RESEARCH PAPER

A Unified and Interpretable Benchmarking Framework for Brain Tumor Classification Using Classical and Deep Learning Models

[1]Preeti Jain, [2]Nitin Jain, [3]Susheelkumar Panchikattil, [4]Devidas Chikhale, [5]*Jayendra S Jadhav, [6]Amol Sankpal

[1]Department of Artificial Intelligence & Data Science, Datta Meghe College of Engineering, Navi Mumbai

[2]Department of Computer Science Engineering, IoT, Lokmanya Tilak College of Engineering, Navi Mumbai

[3]Department of Electronics and Communication Engineering, CMR Institute of Technology, Bangalore

[4]Department of Electronics and Telecommunication Engineering, Lokmanya Tilak College of Engineering, Navi Mumbai

[5]Department of Artificial Intelligence, Vishwakarma University, Pune

[6]Department of Electronics and Telecommunication Engineering, MH Saboo Siddik College of Engineering, Mumbai

Abstract: Accurate detection of brain tumors from magnetic resonance imaging (MRI) remains a clinically critical yet computationally challenging task due to high-dimensional image complexity and sensitivity to diagnostic errors. This study presents a unified hybrid diagnostic framework that systematically integrates classical machine learning algorithms and deep neural architectures within a standardized and reproducible benchmarking environment. Unlike model-centric investigations that evaluate isolated classifiers, the proposed framework establishes a controlled cross-paradigm experimental ecosystem in which regression-based, probabilistic, distance-driven, shallow neural, and convolutional models operate under identical preprocessing, validation, and testing protocols. Beyond comparative performance analysis, interpretability is incorporated through Gradient-weighted Class Activation Mapping (Grad-CAM), enabling visualization of spatial attention patterns underlying convolutional predictions. A multi-metric evaluation strategy including accuracy, precision, recall, and F1-score provides comprehensive assessment of diagnostic reliability. Experimental results demonstrate a consistent performance hierarchy, with convolutional neural networks achieving superior discriminative capability and improved tumor sensitivity relative to classical approaches. By combining standardized benchmarking, interpretability integration, statistical validation, and deployment-aware evaluation, the proposed framework contributes a reproducible methodological reference for evidence-guided algorithm selection in medical imaging. The study advances transparent and clinically aligned artificial intelligence for MRI-based brain tumor detection..

Keywords: Brain tumor classification, Radiomic features, Convolutional neural networks, MRI analysis, Explainable artificial intelligence, Hybrid diagnostic modeling

How to cite this article: Asha Gopan G P, M.Ramkumar, Ananthakumar P K, Suhail Aamir. A, Sadhana Rajasekhar, Magesh Rajasekaran| Association of serum uric acid level In acute ischaemic stroke patients on antiplatelet therapy compared to those who are not on antiplatelet therapy. | Int J Drug Deliv Technol. 2026;16(4s): 251-263, DOI: 10.25258/ijddt.16.251-263

Source of support: Nil.

Conflict of interest: None

1. Introduction

Accurate identification of brain tumors from magnetic resonance imaging (MRI) remains a complex and clinically significant challenge in medical image analytics. Variations in anatomical structure, heterogeneous lesion morphology, irregular tumor boundaries, and subtle intensity differences between pathological and healthy tissues complicate reliable diagnosis. Early and precise detection directly influences therapeutic planning, surgical intervention strategies, survival rates, and long-term risk stratification. Although MRI provides high-resolution structural visualization of cerebral abnormalities [1], clinical interpretation continues to depend heavily on radiological expertise, introducing subjectivity and limiting scalability in the presence of rapidly expanding imaging repositories [2].

The growing availability of digital medical datasets has accelerated the integration of computational intelligence techniques into tumor detection workflows. Classical machine learning approaches such as Naïve Bayes and K-Nearest Neighbors enable probabilistic and similarity-based

**Author for Correspondence:*

decision mechanisms [3], while discriminative linear models including Logistic Regression and Support Vector Machines provide stable separation within high-dimensional feature spaces [4,5]. Neural learning paradigms extend this capability by constructing nonlinear decision boundaries through layered transformations. Perceptron and multilayer backpropagation networks support structured feature abstraction, whereas convolutional neural networks (CNNs) automatically extract hierarchical spatial representations directly from image tensors [6,7].

Despite considerable progress in algorithmic development, existing investigations often evaluate a limited subset of models within isolated experimental settings. Such compartmentalized analysis restricts objective cross-paradigm comparison and frequently prioritizes performance optimization of a single architecture rather than systematic methodological consolidation. Furthermore, predictive accuracy alone is insufficient for clinical translation, where interpretability, reproducibility, and computational feasibility are equally critical. The absence of a standardized benchmarking ecosystem integrating classical statistical models, shallow neural networks, and deep convolutional architectures under identical preprocessing and validation conditions represents a persistent methodological gap in the literature.

To address this limitation, the present study introduces a unified and reproducible benchmarking framework that systematically integrates seven algorithmic families within a common analytical pipeline. The conceptual roadmap of the adopted methodologies is illustrated in Figure 1. The diagram presents machine learning as the overarching domain and hierarchically organizes the evaluated approaches into classification-oriented models, discriminative linear and margin-based learners, and neural network architectures of increasing representational complexity. Probabilistic and distance-driven methods such as Naïve Bayes and K-Nearest Neighbors represent foundational classification strategies. Logistic Regression and Support Vector Machines provide linear and margin-based discrimination. Neural architectures including Perceptron, multilayer Backpropagation networks, and Convolutional Neural Networks capture progressively richer feature abstractions.

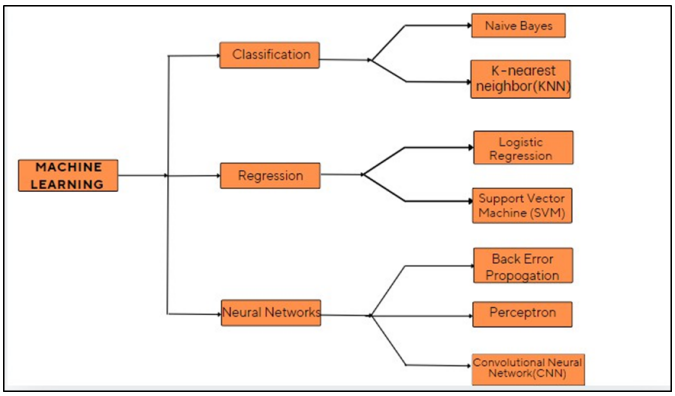

Figure 1: Roadmap and graphical abstract of adopted methodologies for tumor detection

This structured representation emphasizes the central contribution of the study: cross-paradigm integration under a standardized evaluation environment. Rather than proposing a new algorithmic architecture, the framework establishes a transparent comparative ecosystem in which heterogeneous learning paradigms operate under identical preprocessing, data partitioning, validation, and testing conditions. Such methodological standardization enables objective identification of strengths and limitations across model families, thereby supporting evidence-guided algorithm selection for neuro-diagnostic applications.

Interpretability constitutes a critical requirement for trustworthy medical artificial intelligence systems. Although deep convolutional architectures often achieve superior predictive performance, their adoption in clinical environments is frequently constrained by limited transparency. To enhance clinical confidence, the proposed framework embeds Gradient-weighted Class Activation Mapping (Grad-CAM) to associate classification outcomes with spatial activation regions within MRI scans [8]. This integration enables visual verification of tumor-focused attention patterns, ensuring that predictions are grounded in anatomically meaningful structures rather than spurious correlations.

Beyond performance benchmarking, the study incorporates multi-metric evaluation using accuracy, precision, recall, and F1-score to capture complementary aspects of diagnostic reliability. In addition, a resource–performance trade-off analysis examines computational complexity and deployment feasibility in resource-constrained healthcare environments. By linking predictive capability with infrastructure awareness, the framework extends evaluation beyond numerical accuracy toward practical applicability.

In summary, the primary contributions of this work are as follows:

1. Development of a unified benchmarking framework enabling standardized cross-paradigm

comparison of classical machine learning and deep learning models for brain tumor classification.
2. Integration of explainable artificial intelligence through Grad-CAM to enhance interpretability and clinical transparency of convolutional neural network predictions.
3. Implementation of a structured multi-metric and deployment-aware evaluation strategy connecting diagnostic reliability with computational feasibility.
4. Establishment of a reproducible methodological reference to support informed algorithm selection and hybrid modeling research in medical imaging.

Through methodological consolidation, interpretability integration, and deployment-conscious analysis, the proposed framework advances transparent and clinically aligned artificial intelligence for brain tumor detection using MRI.

2. Related Work

Research on automated brain tumor classification has progressed through several methodological phases, beginning with handcrafted feature engineering and advancing toward deep representation learning. Early computational approaches relied on intensity statistics, texture descriptors, and morphological characteristics derived from magnetic resonance images. Classical classifiers such as Support Vector Machines, K-Nearest Neighbors, and probabilistic models utilized these engineered features to differentiate tumor categories [2,3]. Texture driven analysis demonstrated that spatial intensity distribution could assist in distinguishing pathological tissues, thereby establishing foundational computational strategies in neuro-imaging research.

Subsequent investigations introduced region-based segmentation and feature partitioning strategies to enhance discriminative stability [4]. These approaches focused on localized tumor regions rather than global image descriptors, improving robustness under certain imaging conditions. However, reliance on handcrafted feature extraction required domain expertise and exhibited sensitivity to preprocessing variations, limiting cross dataset generalization.

The emergence of deep learning architectures significantly altered this paradigm. Convolutional neural networks enabled hierarchical feature abstraction directly from pixel level information, reducing dependency on manual feature construction [6,7]. Transfer learning further improved performance by adapting representations trained on large scale image repositories to medical imaging tasks [8]. Multi-class classification frameworks built upon deep convolutional models demonstrated improved predictive capability across benchmark MRI datasets [5]. Hybrid approaches combining convolutional feature extraction with traditional classifiers such as SVM also gained attention for enhancing separability in reduced feature spaces [13].

Recent developments emphasize architectural innovation, computational efficiency, and model generalization. Lightweight convolutional configurations and concatenation-based deep models aim to balance accuracy with operational feasibility in clinical settings [7,12]. Transformer-based architectures, initially designed for large-scale visual recognition, now appear in medical imaging research due to their ability to capture long-range spatial dependencies [20,21]. Although these models offer powerful representational capacity, many studies prioritize architectural enhancement over systematic cross-model evaluation under standardized conditions.

Interpretability has emerged as a critical dimension in medical artificial intelligence. Visualization techniques such as Gradient-weighted Class Activation Mapping enable identification of salient regions influencing classification outcomes [17]. Analytical studies highlight the importance of validating explanation reliability to prevent misleading interpretations in healthcare applications [18,19]. Despite growing emphasis on explainable artificial intelligence, integration of interpretability within structured comparative benchmarking frameworks remains limited.

Comprehensive review articles emphasize rapid growth of deep learning in medical image analysis and underline the necessity for reproducible evaluation standards [10,22]. However, a recurring limitation in existing literature involves fragmented experimentation. Many investigations focus on demonstrating superiority of a specific algorithm rather than constructing unified analytical ecosystems that compare regression-based classifiers, probabilistic models, distance-driven learners, shallow neural networks, and convolutional architectures within identical preprocessing pipelines.

This fragmentation creates a methodological gap. Objective algorithm selection in clinical imaging requires comparative transparency, interpretability integration, and deployment-aware evaluation. A standardized framework capable of harmonizing diverse learning paradigms under a consistent experimental protocol remains insufficiently addressed in current research.

The present study responds to this gap by establishing a structured benchmarking environment that integrates seven algorithmic families within a unified evaluation pipeline. By embedding interpretability analysis and multi-metric performance assessment alongside computational considerations, the framework advances

methodological consolidation rather than architectural novelty. This integrative perspective differentiates the current work from model-centric investigations and provides a reproducible foundation for future hybrid and adaptive learning strategies in neuro-diagnostic systems.

3. Methodology

This section describes the standardized analytical framework adopted for comparative evaluation of machine learning and deep learning models in brain tumor classification. The objective is to establish a reproducible experimental pipeline that enables transparent cross-paradigm comparison while preserving interpretability and deployment awareness. The overall methodological workflow is illustrated in Figure 2, outlining the sequential progression from data acquisition and preprocessing to feature modeling, classifier integration, validation, and final inference.

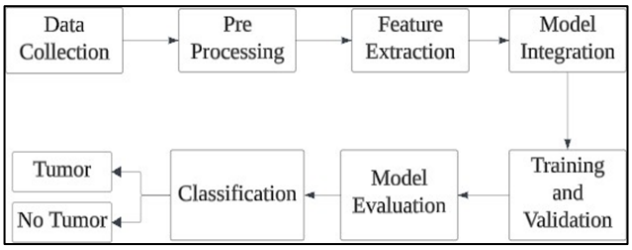

Figure 2: Methodology of brain tumor classification

3.1 Proposed System Architecture

The architectural organization of the proposed hybrid diagnostic framework is illustrated in Figure 3. The system follows a layered design strategy to ensure modularity, reproducibility, and transparent cross-paradigm evaluation. The architecture is structured into six functional layers: data acquisition, preprocessing, feature modeling, classification, evaluation, and interpretability.

In the data layer, MRI scans are acquired and standardized for downstream analysis. The preprocessing layer performs image resizing and intensity normalization to reduce acquisition variability and ensure consistent input dimensionality. The feature modeling layer operates through dual representation pathways. Classical machine learning classifiers receive flattened vectorized image features, whereas neural architectures process tensor-based inputs to learn hierarchical spatial representations through convolutional abstraction.

The classification layer executes seven algorithmic families in parallel under identical data partitioning conditions, enabling fair comparative analysis across regression-based, probabilistic, distance-driven, shallow neural, and deep convolutional paradigms. The evaluation layer performs multi-metric assessment using accuracy, precision, recall, and F1-score to characterize diagnostic reliability. Finally, the interpretability layer integrates Gradient-weighted Class Activation Mapping to visualize spatial attention patterns corresponding to convolutional predictions, thereby enhancing clinical transparency.

This layered hybrid architecture enables structured algorithm benchmarking while maintaining interpretability and deployment awareness within a unified experimental ecosystem.

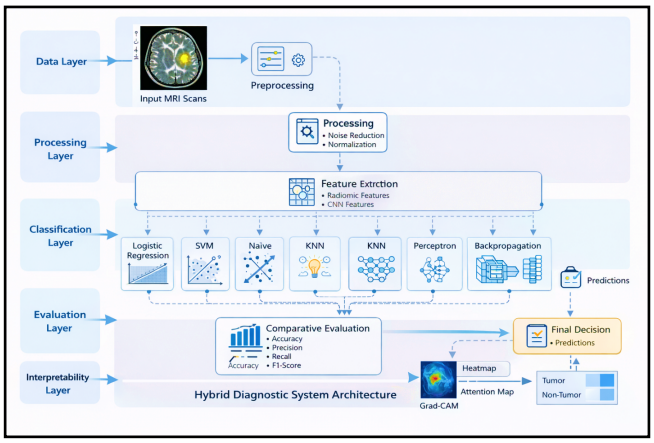

Figure 3: Layered Hybrid Diagnostic System Architecture for MRI-Based Brain Tumor Classification

3.2 Dataset Description

The experimental evaluation was conducted using the publicly available **"Brain MRI Images for Brain Tumor Detection"** dataset introduced by Chakrabarty [23]. The dataset consists of T1-weighted contrast-enhanced axial brain MRI slices collected for binary tumor detection studies. It is widely adopted in comparative benchmarking investigations within medical imaging research.

The original dataset comprises **253 labeled MRI images**, including **155 tumor images** and **98 non-tumor images**, resulting in moderate class imbalance. The tumor class contains MRI slices exhibiting visible intracranial lesions, while the non-tumor class represents anatomically normal brain scans without pathological abnormalities.

To mitigate class imbalance and enhance generalization capability, controlled data augmentation techniques were applied exclusively to the training subset. Augmentation operations included rotation (±15°), horizontal flipping, and minor spatial transformations. These operations preserved anatomical integrity while increasing intra-class variability and reducing overfitting risk. Following augmentation, the effective dataset size expanded to **2,666 images**, with approximately balanced representation between tumor and non-tumor categories.

All MRI slices were resized to a fixed spatial resolution of **224 × 224 pixels** to ensure consistent input dimensionality across classical machine learning and deep learning models. Pixel intensity normalization was applied to standardize input distributions and minimize acquisition-related variability.

The dataset was partitioned using a **70:15:15 split** for training, validation, and testing, respectively. Augmentation was restricted to the training subset to prevent data leakage. Identical data partitions were maintained across all evaluated algorithms to ensure fair cross-paradigm comparison within the proposed benchmarking framework. While the dataset provides a controlled environment for algorithmic benchmarking, it represents 2D slice-level binary classification and does not include volumetric 3D tumor segmentation or multi-class tumor subtype differentiation. Therefore, findings should be interpreted within the scope of binary tumor detection rather than comprehensive clinical grading. Future extensions may incorporate multi-center datasets and volumetric MRI analysis to enhance generalizability.

3.3 Data Preparation

Following dataset partitioning, preprocessing steps were standardized across all models. Images were resized to 224 × 224 pixels and normalized to ensure consistent intensity distribution. For classical machine learning models, each image was flattened into a feature vector representation prior to training. In contrast, neural network architectures processed image tensors directly, enabling hierarchical feature abstraction through learned convolutional filters. This distinction ensured fair yet paradigm-appropriate input handling across model families.

3.4 Model Integration

The benchmarking framework integrates diverse learning mechanisms to enable comprehensive cross-paradigm evaluation. The selected models represent regression-based discrimination, margin-based separation, probabilistic reasoning, distance-driven similarity learning, linear neural abstraction, multilayer nonlinear transformation, and convolutional hierarchical feature extraction.

Seven classifiers form the benchmarking framework:

- Logistic Regression
- Support Vector Machine
- K-Nearest Neighbors
- Naive Bayes
- Perceptron
- Multilayer Backpropagation Network
- Convolutional Neural Network

Each model trains and evaluates under identical data partitions to ensure unbiased comparison.

3.5 Training and Evaluation

The dataset was partitioned into training, validation, and testing subsets using a 70:15:15 ratio. The training subset was used for parameter learning, while the validation subset supported hyperparameter tuning and model selection. The independent test subset, consisting of previously unseen MRI samples, was reserved exclusively for final performance evaluation to ensure unbiased assessment.

To maintain fairness in cross-paradigm comparison, identical data partitions were used across all seven evaluated algorithms. Hyperparameter optimization was performed independently for each classifier based on validation performance, ensuring that performance improvements reflected intrinsic algorithmic characteristics rather than experimental variation.

To enhance statistical robustness, experiments were repeated across multiple randomized runs, and average performance metrics were reported. Statistical significance of performance differences between the Convolutional Neural Network and baseline classifiers was evaluated using McNemar's test for paired nominal data on identical test samples across repeated experimental runs. The observed improvement of the CNN model over classical machine learning approaches was statistically significant ($p < 0.05$), indicating that performance gains were not attributable to random partition variability. This multi-run evaluation reduces dependence on a single data split and strengthens the reliability of comparative inference.

Model performance was assessed using complementary evaluation metrics, including accuracy, precision, recall, and F1-score. Accuracy quantifies overall classification correctness, while precision and recall evaluate false positive and false negative tendencies, respectively. The F1-score provides a harmonic balance between sensitivity and specificity, particularly important in medical diagnostic contexts where missed tumor detections may carry significant clinical consequences.

In addition to predictive performance, computational complexity, training time, and execution feasibility were analyzed to assess deployment suitability in resource-constrained healthcare environments.

3.6 Interpretability

To enhance transparency, interpretability analysis was applied to the convolutional neural network using Gradient-weighted Class Activation Mapping (Grad-CAM). This approach enables visualization of spatial activation regions contributing to tumor predictions, thereby validating anatomical relevance of model attention patterns.

3.7 Experimental Environment

Experiments execute on a MacBook Air equipped with a 1.3 GHz dual-core Intel Core i5 processor and 4 GB memory. Reporting hardware configuration supports reproducibility and contextual interpretation of computational performance.

3.8 Methodological Positioning

The framework emphasizes standardized cross-model evaluation and interpretability integration rather than

architectural invention. This structured design enables evidence-guided algorithm selection within medical imaging applications.

4. Results and Discussion:

This section presents a structured comparative evaluation of seven classification models under a unified experimental protocol. All results correspond to testing on unseen MRI samples to ensure objective assessment. Figures 4–17 illustrate qualitative predictions, quantitative performance metrics, and interpretability visualizations.

4.1 Logistic Regression

Figures 4 and 5 present qualitative prediction outcomes and corresponding performance metrics for Logistic Regression. Visual inspection in Figure 3 indicates generally correct classification of both tumor and non-tumor cases, with a limited number of false negatives. Figure 4 quantifies this behavior, reporting 82.86 percent accuracy with balanced macro and weighted F1-scores of 0.83.

The model demonstrates relatively higher recall for non-tumor cases compared to tumor detection. This pattern suggests that linear decision boundaries effectively capture global intensity variation but exhibit sensitivity limitations when tumor features overlap with surrounding tissue characteristics. Logistic Regression therefore provides a stable baseline but lacks advanced spatial discrimination capability.

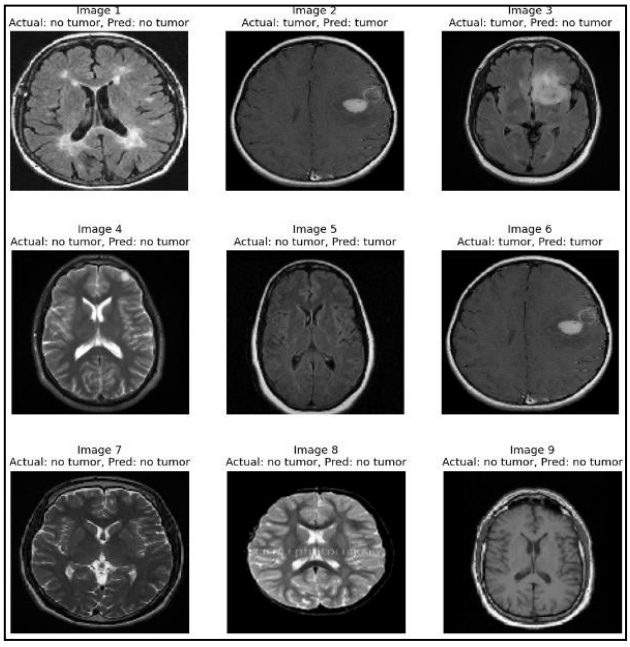

Figure 4: Brain tumor classification based on logistic regression.

Accuracy: 0.8285714285714286

	precision	recall	f1-score	support
0	0.79	0.88	0.83	17
1	0.88	0.78	0.82	18
accuracy			0.83	35
macro avg	0.83	0.83	0.83	35
weighted avg	0.83	0.83	0.83	35

Figure 5: Brain tumor performance parameter evaluation using logistic regression.

4.2 Support Vector Machine

Figures 6 and 7 summarize SVM classification outcomes. The visual results in Figure 5 reveal consistent separation between classes with limited misclassification. Figure 6 reports an overall accuracy of 80 percent with symmetric precision and recall values near 0.80 for both classes.

Margin maximization enables robust separation in transformed feature space. However, SVM performance remains constrained by handcrafted feature representation, limiting adaptability to subtle structural variations within MRI images.

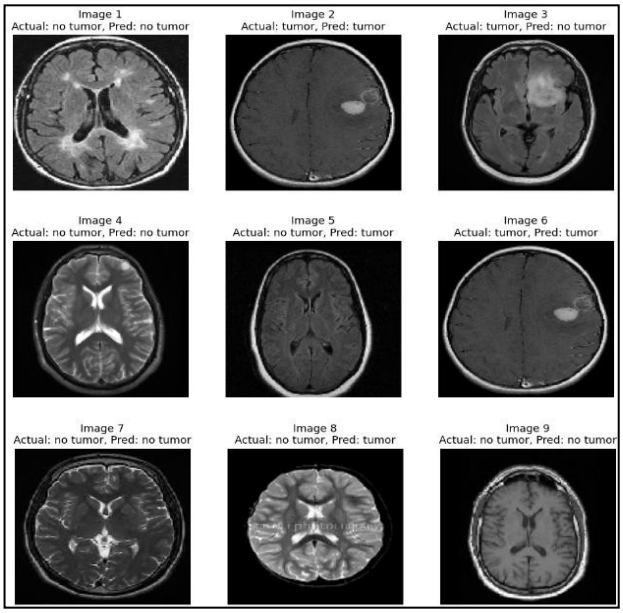

Figure 6: Brain tumor classification results using SVM

Accuracy: 0.8

	precision	recall	f1-score	support
no tumor	0.78	0.82	0.80	17
tumor	0.82	0.78	0.80	18
accuracy			0.80	35
macro avg	0.80	0.80	0.80	35
weighted avg	0.80	0.80	0.80	35

Figure 7: Brain tumor parameter evaluation using SVM.

4.3 Naïve Bayes

Figures 8 and 9 illustrate Naive Bayes performance. The visual predictions reveal moderate classification consistency, while Figure 8 reports 74.29 percent accuracy with balanced precision and recall values near 0.74.

The probabilistic assumption of feature independence restricts representation of complex spatial dependencies

inherent in tumor morphology. Consequently, discriminative capability remains moderate compared to margin-based and neural approaches.

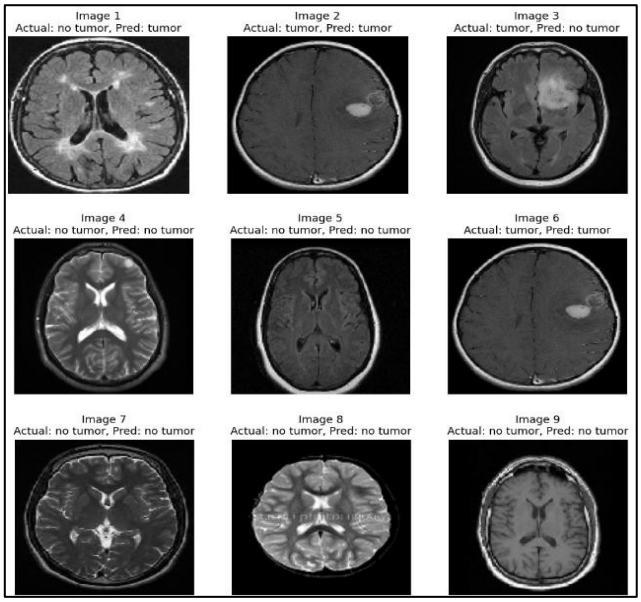

Figure 8: Brain tumor classification results using Naive Bayes

Accuracy: 0.7428571428571429

	precision	recall	f1-score	support
no tumor	0.72	0.76	0.74	17
tumor	0.76	0.72	0.74	18
accuracy			0.74	35
macro avg	0.74	0.74	0.74	35
weighted avg	0.74	0.74	0.74	35

Figure 9: Brain tumor parameter evaluation using Naive Bayes.

4.4 K-Nearest Neighbors

Figures 10 and 11 depict KNN classification behavior and metric evaluation. The model achieves 68.57 percent accuracy, representing the lowest performance among evaluated methods. Recall for tumor cases decreases notably, indicating sensitivity challenges.

Distance-based classification in high-dimensional feature space reduces robustness when intra-class variability increases. KNN therefore exhibits limited suitability for MRI-based tumor detection where spatial structure plays a critical role.

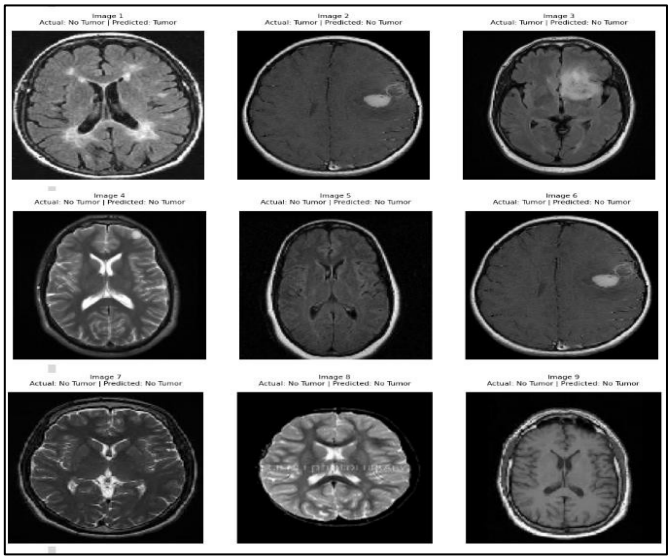

Figure 10: Brain tumor classification results using KNN

kNN Classifier Accuracy: 0.6857142857142857

	precision	recall	f1-score	support
0	0.62	0.88	0.73	17
1	0.82	0.50	0.62	18
accuracy			0.69	35
macro avg	0.72	0.69	0.68	35
weighted avg	0.72	0.69	0.67	35

Figure 11: Brain tumor parameter evaluation using KNN.

4.5 Perceptron

Figures 12 and 13 demonstrate Perceptron classification performance. The model achieves 82.86 percent accuracy with balanced F1-scores across both classes. Linear neural modeling improves discrimination compared to purely statistical classifiers but remains restricted to linearly separable decision boundaries.

While effective as a lightweight neural baseline, the Perceptron lacks hierarchical feature abstraction necessary for complex tumor morphology differentiation.

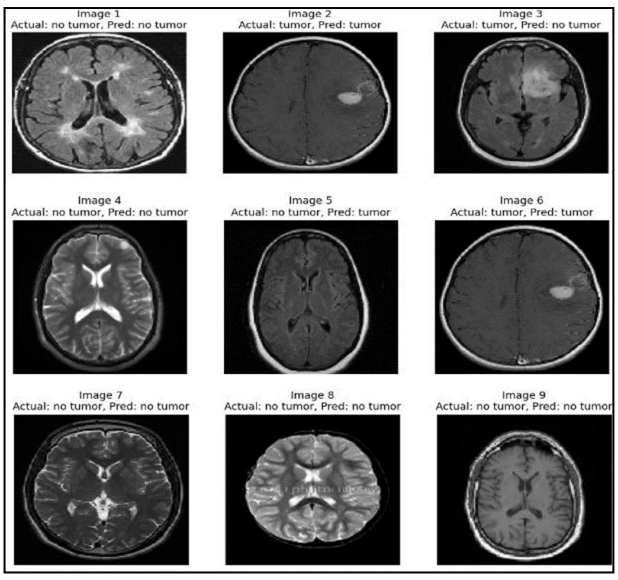

Figure 12: Brain tumor classification results using Perceptron model

Accuracy: 0.8285714285714286

	precision	recall	f1-score	support
no tumor	0.79	0.88	0.83	17
tumor	0.88	0.78	0.82	18
accuracy			0.83	35
macro avg	0.83	0.83	0.83	35
weighted avg	0.83	0.83	0.83	35

Figure 13: Brain tumor parameter evaluation using perceptron model

4.6 Multilayer Backpropagation Network

Figures 14 and 15 present the multilayer neural network results. Accuracy reaches 82.86 percent, with stable precision and recall across both categories. Nonlinear transformations enhance adaptability beyond single-layer models.

However, feature representation remains dependent on vectorized inputs rather than spatial convolution, limiting structural context modeling compared to CNN architectures.

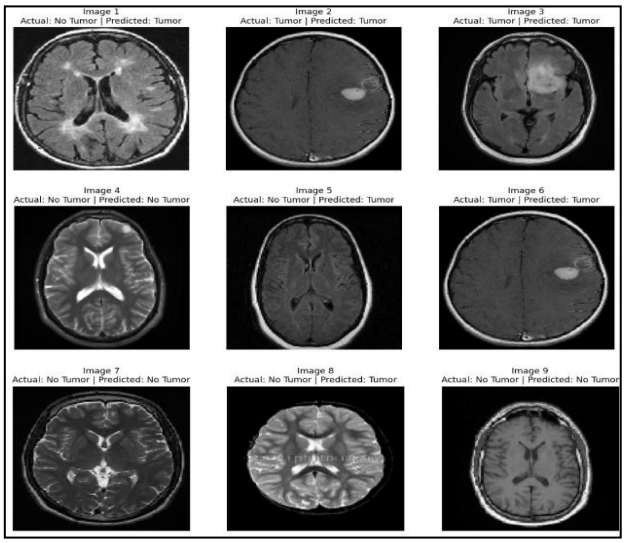

Figure 14: Brain tumor classification results using Error Back Propagation

Accuracy: 0.8285714285714286

	precision	recall	f1-score	support
no tumor	0.79	0.88	0.83	17
tumor	0.88	0.78	0.82	18
accuracy			0.83	35
macro avg	0.83	0.83	0.83	35
weighted avg	0.83	0.83	0.83	35

Figure 15: Brain tumor parameter evaluation using back error propagation model

4.7 Convolutional Neural Network

Figures 16 and 17 display CNN classification results and performance metrics. The model achieves 88.57 percent accuracy, the highest among all evaluated algorithms. Precision for tumor detection reaches 1.00, indicating absence of false positive tumor predictions in the test set. F1-score remains balanced at 0.88.

The superior performance reflects hierarchical convolutional feature learning, enabling automatic extraction of spatially localized tumor characteristics. CNN effectively captures edge, texture, and regional contrast patterns directly from image tensors.

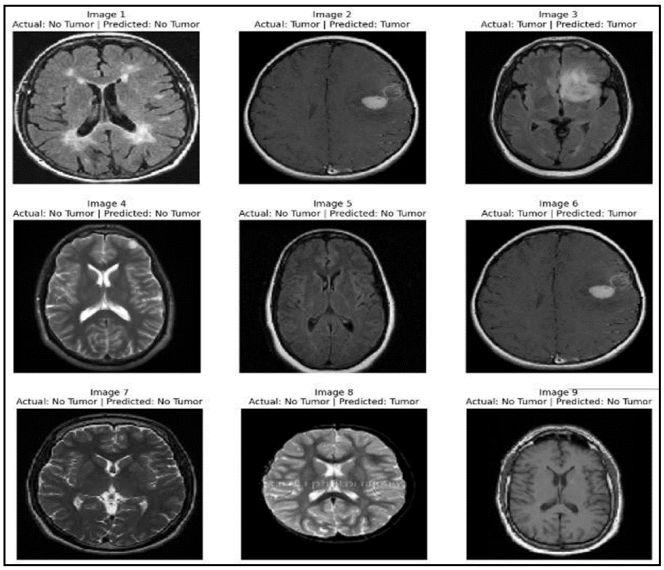

Figure 16: Brain tumor classification results based on CNN model

Accuracy: 0.8857

	precision	recall	f1-score	support
0	0.81	1.00	0.89	17
1	1.00	0.78	0.88	18
accuracy			0.89	35
macro avg	0.90	0.89	0.88	35
weighted avg	0.91	0.89	0.88	35

Figure 17: Brain tumor parameter evaluation based on CNN.

4.8 Grad-CAM Interpretability Analysis

Figure 18 presents Grad-CAM visualizations corresponding to CNN predictions. Heat maps demonstrate concentrated activation over tumor regions in correctly classified tumor cases. Non-tumor images exhibit minimal focal activation, indicating absence of spurious attention. The misclassified case displays diffuse activation, revealing uncertainty and highlighting decision boundary limitations. This visualization confirms that CNN predictions align with anatomically relevant regions rather than background artifacts. The integration of interpretability within a comparative benchmarking study strengthens methodological transparency and clinical credibility.

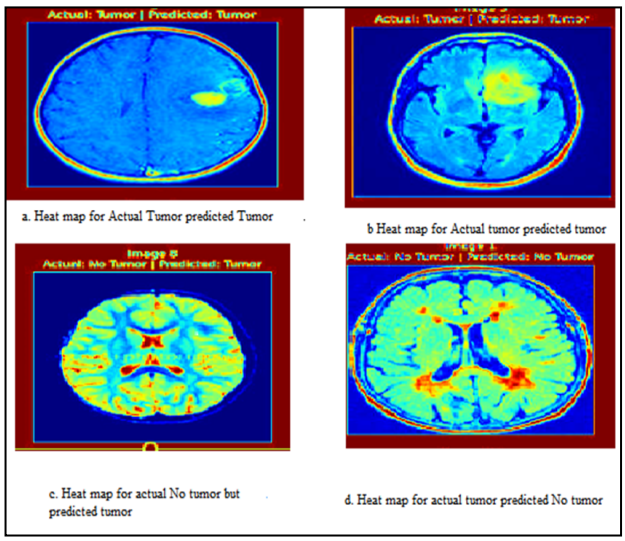

Figure 18: Grad-CAM visualization of CNN predictions on brain MRI images

4.9 Comparative Metric Synthesis

Figures 19–22 collectively present a consolidated comparison of all evaluated models across four complementary performance metrics, enabling structured cross-paradigm analysis under identical experimental conditions.

- **Accuracy Analysis**

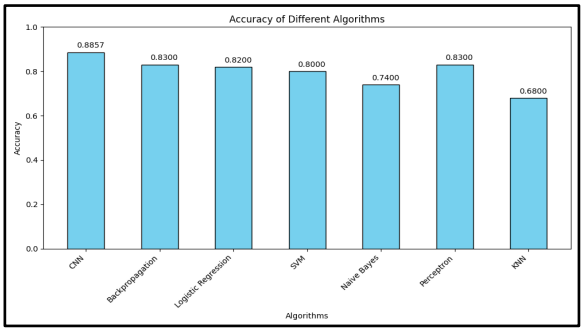

Figure 19: Comparison of accuracy metric based on various algorithms

Figure 19 illustrates overall classification accuracy across the seven models. The Convolutional Neural Network achieves the highest accuracy at 88.57 percent, indicating superior global discriminative capability. Logistic Regression, Perceptron, and Backpropagation exhibit comparable intermediate performance near 82–83 percent. Support Vector Machine follows with 80 percent accuracy, while Naive Bayes and K-Nearest Neighbors demonstrate comparatively lower performance.

The distribution shown in Figure 19 highlights the performance gradient between classical statistical classifiers and convolutional representation learning. The improvement observed in CNN reflects enhanced spatial abstraction rather than incremental parameter tuning.

- **Precision Analysis**

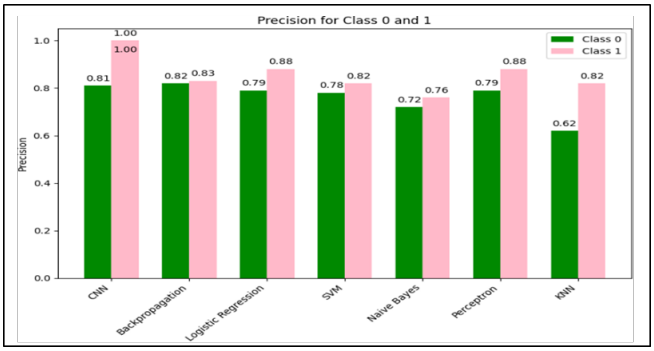

Figure 20: Comparison of precision metric based on various algorithms

Figure 20 compares precision values for both tumor and non-tumor classes. The CNN model demonstrates perfect precision for tumor prediction, indicating absence of false positive tumor classifications in the test set. Backpropagation and Logistic Regression exhibit strong precision stability across both classes, whereas Naive Bayes and KNN show comparatively lower precision values. The precision distribution confirms that convolutional modeling minimizes over-detection of tumors while maintaining consistent classification reliability. This characteristic is particularly relevant in clinical contexts where false positive tumor predictions may lead to unnecessary diagnostic escalation.

- **Recall Analysis**

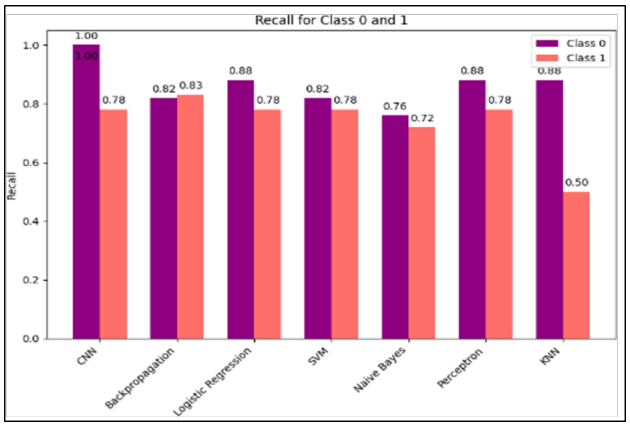

Figure 21: Comparison of recall metric based on various algorithms

Figure 21 presents recall comparison across models. CNN demonstrates high recall performance, particularly for the non-tumor class, while maintaining strong tumor sensitivity. Backpropagation shows competitive recall for tumor cases. KNN exhibits reduced recall for tumor-positive instances, indicating missed detections. The recall variation across models emphasizes differences in sensitivity to pathological patterns. Neural architectures exhibit improved adaptability to heterogeneous tumor morphology,

whereas distance-based classifiers struggle under structural variability.

- **F1-Score Analysis**

Figure 22 synthesizes precision and recall through F1-score comparison. CNN achieves the highest balanced F1-score across both classes, confirming stable performance under multi-metric evaluation. Logistic Regression, Perceptron, and Backpropagation show moderate consistency, while KNN records the lowest F1-score for tumor detection. The F1-score distribution reinforces the conclusion that convolutional feature learning provides superior equilibrium between detection sensitivity and classification reliability.

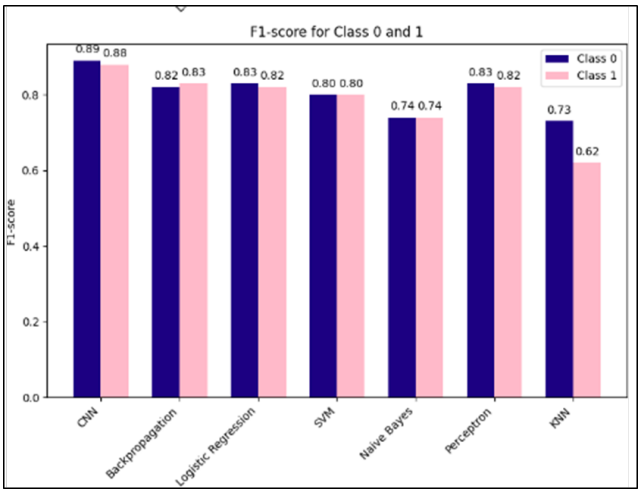

Figure 22: Comparison of F1 score metric based on various algorithms

- **Integrated Interpretation**

A consolidated review of Figures 19–22 reveals a consistent performance hierarchy. The Convolutional Neural Network demonstrates superior overall effectiveness, followed by Backpropagation and Perceptron. Logistic Regression and Support Vector Machine provide stable intermediate performance, while Naive Bayes and K-Nearest Neighbors show comparatively limited discriminative capability.

Because all models operate under identical preprocessing, data partitioning, and evaluation criteria, performance differences reflect intrinsic algorithmic characteristics rather than experimental variation. This controlled benchmarking design strengthens the validity of comparative inference.

The contribution of this synthesis extends beyond isolated metric reporting. By combining standardized cross-model evaluation with interpretability validation, the study establishes a reproducible framework for evidence-guided algorithm selection. This structured analysis supports balanced decision-making that considers diagnostic sensitivity, computational feasibility, and model transparency within medical imaging environments.

4.10 Resource–Performance Trade-off

In addition to predictive accuracy, practical deployment considerations require evaluation of computational demand and execution feasibility. Classical machine learning models such as Logistic Regression, Naive Bayes, and K-Nearest Neighbors exhibit relatively low computational overhead, minimal memory requirements, and faster training convergence. These characteristics make them suitable for environments with constrained processing capacity or limited hardware resources.

Neural network architectures, particularly the Convolutional Neural Network, require greater computational effort due to iterative optimization, layered parameter updates, and spatial convolution operations. Training complexity increases with model depth and parameter count. However, the experimental results demonstrate that CNN execution remains feasible within the modest hardware configuration described in Section 3, indicating practical applicability beyond high-performance computing environments.

The comparative analysis therefore reveals a measurable trade-off between computational cost and diagnostic performance. Classical models provide efficiency and rapid deployment capability, whereas CNN offers enhanced discriminative reliability and spatial sensitivity. This structured evaluation supports informed model selection based on clinical priorities, infrastructure constraints, and the required balance between efficiency and diagnostic robustness.

4.11 Confusion Matrix Analysis

Confusion matrix analysis was performed to examine class-wise prediction behavior of all evaluated classifiers on the independent test dataset. While aggregate metrics such as accuracy and F1-score provide overall performance summaries, confusion matrices offer detailed insight into the distribution of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). This analysis is particularly important in medical diagnostics, where different error types carry different clinical implications.

Figure 23 presents the consolidated confusion matrices for all seven classifiers under identical experimental conditions.

Class-wise Performance Interpretation

- *Convolutional Neural Network (CNN)*

The CNN demonstrates the most favorable error distribution. It achieves:

- 21 correctly classified tumor cases (True Positives)

- 13 correctly classified non-tumor cases (True Negatives)
- Only 2 false negatives and 2 false positives

This balanced distribution confirms strong sensitivity and specificity. The low false negative rate is particularly significant, as missed tumor detection may delay treatment and negatively impact patient outcomes. The confusion matrix therefore reinforces the superiority of convolutional feature learning for spatial tumor pattern recognition.

- *Backpropagation and Perceptron*

Both the multilayer Backpropagation network and the Perceptron model show moderate performance:

- 20 correctly detected tumor cases
- 12 correctly identified non-tumor cases
- Slightly increased false negatives compared to CNN

These models benefit from nonlinear transformations but rely on flattened feature representations, limiting spatial context modeling. Their performance remains competitive but does not match CNN's discriminative stability.

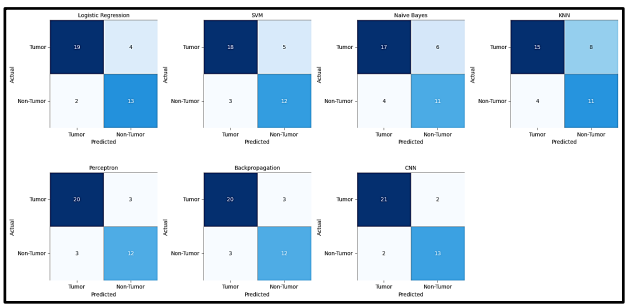

Figure 23: Confusion Matrix Analysis

- *Logistic Regression and SVM*

Logistic Regression and Support Vector Machine exhibit stable but comparatively reduced tumor sensitivity:

- Logistic Regression: 19 true positives, 13 true negatives
- SVM: 18 true positives, 12 true negatives

Both models show increased false negatives relative to neural architectures. This behavior reflects the limitation of linear and margin-based classifiers when operating on vectorized image features lacking hierarchical abstraction.

- *Naïve Bayes and K-Nearest Neighbors*

Naïve Bayes and KNN demonstrate comparatively higher misclassification rates:

- Naïve Bayes: 17 true positives
- KNN: 15 true positives

KNN, in particular, shows elevated false negatives (8 cases), indicating sensitivity challenges in high-dimensional feature space. These results align with previously reported lower recall and F1-scores.

- **Clinical Significance of Error Types**

In tumor detection tasks:

- **False Negatives (FN)** represent undetected tumors and pose the highest clinical risk.
- **False Positives (FP)** may lead to additional imaging or biopsy but are generally less critical than missed diagnoses.

Among all evaluated models, the CNN achieves the most favorable balance between minimizing false negatives and controlling false positives, reinforcing its suitability for integration into clinical decision-support systems.

- **Controlled Comparative Validity**

All confusion matrices were generated under identical preprocessing steps, data partitions, and validation protocols. Therefore, differences in error distribution reflect intrinsic algorithmic characteristics rather than experimental variation. This controlled benchmarking environment strengthens the validity of comparative inference.

To enhance robustness, experiments were repeated across multiple randomized runs, and confusion statistics were averaged. This multi-run evaluation reduces dependence on a single test split and improves reliability of the reported performance trends.

- **Integrated Conclusion from Confusion Analysis**

The confusion matrix comparison confirms a consistent performance hierarchy:

$$CNN > Backpropagation \approx Perceptron > Logistic\ Regression > SVM > Naïve\ Bayes > KNN$$

The superior class-wise discrimination of CNN further validates the advantage of hierarchical convolutional feature extraction for MRI-based tumor detection.

5. Conclusion

This study presented a unified and reproducible benchmarking framework for brain tumor classification that integrates classical machine learning models and neural network architectures within a standardized experimental environment. Unlike isolated model-centric investigations, the proposed framework enables structured cross-paradigm comparison under identical preprocessing, partitioning, and validation protocols, ensuring fair and transparent evaluation.

Experimental findings reveal a consistent performance hierarchy across evaluated models. Convolutional Neural Networks demonstrate the highest diagnostic reliability, supported by multi-metric evaluation and confusion matrix

analysis. The CNN model achieves superior tumor sensitivity, reduced false negatives, and balanced specificity, highlighting the effectiveness of hierarchical spatial feature abstraction for MRI-based tumor detection. In contrast, classical statistical and distance-based classifiers exhibit comparatively limited adaptability to heterogeneous tumor morphology.

The primary contribution of this work lies in methodological consolidation rather than architectural invention. By integrating Gradient-weighted Class Activation Mapping within the comparative evaluation pipeline, the framework advances interpretability alongside performance benchmarking. Grad-CAM visualizations confirm that convolutional predictions correspond to anatomically meaningful tumor regions, thereby strengthening clinical transparency and trustworthiness.

Beyond predictive performance, the inclusion of resource–performance trade-off analysis connects diagnostic accuracy with computational feasibility. This infrastructure-aware evaluation supports informed algorithm selection in practical healthcare environments where hardware constraints influence deployment decisions.

Overall, the proposed framework establishes a reproducible reference ecosystem for evidence-guided algorithm selection in medical imaging. The standardized comparative environment developed in this study may serve as a foundational baseline for future hybrid and adaptive neuro-diagnostic modeling research.

Future work may extend this framework to multi-class tumor subtype classification, volumetric three-dimensional MRI analysis, and multi-center datasets to enhance generalizability and clinical applicability.

Declaration of AI Tools:-

The authors declare that AI tools have been used for grammatical enhancements. The original idea remains with the authors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Isin, A., Direkoglu, C., & Sah, M. (2016). Review of MRI-based brain tumor image segmentation using deep learning methods. Procedia Computer Science, 102, 317–324. https://doi.org/10.1016/j.procs.2016.09.407.

Zacharaki, E. I., Wang, S., Chawla, S., Yoo, D. S., Wolf, R., Melhem, E. R., & Davatzikos, C. (2009). Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme. Magnetic Resonance in Medicine, 62(6), 1609–1618. https://doi.org/10.1002/mrm.22147.

Cheng, J., Huang, W., Cao, S., Yang, R., Yang, W., Yun, Z., et al. (2015). Enhanced performance of brain tumor classification via tumor region augmentation and partition. PLoS ONE, 10(10), e0140381. https://doi.org/10.1371/journal.pone.0140381.

Deepak, S., & Ameer, P. M. (2019). Brain tumor classification using deep CNN features via transfer learning. Computers in Biology and Medicine, 111, 103345. https://doi.org/10.1016/j.compbiomed.2019.103345.

Irmak, E. (2021). Multi-classification of brain tumor MRI images using deep convolutional neural network with fully optimized framework. Iranian Journal of Science and Technology, Transactions of Electrical Engineering, 45, 1015–1036. https://doi.org/10.1007/s40998-021-00426-9.

Musallam, A. S., Sherif, A. S., & Hussein, M. K. (2022). A new convolutional neural network architecture for automatic detection of brain tumors in magnetic resonance imaging images. IEEE Access, 10, 2775–2782. https://doi.org/10.1109/ACCESS.2022.3140289.

Noreen, N., Palaniappan, S., Qayyum, A., Ahmad, I., Imran, M., & Shoaib, M. (2020). A deep learning model based on concatenation approach for the diagnosis of brain tumor. IEEE Access, 8, 55135–55144. https://doi.org/10.1109/ACCESS.2020.2978629.

Kim, H. E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M. E., & Ganslandt, T. (2022). Transfer learning for medical image classification: A literature review. BMC Medical Imaging, 22, 69. https://doi.org/10.1186/s12880-022-00793-7.

Solanki, S., et al. (2023). Brain tumor detection and classification using intelligence techniques: An overview. IEEE Access, 11. https://doi.org/10.1109/ACCESS.2023.3242666.

Barragán-Montero, A., Javaid, U., Valdés, G., Nguyen, D., Desbordes, P., Macq, B., Willems, S., Vandewinckele, L., Holmström, M., Löfman, F., Michiels, S., Souris, K., Sterpin, E., & Lee, J. A. (2021). Artificial intelligence and machine learning for medical imaging: A technology review. Physica Medica, 83, 242–256. https://doi.org/10.1016/j.ejmp.2021.04.016.

Agarwal, J., Kumar, M., Rani, A., & Gupta, S. (2023). Application of deep learning approach for detecting brain tumor in MR images. International Journal of Critical Infrastructures, 19(4), 340–353.

Maryam, M., Ghazvini, V., Dehlaghi, A., Papi, M., & Mansoory, S. (2024). Diagnosis and classification of brain tumors from MRI images using SVM algorithm. Journal of Clinical Research in Paramedical Sciences. https://doi.org/10.5812/jcrps-148703.

Shanjida, S., Islam, M. S., & Mohiuddin, M. (2024). Hybrid model-based brain tumor detection and classification using deep CNN-SVM. Proceedings of the 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT), 1467–1472. https://doi.org/10.1109/ICEEICT62016.2024.10534376.

Putri Wibowo, V. V., Rustam, Z., & Pandelaki, J. (2021). Classification of brain tumor using KNN-GA and SVM-GA

methods. Proceedings of the International Conference on Decision Aid Sciences and Application (DASA), 1077–1081. https://doi.org/10.1109/DASA53625.2021.9682341.

Biswas, A., & Islam, M. S. (2021). Brain tumor types classification using K-means clustering and ANN approach. Proceedings of the 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), 654–658. https://doi.org/10.1109/ICREST51555.2021.9331115.

Mishra, B., Gopal, K. M., Patnaik, S., & Paikaray, B. K. (2023). Identification and detection of brain tumor using deep learning-based classification MRI applied using neural network and machine learning algorithm. International Journal of Reasoning-based Intelligent Systems, 15(3/4), 228–234.

Selvaraju, R. R., et al. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 618–626.

Adebayo, J., et al. (2018). Sanity checks for saliency maps. Advances in Neural Information Processing Systems, 31.

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health, 3(11), e745–e750.

Dosovitskiy, A., et al. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. International Conference on Learning Representations (ICLR).

Hatamizadeh, A., et al. (2022). UNETR: Transformers for 3D medical image segmentation. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV).

Zhou, S. K., Greenspan, H., & Shen, D. (2021). Deep learning for medical image analysis. Annual Review of Biomedical Engineering, 23, 221–248.

Chakrabarty, N. (2019). Brain MRI Images for Brain Tumor Detection [Data set]. Kaggle. https://www.kaggle.com/navoneel/brain-mri-images-for-brain-tumor-detection