaden coverage but create provenance and standardization concerns.
RQ3	Accuracy is disproportionately emphasized. Class-wise sensitivity and specificity, AUC, calibration, patient-level partitioning, external testing, and reproducibility indicators are reported considerably less consistently.
RQ4	Fine-tuning, hybridization, and ensembling can improve predictive performance, but convincing superiority requires within-study comparisons, identical partitions, ablation analyses, patient-level validation, and computational reporting.
RQ5	Persistent barriers include image- or slice-level data leakage, small and imbalanced cohorts, unclear patient counts, inconsistent preprocessing, limited calibration, scarce external validation, and weak reproducibility.
RQ6	Future studies should predefine the intended clinical task, partition data by patient and preferably by site, report class-wise metrics with uncertainty estimates, assess calibration and errors, conduct independent external testing, provide reproducibility materials, and validate interpretability against expert assessment or tumor annotations whenever feasible.

Overall, the findings demonstrate that VGG19 remains a competitive and methodologically useful backbone for MRI brain tumor classification, particularly because it provides a stable platform for transfer learning, controlled architectural modification, and comparison with newer CNN and transformer-based systems. However, the principal

limitation of the existing literature is no longer the absence of very high reported accuracy. Rather, it is the limited quantity of evidence establishing whether such accuracy is reproducible across independent patients, datasets, scanners, institutions, and clinical workflows. The strongest future contribution will therefore not necessarily be another

marginal increase in internal accuracy, but a demonstrably generalizable, calibrated, interpretable, computationally transparent, and externally validated VGG19-based or VGG19-comparator system.

4. Discussion and Conclusion

The present systematic review synthesized 30 empirical studies investigating deep-learning-based brain tumor analysis on MRI, with particular emphasis on the role of VGG19 as a direct classifier, pretrained feature extractor, fine-tuned backbone, hybrid component, ensemble member, or component of explainable and distributed-learning frameworks. The principal finding was that VGG19 remains a competitive and methodologically useful architecture, but the apparent strength of VGG19-based systems depends heavily on how performance is validated. Publication activity increased substantially after 2021, accompanied by a shift from conventional transfer learning toward fine-tuned, hybrid, ensemble, transformer-assisted, explainable, and federated models. Nevertheless, the review demonstrated that increasing architectural sophistication has not been matched uniformly by improvements in validation rigor. All 30 studies reported accuracy or an accuracy-derived outcome, whereas only 25 reported class-wise metrics, 12 used cross-validation, seven explicitly documented patient- or site-level separation, six incorporated external, cross-dataset, or distributed validation, seven used an explainability component, and only four reported code or implementation resources. These findings indicate that the field has progressed rapidly in model development, but evidence of clinical generalizability, reproducibility, and calibration remains comparatively limited.

The observed expansion of deep-learning approaches is consistent with broader developments in medical image analysis, where CNNs have evolved from basic feature-learning systems into increasingly complex transfer-learning, attention, multimodal, transformer, and ensemble frameworks (3, 4). This evolution was also evident within the brain tumor literature. Earlier studies concentrated on tumor-region augmentation, CNN-based classification, and conventional transfer learning (5, 6), whereas more recent research increasingly emphasized architectural fusion, fine-tuning, explanation, and robustness (9, 11, 18). The present

findings therefore suggest that VGG19 has persisted not because it represents the newest architectural paradigm, but because its regular layer structure, mature pretrained weights, and compatibility with downstream classifiers make it a stable benchmark against which newer approaches can be evaluated. This interpretation is supported by studies in which VGG19 remained a strong comparator even when MobileNet, Inception-based networks, transformer models, or hybrid architectures achieved higher headline performance (9, 11, 17).

Another important finding was the dominance of three-class classification of glioma, meningioma, and pituitary tumors using contrast-enhanced T1-weighted MRI, particularly from Figshare-derived datasets. This finding reflects the long-standing influence of public benchmark datasets on brain tumor classification research. Publicly available MRI collections have played an important role in accelerating method development and enabling direct algorithmic comparison, while benchmark initiatives such as BraTS have supported more standardized glioma-focused and multimodal analysis (1, 2). However, the review also showed that repeated reliance on the same or derivative datasets creates a risk of overestimating generalizability. High performance on a familiar public repository may indicate strong dataset-specific optimization rather than robustness to independent clinical populations. This concern is especially important when repositories contain repackaged or overlapping images, because apparently independent external datasets may not be fully independent in practice. The present results therefore reinforce the distinction between benchmark performance and true clinical validation.

The review also found that fine-tuning and hybridization frequently produced higher reported performance than simpler frozen-feature approaches, but these gains were not consistently supported by equally rigorous validation designs. Studies such as those by Noreen et al. and Rasa et al. demonstrated the potential value of fine-tuned pretrained models for brain tumor classification (8, 10), whereas Khan et al. showed how ROI-based preprocessing and synthetic augmentation could support VGG19 adaptation (12). These findings are consistent with the theoretical advantage of fine-tuning: higher-level convolutional representations can be adapted from natural-image features toward MRI-specific

morphology and texture. However, fine-tuning also increases the number of trainable parameters and can exacerbate overfitting when datasets are small. Thus, the present findings suggest that fine-tuning should be interpreted as a potentially effective adaptation mechanism rather than an inherently superior strategy. Its value depends on adequate regularization, appropriate learning-rate control, patient-independent validation, and a sufficiently representative dataset.

Hybrid and ensemble models represented another major trend. Aurna et al. reported strong performance using a two-stage feature-level ensemble across multiple public datasets (15), while Rasool et al. combined deep feature extraction with machine-learning classification (13). Veeramuthu et al. integrated image-level and feature-level classification (14), and Nassar et al. used majority voting within a hybrid deep-learning framework (16). These results align with the present review's observation that ensemble and hybrid systems often achieve some of the highest reported accuracies. The likely explanation is that combining complementary feature representations or decision functions can reduce model-specific variance and improve discrimination. Nevertheless, the gains come at the cost of additional parameters, greater training complexity, increased inference latency, and more difficult interpretation. Therefore, the clinical significance of a small increase in accuracy should be weighed against the computational burden and reproducibility requirements introduced by the additional components.

The findings also highlighted the growing role of transformer-based and multimodal systems. Vision Transformer ensembles have demonstrated high multiclass classification performance (17), while Hossain et al. combined transformer architectures, ensemble learning, transfer learning, and explainable AI within a unified brain tumor classification framework (18). Ullah et al. further extended the field toward joint multimodal segmentation and classification with information fusion and explainability (20). These studies support the interpretation that the field is gradually moving from isolated image classification toward richer modeling of tumor structure, multimodal context, and model transparency. However, the present review indicates that this architectural progress does not remove the need for rigorous evaluation. A highly complex system trained and tested on non-independent samples remains

methodologically weaker than a simpler system validated on independent patients or sites.

A particularly important result concerned the risk of data leakage. Only seven studies explicitly documented patient- or site-level separation, despite the fact that many brain MRI datasets contain multiple slices from the same patient. Random image-level splitting can therefore result in anatomically similar or nearly identical slices being distributed across training and test sets, producing artificially optimistic performance. This concern is especially relevant for studies involving aggressive augmentation, ROI extraction, or synthetic data generation. Cheng et al. demonstrated the importance of tumor-region augmentation and partitioning in improving classification performance (5), and later studies used increasingly sophisticated augmentation pipelines (7, 12). The present review suggests that such procedures are valid only when they are performed after appropriate patient-level partitioning. Otherwise, transformed versions of the same underlying image may inadvertently appear in both development and test subsets.

The predominance of accuracy as the primary performance measure was another major finding. Although 25 studies reported class-wise metrics, accuracy remained the universal headline outcome. This is problematic because overall accuracy is sensitive to class imbalance and does not reveal whether clinically important minority classes are being detected reliably. Studies that reported precision, recall, sensitivity, specificity, F1-score, and AUC provided a more informative picture of model behavior (10, 14, 24). The review therefore supports the broader principle that performance reporting should correspond to the clinical task rather than relying on a single aggregate metric. This position is consistent with reporting frameworks emphasizing transparent specification of outcome definitions, validation procedures, and model performance (27, 29).

The review further identified limited external validation as a central weakness. Only six studies incorporated external, cross-dataset, or distributed evaluation. This is concerning because MRI data are highly sensitive to scanner vendor, acquisition protocol, field strength, sequence parameters, reconstruction, and institutional workflow. Cross-dataset studies such as Islam et al. and Rasa et al.

provide stronger evidence of generalization than single-dataset hold-out experiments (9, 10). Similarly, the distributed approach reported in more recent federated research illustrates the potential value of testing models under non-identically distributed conditions. Such designs are especially relevant because trustworthy clinical AI must perform across heterogeneous environments rather than only within the dataset used for development. This interpretation aligns with the broader emphasis of FUTURE-AI on robustness, generalizability, traceability, and deployability in healthcare AI (30).

Explainability was present in seven studies and represented an emerging, but still limited, dimension of clinical-readiness evidence. Grad-CAM and related gradient-based techniques can identify image regions associated with neural-network predictions (21), while attention mechanisms such as convolutional block attention can emphasize informative channels and spatial regions (22). Recent studies have incorporated LIME, gradient explainers, and other XAI approaches into brain tumor classification (23-25). These developments are encouraging because explanation maps can support plausibility assessment and error analysis. However, the present review reinforces that visual explanation should not be equated with causal reasoning or clinical validity. Explanation quality should ideally be compared with tumor masks, radiologist assessment, or predefined anatomical criteria before being interpreted as evidence of trustworthy model behavior.

The findings also revealed that reproducibility remains underdeveloped. Only four studies reported code or implementation materials, and reported links were not independently tested. This is important because complex hybrid, ensemble, and transfer-learning pipelines can be difficult to reproduce without explicit documentation of preprocessing, augmentation, initialization, optimization, training schedules, model-selection criteria, and hardware. The CLAIM 2024 update emphasizes the importance of comprehensive reporting in AI-based medical imaging (29), while PRISMA 2020 provides a corresponding framework for transparent systematic-review reporting (28). The present review therefore indicates that future progress should be judged not only by predictive accuracy but also by whether methods can be independently understood, reproduced, and tested in new settings.

Taken together, the results support a three-axis interpretation of the literature: architectural adaptation, validation rigor, and clinical readiness. VGG19-based models can be highly competitive, particularly when fine-tuned or incorporated into hybrid systems, but architectural improvement does not automatically imply stronger evidence. A model can be technically sophisticated and still provide weak evidence if it relies on image-level splitting, incomplete provenance, internal hold-out testing, and absent calibration. Conversely, a simpler architecture may provide more clinically informative evidence if it is evaluated with patient-level separation, external datasets, class-wise metrics, and transparent implementation. This distinction is central to interpreting the current state of VGG19-based brain tumor classification and supports a shift from performance-centric reporting toward validation-centered evaluation.

The limitations of this review should be acknowledged. The final evidence base was reconstructed and audited at the reference level because the original database export files were not retained; therefore, database-specific retrieval, duplicate-removal, and exclusion counts could not be reported as factual historical quantities. Considerable heterogeneity existed across tasks, class definitions, MRI sequences, datasets, preprocessing procedures, sample sizes, partitioning strategies, and outcome metrics, preventing quantitative meta-analysis. In addition, the appraisal relied on information explicitly reported in the included articles; absence of reporting does not necessarily indicate absence of the corresponding methodological procedure. The review was restricted to English-language peer-reviewed studies published within the predefined period, which may have excluded relevant work in other languages or unpublished technical reports. Finally, implementation links and source-code availability were recorded as reported but were not independently tested for functionality or reproducibility.

Future research should prioritize validation design as strongly as architectural development. Studies should use patient-level and, where possible, site-level data partitioning before augmentation or model fitting, and should clearly report the number of patients in addition to the number of images. External validation across independent institutions, scanner vendors, and acquisition protocols should become routine for claims of generalizability. Researchers should

report class-wise sensitivity, specificity, precision, recall, F1-score, AUC, confidence intervals, and calibration metrics rather than relying predominantly on accuracy. Controlled ablation studies should isolate the contribution of fine-tuning, augmentation, attention, feature fusion, and ensemble components. Prospective multicenter studies, robust duplicate detection across public repositories, standardized benchmarking datasets with subject identifiers, and expert evaluation of explanation maps would substantially strengthen the evidence base. Public release of code, preprocessing pipelines, trained weights, and complete implementation settings should also be encouraged whenever ethically and legally feasible.

For practice, VGG19-based systems should currently be viewed primarily as research and decision-support tools rather than autonomous diagnostic systems. Clinical teams considering such models should prioritize systems that have been validated on independent patients and institutions, provide class-specific performance and calibrated probabilities, document data provenance, and include transparent information about failure modes and computational requirements. Explanation maps may assist clinicians in reviewing whether predictions are anatomically plausible, but they should supplement rather than replace expert interpretation. Before implementation, each model should be evaluated on local MRI protocols, patient populations, scanner characteristics, and workflow conditions, with continuous monitoring for performance drift and unexpected errors after deployment.

Authors' Contributions

All authors contributed substantially to the study and to manuscript development, and all approved the final version.

Declaration

The authors declare that artificial intelligence tools were used only to assist with language editing, translation, and improvement of the manuscript's readability. All conceptualization, study design, data collection, data analysis, interpretation of findings, and final approval of the manuscript were performed by the authors. The authors take full responsibility for the accuracy, integrity, and originality of the content.

Transparency Statement

Data are available for research purposes upon reasonable request to the corresponding author.

Acknowledgments

The authors thank the participating centers, clinicians, adolescents, and guardians who made the study possible.

Declaration of Interest

The authors report no conflict of interest.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Ethics Considerations

The study placed a high emphasis on ethical considerations. Informed consent obtained from all participants, ensuring they are fully aware of the nature of the study and their role in it. Confidentiality strictly maintained, with data anonymized to protect individual privacy. The study adhered to the ethical guidelines for research with human subjects as outlined in the Declaration of Helsinki.

References

1. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (Brats). *IEEE Transactions on Medical Imaging*. 2015;34(10):1993-2024. [PMID: 25494501] [DOI]
2. Bakas S, Akbari H, Sotiras A, Bilello M, Rozycki M, Kirby JS, et al. Advancing The Cancer Genome Atlas Glioma MRI Collections With Expert Segmentation Labels And Radiomic Features. *Scientific Data*. 2017;4:170117. [PMID: 28872634] [DOI]
3. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A Survey On Deep Learning In Medical Image Analysis. *Medical Image Analysis*. 2017;42:60-88. [PMID: 28778026] [DOI]
4. Zhou SK, Greenspan H, Davatzikos C, Duncan JS, Van Ginneken B, Madabhushi A, et al. A Review Of Deep Learning In Medical Imaging: Imaging Traits, Technology Trends, Case Studies With Progress Highlights, And Future Promises. *Proceedings of the IEEE*. 2021;109(5):820-38. [PMID: 37786449] [PMCID: PMC10544772] [DOI]
5. Cheng J, Huang W, Cao S, Yang R, Yang W, Yun Z, et al. Enhanced Performance Of Brain Tumor Classification Via Tumor Region Augmentation And Partition. *PLoS ONE*. 2015;10(10):e0140381. [PMID: 26447861] [DOI]

6. Khan HA, Jue W, Mushtaq M, Mushtaq MU. Brain Tumor Classification In MRI Image Using Convolutional Neural Network. *Mathematical Biosciences and Engineering*. 2020;17(5):6203-16. [PMID: 33120595] [DOI]
7. Alanazi MF, Ali MU, Hussain SJ, Zafar A, Mohatram M, Irfan M, et al. Brain Tumor/Mass Classification Framework Using Magnetic-Resonance-Imaging-Based Deep Learning. *Sensors*. 2022;22(1):372. [PMID: 35009911] [PMCID: PMC8749789] [DOI]
8. Noreen N, Palaniappan S, Qayyum A, Ahmad I, Allassafi MO. Brain Tumor Classification Based On Fine-Tuned Models And The Ensemble Method. *Computers, Materials & Continua*. 2021;67(3):3967-82. [DOI]
9. Islam MM, Barua P, Rahman M, Ahammed T, Akter L, Uddin J. Transfer Learning Architectures With Fine-Tuning For Brain Tumor Classification Using Magnetic Resonance Imaging. *Healthcare Analytics*. 2023;4:100270. [DOI]
10. Rasa SM, Islam MM, Talukder MA, Uddin MA, Khalid M, Kazi M, et al. Brain Tumor Classification Using Fine-Tuned Transfer Learning Models On Magnetic Resonance Imaging (MRI) Images. *Digital Health*. 2024;10. [PMID: 39381813] [PMCID: PMC11459499] [DOI]
11. Rastogi D, Johri P, Donelli M, Kumar L, Bindewari S, Raghav A, et al. Brain Tumor Detection And Prediction In MRI Images Utilizing A Fine-Tuned Transfer Learning Model Integrated Within Deep Learning Frameworks. *Life*. 2025;15(3):327. [PMID: 40141673] [PMCID: PMC11944010] [DOI]
12. Khan AR, Khan S, Harouni M, Abbasi R, Iqbal S, Mehmood Z. Brain Tumor Segmentation Using K-Means Clustering And Deep Learning With Synthetic Data Augmentation For Classification. *Microscopy Research and Technique*. 2021;84(7):1389-99. [PMID: 33524220] [DOI]
13. Rasool M, Ismail NA, Boulila W, Ammar A, Samma H, Yafooz WMS, et al. A Hybrid Deep Learning Model For Brain Tumour Classification. *Entropy*. 2022;24(6):799. [PMID: 35741521] [PMCID: PMC9222774] [DOI]
14. Veeramuthu A, Meenakshi S, Mathivanan G, Kotecha K, Saini JR, Vijayakumar V, et al. MRI Brain Tumor Image Classification Using A Combined Feature And Image-Based Classifier. *Frontiers in Psychology*. 2022;13:848784. [PMID: 35310201] [PMCID: PMC8931531] [DOI]
15. Aurna NF, Yousuf MA, Taher KA, Azad AKM, Moni MA. A Classification Of MRI Brain Tumor Based On Two Stage Feature Level Ensemble Of Deep Cnn Models. *Computers in Biology and Medicine*. 2022;146:105539. [PMID: 35483227] [DOI]
16. Nassar SE, Yasser I, Amer HM, Mohamed MA. A Robust MRI-Based Brain Tumor Classification Via A Hybrid Deep Learning Technique. *The Journal of Supercomputing*. 2024;80:2403-27. [DOI]
17. Tummala S, Kadry S, Bukhari SAC, Rauf HT. Classification Of Brain Tumor From Magnetic Resonance Imaging Using Vision Transformers Ensembling. *Current Oncology*. 2022;29(10):7498-511. [PMID: 36290867] [DOI]
18. Hossain S, Chakrabarty A, Gadekallu TR, Alazab M, Piran MJ. Vision Transformers, Ensemble Model, And Transfer Learning Leveraging Explainable Ai For Brain Tumor Detection And Classification. *IEEE Journal of Biomedical and Health Informatics*. 2024;28(3):1261-72. [PMID: 37043319] [DOI]
19. Rohini A, Praveen C, Mathivanan SK, Muthukumaran V, Mallik S, Alqahtani MS, et al. Multimodal Hybrid Convolutional Neural Network Based Brain Tumor Grade Classification. *BMC Bioinformatics*. 2023;24:382. [PMID: 37817066] [PMCID: PMC10566188] [DOI]
20. Ullah MS, Khan MA, Albarakati HM, Damaševičius R, Alsenan S. Multimodal Brain Tumor Segmentation And Classification From MRI Scans Based On Optimized Deeplabv3+ And Interpreted Networks Information Fusion Empowered With Explainable Ai. *Computers in Biology and Medicine*. 2024;182:109183. [PMID: 39357134] [DOI]
21. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-Cam: Visual Explanations From Deep Networks Via Gradient-Based Localization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)2017*. p. 618-26. [DOI]
22. Woo S, Park J, Lee JY, Kweon IS. Cbam: Convolutional Block Attention Module. *Proceedings of the European Conference on Computer Vision (ECCV)2018*. p. 3-19. [DOI]
23. Haque R, Hassan MM, Bairagi AK, Islam SMS. Neuronet19: An Explainable Deep Neural Network Model For The Classification Of Brain Tumors Using Magnetic Resonance Imaging Data. *Scientific Reports*. 2024;14:1524. [PMID: 38233516] [PMCID: PMC10794704] [DOI]
24. Alshuhail A, Thakur A, Chandramma R, Mahesh TR, Almusharraf A, Vinoth Kumar V, et al. Refining Neural Network Algorithms For Accurate Brain Tumor Classification In MRI Imagery. *BMC Medical Imaging*. 2024;24:118. [PMID: 38773391] [PMCID: PMC11110259] [DOI]
25. Kothadiya D, Rehman A, Alghofaily B, Bhatt C, Ayesha N, Saba T. Vgg19-Based Gradient Explainer Interpretable Architecture For Brain Tumor Detection In Microscopy Magnetic Resonance Imaging (Mmri). *Microscopy Research and Technique*. 2025;88(5):1544-54. [PMID: 39825619] [DOI]
26. Nawaz M, Nazir T, Masood M, Mehmood A, Mahum R, Khan MA, et al. Analysis Of Brain MRI Images Using Improved Cornernet Approach. *Diagnostics*. 2021;11(10):1856. [PMID: 34679554] [PMCID: PMC8535141] [DOI]
27. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting Of A Multivariable Prediction Model For Individual Prognosis OR Diagnosis: Tripod. *Annals of Internal Medicine*. 2015;162(1):55-63. [PMID: 25560714] [DOI]
28. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The Prisma 2020 Statement: An Updated Guideline For Reporting Systematic Reviews. *BMJ*. 2021;372:n71. [PMID: 33782057] [DOI]
29. Tejani AS, Klontzas ME, Gatti AA, Mongan JT, Moy L, Park SH, et al. Checklist For Artificial Intelligence In Medical Imaging (Claim): 2024 Update. *Radiology: Artificial Intelligence*. 2024;6(4):e240300. [DOI]
30. Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, et al. Future-Ai: International Consensus Guideline For Trustworthy And Deployable Artificial Intelligence In Healthcare. *BMJ*. 2025;388:e081554. [DOI]