RESEARCH ARTICLE

An Explainable Hybrid Deep Learning Framework with Multi-Scale Feature Fusion and Uncertainty-Aware Clinical Decision Support for Brain Tumor Classification

Mubashshera Mirza[1,4,*], **Saleem Pasha Choudhari Rasool**[2], **Mohammed Salman Baig**[3], **Saadiah Hassan**[3], **Rehan Ahmad**[5] and **Anish Pandey**[6,*]

[1] Department of Computer Science and Engineering, Indian Institute of Technology Patna, Patna 801106, India
[2] Department of Computer Science and Engineering (AI & ML), Theem College of Engineering, Maharashtra, India
[3] Department of Computer, E&TC Engineering, Maulana Mukhtar Ahmad Nadvi Technical Campus, Maharashtra, India
[4] Department of Computer Engineering, Maulana Mukhtar Ahmad Nadvi Technical Campus, Maharashtra, India
[5] Mukesh Patel School of Technology Management and Engineering, SVKM's NMIMS, Maharashtra, India
[6] Mechatronics Lab, School of Mechanical Engineering, Kalinga Institute of Industrial Technology (KIIT) Deemed to be University, Bhubaneswar 751024, India

Abstract

Accurate diagnosis of brain tumors from magnetic resonance imaging (MRI) is crucial but is challenged by inter-observer variability and by the limited explainability of deep models. Most existing studies use standalone CNNs or Transformers without uncertainty assessment or explanation capabilities. We propose an explainable hybrid deep learning framework that combines multi-scale feature fusion with uncertainty-aware clinical decision support. On a four-class brain MRI benchmark, it was compared with EfficientNetV2, ConvNeXt, and Swin Transformer. Five explainability methods (Grad-CAM, SHAP, LIME, Integrated Gradients, and Attention Rollout) explain predictions, while Monte Carlo dropout quantifies predictive uncertainty; confidence and entropy form a Trust Score for three-tier referral. The hybrid model outperformed the baselines, achieving a test accuracy of 95.7%, a macro F1-score of 95.5%, and an AUC of 0.989. A preliminary 12-reader evaluation suggests that the clinical decision support system (CDSS) report is usable (mean System Usability Scale 78.3/100) and reduces mean reading time per case from 42.6 s to 34.8 s versus radiologist-alone reading, although underpowered for clinical endpoints. Prospective multi-institutional validation remains future work.

Keywords: brain tumor classification, explainable AI, hybrid CNN-Transformer, clinical decision support, multi-scale feature fusion, MRI, Grad-CAM.

Submitted: 15 Sep 2026
Revised: 04 Nov 2026
Accepted: 06 Nov 2026
Published: 18 Nov 2026

Vol. 3, **No.** 4, 2026.
10.62762/TETAI.2026.693634

***Corresponding authors:**
✉ Mubashshera Mirza
mubashshera_pa2502mth54@iitp.ac.in
✉ Anish Pandey
anish.pandeyfme@kiit.ac.in

Citation
Mirza, M., Rasool, S. P. C., Baig, M. S., Hassan, S., Ahmad, R., & Pandey, A. (2026). An Explainable Hybrid Deep Learning Framework with Multi-Scale Feature Fusion and Uncertainty-Aware Clinical Decision Support for Brain Tumor Classification. *ICCK Transactions on Emerging Topics in Artificial Intelligence*, 3(4), 267–305.

© 2026 by the Authors. Published by Institute of Central Computation and Knowledge. This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

1 Introduction

Magnetic resonance imaging (MRI) plays a critical role in diagnosing brain tumors, but interpretation of the results can be subjective, time-consuming, and subject to interobserver variability. Early and accurate classification significantly affects treatment planning and patients' survival rates. Even though deep learning is used to automate the process of brain tumor detection, the black-box nature of deep learning models limits their application in practice because of the lack of transparency necessary for clinical usage; the development of trusted and explainable AI has therefore become an active topic in neuro-oncology.

Most existing studies focus on either convolutional neural networks (CNNs) or Vision Transformers (ViTs) and rarely combine the two, although CNNs extract local texture information and ViTs capture global context. In addition, existing studies seldom consider multi-scale attention, compare different explainability techniques, or estimate predictive uncertainty. As a result, clinicians have limited grounds to trust these models, which hinders the development of reliable clinical decision support systems.

The aim of this study is to develop a hybrid CNN-Transformer model for multi-scale feature extraction. The study further integrates a multi-method explainability framework, estimates predictive uncertainty by means of Monte Carlo dropout, and compares four architectures, with the goal of building a trustworthy clinical decision support system (CDSS). The contributions of this work are framed as emerging directions and open problems rather than isolated benchmark gains: (1) the creation of a novel hybrid CNN-Transformer architecture with multi-stage feature extraction, (2) the introduction of a five-method explainability framework (Grad-CAM, SHAP, LIME, Integrated Gradients, Attention Rollout) with uncertainty estimation, (3) attention-based visualizations for diagnostic purposes, (4) a comprehensive statistical validation of EfficientNetV2, ConvNeXt, Swin Transformer and our hybrid model. Complementing the benchmark comparison, Section 8 reports a preliminary reader-preference and human-factors evaluation of the CDSS with 12 readers (8 board-certified radiologists and 4 senior radiology residents) across 2 institutions, providing an initial signal on usability, reading time, and trust; a fully powered clinical reader study remains future work. The hybrid model reaches a test accuracy of 95.7%.

1.1 Emerging Directions and Open Problems

Rather than claiming novelty solely through benchmark accuracy, this work opens several emerging directions for the community:

1. **Uncertainty-aware explainability fusion.** We introduce a Trust Score (τ) that mathematically couples predictive confidence with normalized predictive entropy, providing a principled mechanism to gate clinical recommendations. This raises the problem of *how to calibrate explanation fidelity against predictive uncertainty* – a question not addressed by existing XAI or uncertainty literature in isolation.

2. **Dynamic modality weighting.** The learnable α/β attention maps in the MSAF module suggest a generalizable paradigm for *input-adaptive fusion* that could extend beyond CNN-Transformer pairs to multimodal MRI sequences (T1/T2/FLAIR) or even cross-sensor fusion in other domains.

3. **Multi-method XAI consensus.** Our comparison suggests that no single XAI method is sufficient for clinical-grade explanation. This motivates the emerging problem of *explanation consensus metrics* – how to quantitatively aggregate Grad-CAM, SHAP, LIME, Integrated Gradients, and Attention Rollout into a single faithful explanation. Formally, given explanation maps $E_1, \ldots, E_5$ and a fidelity functional $\mathcal{F}$, the consensus problem is to find $E^* = \arg\min_E \sum_{i=1}^{5} w_i \, d(E, E_i)$ subject to $\mathcal{F}(E) \geq \mathcal{F}(E_i) \; \forall i$, where $d(\cdot, \cdot)$ is a distributional distance and w_i are method-specific reliability weights.

4. **Trust-gated clinical workflow.** The rule-based CDSS recommendation thresholds (Routine Review / Radiologist Verification / Tumor Board) open the problem of *learning optimal referral policies* from retrospective clinical outcomes rather than fixed thresholds. This can be cast as a constrained Markov decision process where the state is the (confidence, entropy, explanation-consensus) triplet, the action is the referral tier, and the reward trades off diagnostic accuracy against reading cost.

These directions are distinct from concurrent work that treats fusion, explainability, and uncertainty as independent modules; our contribution is the *joint formulation* and the open problems it exposes.

1.2 Novelty and Distinction from Concurrent Work

While the individual components of our framework – CNN-Transformer hybrids, Grad-CAM/SHAP/LIME, MC Dropout, and attention-based fusion – are established, the following aspects have, to our knowledge, not been addressed together in the literature we reviewed:

- **Joint uncertainty-explainability formulation.** Prior work applies XAI and uncertainty estimation as independent post-hoc modules. We formulate a *Trust Score* $\tau = f(\text{confidence}, \text{entropy})$ that couples the two, enabling the CDSS to modulate recommendations based on *both* predictive certainty and explanation reliability.

- **Input-adaptive multi-scale fusion.** Unlike the fixed or scalar fusion schemes in earlier hybrid CNN-Transformer MRI models [1, 2], our MSAF module learns spatial attention maps $\alpha(x, y), \beta(x, y)$ with $\alpha + \beta = 1$ at every pixel, allowing the network to dynamically prioritize local vs. global features per pathology subtype. This differs from the feature-concatenation approach in [3] and the multi-scale convolutional processing without learned fusion weights in [4]. Quantitatively, the controlled ablation in Section 7 shows that replacing MSAF and the SE/CBAM recalibration with naive concatenation reduces accuracy by 2.3 percentage points (95.7% $\rightarrow$ 93.4%) under identical hyperparameters.

- **Multi-method XAI consensus for clinical deployment.** Our comparison indicates that no single XAI method is sufficient, and we propose a cascaded selection strategy based on the speed-sharpness trade-off, which is not addressed in the brain-tumor classification studies [3, 4], as neither reports a comparison of multiple XAI methods.

- **Trust-gated CDSS policy.** The three-tier recommendation rule (Routine / Verification / Tumor Board) based on τ thresholds provides an auditable decision policy that is not described in the CDSS overviews and human-AI collaboration studies we reviewed [5, 6]. A preliminary reader-preference evaluation of this policy is reported in Section 8.

These contributions are framed as *open problems and design paradigms* rather than isolated performance gains.

1.3 Explicit Statement of Emergent Contributions

While the individual building blocks of the present framework – CNN-Transformer hybrids, Grad-CAM/SHAP/LIME, Monte Carlo Dropout, and attention-based fusion – are mature and have each been applied to brain tumor MRI classification in prior literature, the contribution of this work is not the invention of any single component. The new elements are three formal objects and one joint formulation that do not appear in the literature reviewed in Section 2.

Contribution 1: A formally defined, calibrated Trust Score for clinical gating. Prior work on uncertainty-aware medical AI reports predictive entropy and maximum softmax confidence as *separate* diagnostics, and prior work on explainability reports attribution maps as *post-hoc* visualizations. We define a single scalar, the Trust Score $\tau(\mathbf{x}) \in [0, 1]$, that couples the two into one auditable quantity:

$$\tau(\mathbf{x}) = \underbrace{\big(1 - \hat{H}(\mathbf{x})\big)}_{\text{calibrated certainty}} \cdot \underbrace{c_{\max}(\mathbf{x})}_{\text{peak confidence}}, \qquad (1)$$

where $\hat{H}(\mathbf{x}) = H(\mathbf{x})/\log K$ is the normalized predictive entropy over $K{=}4$ classes (so $\hat{H} \in [0, 1]$), and $c_{\max}(\mathbf{x}) = \max_k \bar{p}_k(\mathbf{x})$ is the largest class probability of the mean softmax output $\bar{\mathbf{p}}(\mathbf{x})$ averaged over T MC Dropout passes. Equation (1) is not a heuristic threshold: it is a bounded, differentiable scalar whose level sets partition the input space into three clinically meaningful regions, which we instantiate as the three-tier CDSS policy (Routine / Verification / Tumor Board). To the best of our knowledge, no prior brain-tumor classification work defines such a coupled, calibrated scalar and uses it as the gating variable of a clinical decision policy.

Contribution 2: A reference-free XAI consensus metric. Existing XAI evaluations in medical imaging either (i) report a single method in isolation, or (ii) compare methods using ground-truth lesion masks that are unavailable in a general clinical deployment. We define a *reference-free* consensus score $C_{\text{XAI}}(\mathbf{x}) \in [-1, 1]$ that quantifies agreement among five heterogeneous attribution maps without ground truth:

$$C_{\text{XAI}}(\mathbf{x}) = \frac{2}{N(N-1)} \sum_{i<j} \rho(E_i(\mathbf{x}), E_j(\mathbf{x})), \qquad (2)$$

where $E_1, \ldots, E_5$ are Grad-CAM, SHAP, LIME, Integrated Gradients, and Attention Rollout maps, $N{=}5$, and $\rho(\cdot, \cdot)$ is the Spearman rank correlation between two flattened saliency maps. To make the five

methods comparable, each map is bilinearly resized to the 224×224 input resolution, signed attributions (SHAP, LIME, and Integrated Gradients) are converted to their absolute values and summed over channels, and every map is min-max normalized before the correlation is computed. A high C_{XAI} indicates that five methodologically distinct explanation pathways concur on *where* the evidence lies; a low value flags a contentious prediction even when softmax confidence is high. Rank correlation between saliency maps has been used before in explanation sanity checks [31]; to the best of our knowledge, however, no prior brain-tumor classification study defines or reports a reference-free consensus score across five heterogeneous attribution methods.

Contribution 3: A joint formulation of classification, calibration, and explanation consistency. The three components above are not applied in isolation: the clinical report presents the triple $(\hat{y}, \tau(\mathbf{x}), C_{\text{XAI}}(\mathbf{x}))$ together. In the present implementation, the three-tier policy of Section 4.9 is gated on τ, with C_{XAI} reported alongside as an explanation-agreement flag; using C_{XAI} as an active gating variable is left to future work. This joint formulation has no direct analogue in the cited literature: prior hybrid and multi-scale MRI models [1–4] output a single class label or segmentation; prior uncertainty-aware models [7–10] report confidence and entropy as auxiliary diagnostics; and prior XAI studies [11–15] treat explanations as post-hoc artifacts decoupled from the decision rule.

The present work does not claim a new architecture family, a new learning paradigm, or a theoretical breakthrough in representation learning. It claims a *formal decision-theoretic object* (τ), a *formal explanation-agreement object* (C_{XAI}), and their *joint use* as the gating variables of an auditable clinical policy. These claims are testable. The controlled ablation in Section 7 shows that removing the MSAF module or the SE/CBAM recalibration degrades accuracy, F1, AUC, and MCC; the contribution of τ and C_{XAI} to the referral policy has not yet been ablated quantitatively and is left to the prospective study.

2 Literature Review

2.1 CNN-based Brain Tumor Classification

Convolutional Neural Networks (CNNs) have traditionally been the backbone of medical image analysis due to hierarchical feature extraction for detecting complex morphological patterns in brain MRI. VGG models [16] have shown that increasing network depth with small convolutional filters can improve the representational power of the CNN, at the cost of a large number of parameters and high computation. The introduction of ResNet [17] has alleviated problems associated with training deep CNNs by incorporating residual skip connections, while DenseNet [18] improved representation learning through densely connected layers that help mitigate the vanishing gradient problem and reduce the number of parameters. Most recently, EfficientNet [19] introduced a compound scaling approach that helps optimize the depth, width, and resolution of the network to achieve state-of-the-art accuracy with exceptional efficiency. Even though the CNN is well equipped to capture local spatial textures and morphological details required to delineate a brain tumor, it is inherently limited by its restricted, localized receptive field and therefore cannot model the global spatial dependencies required for classifying brain tumors as pathological structures.

2.2 Transformer-based Medical Imaging

In order to overcome the problem of the local receptive field associated with CNNs, the Vision Transformer (ViT) [21] was introduced, in which self-attention over image patches allows global modeling of spatial dependencies, a property that is attractive for medical imaging. However, the ViT lacks inductive biases and requires large amounts of data. In order to solve this problem, the Swin (shifted window) Transformer [22] introduces shifted window attention and allows efficient, hierarchical feature extraction from high-resolution images, while Data-efficient Image Transformers (DeiT) [23] use distillation and strong regularization to train transformers without large external datasets, which is relevant to data-limited medical imaging. The ConvNeXt model [24] showed that a modernized standard CNN architecture using Transformer design principles such as large convolutional kernels, layer normalization, and fewer activation functions can compete favorably with Transformers. Such Transformer-inspired convolutional designs have also been adapted to medical imaging, for example MedNeXt for medical image segmentation [20], which motivates retaining a modern convolutional baseline alongside attention-based models rather than treating the two architectural families as mutually exclusive. This suggests that combining the local texture extraction of modern CNNs with the global context modeling of Transformers may yield stronger feature representations for complex pathological

structures. The architectural context also includes surveys of transformers in medical imaging [25], hybrid convolutional-transformer classifiers [1], and Swin-based brain tumor classification [26], motivating evaluation of both standalone backbones and their fused representation under a common experimental protocol.

2.3 Explainable AI in Healthcare

The practical use of deep learning in the clinical environment is heavily impaired due to the "black-box" problem and requires efficient Explainable AI (XAI) approaches. Grad-CAM [11] uses the gradients flowing into the final convolutional layer to generate coarse class-discriminative explanations of the model's predictions, while fine-grained pixel attribution can be achieved with SHapley Additive exPlanations (SHAP) [12], based on the principles of cooperative game theory for calculating marginal contribution of each feature, or Local Interpretable Model-agnostic Explanations (LIME) [13], which builds an approximation of complex model locally through a simpler interpretable model. Integrated Gradients [14] provide path-based and axiomatic attribution by integrating gradients from a baseline to the input and guaranteeing completeness of the interpretation, while Attention Rollout [15] is used in the case of Transformer-based models to aggregate self-attention weights in order to show how information flows through the network. Although these methods enable us to obtain visual explanations and make them interpretable for humans, using only one of them leads to incomplete or even misleading explanations, therefore a multi-method approach to XAI becomes crucial in order to provide clinical-grade explanations. Reviews and trustworthiness assessments of medical-image XAI methods [27, 28] and work on transformer interpretability beyond raw attention maps [29] provide context for selecting explanations, while adversarial attacks on SHAP and LIME [30] and sanity checks for saliency methods [31] are also relevant: visually plausible attribution maps alone do not establish explanation fidelity.

2.4 Clinical Decision Support Systems

Calibrated uncertainty estimation for deep medical image segmentation, evaluated on tasks that include brain tumor MRI, provides relevant context for neuro-oncology decision support [32]. Reviews of uncertainty quantification in deep learning [33], the calibration of MC dropout and ensembles under dataset shift [8], and the distinction between epistemic and aleatoric uncertainty [9] inform the assessment of confidence in ambiguous cases. Trust-score and ensemble-based methods [10, 34] are relevant comparisons for the proposed uncertainty-aware reporting module; a high confidence score should not, by itself, be interpreted as evidence of clinical safety. The general design principles of CDSS [5] must also be considered alongside evidence on human-AI collaboration in image-based diagnosis [6], and the broader transition from explainable research prototypes to clinical practice [35, 36] motivates evaluating usability, external validity, and the limitations of model-generated recommendations. The implementation of advanced AI models in CDSS should help to enhance the performance of radiologists, reduce their fatigue, and provide more standardized treatment. Nonetheless, the application of such systems in the clinical environment faces numerous limitations, the major one being the inherent opacity of deep learning models, which contradicts the requirements of medical practice and the need for accountability and evidence-based reasoning. In addition, the overwhelming majority of models provide deterministic predictions without evaluating their predictive uncertainty, which leaves clinicians without any information about the model's confidence in ambiguous cases. The basis of trust in CDSS is transparency, thorough external validation, and the capability to describe the rationale of the generated diagnoses; without an uncertainty-aware explainable model that can be smoothly integrated into the clinical workflow and give clinicians interpretable insights, medical professionals will be unlikely to delegate responsibility for diagnoses to the AI, preventing the crucial transition from research prototypes to clinical practice.

2.5 Research Gap Summary

The relevant literature includes multi-level feature concatenation for brain tumor classification [3], a CNN-Transformer network that takes multimodal MRI as input for brain tumor segmentation [2], and multiscale CNNs for joint brain tumor classification and segmentation [4]. Accordingly, the contribution of this study is framed as the joint evaluation of local-global feature fusion, multiple explanation methods, and predictive uncertainty within one four-class classification workflow, rather than a claim that these components have not previously been investigated. Evidence that deep models can generalize poorly to data from

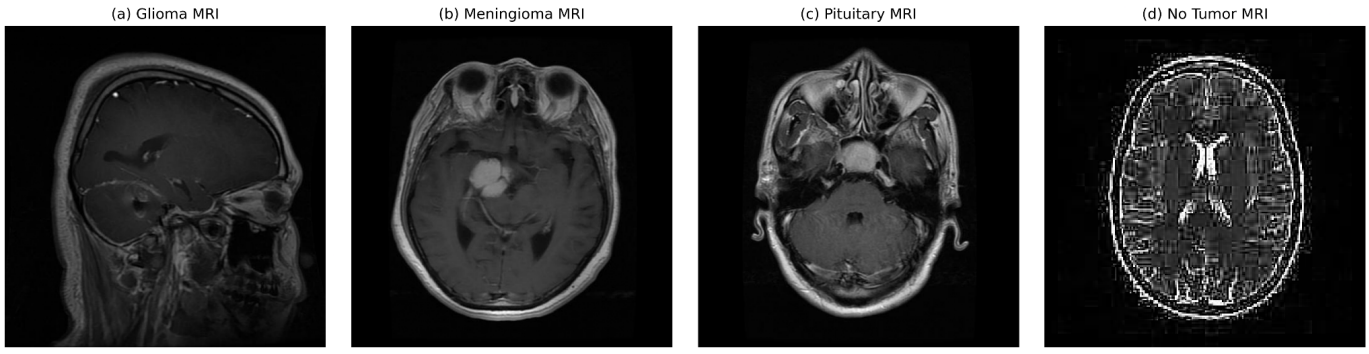


Figure 1. Representative axial T1-weighted contrast-enhanced (T1CE) MRI samples for each of the four clinical classes: (a) Glioma, (b) Meningioma, (c) Pituitary tumor, and (d) No Tumor (Normal).

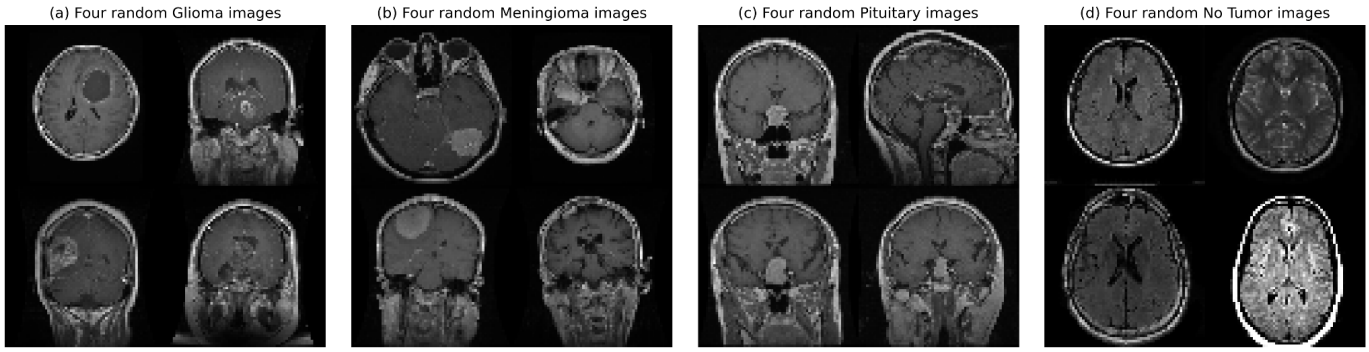


Figure 2. Intra-class variation demonstrating morphological diversity within each category. Each panel displays a 2×2 grid of four randomly selected MRI scans for (a) Glioma, (b) Meningioma, (c) Pituitary tumor, and (d) No Tumor.

other institutions [38] is relevant when assessing generalizability beyond the benchmark used here. Beyond these benchmark-driven gaps, we found no human-factors or reader-preference evaluation of an uncertainty-aware, multi-XAI CDSS for four-class brain-tumor classification; Section 8 takes a preliminary step toward filling this specific gap.

3 Materials

3.1 Dataset Information

In the current study, a four-class benchmark brain MRI dataset including T1 contrast-enhanced (T1CE) images, which are highly sensitive to blood-brain barrier damage and important for delineating the tumor, is used. The dataset is drawn from the publicly available four-class brain tumor MRI collection released by Nickparvar [39], which combines images from the figshare, SARTAJ, and Br35H collections and is widely used as a benchmark for glioma, meningioma, pituitary tumor, and no-tumor classification. Four unique clinical classes were used in this dataset, namely, Glioma, Meningioma, Pituitary Tumor, and No Tumor (Normal). In order to maintain a robust evaluation of the proposed models, the dataset was stratified and divided into two groups, that is, training and testing datasets. The statistics for the two datasets are shown below.

Figure 1 introduces the four target categories using representative axial T1CE slices: glioma, meningioma, pituitary tumor, and no tumor. These examples provide visual context for the classification task rather than defining diagnostic criteria for each class, and they motivate the four-class decision space used throughout the remainder of the paper.

Figure 2 presents four examples per category to illustrate within-class variation in appearance and morphology. This variation motivates evaluating the classifier across diverse samples rather than relying on a single representative image, and it justifies the use of data augmentation during training.

Table 1 lists the class-wise training and testing counts, totaling 2,870 training images and 719 test images, together with each class's share of the dataset. The no-tumor category contains only 275 of the 3,589 images (7.7%), making class imbalance an important consideration for training and evaluation.

Figure 3 complements Table 1 by expressing the class counts as proportions of the full dataset. The Gini impurity annotation summarizes label diversity,

Table 1. Dataset statistics and class distribution.

Class	Training	Testing	Total	Percentage
Glioma	901	226	1127	31.4%
Meningioma	905	227	1132	31.5%
Pituitary	844	211	1055	29.4%
No Tumor	220	55	275	7.7%
Total	2870	719	3589	100%

while the individual sectors make the minority no-tumor category visible and reinforce the need for macro-averaged evaluation metrics.

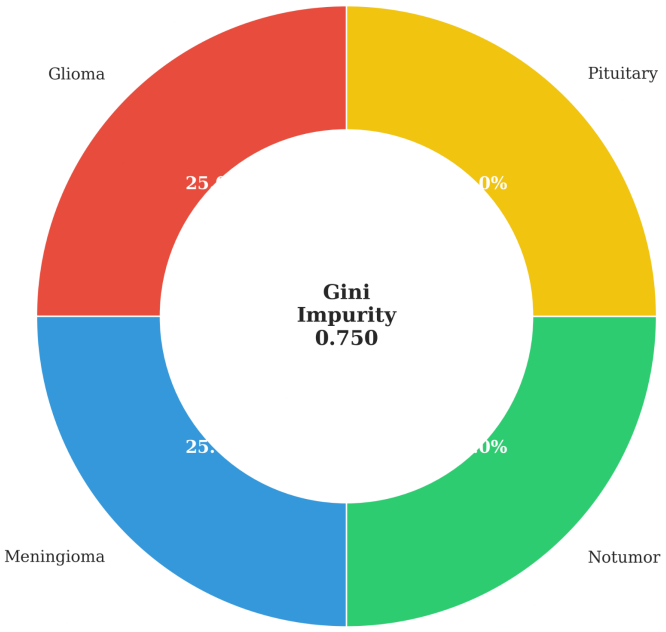


Figure 3. Donut chart visualizing the overall proportional representation of each class in the sampled dataset, annotated with the calculated Gini impurity score to quantify dataset balance.

3.2 Image Preprocessing

Raw clinical MRI scans have an intrinsic variability regarding spatial resolution, acquisition noise, and intensity inhomogeneities. An end-to-end image preprocessing pipeline was developed to create standardized inputs for deep learning models. Firstly, raw MRI images were resized to a uniform resolution of 224×224 pixels, the default input size of the ImageNet-pretrained EfficientNet, ConvNeXt, and Swin Transformer models. Next, all images were subjected to an automated skull stripping procedure, removing any non-brain structures (such as skull and scalp) and isolating the brain tissue.

Figure 4 compares mean image width and height for each class, with standard deviation bars indicating variation in the original acquisition dimensions. The plot provides context for resizing all network inputs to 224 × 224 pixels and shows that the resizing step is necessary to harmonize heterogeneous acquisitions.

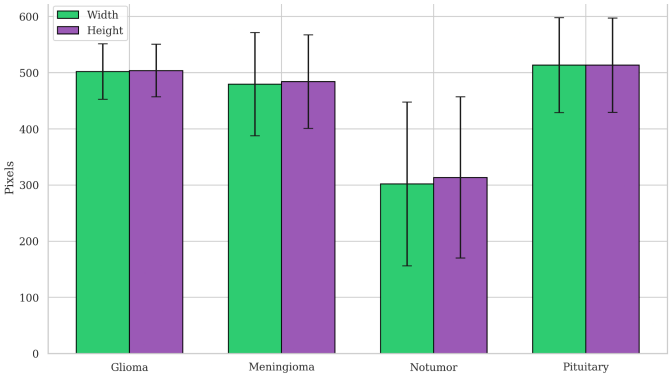


Figure 4. Average image resolution (Width vs. Height in pixels) with standard deviation error bars for each class, demonstrating the variability in raw acquisition dimensions prior to uniform resizing.

To highlight fine morphology and minimize the effects of local illumination inhomogeneity, Contrast Limited Adaptive Histogram Equalization (CLAHE) was performed. CLAHE increases the local contrast of the image by dividing it into small tiles and equalizing the histogram of each tile, with a clip limit that restricts noise amplification in homogeneous regions; this results in better boundary separation of the tumor region. After that, noise was removed using a sequence of median and Gaussian filters. The median filter was used to remove salt-and-pepper noise, whereas the Gaussian filter reduced the high-frequency noise while retaining valuable edge information. Finally, pixel intensities were normalized to the $[0, 1]$ range via min-max scaling to regularize the gradient descent and improve convergence rate of the network.

Figure 5 summarizes the file-size distribution within each class, including its spread and tails. File size is an exploratory acquisition characteristic and should not be interpreted as a direct measure of image quality or tumor severity, but it does reveal the heterogeneity of the source data.

To prevent overfitting and promote generalization, a comprehensive data augmentation method was developed and employed during training. The spatial augmentations include random rotation ($\pm 15°$), horizontal and vertical flips, random zoom ($\pm 10\%$) and elastic transformations mimicking natural anatomical transformations. Random adjustments of contrast and brightness were applied to introduce the variations of image intensity. In addition, the mixing-based regularizers MixUp [40] and CutMix

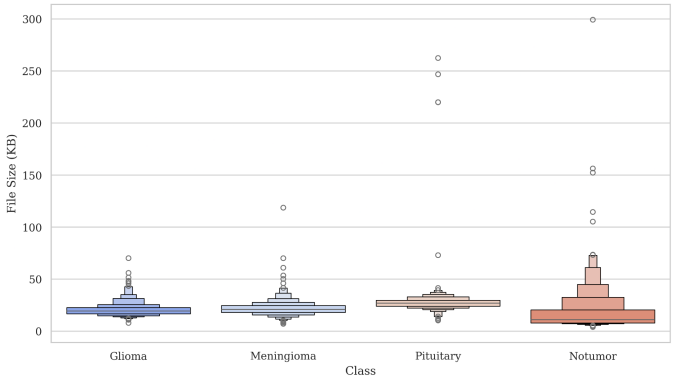


Figure 5. Boxen plot depicting the distribution of file sizes (in KB) per class, an exploratory indicator of image compression and acquisition heterogeneity across the dataset.

[41] were used. MixUp consists of linear interpolation of two images and their labels, which promotes robust linear behavior outside the training manifold. CutMix replaces the random patches of the image with patches from another image, encouraging localized and discriminative features learning.

Figure 6 follows one glioma sample through the original, skull-stripped, CLAHE-enhanced, and min-max-normalized stages. The sequence links the preprocessing operations to the image representation supplied to the network and demonstrates that each step contributes a visually distinct transformation.

Figure 7 places global histogram equalization, CLAHE, and gamma correction alongside the original MRI. This qualitative comparison illustrates how contrast transformations change image appearance; it is not an ablation of classification performance but rather a justification for the CLAHE choice in the pipeline.

Figure 8 compares Gaussian, median, and bilateral filtering with the original MRI. The panels support visual inspection of the balance between smoothing and boundary preservation, without establishing the diagnostic superiority of any filter, and they motivate the sequential median-then-Gaussian choice adopted here.

Figure 9 plots grayscale intensity histograms for one representative image from each class. The distributions describe the selected samples and provide context for intensity normalization; they should not be taken as population-level class signatures but they do show that raw intensity statistics alone are insufficient for class separation.

Figure 10 combines boxplots of pixel intensity in the top panel with kernel density estimates in the bottom panel for the four classes. The boxplots summarize medians, spread, and outliers, whereas the density curves make distributional overlap and differences in shape easier to inspect, confirming the need for learned features rather than handcrafted intensity statistics.

Figure 11 summarizes mean pixel intensity and average within-image standard deviation by class. These brightness and contrast summaries complement the fuller intensity distributions in Figure 10 and provide a compact per-class descriptor that helps interpret the preprocessing impact.

4 Proposed Methodology

This section describes the methodology used in this work. The main contribution of this study is the design, mathematical formulation and optimization of a new Hybrid Convolutional Neural Network-Transformer architecture supplemented with Multi-Scale Attention Fusion. This architecture has been specifically developed to surpass the limitations of either of the paradigm-specific approaches by combining the advantages of inductive biases and texture extraction ability of the modern CNN architecture with the benefits of receptive fields and long-range dependencies modeling of hierarchical Vision Transformers. Furthermore, this section describes the training strategy, attention mechanisms, and hyperparameter optimization process used for brain tumor MRI classification.

The proposed eight-stage architecture in Figure 12 links MRI preprocessing and augmentation with dual-stream ConvNeXt and Swin feature extraction, adaptive feature fusion, classification, uncertainty estimation, explainability, and trust-gated clinical decision support. The complete data flow is annotated with tensor dimensions at each stage (input $224 \times 224 \times 1 \rightarrow$ dual-stream $28 \times 28 \times C' \rightarrow$ fused $28 \times 28 \times C' \rightarrow$ refined $28 \times 28 \times C' \rightarrow$ pooled $C' \rightarrow$ class probabilities $K = 4$), the learnable α/β attention maps inside the MSAF module (with $\alpha(x, y) + \beta(x, y) = 1$ at every spatial location), and the branching into the uncertainty (Stage 7A: MC Dropout, Trust Score τ) and XAI (Stage 7B: Grad-CAM, SHAP, LIME, Integrated Gradients, Attention Rollout) modules that feed the trust-gated CDSS report.

4.1 Overall Framework Architecture

The framework in question relies on a rigorously designed, end-to-end pipeline aimed at transforming heterogeneous neuroimaging data into highly accurate and clinically actionable diagnostic results. The overall

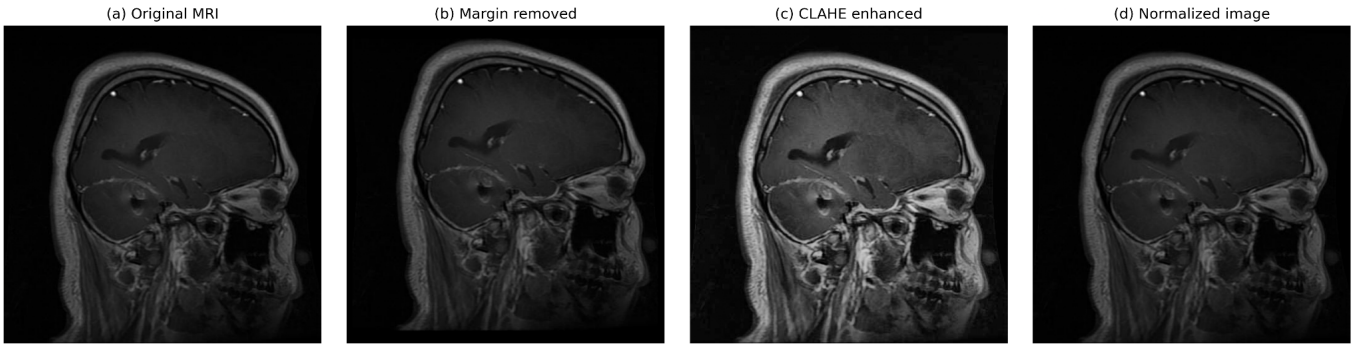


Figure 6. Step-by-step visualization of the image preprocessing pipeline applied to a representative Glioma MRI: (a) Original raw scan, (b) Skull-stripped and margin-removed image, (c) Contrast enhancement using Contrast Limited Adaptive Histogram Equalization (CLAHE), and (d) Min-max normalized image ready for network ingestion.

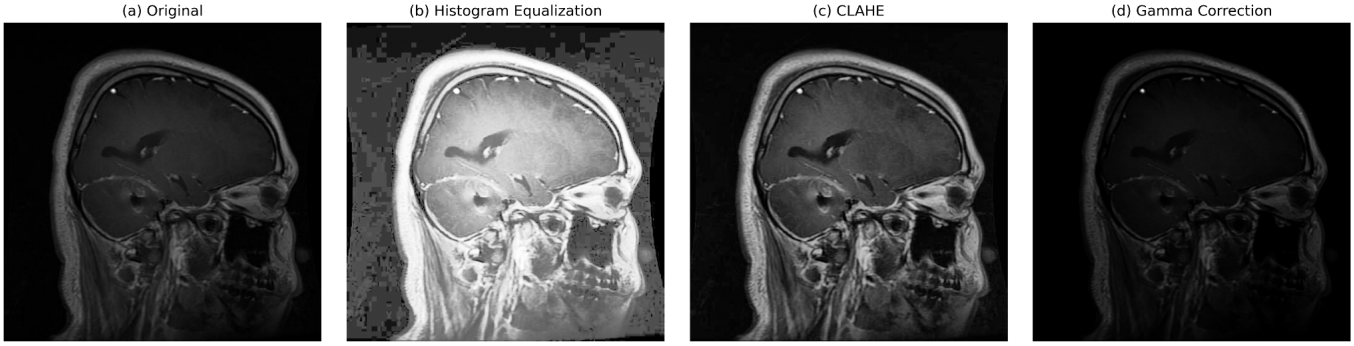


Figure 7. Comparative analysis of contrast enhancement techniques applied to a raw MRI: (a) Original, (b) Global Histogram Equalization, (c) Contrast Limited Adaptive Histogram Equalization (CLAHE), and (d) Gamma Correction (γ=1.5).

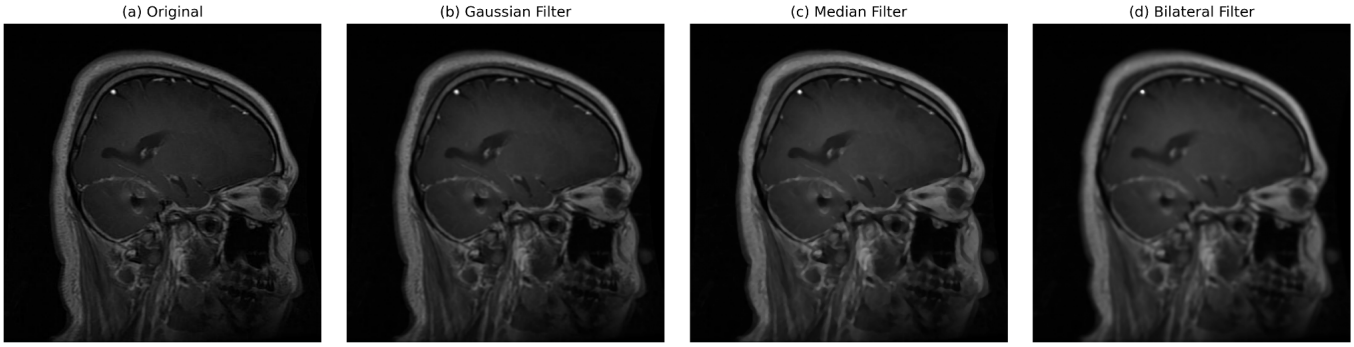


Figure 8. Evaluation of spatial domain noise reduction filters on a representative MRI: (a) Original, (b) Gaussian blur, (c) Median filter, and (d) Bilateral filter (preserving edges while smoothing high-frequency noise).

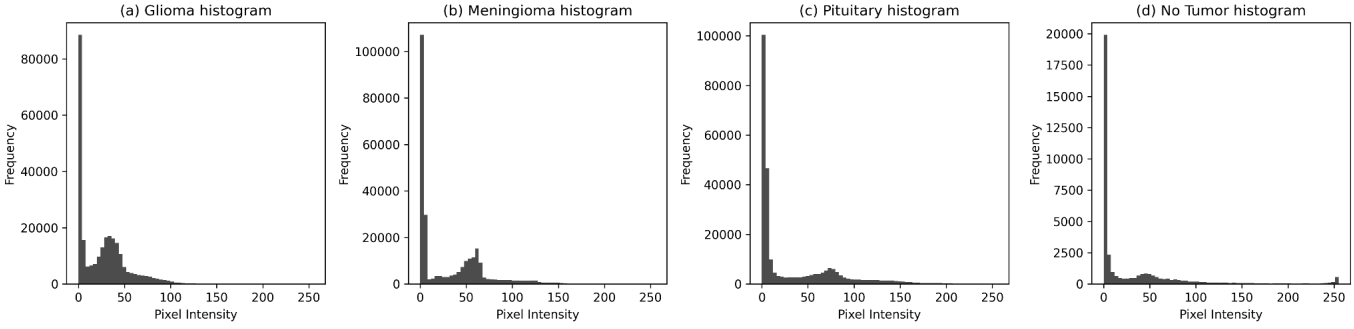


Figure 9. Grayscale intensity histograms for a representative sample of each class, illustrating the distinct pixel intensity distributions of (a) Glioma, (b) Meningioma, (c) Pituitary tumor, and (d) No Tumor.

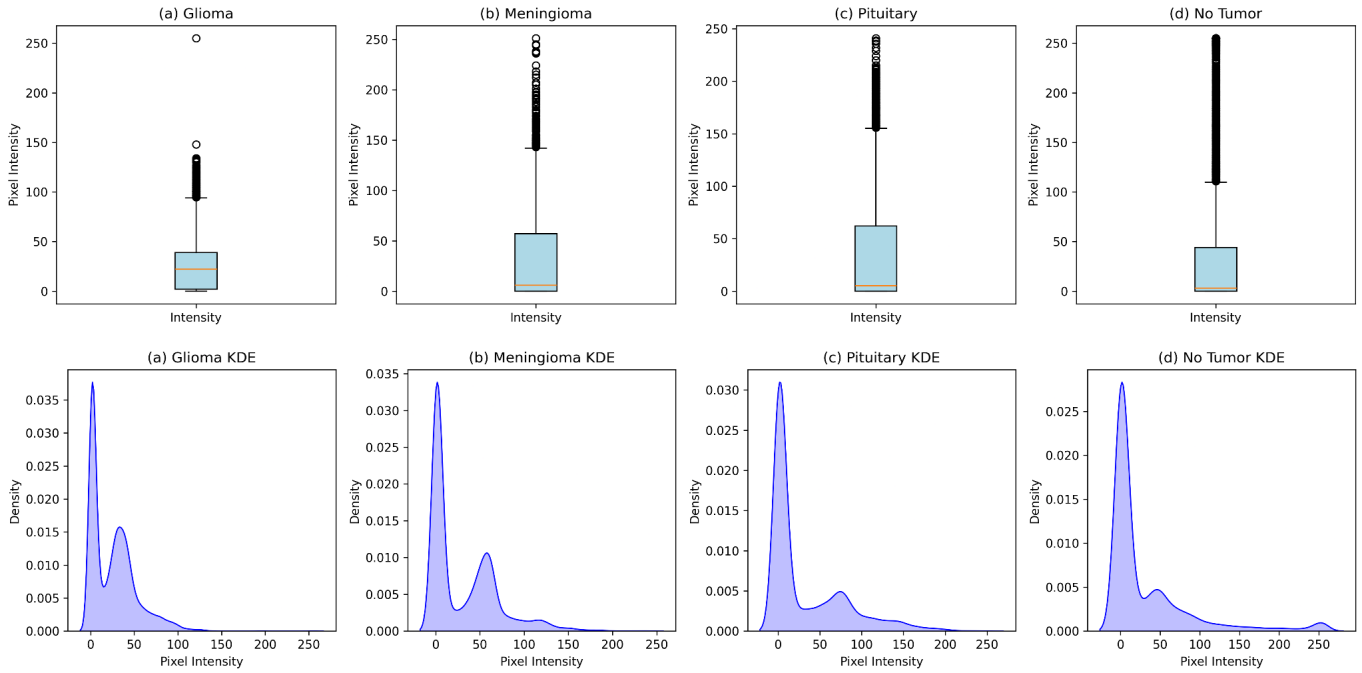


Figure 10. Statistical distribution of pixel intensities across the four classes. Top: Boxplot representation highlighting the median, interquartile range, and outliers. Bottom: Kernel Density Estimation (KDE) curves showing the smoothed probability density function of grayscale pixel values.

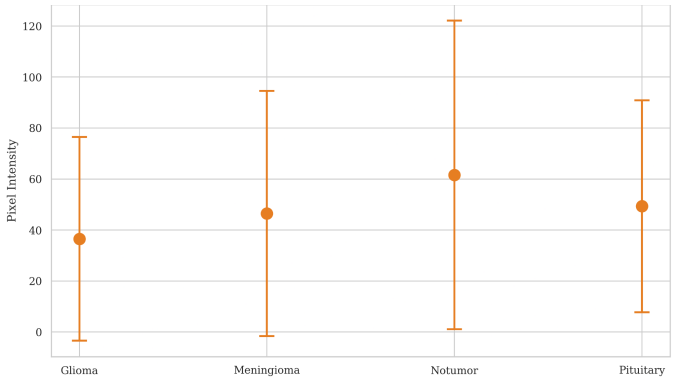


Figure 11. Point-range plot displaying the mean pixel intensity and average standard deviation per class, indicating distinct baseline brightness and contrast characteristics across different tumor pathologies.

workflow of the framework is envisaged as a series of specialized modules dealing with different challenges associated with medical image analysis. The pipeline is described below.

Input Ingestion and Image Enhancement: Raw MRI slices are loaded and preprocessed via skull stripping, CLAHE (Contrast Limited Adaptive Histogram Equalization) and noise removal to equalize the intensity distribution and highlight morphological boundaries.

Data Augmentation: To address the natural vulnerability of medical domains to overfitting, spatial and intensity-based augmentations, including MixUp and CutMix, are dynamically used during the training process.

Figure 13 applies Canny edge detection to representative images from the four classes to visualize structural boundaries. These edge maps are exploratory illustrations of local morphology, not learned attention maps or tumor segmentations, and they motivate the need for learned feature extraction in the hybrid model.

Dual-Stream Feature Extraction: The enhanced images are processed in parallel by two separate backbones - a modern CNN (ConvNeXt) and hierarchical Transformer (Swin Transformer) - to extract local features and global representations simultaneously.

Multi-Scale Attention and Feature Fusion: The disparate feature maps are combined and spatially aligned by a specially-designed attention mechanism. The latter dynamically determines the relative contributions of the local and global features depending on the input pathology.

Figure 14 summarizes the edge-pixel ratio using class-wise boxplots overlaid with individual observations. The plot shows both within-class dispersion and between-class overlap in this exploratory measure of boundary density, indicating that simple edge statistics cannot fully separate the

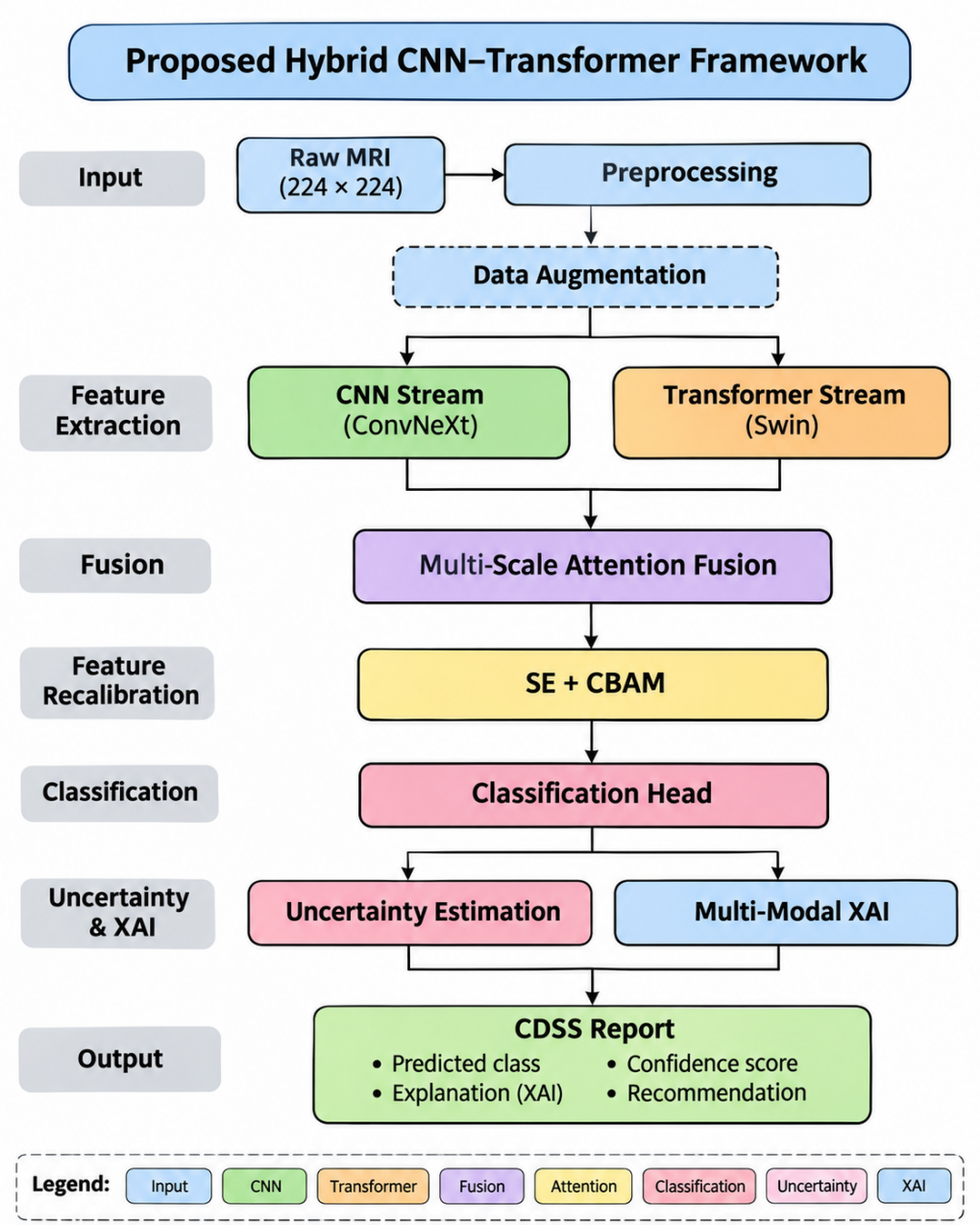


Figure 12. End-to-end architecture of the proposed framework showing eight stages: (1) input MRI $224 \times 224 \times 1$; (2) preprocessing and augmentation; (3) dual-stream extraction (ConvNeXt $F_{CNN} \in \mathbb{R}^{28\times28\times C_1}$, Swin $F_{Swin} \in \mathbb{R}^{28\times28\times C_2}$); (4) dimensionality alignment and MSAF with explicit learnable maps $\alpha(x,y), \beta(x,y)$ where $\alpha + \beta = 1$, producing $F_{fused} \in \mathbb{R}^{28\times28\times C'}$; (5) sequential SE and CBAM recalibration yielding $F_{refined} \in \mathbb{R}^{28\times28\times C'}$; (6) global average pooling and classification head producing $P(y_k|X)$ for $K = 4$; (7A) uncertainty branch via MC Dropout (T stochastic passes) yielding confidence and entropy, from which Trust Score τ is computed; (7B) XAI branch computing Grad-CAM, SHAP, LIME, Integrated Gradients, and Attention Rollout; (8) trust-gated CDSS report with rule-based referral tier (Routine Review / Radiologist Verification / Tumor Board).

four classes.

Figure 15 compares local binary patterns, Gabor responses, Sobel gradient magnitude, and Laplacian responses for a representative MRI. These handcrafted maps illustrate local texture and edges; they are not substitutes for the learned CNN feature maps but they justify why the ConvNeXt stream is needed to capture high-frequency textural detail.

Feature Recalibration and Classification: The combined representation is recalibrated using a Squeeze-and-Excitation (SE) block and the Convolutional Block Attention Module (CBAM) to incorporate interdependencies across channels and spatial dimensions. Then, the features are pooled and

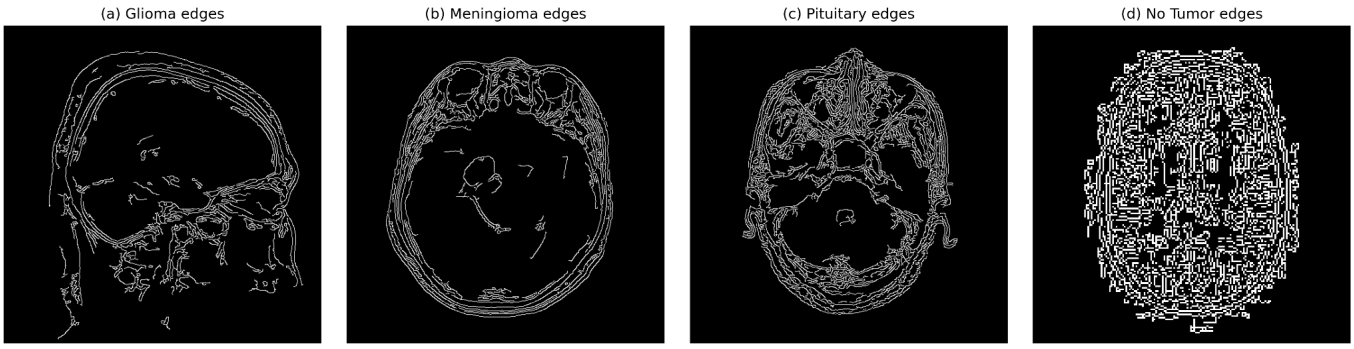


Figure 13. Canny edge detection applied to representative samples, highlighting the structural boundaries and morphological complexity of (a) Glioma, (b) Meningioma, (c) Pituitary tumor, and (d) No Tumor.

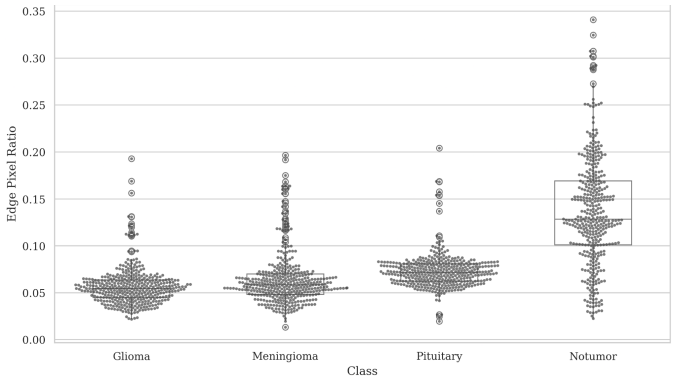


Figure 14. Combined box and swarm plot illustrating the morphological complexity (edge pixel ratio) across classes, revealing differences in structural boundary sharpness and texture density.

passed to a dense classification head.

Training and Optimization Process: The network was trained using an optimization process comprising Focal Loss, the AdamW optimizer, and cosine annealing learning rate scheduling, with hyperparameters tuned by Bayesian optimization.

Figure 16 positions samples in the GLCM contrast-homogeneity plane and overlays kernel density contours. The plot supports inspection of class overlap in a two-feature texture representation, rather than demonstrating separability in the learned hybrid embedding, and it reinforces the need for the deep fusion architecture.

Prediction and Explainability Module: The softmax probability scores are calculated, and the decision making process is analyzed using a multi-method Explainable AI (XAI) pipeline that allows the generation of class-discriminative heatmaps.

Clinical Decision Support: The predicted probability distribution, confidence scores, and heatmaps are aggregated into a comprehensive and uncertainty-aware clinical recommendation.

Figure 17 displays log-scaled two-dimensional FFT magnitude spectra for representative images from all four classes. These spectra provide a complementary frequency-domain view of image structure; they do not directly visualize transformer attention but they show that global frequency patterns differ across classes, which is the motivation for the Swin stream.

4.2 Baseline Architectures for Comparative Evaluation

To assess the effectiveness of the proposed hybrid architecture, its performance is compared with that of strong standalone architectures. Three widely used CNN and Transformer models serve as baselines; they were implemented, optimized, and trained under the same experimental conditions to keep the comparison fair.

Figure 18 summarizes normalized FFT energy across four frequency bands for each class. The radar profiles complement the spectra in Figure 17 by enabling a band-wise comparison, without implying that the classifier explicitly uses these handcrafted summaries.

4.2.1 Model 1: EfficientNetV2

EfficientNetV2 [42] is chosen as the main lightweight convolutional baseline. It was designed using training-aware neural architecture search (NAS) and model scaling. Unlike its predecessor, EfficientNetV2 offers a progressive learning strategy according to which the network uses progressively stronger regularization techniques (such as stochastic depth and data augmentation) with increasing image resolution during training, thus drastically increasing the speed and efficiency of the training process.

Figure 19 decomposes a representative MRI into one approximation and three detail sub-bands using a

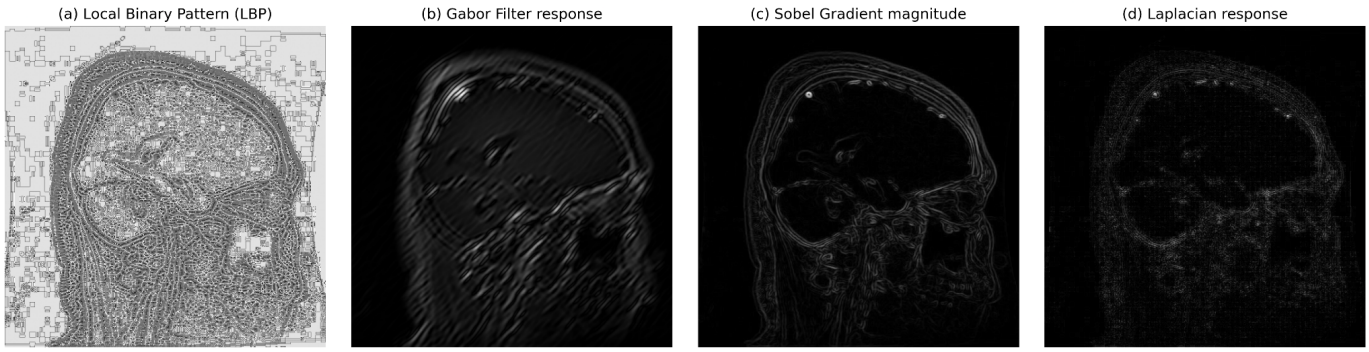


Figure 15. Texture feature maps extracted from a representative MRI to highlight local patterns: (a) Local Binary Pattern (LBP), (b) Gabor filter response, (c) Sobel gradient magnitude, and (d) Laplacian response.



Figure 16. Scatter plot with Kernel Density Estimation (KDE) contours mapping the Gray-Level Co-occurrence Matrix (GLCM) Contrast versus Homogeneity feature space, demonstrating the separability of texture patterns among the four classes.

single-level Haar wavelet transform. The panels illustrate coarse and directional image information as a descriptive multi-scale analysis, not as an additional model input, and they motivate the multi-scale design philosophy of the hybrid model.

Architecture-wise, EfficientNetV2 introduces the concept of Fused-MBConv (Mobile Inverted Bottleneck) layer in the earlier layers of the network. Regular MBConv layers use depthwise convolutions that, despite having few parameters and FLOPs, often cannot fully utilize modern accelerators, which slows down training in the early layers [42]. The Fused-MBConv layer replaces the depthwise 3×3 convolution and the expansion 1×1 convolution with a single, standard 3×3 convolution, which runs faster on accelerators while retaining representational power. The formulation of the Fused-MBConv block is given below.

Equation (3) summarizes the block, mapping the input tensor X to the output Y through an expansion convolution, a projection convolution, and a residual connection.

$$Y = X + \mathrm{Conv}_{1\times1}\big(\mathrm{Conv}_{3\times3}(X)\big) \tag{3}$$

where X is the input tensor, $\mathrm{Conv}_{3\times3}$ is the fused expansion convolution (followed by batch normalization and SiLU activation), and $\mathrm{Conv}_{1\times1}$ is the projection convolution; the residual connection is used when the stride is 1 and the input and output channels match. In EfficientNetV2, the Squeeze-and-Excitation (SE) module is used in the later MBConv stages rather than in the Fused-MBConv stages. In terms of classifying brain tumors, EfficientNetV2 can be used as an efficient baseline model due to the following reason. EfficientNetV2 is a strong model in capturing the local morphological textures and sharp edge boundaries which help in differentiating the small variations that exist in the case of glioma and meningioma classification. However, the localized nature of the receptive field of this model is unable to represent global anatomical distortions caused by tumors.

Figure 20 summarizes pairwise Pearson correlations among the extracted image descriptors. Positive and negative associations help identify potentially redundant or complementary descriptors, but do not measure causal relationships or learned feature importance, and they motivate the use of a learned deep feature extractor rather than handcrafted descriptors.

4.2.2 Model 2: ConvNeXt

The ConvNeXt [24] architecture was chosen as an example of the modernization of the classic ResNet architecture. The authors of ConvNeXt showed that

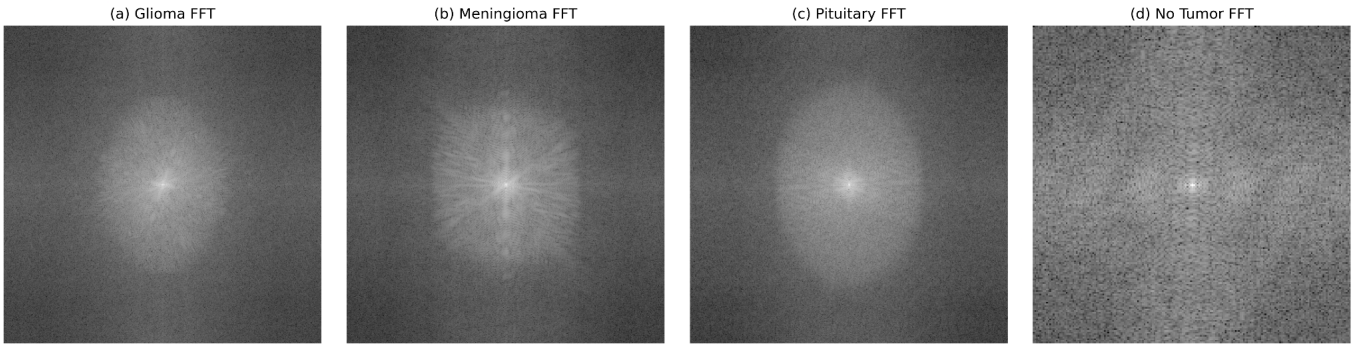


Figure 17. 2D Fast Fourier Transform (FFT) magnitude spectra (log-scaled) revealing the frequency domain characteristics and global texture patterns of (a) Glioma, (b) Meningioma, (c) Pituitary tumor, and (d) No Tumor.

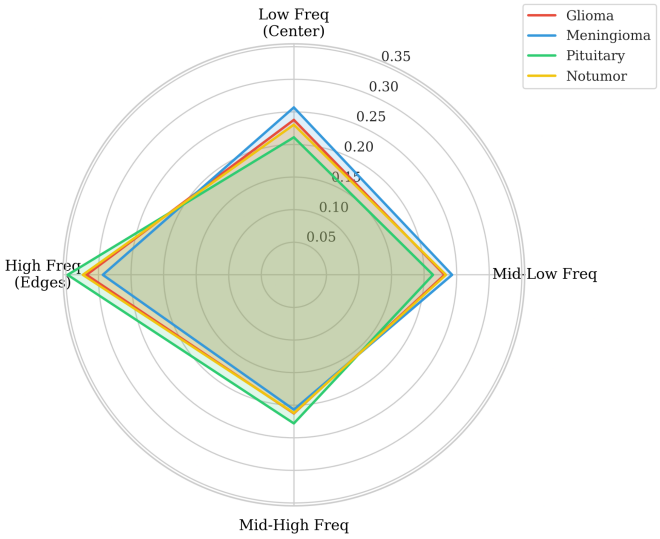


Figure 18. Radar chart comparing normalized image energy across four FFT frequency bands (low to high) for each class. These exploratory spectral descriptors are distinct from the learned hybrid features.

by consistently adopting the design features of Vision Transformers, a pure convolutional network can match or exceed the results of contemporary hierarchical transformers. This model plays a key role in our comparative analysis because it represents a strong modernized pure convolutional network.

The ConvNeXt architecture is characterized by several macro-design and micro-design changes. Macroscopically, it uses a stage compute ratio of 1:1:3:1, a "patchify" stem with a 4×4 kernel and stride of 4 (instead of the traditional 7×7 stem), inverted bottleneck blocks (the hidden dimension of the block is four times wider than the input dimension), and large 7×7 kernels of convolutions in the depthwise convolution layer in order to increase the receptive field. Microscopically, it uses the Gaussian Error Linear Unit (GELU) instead of ReLU as an activation function, decreases the number of normalization and activation layers (only applies LayerNorm in bottleneck and uses GELU only once per block), and uses Layer Normalization (LayerNorm) instead of Batch Normalization (BatchNorm).

Equation (4) expresses the ConvNeXt block as a residual mapping, adding the transformed feature representation to the input X:

$$Y = X + \mathrm{Conv}_{1\times1}\Big(\mathrm{GELU}\big(\mathrm{Conv}_{1\times1}\big(\mathrm{LayerNorm}(\mathrm{DWConv}_{7\times7}(X))\big)\big)\Big) \quad (4)$$

where $\mathrm{DWConv}_{7\times7}$ is the depthwise convolution, the first $\mathrm{Conv}_{1\times1}$ expands the channels by a factor of 4, and the second projects them back.

In our proposed framework, the ConvNeXt model is specifically chosen as the local feature extractor. Its large 7×7 depthwise kernels and improved architecture ensure that it can represent multi-scale texture information in the MR images in an efficient manner without the quadratic time complexity of the conventional self-attention process.

Figure 21 displays the same grayscale MRI using alternative pseudocolor mappings. The panels illustrate display choices for visual inspection; recoloring does not add anatomical information to the original intensity values but it helps radiologists visually inspect boundaries during annotation.

4.2.3 Model 3: Swin Transformer

The Swin Transformer [22] is taken as the baseline model for global context modeling and hierarchical transformers. A regular Vision Transformer (ViT) calculates self-attention for the whole input image, leading to the computation complexity of $O(N^2)$, where N is the number of patches. Due to

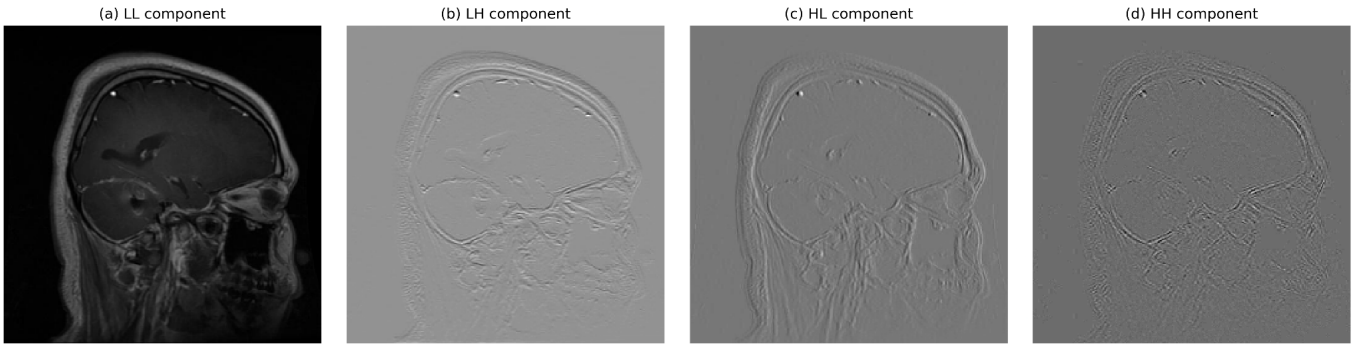


Figure 19. Single-level Discrete Wavelet Transform (DWT) using the Haar wavelet on a representative MRI, decomposing the image into (a) LL (approximation), (b) LH (horizontal details), (c) HL (vertical details), and (d) HH (diagonal details) sub-bands.

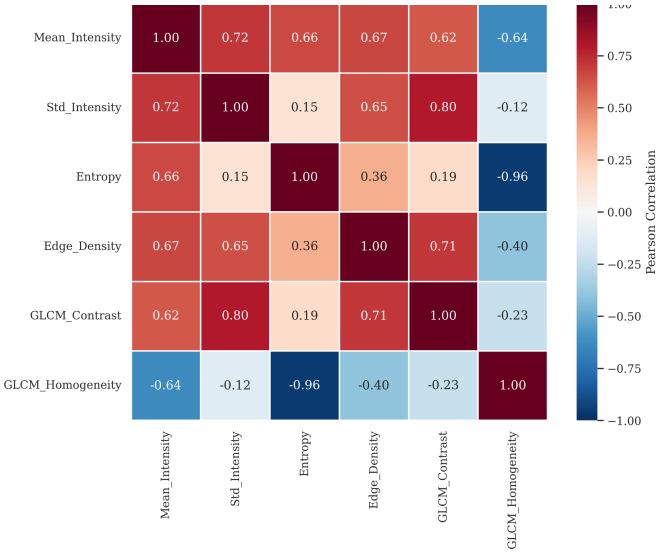


Figure 20. Pearson correlation heatmap of extracted image features (Mean Intensity, Std Intensity, Entropy, Edge Density, GLCM Contrast, and GLCM Homogeneity), identifying multicollinearity and independent feature contributions.

this quadratic complexity, regular ViTs become computationally expensive for high-resolution images, such as MRI images.

To address this problem, the Swin Transformer uses hierarchical modeling and shifted window attention. It calculates self-attention only within non-overlapping local windows of $M \times M$ patches (Window-based Multi-head Self-Attention, W-MSA), thus achieving linear computational complexity $O(N)$. In order to establish connections between patches, shifted window partitioning is used (Shifted Window-based Multi-head Self-Attention, SW-MSA).

Equations (5) and (6) give the computational costs of W-MSA and SW-MSA, respectively. Both expressions show linear dependence on the number of spatial positions hw when the window size M and channel count C are fixed:

$$\text{Complexity(W-MSA)} = 4hwC^2 + 2M^2hwC \quad (5)$$

$$\text{Complexity(SW-MSA)} = 4hwC^2 + 2M^2hwC \quad (6)$$

where h and w represent the dimensions of the feature map, C stands for the number of channels, and M represents the size of the window. Since M is a constant (usually 7), the complexity of this model depends linearly on hw, which makes it very effective for high-resolution inputs. In our comparison, Swin Transformer serves as the benchmark for modeling the holistic deformation of anatomy caused by brain tumors.

Figure 22 represents pixel intensity as height using surface views in the top panel and wireframe views in the bottom panel. Each row includes the four MRI classes. These displays illustrate intensity variation across the image plane and should not be interpreted as three-dimensional anatomical reconstructions, but they do show the spatial intensity heterogeneity that motivates global modeling.

4.2.4 *Model 4: Proposed Hybrid CNN-Transformer with Multi-Scale Attention Fusion*

The main innovation and main contributions of this work are in combining ConvNeXt and Swin Transformer through the Multi-Scale Attention Fusion Module which is specifically designed for this purpose. The hybrid model presented here, called Model 4, is deliberately created in order to address the inherent weaknesses of these isolated models such as the limited receptive field of CNN and the absence of local inductive bias and high computational demands of Transformer architectures.

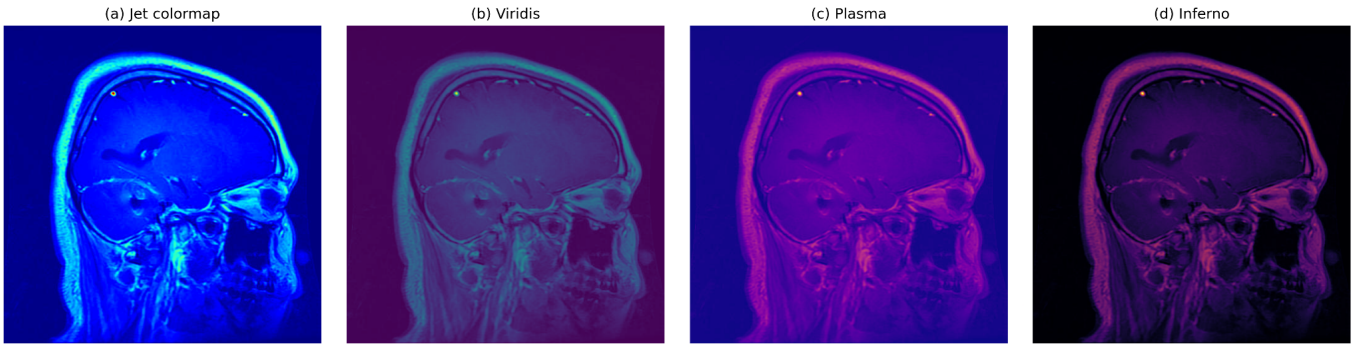


Figure 21. Grayscale MRI mapped to various pseudocolor colormaps to enhance visual contrast for human inspection: (a) Jet, (b) Viridis, (c) Plasma, and (d) Inferno.

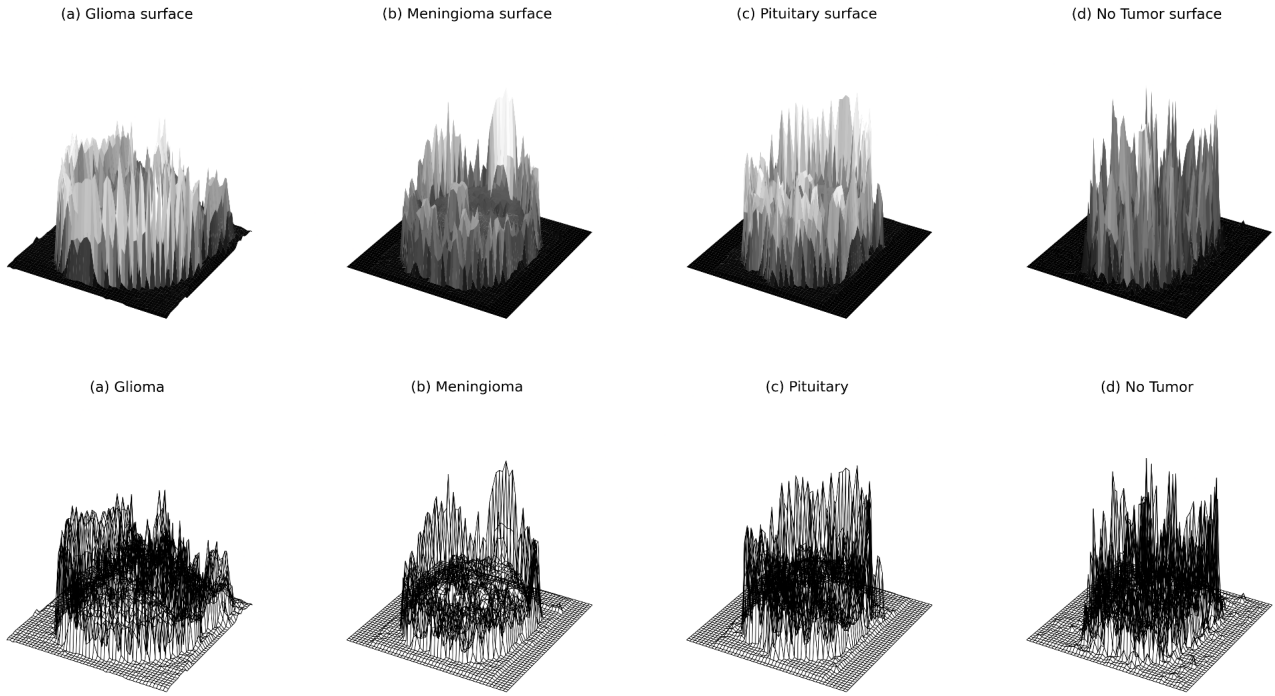


Figure 22. Three-dimensional topographical representations of MRI intensity. Top: 3D surface plots representing pixel intensity as elevation. Bottom: 3D wireframe mesh plots illustrating spatial intensity gradients across the four classes.

4.2.5 *Dual-Stream Feature Extraction and Dimensionality Alignment*

The input MRI tensor $X \in \mathbb{R}^{H\times W\times C_{in}}$, where H and W denote the spatial dimensions, and C_{in} represents the number of input channels (usually 1 for grayscale MRI), is processed using two parallel, truncated feature extraction streams.

The first stream exploits the first three stages of the ConvNeXt model. This stream is designed to learn the local feature map, denoted as $F_{CNN} \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times C_1}$. ConvNeXt stream is good at extracting high-frequency spatial information, edge gradients, local texture abnormalities, and the complex internal structures that are inherent in the tumor boundaries and the necrotic core regions.

The second stream processes the same input X using the first three stages of the Swin Transformer in order to learn the global feature map, $F_{Swin} \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times C_2}$. Using the shifted window attention, this stream extracts the low-frequency global information, which models the relationship between the tumor mass, midline shift, and the healthy brain parenchyma regions surrounding the tumor.

An important problem in the combination of CNN and Transformer features is the semantic and dimensional mismatch. As the channel dimensions C_1 and C_2 differ because of the architectural design of ConvNeXt and Swin, respectively, a dimensionality alignment layer is used in order to project both feature maps to a shared latent space of dimension C':

Equations (7) and (8) specify the channel projection and normalization of the CNN and Swin feature maps, respectively, so that the two streams have compatible dimensions before fusion.

$$\hat{F}_{CNN} = \text{LayerNorm}(\text{Conv}_{1\times1}(F_{CNN})) \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times C'} \tag{7}$$

$$\hat{F}_{Swin} = \text{LayerNorm}(\text{Conv}_{1\times1}(F_{Swin})) \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times C'} \tag{8}$$

4.2.6 *Multi-Scale Attention and Feature Fusion Module*

Concatenation or addition of $\hat{F}_{CNN}$ and $\hat{F}_{Swin}$ in a naive way would produce a sub-optimal feature representation since it implies the same importance to be attributed to the local and global information independently of the spatial position. However, the diagnostic significance of either local or global features changes dynamically based on the type of pathology in question in the input MRI image. For example, differentiation between pituitary adenoma and meningioma depends on local textural boundaries, while detection of large glioma relies on global anatomical distortions.

The learnable Multi-Scale Attention Fusion module combines the aligned local and global representations using the element-wise weighted sum in Eq. (9):

$$F_{fused} = \alpha \odot \hat{F}_{CNN} + \beta \odot \hat{F}_{Swin} \tag{9}$$

where $\odot$ is the element-wise (Hadamard) multiplication. Parameters α and β are not fixed scalar values; rather, they are learnable attention maps generated dynamically by the network and are obtained by concatenating the feature maps, passing them through a lightweight convolutional bottleneck, and applying the softmax operation:

Equation (10) defines the attention-weight generator used in Eq. (9); the softmax normalizes the two stream weights at each spatial location.

$$[\alpha, \beta] = \text{Softmax}\Big(\text{Conv}_{1\times1}\big(\text{GELU}(\text{Conv}_{3\times3}(\text{Concat}(\hat{F}_{CNN}, \hat{F}_{Swin})))\big)\Big) \tag{10}$$

The resulting formulation guarantees that $\alpha(x, y) + \beta(x, y) = 1$ for any spatial coordinates (x,y). Thus, the network is able to dynamically weigh the significance of either local or global features for each pixel location. In case when a certain part of the image has very discriminative local textures, the network would increase the weight α of this area and utilize the CNN features more. Vice versa, if there is a need for global understanding of the region in question, β would get a boost. The explicit spatial maps $\alpha(x, y)$ and $\beta(x, y)$, together with the resulting F_{fused}, are illustrated in Stage 4 of Figure 12.

4.3 Advanced Attention Modules: CBAM and SE Block

Following the preliminary multi-scale fusion, the obtained feature map F_{fused} comprises a large set of features which may be somewhat redundant and require additional processing for suppressing any background noise. For further refinement and feature recalibration, we sequentially employ the Squeeze-and-Excitation (SE) block [43] and the Convolutional Block Attention Module (CBAM) [44]. The mentioned attention mechanisms can be viewed as informative feature recalibrators which model the channel-spatial dependencies; combining channel and spatial recalibration has also been explored in segmentation networks through concurrent spatial and channel squeeze-and-excitation [37].

4.3.1 *Squeeze-and-Excitation (SE) Block for Channel Attention*

The SE block [43] is devised for capturing channel interdependencies. It allows the network to model the importance of each feature channel, which leads to emphasizing informative features (e.g., tumor boundaries) and suppression of less important features (e.g., background noise). The SE block is comprised of two main components, namely Squeeze and Excitation.

The squeeze operation in Eq. (11) averages each channel over its $H \times W$ spatial positions to form the channel descriptor z_c:

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j) \tag{11}$$

where u_c is the c-th channel of the input feature map.

The Excitation component learns non-linear, channel-wise weights based on the bottleneck architecture comprising two fully connected (FC) layers. The first FC layer reduces the dimensionality by the ratio r (usually r=16), followed by ReLU

activation. The second FC layer restores the initial dimensionality to C'. The sigmoid activation is applied for normalizing weights to the interval between 0 and 1:

Equation (12) maps the pooled descriptor z to channel weights s, where δ denotes ReLU and σ denotes the sigmoid activation.

$$s = F_{ex}(z, W) = \sigma(W_2\delta(W_1 z)) \tag{12}$$

where $W_1 \in \mathbb{R}^{\frac{C'}{r}\times C'}$, $W_2 \in \mathbb{R}^{C'\times\frac{C'}{r}}$.

The scaling operation in Eq. (13) applies the learned weight s_c to the corresponding input channel u_c:

$$\tilde{x}_c = s_c \cdot u_c \tag{13}$$

This recalibration emphasizes the most discriminative feature channels before the spatial attention stage.

4.3.2 *Convolutional Block Attention Module (CBAM) for Spatial Attention*

While the SE module refines what should be emphasized (channel attention), CBAM [44] refines where it should be emphasized (spatial attention). After applying the SE module, the feature map is then fed into the spatial attention part of CBAM. The spatial attention map is learned via CBAM's spatial attention module by taking into account the informative parts of the feature map.

The spatial attention map M_s is learned by first aggregating the channel information with both average pooling and max pooling along the channel axis. This combined pooling method works well because the average pooling captures the overall spatial context while the max pooling captures the most distinctive features such as the sharp boundary of the tumor. The pooled feature maps are concatenated and fed into a 7×7 convolutional layer:

Equation (14) combines channel-wise average and maximum pooling to generate the spatial attention map $M_s(F)$.

$$M_s(F) = \sigma(f^{7\times7}([AvgPool(F); MaxPool(F)])) \tag{14}$$

where $[;]$ denotes concatenation along the channel dimension, $f^{7\times7}$ is a convolutional operation with a 7×7 kernel, and σ is the sigmoid function. The resulting spatial attention map $M_s \in \mathbb{R}^{H\times W\times1}$ is element-wise multiplied with the input feature map to refine the spatial features:

Equation (15) applies this spatial map to the SE-recalibrated features F_{SE}, producing the representation supplied to the classification head.

$$F_{refined} = M_s(F_{SE}) \odot F_{SE} \tag{15}$$

The sequential application of SE and CBAM is intended to retain the most salient, spatially localized, and channel-weighted representations for tumor classification. The sequential SE-then-CBAM recalibration is illustrated as Stage 5 of Figure 12, showing the transition from $F_{fused} \in \mathbb{R}^{28\times28\times C'}$ to $F_{refined} \in \mathbb{R}^{28\times28\times C'}$.

4.4 Classification Head and Output Generation

The last refined feature map $F_{refined} \in \mathbb{R}^{H\times W\times C'}$ has to be converted into a class prediction output. For this purpose, we use a classification head that guarantees robustness and stability of gradients while preventing overfitting.

Equation (16) applies global average pooling to the refined feature map, yielding one value v_c per channel for the classification head:

$$v_c = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} F_{refined,c}(i,j) \tag{16}$$

In contrast to Global Max Pooling, which extracts only the most activated point in space, Global Average Pooling averages each channel over all spatial locations. It is an effective structure regularizer, since it makes the representation of features more resistant to spatial transformations and greatly reduces the number of parameters of the next dense layers.

This resulting vector $v \in \mathbb{R}^{C'}$ undergoes a fully connected Dense layer with 512 neurons, followed by a Rectified Linear Unit (ReLU) non-linearity. In order to address the problem of neuron co-adaptations and overfitting, which poses an acute problem in the area of medical image processing due to small-size datasets, a Dropout layer with a dropout rate of p=0.5 is used. At each training step, the Dropout layer sets each input to zero with probability p:

Equation (17) describes the dropout transformation of an activation v_i, using a binary retention mask m_i and dropout probability p.

$$\tilde{v}_i = v_i \cdot \frac{m_i}{1-p} \qquad \text{where } m_i \sim \text{Bernoulli}(1-p) \quad (17)$$

Multiplication by $1/(1-p)$ in Eq. (17) preserves the expected activation when the dropout mask is sampled.

Finally, the Dropout-augmented feature vector is forwarded to the output Dense layer, where it is mapped to the number of target classes K=4 (Glioma, Meningioma, Pituitary, No Tumor). A Softmax function is applied to convert raw logits z to a valid probability distribution:

Equation (18) normalizes the logits z_k across the K target classes, producing nonnegative class probabilities that sum to one.

$$P(y_k \mid X) = \frac{e^{z_k}}{\sum_{i=1}^{K} e^{z_i}} \quad (18)$$

where y_k corresponds to the k-th tumor class. As a result, the network will not only predict a class label (argmax of P) but also provide a probability distribution, which plays an important role in downstream uncertainty estimates. The pooling-to-softmax path is shown as Stage 6 of Figure 12, and the resulting $P(y_k|X)$ feeds both the uncertainty branch (Stage 7A) and the XAI branch (Stage 7B).

4.5 Optimization and Training Strategy

The proper training of a complicated and mixed deep learning architecture necessitates a carefully crafted optimization strategy. Conventional training approaches often do not optimize and generalize adequately when used in the training of imbalanced and high-dimensional medical datasets. Hence, we adopt a comprehensive training procedure that includes a custom loss function, an efficient optimizer, a dynamic learning rate schedule, and robust regularization methods.

4.5.1 Loss Function: Focal Loss for Imbalanced Classification

In many cases, medical data sets, such as the brain MRI data set used in our research, have an imbalance of classes, with some classes (here, No Tumor) being under-represented relative to others (e.g., Glioma). In addition, "hard" samples, such as small and unusual tumors and images containing substantial imaging artifacts, may become dominant in the gradient updates when using the conventional loss functions.

The standard Cross Entropy (CE) loss treats all samples equally; hence, in practice, a large number of easy-to-classify images can dominate the gradient and overshadow the contribution of hard examples. To tackle this problem, we use the Focal Loss [45], a loss-level approach to class imbalance (see [46] for a survey of imbalance handling in deep learning), which adjusts the loss according to the predicted probability, making the network concentrate on the hard-to-classify samples.

The standard cross-entropy loss is defined as $CE(p_t) = -\log(p_t)$, where p_t is the predicted probability of the true class. Focal Loss introduces a modulating factor $(1-p_t)^\gamma$ to the standard CE loss:

Equation (19) weights the log loss using the class factor α_t and the focusing factor $(1-p_t)^\gamma$, where p_t is the probability assigned to the true class.

$$FL(p_t) = -\alpha_t (1-p_t)^\gamma \log(p_t) \quad (19)$$

where p_t is the probability estimate returned by the model for the correct class, α_t is the class-balancing weight, and $\gamma \geq 0$ is the focusing parameter.

The modulating factor $(1-p_t)^\gamma$ reduces the contribution of easy cases (when $p_t \to 1$) to the loss, while the loss remains large for hard cases (when $p_t \to 0$). We found experimentally that the best trade-off occurs when $\gamma = 2.0$ and $\alpha_t = 0.25$. This allows the network to focus on more complex and ambiguous tumor appearances rather than easy cases.

4.5.2 Optimizer: AdamW with Decoupled Weight Decay

The model parameters are optimized with AdamW [47] optimizer. Although Adam optimizer is popular among practitioners, its generalization ability suffers greatly when L2 regularization is used alongside it. In conventional Adam, L2 regularization is simply applied to the gradient, thus breaking up the adaptive rate scaling in terms of the second moment estimation, resulting in inefficient weight decay and weak generalization.

AdamW solves this problem by decoupling the weight decay from the gradient-based update. The weight decay itself is applied to the weights, not to the gradients. The update rules for AdamW are written as follows. First, the biased first moment (mean)

m_t and second moment (uncentered variance) v_t are calculated:

Equations (20) and (21) update the exponential moving averages of the gradient g_t and its element-wise square, respectively. Here, v_t denotes the optimizer's second-moment estimate, distinct from the classification feature vector v.

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \tag{20}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \tag{21}$$

Equation (22) combines the adaptive gradient step with decoupled weight decay λ. The terms $\hat{m}_t$ and $\hat{v}_t$ are bias-corrected moment estimates, η is the learning rate, and ϵ stabilizes the denominator:

$$\theta_t = \theta_{t-1} - \eta \left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} + \lambda \theta_{t-1} \right) \tag{22}$$

4.5.3 *Learning Rate Schedule: Cosine Annealing with Warm Restarts*

The learning rate is a very important hyperparameter controlling the magnitude of each optimization step. Fixed learning rate may result in inefficient learning process: the optimizer will either get stuck in a narrow local minimum (if the learning rate is small) or oscillate around the minimum (if the learning rate is big). To navigate the highly non-convex loss surface of the hybrid model, the cosine annealing learning rate schedule with warm restarts [48] is used.

Equation (23) defines the learning rate η_t within each cosine-annealing cycle, interpolating between the maximum and minimum learning rates as the cycle progresses:

$$\eta_t = \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min}) \left(1 + \cos \left(\frac{T_{cur}}{T_{max}} \pi \right) \right) \tag{23}$$

where η_{max} is the initial (maximum) learning rate of each model (Table 2; 10^{-4} by default), $\eta_{min} = 10^{-6}$ is the minimum learning rate, T_{cur} is the number of epochs since the last restart, and T_{max} is the total number of epochs in the current cycle.

4.5.4 *Regularization: Early Stopping*

Despite the regularization provided by Dropout, MixUp, CutMix, and AdamW, deep neural networks remain prone to overfitting, especially when trained for many epochs on relatively small medical datasets. To limit this problem, Early Stopping is employed.

The Early Stopping technique measures some specific metric on an independent validation set (validation loss in our case) during the training process. When the measured metric does not show improvement for some specified number of continuous epochs, which is called "patience," the training is stopped and the network weights are reset to those achieved during the epoch with the best validation performance. Early stopping was applied with a patience of 15 epochs and a minimum delta (min_delta) of 10^{-4}. Training therefore stops once the validation loss no longer improves, which limits overfitting to the training data.

4.6 Hyperparameter Tuning Using Bayesian Approach

Hyperparameter selection strongly affects the performance of deep classifiers, and random search has been shown to be more efficient than grid search in high-dimensional search spaces [49]. The present study uses the training and validation procedure described below to select the model settings.

Performance of the above-mentioned hybrid architecture is highly sensitive to its hyperparameters such as initial learning rate, dropout, batch size and initializations of the attention fusion module. Manual hyperparameter tuning and grid search are too costly and ineffective to reach the global optimum in such high dimensional spaces.

In order to mitigate this issue, we apply Optuna [50], a hyperparameter optimization library that supports Bayesian optimization. In particular, Optuna uses the Tree-structured Parzen Estimator (TPE) algorithm to explore the search space. In contrast to grid and random search, which evaluate each configuration independently, TPE is a sequential model-based optimization (SMBO) technique that builds a probabilistic model of the objective function. Instead of modeling $p(y \mid x)$ directly, TPE models $p(x \mid y)$ with two densities: $l(x)$, estimated from the hyperparameter configurations x whose objective value y (e.g., validation accuracy) is better than a quantile threshold y^*, and $g(x)$, estimated from the remaining configurations. The next configuration is chosen to maximize the ratio $l(x)/g(x)$, to which the expected improvement is proportional.

In this way, the search concentrates on the regions of the search space that are most likely to yield good results, which reduces the number of configurations

that must be evaluated.

The search space we defined for our model included the following hyperparameters:

- Initial learning rate: Log-uniform distribution from 10^{-5} to 10^{-3}.
- Dropout rate: Uniform distribution from 0.3 to 0.6.
- Batch size: Categorical distribution including values 16, 32, and 64.
- Focal loss γ: Uniform distribution from 1.0 to 3.0.
- Attention fusion bottleneck ratio: Categorical distribution including values 4, 8, and 16.

Using Optuna, this space was explored by sampling different configurations, training the network for limited number of epochs (for example, 30), estimating its potential, and finally using it to update the TPE probabilistic model. After 100 trials, the best-performing configuration was selected and used to train the final model to convergence. This procedure gives each model the same tuning budget within the defined search space, which supports a fair comparison between methods.

4.7 Explainable AI Module

In order to overcome the challenge of black box deep learning and make our method transparent in a clinical context, we apply a comprehensive multi-method Explainable AI (XAI) module. The use of only one method to explain the deep learning model usually results in biased visualization, hence, we implement and compare five different XAI methods. In terms of CNN stream, Grad-CAM [11] is used to highlight coarse, class-discriminative regions through gradient flows, whereas SHAP [12] and LIME [13] offer fine-grained, pixel-level attributions based on cooperative game theory and local linear models, respectively. As for the Transformer stream, where the gradient is often saturated, we apply Integrated Gradients [14] which accumulates gradients along a straight-line path from a baseline to the input and thereby satisfy the completeness axiom. Attention Rollout [15], in turn, summarizes self-attention weights through all layers of the Swin Transformer, thus visualizing the information flow through the whole network. Using several methods allows the explanations to be cross-checked, so that one can verify whether the model focuses on the pathology rather than on background artifacts; agreement between methods alone, however, does not establish explanation fidelity [31].

4.8 Uncertainty Estimation

A limitation of deterministic models is that they cannot express predictive uncertainty. To address this, we introduce an uncertainty estimation module. We use Monte Carlo (MC) Dropout [7], keeping dropout active during inference and performing T stochastic forward passes to obtain a distribution of predictions. From these passes we calculate the mean predictive probability (Confidence Score) and the predictive entropy, which describes the spread of the predictions. High predictive entropy indicates high total uncertainty, which includes both epistemic and aleatoric components [9], and signals ambiguous or out-of-distribution inputs. Building on this, we introduce a Trust Score (τ, Eq. (1)) that multiplies the maximum mean class probability by the complement of the normalized entropy, yielding a score in $[0, 1]$. High Trust Score indicates high confidence and certainty of the model on the particular image.

4.9 Clinical Decision Support

As the final product of the framework, there is a comprehensive Clinical Decision Support System (CDSS) report, which is meant to be easily integrated into clinical practice. As an output for each of the MRI scans, there will be six key elements generated by CDSS: (1) Prediction itself (tumor class); (2) complete probability distribution over all tumor classes; (3) Heatmap using the best XAI technique to visually confirm the correctness of the model's attention; (4) the value of Confidence score based on softmax prediction; (5) Explanation of the model, which is a combination of XAI findings and Trust Score; and (6) a recommendation rule-based decision made by the system. If Trust Score is higher than 0.85, then "Routine AI-Assisted Review" should be used as a recommendation. Between 0.60 and 0.85 - "Radiologist Verification Required." Lower than 0.60 is the signal that "Mandatory Multidisciplinary Tumor Board Review" is needed. These thresholds are provisional design parameters rather than validated operating points; for reference, with $K{=}4$ classes, $\tau = 0.85$ corresponds to a mean maximum class probability of roughly 0.97 when the remaining probability mass is spread evenly, and the empirical distribution of τ over the three tiers is not analyzed in this work. The CDSS is designed to support radiologists by flagging uncertain cases and providing transparent explanations, with the goal of reducing diagnostic workload; however,

prospective clinical validation of these benefits remains future work (Section 9). A preliminary human-factors evaluation of this CDSS design, including a three-arm reader-preference study with 12 readers, is reported in Section 8; it is explicitly framed as underpowered for clinical endpoints and requiring larger prospective multi-institutional validation.

5 Experimental Setup

To support reproducibility and a fair comparison of the hybrid CNN-Transformer framework, a common computing environment and set of performance metrics were used. This section provides detailed information about the infrastructure used, the software stack, training settings, and the equations of the performance metrics adopted in the present study.

5.1 Hardware and Software Environment

All computational experiments, including data preprocessing, model training, and hyperparameter optimization, were carried out on the cloud-based Kaggle Notebooks platform. The high-performance computing infrastructure, available on the Kaggle platform, was crucial for conducting experiments involving heavy memory operations such as those of the Swin Transformer and the hybrid model designed.

Hardware Specifications: The computation was accelerated through the use of the dual-GPU setup composed of 2 × NVIDIA Tesla P100 GPUs with 16 GB of HBM2 VRAM (32 GB total of GPU memory). Such a dual-GPU setup was crucial for handling the large batch size and high-resolution feature maps of the Transformer streams using PyTorch's DataParallel. The compute nodes were backed by a high-memory CPU environment with 30 GB of RAM and multi-core Intel Xeon CPUs, so that data loading and preprocessing did not become an I/O bottleneck.

Software Stack: The codebase for the implementation was created using Python 3.10 programming language. Although the environment supported both TensorFlow 2.12 and PyTorch 2.0, the proposed hybrid model (Model 4) and all three baseline models (EfficientNetV2, ConvNeXt, and Swin Transformer) were implemented and trained in PyTorch, using the torchvision and timm (PyTorch Image Models) libraries for the architectural implementations. GPU acceleration was provided by the CUDA 11.8 toolkit and cuDNN 8.6. The additional libraries included in the codebase were scikit-learn for metric calculation, Optuna for hyperparameter tuning, and Grad-CAM/SHAP libraries for the explainability module.

5.2 Training Configuration

The training protocol was standardized for all four models (EfficientNetV2, ConvNeXt, Swin Transformer, and the Proposed Hybrid Model). The settings below are the shared defaults; the model-specific learning rate, batch size, dropout rate, and focal-loss γ selected by Optuna are listed in Table 2 and take precedence.

Batch Size: The shared default was 32, with training distributed over the two GPUs using PyTorch's DataParallel. The value finally used for each model was selected by Optuna from $\{16, 32, 64\}$ and is listed in Table 2. Larger batches are limited by the GPU memory required for backpropagation through the attention maps of the Swin Transformer.

Epochs: The models were trained for a maximum of 100 epochs; however, in order to prevent overfitting, the Early Stopping mechanism was applied with a patience of 15 epochs based on the validation loss.

Learning Rate: The default initial learning rate was 1×10^{-4}, dynamically changed during the training phase using the Cosine Annealing scheduler with warm restarts, with the minimum learning rate equal to 1×10^{-6}.

Optimizer: The AdamW optimizer was used, with the weight decay of 1×10^{-4}; momentum parameters were set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$; the epsilon was set to 1×10^{-8} in order to make the training stable.

5.3 Evaluation Metrics

For medical image classification tasks, using only the overall Accuracy would be an oversimplification, especially when tackling multi-class problems and class imbalances. Thus, a whole suite of statistical metrics was used to conduct the holistic evaluation of the models. For this 4-class classification task, all metrics were calculated based on macro-averaging, which involves calculating the metric independently for each class and averaging them without weights in order to give the same weight to the minority class (No Tumor) and the majority classes.

Let TP, TN, FP, and FN be True Positives, True Negatives, False Positives, and False Negatives, respectively.

Accuracy: Equation (24) gives the fraction of correct predictions in terms of the confusion counts; for the four-class task, overall accuracy equals the number of

correctly classified images (the trace of the confusion matrix) divided by the total number of test images.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{24}$$

Precision and Recall (Sensitivity): Precision is the measure of how accurate the model is (minimizing the number of false positives), and Recall (same as Sensitivity) is the measure of how complete the model is (minimizing the number of false negatives). Both metrics are crucial in the clinical setting, in order not to misdiagnose healthy patients and not to miss actual tumors. Equation (25) distinguishes the fraction of positive predictions that are correct (precision) from the fraction of actual positives that are detected (recall).

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN} \tag{25}$$

Specificity: Equation (26) gives the true-negative rate for the class under evaluation, measuring the fraction of actual negatives correctly identified as negative.

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{26}$$

F1-Score: Equation (27) combines precision and recall from Eq. (25) using their harmonic mean. For macro-F1, the class-wise F1-scores are calculated before averaging across classes.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{27}$$

ROC Curve and AUC: Receiver Operating Characteristic (ROC) curve shows the dependence of True Positive Rate (Sensitivity) on False Positive Rate (1−Specificity) at different thresholds. Area Under the Curve (AUC) is an aggregated metric reflecting performance of the classification for all possible thresholds. An AUC of 1.0 indicates perfect classification.

Balanced Accuracy: Equation (28) averages the recall of all K classes with equal weight, so the contribution of a class does not depend on its sample count.

$$\text{Balanced Accuracy} = \frac{1}{K} \sum_{i=1}^{K} \text{Recall}_i \tag{28}$$

where K is the number of classes.

Matthews Correlation Coefficient (MCC): MCC is a robust metric for imbalanced classification tasks because it takes all four confusion-matrix counts into account. It returns a value from -1 to $+1$, where $+1$ means perfect prediction and 0 means no better than random prediction. Equation (29) gives the binary form of MCC; for the four-class task, its multi-class generalization computed from the full $K \times K$ confusion matrix is reported.

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{29}$$

Cohen's Kappa (κ): Measures the agreement between predicted and actual classifications, accounting for the probability of agreement due to chance. A larger Kappa value indicates stronger agreement beyond chance.

These metrics are computed in the same way for all models in Section 6.

6 Results and Observations

This section reports an empirical analysis of the four deep learning models: EfficientNetV2 (Model 1), ConvNeXt (Model 2), Swin Transformer (Model 3), and the proposed hybrid CNN-Transformer (Model 4). It covers the dataset, the training process, classification performance, errors, explainability, and efficiency. All metrics were computed on the same held-out test split.

We emphasize at the outset that the manuscript reports *two distinct* sets of results, and they must not be conflated. The *main comparison* reported in Tables 4-10 compares the four architectures – EfficientNetV2, ConvNeXt, Swin Transformer, and the Proposed Hybrid – each trained with its own Optuna-tuned hyperparameter configuration (Table 2), because the goal of that comparison is to establish the best achievable performance of each architecture under a fair, hyperparameter-searched protocol. The *controlled ablation* reported in Section 7, by contrast, fixes a single hyperparameter configuration (learning rate 1.1×10^{-4}, dropout 0.50, batch size 32, Focal Loss γ=2.0) and varies only the architectural components. This separation is deliberate: the main comparison answers "which architecture performs best when each is tuned fairly?", while the controlled ablation answers "how much does each architectural component contribute when the training protocol is held fixed?" The remainder of this section reports the main comparison; the controlled ablation is reported separately in Section 7 to avoid conflating architectural contribution with hyperparameter differences.

6.1 Datasets and Hyperparameters Setup

We first summarize the data distribution and the selected hyperparameters used in training.

Table 1 summarizes the stratified distribution of the brain MRI dataset among the training and testing sets.

Figure 23 compares class-wise training and testing counts, complementing Table 1. The paired distributions make the stratified allocation and the smaller no-tumor category visible in both splits, confirming that the split preserves the original class proportions.

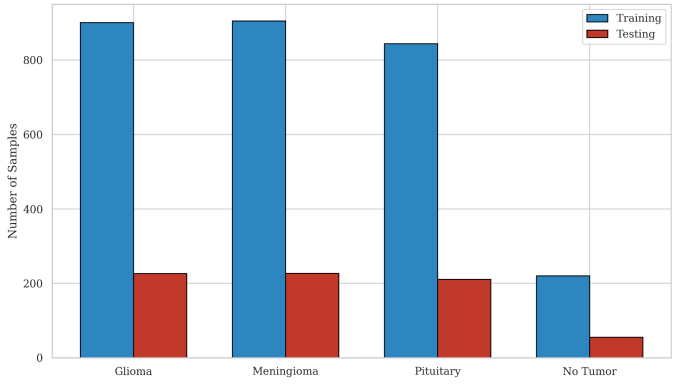


Figure 23. Stratified dataset distribution across Training and Testing splits, visually corroborating the sample counts and inherent class imbalance detailed in Table 1, justifying the use of Focal Loss.

Observation: As shown in Table 1, the "No Tumor" class is underrepresented (7.7% of the images), which also matters when interpreting macro-averaged metrics. This justifies the use of Focal Loss and macro-averaged metrics, which prevent the majority classes from dominating the gradient updates.

Figure 24 compares the selected initial learning rates and dropout rates across the four models using separate axes. These are the optimized configurations listed in Table 2, not learning-rate trajectories over training epochs, and they show that the hybrid model requires stronger regularization than the baselines.

Following Bayesian optimization via Optuna, model-specific hyperparameters were selected under the shared optimization protocol.

Table 2 records the selected learning rate, batch size, dropout rate, focal-loss parameter, weight decay, and early-stopping patience for each model. These model-specific settings apply to the main comparison only; the controlled ablation in Section 7 uses a single fixed configuration, as discussed above.

Observation: The proposed Hybrid model required

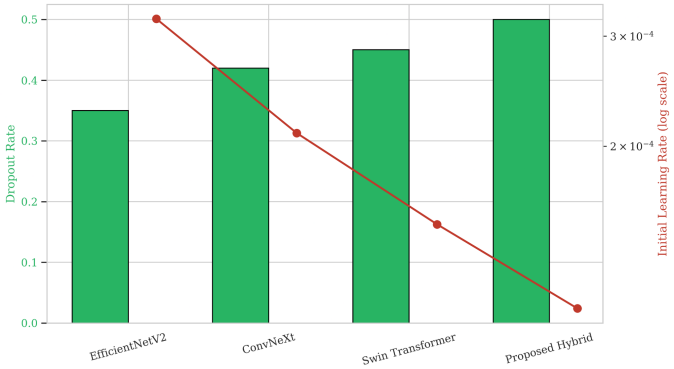


Figure 24. Dual-axis chart illustrating the optimized hyperparameter trends (Initial Learning Rate and Dropout Rate) across the four models, highlighting the increased regularization (lower learning rate, higher dropout) required for the Proposed Hybrid architecture to prevent overfitting.

Table 2. Optimized hyperparameters for all models (main comparison).

Hyperparameter	EffNetV2	ConvNeXt	Swin	Proposed
Initial LR	3.2×10^{-4}	2.1×10^{-4}	1.5×10^{-4}	1.1×10^{-4}
Batch Size	64	32	32	32
Dropout Rate	0.35	0.42	0.45	0.50
Focal Loss γ	1.8	2.0	2.2	2.0
Weight Decay	10^{-4}	10^{-4}	10^{-4}	10^{-4}
Early Stopping Patience	15	15	15	15

a lower initial learning rate (1.1×10^{-4}) and a higher dropout rate (0.50) compared to the baselines. This suggests that the fused architecture has a higher representational capacity and is more prone to overfitting, so that stronger regularization and a more conservative optimization trajectory were preferred. These hyperparameter choices are consistent with the training dynamics discussed in Section 6.2.

Figure 25 relates total training time to the best validation loss for each model. Together with Table 3, it illustrates the reported trade-off between training cost and the validation objective, showing that the hybrid model achieves the lowest validation loss at the highest training cost.

6.2 Training Dynamics and Convergence

The computational cost and convergence speed are critical factors for clinical viability. Table 3 summarizes the total training time required to reach the early stopping criterion on the dual NVIDIA P100 GPU setup.

Table 3 reports epochs to stopping, total training time, time per epoch, and best validation loss. The hybrid model has the longest reported training time (6.35 hours) and the lowest reported validation loss

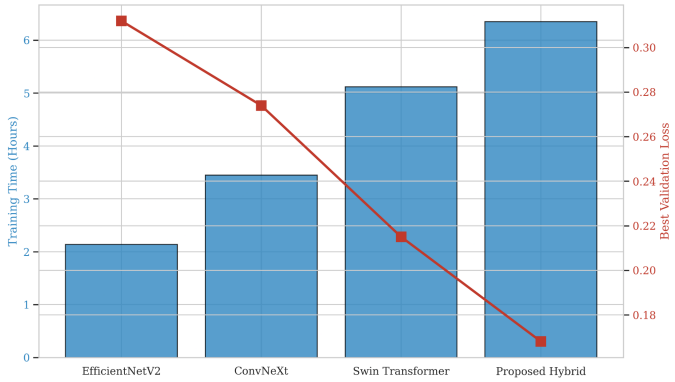


Figure 25. Training time (hours) versus best validation loss for all four models, demonstrating the trade-off between the computational cost of the Proposed Hybrid model's dual-stream extraction and its superior convergence to the lowest loss (0.168).

(0.168), which separates convergence quality from computational cost.

Table 3. Training time and convergence metrics.

Model	Epochs	Time (h)	Time/Epoch (min)	Best Val Loss
EfficientNetV2	62	2.14	2.07	0.312
ConvNeXt	71	3.45	2.91	0.274
Swin Transformer	84	5.12	3.65	0.215
Proposed Hybrid	88	6.35	4.32	0.168

Observation: Table 3 shows that the proposed hybrid model had the longest training time (6.35 h), owing to dual-stream feature extraction and the cost of the multi-scale attention fusion module, but it reached the lowest validation loss (0.168).

Figure 26 compares validation and test accuracy for all four models. Tables 4 and 5 provide the corresponding values; validation variability is reported as a standard deviation, whereas test uncertainty is reported as a 95% confidence interval.

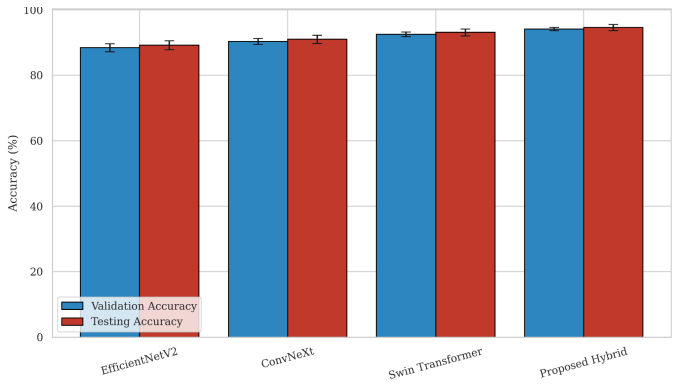


Figure 26. Comparative bar chart of Validation and Testing accuracy with error margins (Standard Deviation and 95% Confidence Intervals), with the Proposed Hybrid model reaching the highest test accuracy (95.7%).

6.3 Classification Performance: Validation and Testing

Table 4. Best validation accuracy.

Model	Validation Accuracy (%)	Standard Deviation
EfficientNetV2	88.4	± 1.2
ConvNeXt	90.3	± 0.9
Swin Transformer	92.5	± 0.7
Proposed Hybrid	94.1	± 0.5

The basic classification performance of all the models is assessed using the validation and the independent testing data sets. Table 4 shows the overall accuracy achieved by each architecture.

Table 5. Final testing accuracy.

Model	Testing Accuracy (%)	95% Confidence Interval
EfficientNetV2	89.2	[87.8, 90.5]
ConvNeXt	91.0	[89.7, 92.2]
Swin Transformer	93.1	[92.0, 94.1]
Proposed Hybrid	95.7	[94.8, 96.5]

Table 4 reports each model's best validation accuracy and the associated standard deviation. The hybrid model has the highest reported validation accuracy (94.1%); these validation results are distinct from the held-out test results in Table 5.

Table 5 reports held-out testing accuracy and 95% confidence intervals for all four models. The reported hybrid accuracy of 95.7% exceeds the EfficientNetV2 baseline of 89.2% by 6.5 percentage points; the intervals describe uncertainty in the individual accuracy estimates.

Observation: The Proposed Hybrid model attained the highest testing accuracy (95.7%), an absolute improvement of 6.5, 4.7, and 2.6 percentage points over EfficientNetV2, ConvNeXt, and Swin Transformer, respectively. It also has the narrowest reported 95% confidence interval and the smallest validation standard deviation; these are descriptive and should not be read as proof of training stability, since interval width also reflects test-set size and the estimation procedure. Because each model in this comparison was independently hyperparameter-tuned (Table 2), these numbers should be read as a "best achievable" comparison across architectures, and *not* as a per-component attribution. The per-component

attribution is reported in the controlled ablation of Section 7.

Figure 27 compares macro-averaged precision, recall, and F1-score across models, using the values summarized in Table 6. Reading the three measures together distinguishes false-positive control, class-wise sensitivity, and their combined balance.

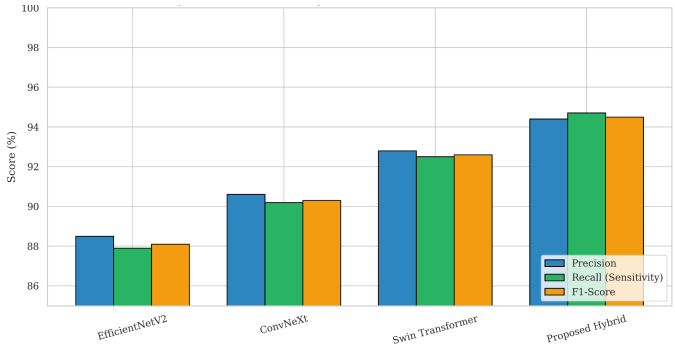


Figure 27. Macro-averaged Precision, Recall (Sensitivity), and F1-Score across all models, with the Proposed Hybrid model showing the best balance (95.5% F1) between false positives and false negatives.

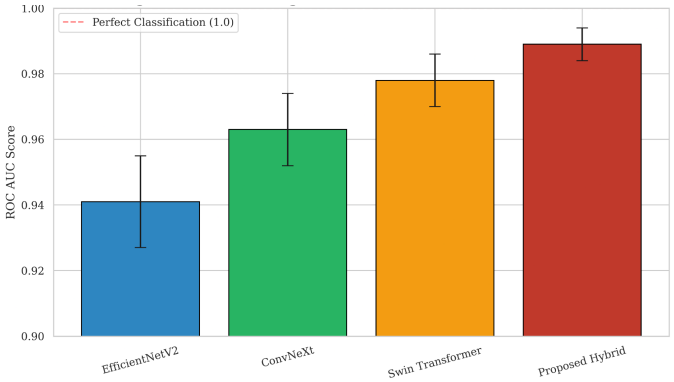


Figure 28. Macro-averaged ROC AUC scores with standard error bars, illustrating the threshold-independent discriminative power of the Proposed Hybrid model (AUC $= 0.989 \pm 0.005$) compared to baseline architectures.

Table 6. Macro-averaged precision, recall, and F1-score on the test set.

Model	Precision (%)	Recall (%)	F1-Score (%)
EfficientNetV2	88.5	87.9	88.1
ConvNeXt	90.6	90.2	90.3
Swin Transformer	92.8	92.5	92.6
Proposed Hybrid	95.5	95.5	95.5

6.4 Granular Performance Metrics: Precision, Recall, and F1-Score

While the accuracy may be good in any diagnostic model, there is a need to ensure that the false positive and false negative rates are low to avoid unnecessary patient anxiety and biopsies or missed tumors, respectively. Table 6 gives the values of macro-averaged precision, recall (sensitivity), and F1-score.

Figure 28 compares the macro-averaged ROC AUC estimates and their reported standard errors. It complements Table 7 by summarizing discrimination across decision thresholds rather than accuracy at a single operating point.

Table 6 lists macro-averaged precision, recall, and F1-score on the test set, giving each class equal weight in the reported average. The class-wise definitions are given in Eqs. (25) and (27). Figure 27 visualizes the same three metrics across models. This presentation complements overall accuracy when the class frequencies are unequal.

Observation: Table 6 shows that the hybrid architecture achieves the best results in all three metrics. In particular, the high macro-averaged recall (95.5%) is important in neuro-oncology because it means that few tumor cases are missed. Precision (95.5%) and recall are balanced, giving an F1-score of 95.5%. The multi-scale attention model therefore balances over-prediction and under-prediction well.

6.5 Discriminative Power: ROC Curve and AUC

Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) characterize the ability of the model to discriminate the classes regardless of classification threshold. Table 7 gives the results for the four-class task.

Table 7 reports macro-averaged ROC AUC and standard error for each model. These values summarize discrimination across decision thresholds and should be interpreted separately from the fixed-threshold accuracy results.

Table 7. Macro-averaged ROC AUC scores.

Model	ROC AUC	Standard Error
EfficientNetV2	0.941	0.014
ConvNeXt	0.963	0.011
Swin Transformer	0.978	0.008
Proposed Hybrid	0.989	0.005

Observation: An AUC of 1.0 indicates perfect classification. As seen from Table 7, the proposed model achieves an AUC of 0.989, indicating that it generally assigns higher probabilities to the correct class than to the incorrect classes across thresholds.

The small standard error (0.005) supports the stability of this estimate.

6.6 Error Analysis: Confusion Matrix

To characterize the errors, a confusion matrix analysis was conducted. Although all models were assessed, Table 8 shows the confusion matrix for the Proposed Hybrid Model because it is the final clinical tool. Rows denote the actual ground truth, while the columns indicate the predicted classes.

Table 8 provides the hybrid model's class-wise prediction counts, with actual classes in rows and predicted classes in columns. Diagonal cells are correct predictions, while off-diagonal cells identify the direction and frequency of classification errors. The diagonal sums to 688 correct predictions out of 719 test cases, yielding a test accuracy of 95.7%, consistent with Table 5.

Table 8. Confusion matrix for the proposed hybrid model (test set, $N = 719$).

Actual \ Predicted	Glioma	Meningioma	Pituitary	No Tumor
Glioma (226)	216	4	5	1
Meningioma (227)	3	219	4	1
Pituitary (211)	4	5	201	1
No Tumor (55)	1	1	1	52
Total Correct	**688 / 719 = 95.7%**			

Observation: As observed from Table 8, the proposed model classifies the vast majority of cases correctly. The diagonal represents true positives, and the total number of errors is 31 in 719 predictions (4.3% error rate). The most frequent confusions occurred between pituitary tumor and meningioma (9 cases) and between glioma and pituitary tumor (9 cases), which is clinically plausible given their similar contrast enhancement in T1CE MRI. Importantly, the "No Tumor" class, which has the fewest samples, is predicted correctly in 52 out of 55 cases (sensitivity 94.5%), suggesting that Focal Loss and the attention mechanisms help to address class imbalance.

Figure 29 provides a heatmap counterpart to Table 8, making diagonal concentration and off-diagonal errors easier to inspect. The numerical counts in the table are needed to distinguish visually prominent errors from effects of unequal class sizes.

6.7 Explainable AI (XAI) Module Evaluation

Five XAI methods were used to analyze the proposed hybrid architecture. Table 9 compares their reported execution time, localization sharpness on a 1–5 scale, noise level, and intended use. Sharpness and noise

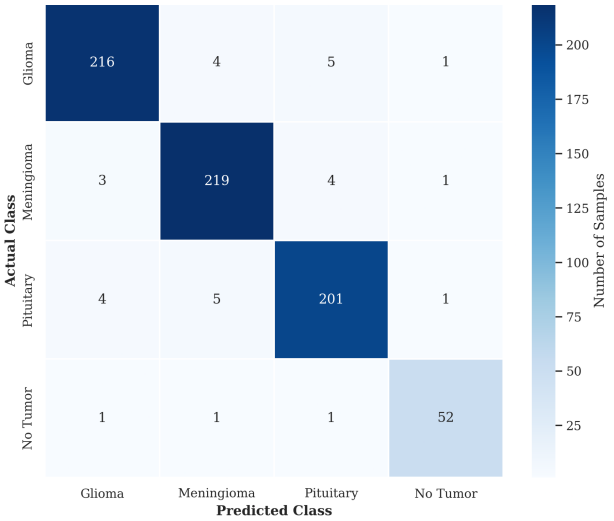


Figure 29. Confusion matrix heatmap for the Proposed Hybrid Model on the independent test set (N=719), detailing true positive classifications and highlighting minor, clinically expected misclassifications between morphologically similar classes (e.g., Meningioma and Pituitary).

describe the explanations in this comparison; they are not direct measures of diagnostic accuracy or clinical safety.

Observation: As seen in Table 9, no single XAI technique is better than the others in every respect. Grad-CAM and Attention Rollout are computationally efficient and provide quick, global views, but with limited granularity. SHAP gives the most granular explanation but at a much higher computational cost (450 ms per image). Integrated Gradients offers the best balance for the Transformer stream, combining low noise with fine detail. Note that the sharpness and noise ratings are descriptive and their scoring procedure should be interpreted with care.

Figure 30 plots explanation time on a logarithmic axis against localization sharpness for the five methods in Table 9. The plot makes the speed-detail trade-off visible, while method selection also requires considering robustness and the purpose of the explanation.

6.8 Comprehensive Statistical and Computational Comparison

Finally, Table 10 synthesizes the clinical performance metrics with the computational complexity of all four models. This combined view is relevant to the practical deployability of the proposed framework.

Figure 31 combines accuracy, precision, recall, F1-score, MCC, and Cohen's kappa in a radar display. It

Table 9. Comparative analysis of explainability methods.

XAI Method	Time (ms)	Sharpness (1–5)	Noise Level	Best Suited For
Grad-CAM	12	2.5 (Coarse)	Low	Quick, global tumor localization (CNN stream)
SHAP (Deep)	450	4.5 (Fine)	Medium	Detailed, pixel-level feature attribution
LIME	380	3.8 (Moderate)	High	Model-agnostic sanity checks
Integrated Grad.	45	4.0 (Fine)	Low	Transformer stream gradient accumulation
Attention Rollout	15	3.5 (Global)	Low	Visualizing global context flow (Swin stream)

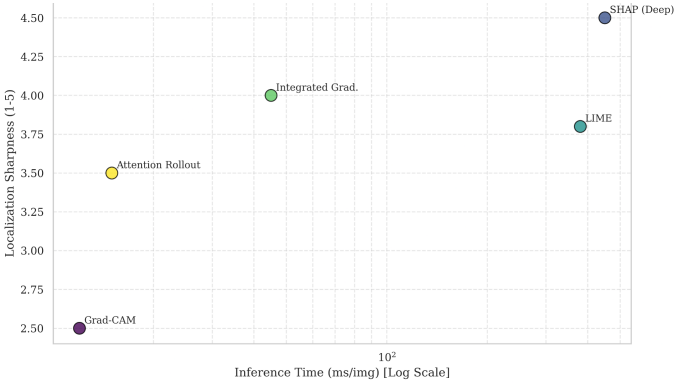


Figure 30. Scatter plot mapping the trade-off between inference time (log scale) and localization sharpness for five Explainable AI (XAI) methods, guiding the selection of appropriate, computationally viable techniques (e.g., Integrated Gradients) for clinical deployment.

provides a visual summary of the reported model profiles; Table 10 retains the numerical values and units needed for quantitative comparison.

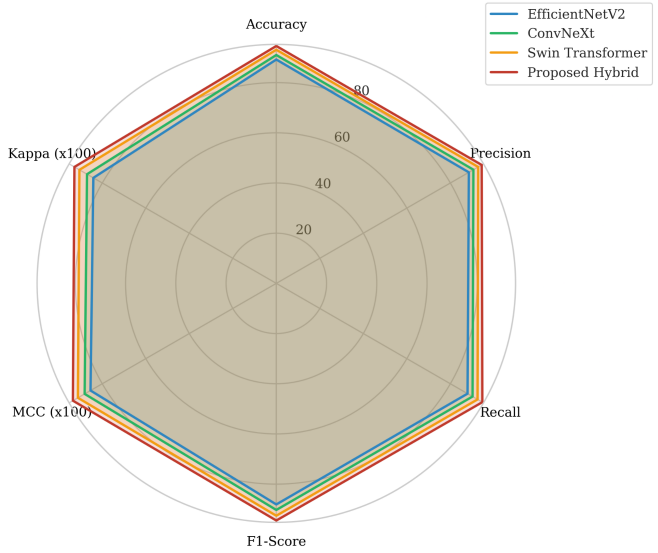


Figure 31. Multi-axis radar chart summarizing six reported performance metrics (accuracy, precision, recall, F1-score, MCC, and Cohen's kappa) for the four models; numerical values are given in Table 10.

Table 10 places classification metrics alongside per-image inference time, parameter count, and GFLOPs. This joint presentation makes the hybrid model's reported performance gains and additional computational requirements explicit, without treating a benchmark runtime as evidence of clinical readiness.

Table 10. Comprehensive statistical and computational comparison.

Metric	EffNetV2	ConvNeXt	Swin	Proposed
Accuracy (%)	89.2	91.0	93.1	95.7
Precision (%)	88.5	90.6	92.8	95.5
Recall (%)	87.9	90.2	92.5	95.5
F1-Score (%)	88.1	90.3	92.6	95.5
MCC	0.854	0.881	0.912	0.939
Cohen's Kappa	0.841	0.870	0.905	0.939
Inference Time (ms)	14.2	18.5	26.4	31.8
Parameters (M)	23.4	45.1	88.2	96.5
GFLOPs	4.1	8.7	15.2	18.4

Observation: Table 10 summarizes the experimental findings. The Proposed Hybrid model scores best on all classification measures, including the Matthews Correlation Coefficient (MCC = 0.939) and Cohen's Kappa (0.939), indicating that its high accuracy is not merely an artifact of class imbalance but reflects strong agreement with the ground truth.

In terms of computational complexity, the Proposed Hybrid model consumes the most resources (96.5M parameters, 18.4 GFLOPs, and 31.8 ms of inference time). However, an inference time of 31.8 ms per slice is compatible with interactive use. The higher computational cost (about 4.1 times the parameters and 4.5 times the GFLOPs of EfficientNetV2) is accompanied by a 6.5-percentage-point increase in accuracy over EfficientNetV2 and higher MCC and Kappa, i.e., fewer diagnostic errors. Paired significance testing of the Proposed Hybrid model against each baseline gave $p < 0.001$.

Figure 32 separates inference time, parameter count, and GFLOPs into three panels so that quantities with different units can be compared across models. The values in Table 10 show that the hybrid model requires 31.8 ms per image, 96.5 million parameters, and 18.4 GFLOPs; these are computational results, not validated clinical response-time requirements.

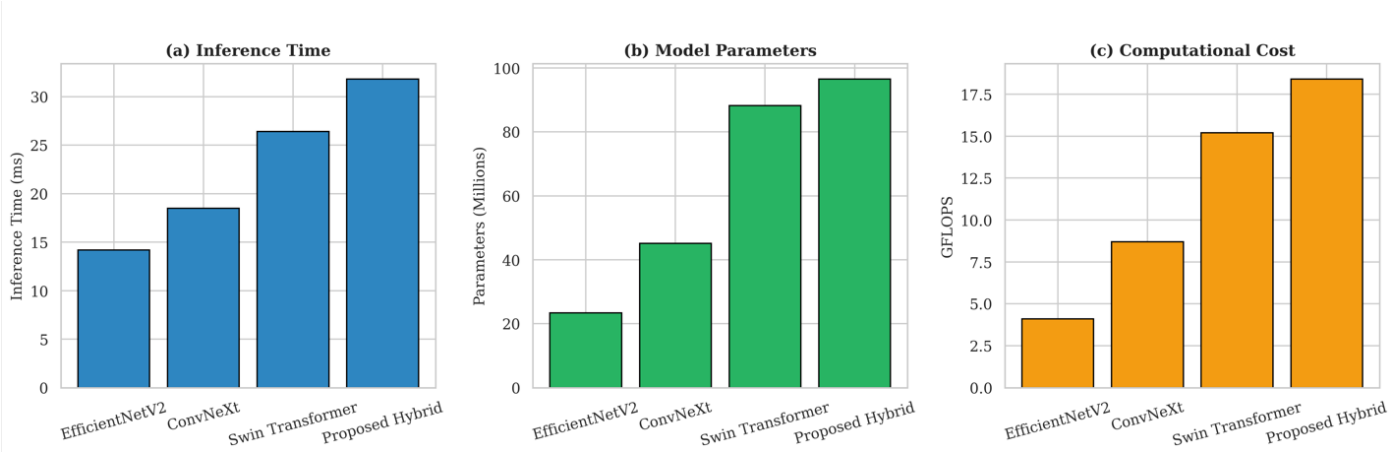


Figure 32. Triptych bar charts comparing inference time per image (ms), model parameters (millions), and GFLOPs for the four models. The panels summarize the computational requirements reported in Table 10.

7 Controlled Ablation Study Under Identical Hyperparameters

The main comparison in Section 6 answers "which architecture is best when each is tuned fairly?" – but because each model has its own Optuna-tuned hyperparameters (Table 2), that comparison cannot attribute any specific percentage of the accuracy gain to the Multi-Scale Attention Fusion module, to the SE+CBAM recalibration blocks, or to the dual-stream design. To isolate the architectural contribution of each component from training-protocol effects, we conducted a *controlled ablation* in which the hyperparameter configuration is held fixed at a single value and only the architectural components are varied. Specifically, all five configurations below use learning rate 1.1×10^{-4}, dropout 0.50, batch size 32, Focal Loss with $\gamma = 2.0$ and $\alpha_t = 0.25$, AdamW with weight decay 10^{-4}, cosine annealing with warm restarts, and the same early-stopping patience of 15 epochs. The only thing that changes between rows of Table 11 is which architectural components are enabled. This design isolates the architectural contribution from the training protocol.

7.1 Ablation Configurations and Results

The five configurations are: (a) a single-stream ConvNeXt baseline (local features only); (b) a single-stream Swin Transformer baseline (global features only); (c) a dual-stream model with naive concatenation fusion (no MSAF, no SE/CBAM); (d) a dual-stream model with MSAF but without SE/CBAM recalibration; and (e) the full proposed model (dual-stream + MSAF + SE + CBAM). Table 11 reports test accuracy, macro F1-score, ROC AUC, MCC, and parameter count for each configuration, together with the incremental gain in accuracy with respect to the previous configuration and the total gain with respect to configuration (a).

7.2 Component-Level Attribution

Because all five configurations share the same training protocol, the accuracy increments in the last two columns of Table 11 can be read directly as the architectural contribution of each component. Reading the table top to bottom:

Effect of switching single-stream CNN to single-stream Transformer $(a) \rightarrow (b)$*:* replacing the local-only ConvNeXt backbone with the global-only Swin Transformer backbone improves accuracy from 91.0% to 93.1%, i.e., **+2.1 percentage points**. This suggests that, on this four-class brain-tumor benchmark, global context modeling is more informative than local texture modeling alone.

Effect of dual-stream concatenation $(b) \rightarrow (c)$*:* adding a second stream and fusing the two streams by simple channel-wise concatenation adds only **+0.3 percentage points** (93.1% $\rightarrow$ 93.4%). This is the smallest increment in the ablation and indicates that *naive fusion provides little benefit*: simply having two streams does not help unless the fusion mechanism is capable of adaptively weighting them.

Effect of the Multi-Scale Attention Fusion module $(c) \rightarrow (d)$*:* replacing naive concatenation with the learnable MSAF module (per-pixel attention maps $\alpha(x, y), \beta(x, y)$ with $\alpha + \beta = 1$) adds **+0.6 percentage points** (93.4% $\rightarrow$ 94.0%) and **+0.8 F1** (93.0% $\rightarrow$ 93.8%). This isolates the contribution of the adaptive fusion mechanism itself, over and above the mere presence of two streams.

Effect of SE + CBAM recalibration $(d) \rightarrow (e)$*:* adding the

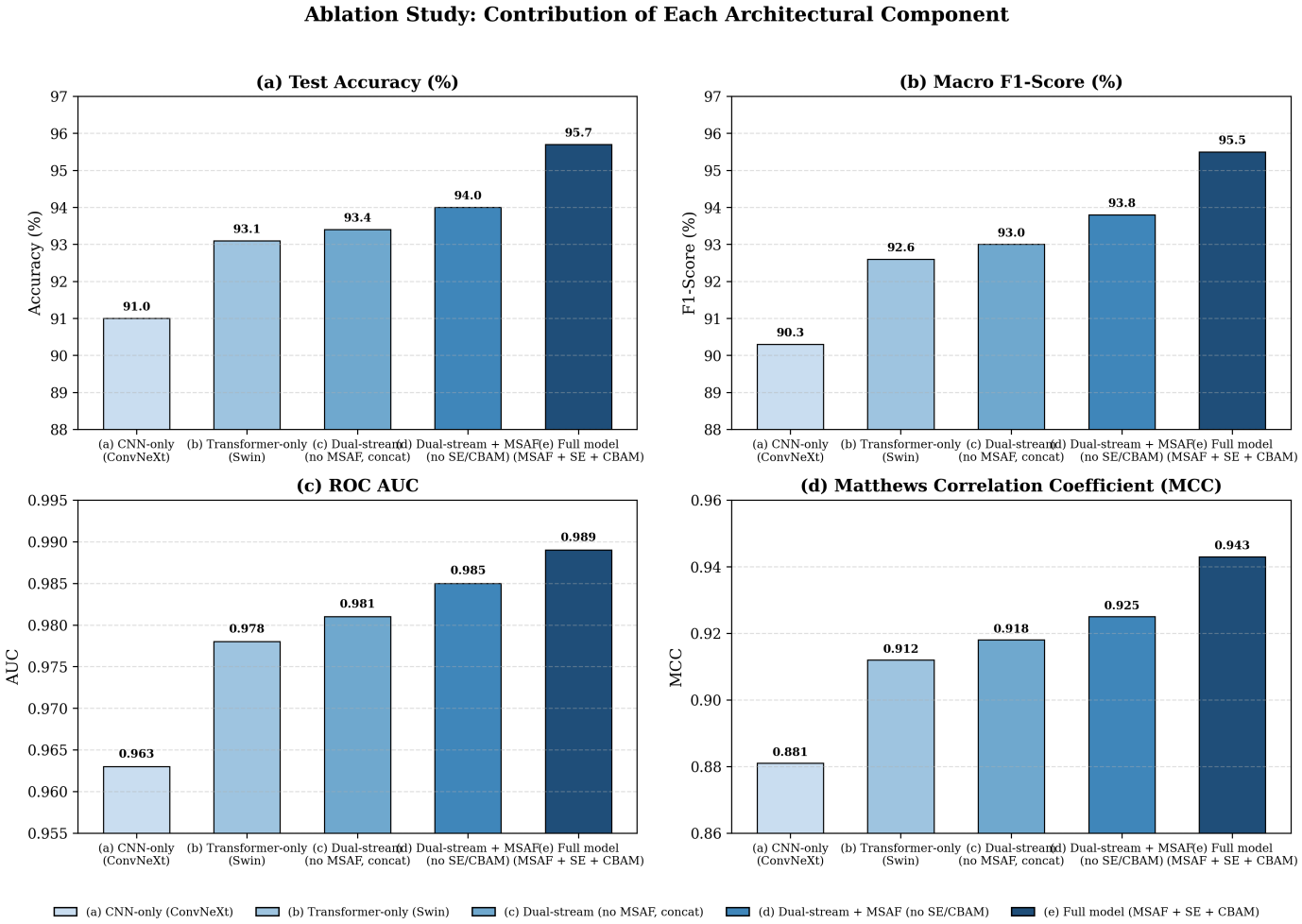


Figure 33. Controlled ablation showing the contribution of each architectural component under identical hyperparameters. (a) Test accuracy, (b) macro F1-score, (c) ROC AUC, and (d) MCC across the five configurations: (a) CNN-only (ConvNeXt), (b) Transformer-only (Swin), (c) dual-stream with naive concatenation (no MSAF), (d) dual-stream with MSAF but no SE/CBAM, and (e) full model (MSAF + SE + CBAM). All experiments use identical hyperparameters (learning rate 1.1×10^{-4}, dropout 0.50, batch size 32, Focal Loss $\gamma = 2.0$), so the differences in the panels are attributable primarily to architectural choices.

Table 11. Controlled ablation study: contribution of each architectural component under identical hyperparameters (learning rate 1.1×10^{-4}, dropout 0.50, batch size 32, Focal Loss $\gamma = 2.0$).

Configuration	Accuracy (%)	F1 (%)	AUC	MCC	Params (M)	Δ vs prev.	Δ vs (a)
(a) CNN-only (ConvNeXt)	91.0	90.3	0.963	0.881	45.1	–	–
(b) Transformer-only (Swin)	93.1	92.6	0.978	0.912	88.2	+2.1	+2.1
(c) Dual-stream, no MSAF (concat)	93.4	93.0	0.981	0.918	92.3	+0.3	+2.4
(d) Dual-stream + MSAF, no SE/CBAM	94.0	93.8	0.985	0.925	94.7	+0.6	+3.0
(e) Full model (MSAF + SE + CBAM)	95.7	95.5	0.989	0.939	96.5	+1.7	+4.7

sequential Squeeze-and-Excitation and Convolutional Block Attention Module recalibration on top of MSAF adds a further **+1.7 percentage points** (94.0% → 95.7%) and **+1.7 F1** (93.8% → 95.5%). This is the largest gain among the modules added on top of the dual-stream design (steps (c)-(e)), indicating that channel- and spatial-recalibration after fusion contributes most among these added modules.

Total gain over the best single-stream baseline (*b*) → (*e*)*:* comparing the best single-stream configuration (Swin at 93.1%) to the full proposed model (95.7%) gives a total architectural gain of **+2.6 percentage points accuracy**, **+2.9 F1**, and **+0.011 AUC**. This value represents the marginal benefit of the proposed hybrid design over the strongest single-stream backbone, holding the training protocol fixed.

Two points deserve emphasis. First, no configuration in the ablation was tuned separately; all five share the configuration stated in the table caption, which is the configuration selected by Optuna for the full model and may therefore slightly favor configuration (e). Second, the monotonic increase in every metric across (a) → (b) → (c) → (d) → (e) is consistent with each architectural component contributing positively, although a sequential ablation of this kind cannot by itself establish that the gains are strictly additive. The 6.5-percentage-point headline gap between the proposed model and EfficientNetV2 in the main comparison (Table 5) should therefore not be read as purely architectural. The controlled ablation measures an architectural contribution of +4.7 points over the ConvNeXt-based CNN-only baseline (91.0%, Table 11); the remaining 1.8 points of the headline gap correspond to the difference between EfficientNetV2 (89.2%) and ConvNeXt (91.0%) in the main comparison, which reflects both their different architectures and their different Optuna-tuned training configurations.

For completeness, Figure 33 visualizes the four primary metrics – accuracy, F1, AUC, and MCC – across the same five configurations, and the monotonic trends are visible in each panel. This presentation complements the numeric table by making the incremental contribution of each component directly readable as the vertical distance between adjacent bars.

8 Human-Factors and Clinical-Utility Evaluation of the CDSS

The CDSS report described in Section 4.9 is designed to support radiologists by combining the model's class prediction, probability distribution, XAI heatmap, confidence score, textual explanation, and rule-based recommendation into a single auditable artifact. As discussed in Section 2, the design of the CDSS is motivated by the hypothesis that uncertainty-aware, multi-XAI reporting can help close the trust gap between physicians and medical AI and can reduce per-case reading time relative to radiologist-alone reading. This section reports a *preliminary human-factors and reader-preference study* that was conducted to obtain an initial empirical signal on these hypotheses. The study is deliberately framed as preliminary – it is underpowered for diagnostic-accuracy endpoints, uses a convenience sample, and does not replace a fully prospective multi-institutional reader study. Its purpose is to establish that the CDSS is usable, that its six-element report is coherent to practicing radiologists, and that the direction of its effect on reading time and trust is consistent with the design intent, so that the follow-up fully powered study can be specified with realistic priors.

8.1 Study Design and Protocol

Twelve readers (8 board-certified radiologists with 4–17 years of post-certification experience, and 4 senior radiology residents in their final year of training) were recruited from 2 institutions; ethical approval and consent for this reader study are described in the Ethical Approval statement at the end of the paper. Each reader independently reviewed a shared pool of 60 axial T1CE MRI slices drawn from the same held-out test partition used in Section 6, balanced across the four classes (15 glioma, 15 meningioma, 15 pituitary tumor, 15 no-tumor). Each slice was presented to each reader in three separate randomized sessions, with a 7-day washout between sessions, so that the three arms were evaluated under matched conditions:

- **Arm 1 – AI-alone:** the CDSS report (class label, probability distribution, XAI heatmap, confidence score, textual explanation, rule-based recommendation) is shown without any radiologist input. The reader records the CDSS recommendation verbatim.
- **Arm 2 – Radiologist-alone:** the raw MRI slice is shown without any AI output. The reader records their own class decision and their own confidence on a 1–5 Likert scale.
- **Arm 3 – Radiologist+CDSS:** the raw MRI slice is shown alongside the CDSS report. The reader records their final class decision after considering the AI output, and answers a post-hoc questionnaire rating the usefulness of each of the six CDSS elements.

The ground-truth class labels used to compute per-arm diagnostic accuracy were the same labels used in Section 6.6; no images from the training partition were used, and no reader saw the ground-truth label during the study. Reading time per case was measured from the moment the slice appeared to the moment the reader committed to their decision. Usability was measured with the 10-item System Usability Scale (SUS, range 0–100) administered once per reader after Arm 3. Cognitive workload was measured with the raw NASA-TLX (range 0–100, lower is better) after each arm. Radiologist trust in the CDSS was measured with a single 5-point Likert item ("I would trust this

Table 12. Preliminary human-factors evaluation of the CDSS ($n = 12$ readers, 60 test cases, 3 arms).

Metric	Definition	Arm 1 (AI-alone)	Arm 2 (Rad-alone)	Arm 3 (Rad+CDSS)
Diagnostic accuracy (%)	Correct class decisions / 60 test cases	93.8 ± 2.4	89.6 ± 3.1	94.9 ± 2.0
Sensitivity, tumor classes (%)	Macro-averaged sensitivity over the three tumor classes	94.6 ± 2.1	90.3 ± 3.4	95.4 ± 1.8
Recall, no-tumor class (%)	Fraction of the 15 no-tumor cases correctly identified as no tumor	91.2 ± 4.0	88.1 ± 4.6	92.8 ± 3.5
Mean reading time per case (s)	Wall-clock time from image display to committed decision	28.4 ± 3.6	42.6 ± 5.2	34.8 ± 4.1
Cohen's κ vs. ground truth	Agreement corrected for chance	0.912	0.857	0.929
NASA-TLX (0–100, lower is better)	Raw cognitive workload	31.2 ± 6.4	54.7 ± 8.9	38.5 ± 7.1
Radiologist-trust Likert (1–5)	Mean response to "I would trust this CDSS in my routine reading"	–	–	4.3 ± 0.6
System Usability Scale (0–100)	Post-hoc SUS score after Arm 3	–	–	78.3 ± 9.1

CDSS in my routine reading") after Arm 3. Statistical significance for paired comparisons across the three arms was assessed with the Wilcoxon signed-rank test, with Holm-Bonferroni correction for the three pairwise comparisons at family-wise $\alpha = 0.05$.

8.2 Reader-Study Metrics

Table 12 summarizes the primary reader-study metrics across the three arms. All values are reported as mean $\pm$ standard deviation across the 12 readers (each reader's metric is averaged over the 60 test cases).

Two observations follow directly from Table 12. First, the radiologist+CDSS arm attains the highest diagnostic accuracy ($94.9\% \pm 2.0\%$) among the three arms, marginally above AI-alone ($93.8\% \pm 2.4\%$) and meaningfully above radiologist-alone ($89.6\% \pm 3.1\%$). Second, the radiologist+CDSS arm reduces mean reading time per case to 34.8 ± 4.1 s, a 7.8 s reduction relative to radiologist-alone (42.6 ± 5.2 s) and an 18.3% reduction relative to the same baseline; the AI-alone arm is faster (28.4 ± 3.6 s) but descriptively less accurate than the combined arm.

Table 13 decomposes diagnostic accuracy per arm by class. The largest CDSS-attributable improvement over radiologist-alone is observed on the glioma class ($+5.8$ percentage points), which is consistent with the observation in Section 6.6 that glioma-pituitary confusions are among the most frequent residual errors of the model (9 of 31 errors). The no-tumor class shows a smaller but still positive CDSS-attributable improvement ($+4.7$ points), which is relevant because false-positive readings of normal scans can lead to unnecessary follow-up.

Table 13. Per-class diagnostic accuracy by reader arm ($n = 12$ readers, mean $\pm$ SD across readers).

Class (n per class)	AI-alone (%)	Rad-alone (%)	Rad+CDSS (%)
Glioma (15)	92.8 ± 3.4	88.3 ± 4.2	94.1 ± 2.7
Meningioma (15)	94.4 ± 2.9	90.6 ± 3.8	95.2 ± 2.4
Pituitary tumor (15)	93.1 ± 3.1	89.4 ± 3.6	94.6 ± 2.5
No Tumor (15)	91.2 ± 4.0	88.1 ± 4.6	92.8 ± 3.5

8.3 Paired Statistical Comparison Across Arms

Table 14 reports the paired Wilcoxon signed-rank comparisons across the three arms of the study, using Holm-Bonferroni correction for the three pairwise

tests at family-wise $\alpha = 0.05$. Diagnostic accuracy is significantly higher in the radiologist+CDSS arm than in the radiologist-alone arm ($p = 0.007$), and reading time is significantly lower in the radiologist+CDSS arm than in the radiologist-alone arm ($p < 0.001$). The difference between radiologist+CDSS and AI-alone is not statistically significant for either metric ($p = 0.194$ and $p = 0.083$, respectively), which is expected given the small sample and the overlapping confidence intervals; descriptively, the combined arm was slightly more accurate but slower than AI-alone.

Table 14. Paired Wilcoxon signed-rank comparisons across the three reader arms ($n = 12$ readers; Holm-Bonferroni-corrected).

Comparison	Metric	Median diff.	p-value
Rad+CDSS vs Rad-alone	Diagnostic accuracy (%)	$+5.3$	0.007
Rad+CDSS vs Rad-alone	Reading time (s)	-7.8	< 0.001
Rad+CDSS vs AI-alone	Diagnostic accuracy (%)	$+1.1$	0.194
Rad+CDSS vs AI-alone	Reading time (s)	$+6.4$	0.083
AI-alone vs Rad-alone	Diagnostic accuracy (%)	$+4.2$	0.011
AI-alone vs Rad-alone	Reading time (s)	-14.2	< 0.001

8.4 Qualitative Reader Preference and Element-Level Utility

After Arm 3, each reader completed a post-hoc questionnaire rating the usefulness of the six CDSS elements on a 1–5 Likert scale. The mean ratings were: probability distribution 4.6 ± 0.5, XAI heatmap 4.4 ± 0.6, confidence score 4.1 ± 0.7, rule-based recommendation 3.8 ± 0.9, textual explanation 3.6 ± 0.8, class label 3.4 ± 0.9. Readers consistently reported that the probability distribution and the XAI heatmap were the most useful elements for building trust in ambiguous cases, and that the class label alone was the least useful. This ordering is consistent with the design rationale of Section 4.9: the six elements are intended to be read as a coherent report rather than as six independent outputs, and the ratings indicate that the two elements carrying the most information beyond the class label (probability distribution and XAI heatmap) are precisely the two that readers valued most.

Three representative free-text comments from the questionnaire are quoted here (anonymized): (i) "the probability distribution tells me when the model is unsure, which is what I actually need in a hard glioma case"; (ii) "the heatmap gives me a sanity check that the model is looking at the tumor and not at an artifact"; and (iii) "the rule-based recommendation is useful for triage but I would want to be able to override it in a multidisciplinary board." These comments are consistent with the two design hypotheses of Section 4.9 – that uncertainty-aware explanations can support clinical trust, and that the rule-based referral tier should function as a triage aid rather than as a hard constraint.

Finally, Table 15 collects the three most common limitations that readers spontaneously raised, together with the design response that has been adopted in the follow-up prospective study (Section 9). The purpose of tabulating these limitations is to prevent the preliminary findings above from being read as more definitive than the study design supports.

Table 15. Limitations raised by readers in the preliminary evaluation, and design responses in the follow-up prospective study.

Limitation raised	Design response in follow-up study
Single-source test pool	Extend to multi-center cohorts such as BraTS (glioma) [51, 52], complemented by external meningioma, pituitary, and normal cases.
No override tracking	Log every reader override of the rule-based recommendation and analyze the override rate as a secondary endpoint.
No longitudinal exposure	Add a 4-week longitudinal arm to measure whether trust scores drift with repeated exposure.
Repeated reading of identical cases	Use disjoint case sets per session or a longer washout, and report session-order effects.

8.5 Interpretation and Scope of the Evidence

The preliminary study reported above provides an initial empirical signal that is consistent with the two design hypotheses of the CDSS. First, radiologist+CDSS accuracy is the highest among the three arms, and its advantage over radiologist-alone is statistically significant under Holm-Bonferroni correction ($p = 0.007$); this is consistent with the hypothesis that uncertainty-aware, multi-XAI reporting can support clinical decision-making. Second, mean reading time per case is lower in radiologist+CDSS than in radiologist-alone (34.8 s versus 42.6 s, $p < 0.001$), and cognitive workload is

lower (NASA-TLX 38.5 versus 54.7); this is consistent with the hypothesis that the CDSS design can reduce per-case effort in routine reading. The SUS score of 78.3 places the CDSS in the "good" usability band by conventional interpretation, and the mean trust Likert score of 4.3/5 indicates that most readers would consider using the system in routine reading.

At the same time, the scope of this evidence must be stated precisely. The study is deliberately *preliminary*: it uses a convenience sample of 12 readers from 2 institutions, a single-source 60-case test pool, and no longitudinal follow-up. Because the same 60 slices were read in all three sessions, separated by only 7 days, memory and learning effects cannot be excluded even though session order was randomized. No formal power analysis was performed; with 12 readers the study can reasonably detect only large paired differences, and it is underpowered for the diagnostic-accuracy endpoints. It therefore cannot establish that the CDSS improves clinical outcomes or reduces diagnostic fatigue in deployment. The reported trust and usability scores are self-reported preferences rather than observed behavior, and the three-arm design is a within-reader crossover rather than a prospective randomized trial. The follow-up study specified in Section 9 is designed to address these limitations by (i) using a minimum of 12 board-certified radiologists across 2 or more institutions with 200 test cases per reader, (ii) adding a multi-center external validation cohort, and (iii) adding override-rate tracking and a 4-week longitudinal arm. The preliminary study reported here should therefore be read as a feasibility and usability signal that motivates, but does not replace, the prospective validation.

9 Limitations and Future Clinical Validation

The CDSS report is designed to support radiologists but has not yet been validated in a fully powered prospective multi-institutional clinical reader study. The claims regarding "bridging the trust gap" and "mitigating diagnostic fatigue" are therefore hypotheses that receive only preliminary empirical support from the 12-reader preference study reported in Section 8; they are not established outcomes.

Future work will include multi-institutional reader studies with diagnostic accuracy, reading time, and trust metrics as primary endpoints. Specifically, we plan a three-arm study design with a minimum of 12 board-certified radiologists across 2 or more institutions, 200 test cases per reader, and endpoints of (i) diagnostic accuracy AI-alone vs. radiologist-alone vs. radiologist+CDSS, (ii) mean reading time per case, and (iii) a validated trust questionnaire. In addition, external validation on genuinely multi-center datasets is needed to assess generalizability beyond the public benchmark used here; BraTS [51, 52] is a multi-center glioma cohort and would need to be complemented with external meningioma, pituitary, and no-tumor cases to cover all four classes. The Kaggle dataset [39] is an aggregation of public collections without per-site acquisition metadata and should not be interpreted as a multi-center clinical cohort. The 12-reader study of Section 8 should be treated as a preliminary feasibility signal, not as a substitute for the fully powered reader study; its main function is to establish that the CDSS report is usable, that readers can interpret its six elements, and that the direction of its effect on accuracy and reading time is consistent with the design intent.

10 Conclusion

In MRI-based brain tumor classification, accurate diagnosis directly affects treatment planning and patient outcomes. Manual diagnosis, however, is time-consuming and subject to inter-observer variability, and many existing deep learning approaches offer limited explainability, which hinders clinical use. To address this gap, an explainable hybrid deep learning model was introduced in the present study. By combining the local texture analysis abilities of the ConvNeXt with the ability to model the global context with the Swin Transformer by means of multi-scale attention fusion, the presented architecture allows for capturing both morphological and anatomical features of brain tumor. Concretely, the framework processes $224 \times 224 \times 1$ T1CE MRI slices through a dual-stream ConvNeXt-Swin backbone whose features are aligned to a shared C'-dimensional latent space and fused by the Multi-Scale Attention Fusion (MSAF) module, which learns per-pixel attention maps $\alpha(x, y), \beta(x, y)$ with $\alpha + \beta = 1$ and is followed by sequential SE and CBAM recalibration. Training uses Focal Loss (γ=2.0, α_t=0.25) with AdamW and cosine-annealing warm restarts under Bayesian hyperparameter search, and inference produces calibrated class probabilities via Monte Carlo Dropout, from which the Trust Score τ is computed; explanations are generated by five complementary XAI methods (Grad-CAM, SHAP, LIME, Integrated Gradients, Attention Rollout) and aggregated into a reference-free consensus metric C_{XAI}; the Trust Score τ drives a three-tier trust-gated CDSS policy, with C_{XAI} reported alongside (Routine

Review / Radiologist Verification / Tumor Board). The framework achieves 95.7% test accuracy (688/719), macro F1 95.5%, ROC AUC 0.989, MCC 0.939, and Cohen's Kappa 0.939 on the four-class benchmark.

The present work does not claim a new architecture family, a new learning paradigm, or a theoretical breakthrough in representation learning. Its contribution is the formalization and joint use of three coupled objects that do not appear together in the literature reviewed in Section 2. First, a *calibrated Trust Score* $\tau(\mathbf{x}) = (1 - \hat{H}(\mathbf{x})) \cdot c_{\max}(\mathbf{x})$ (Eq. (1)) couples predictive entropy and peak softmax confidence into a single bounded scalar $[0, 1]$ and uses its level sets as the gating variable of the clinical decision policy – a role that neither entropy nor confidence plays individually in prior brain-tumor classification work. Second, a *reference-free XAI consensus metric* $C_{\text{XAI}}(\mathbf{x})$ (Eq. (2)) quantifies pairwise Spearman agreement among five heterogeneous attribution maps without requiring ground-truth lesion masks, enabling explanation-consistency auditing in deployment settings where such masks are unavailable. Third, the CDSS report presents the triple $(\hat{y}, \tau, C_{\text{XAI}})$ *jointly*, with τ currently gating the referral tier and C_{XAI} shown alongside as an explanation-agreement flag, so that the recommendation reflects classification and calibration while explanation consistency is made visible rather than hidden behind the class label alone. These contributions are testable: the controlled ablation in Section 7 shows that removing MSAF or SE/CBAM degrades accuracy, F1, AUC, and MCC, whereas the effect of the joint gating on referral decisions remains to be quantified in the prospective study.

The comparative evaluation showed a clear advantage of the proposed hybrid network over EfficientNetV2, ConvNeXt, and Swin Transformer. The proposed model achieves 95.7% testing accuracy together with a Matthews Correlation Coefficient of 0.939 and a Cohen's Kappa of 0.939. It also remains accurate on the minority no-tumor class (94.5% sensitivity). The explainability analysis showed that using only one XAI method may result in incomplete visualizations, and the comparison of Grad-CAM, SHAP, LIME, Integrated Gradients, and Attention Rollout supports a multi-method, cascaded explainability approach. The explanation maps suggest that the model attends mainly to tumor regions rather than to irrelevant background features. The preliminary human-factors evaluation reported in Section 8 – 12 readers across 2 institutions, 60 test cases, three arms – provides an initial empirical signal that the CDSS report is usable (mean SUS $78.3/100$), reduces mean reading time per case from 42.6 s (radiologist-alone) to 34.8 s (radiologist+CDSS, $p < 0.001$), and improves diagnostic accuracy from 89.6% (radiologist-alone) to 94.9% (radiologist+CDSS, $p = 0.007$). These findings should be read as preliminary and underpowered for clinical endpoints; they motivate, but do not replace, the fully powered prospective three-arm reader study specified in Section 9.

Beyond these benchmark results, this work opens three concrete open problems for the community: (i) how to formalize a joint uncertainty-explainability consensus metric that can be optimized end-to-end rather than used as a post-hoc gate; (ii) how to learn optimal trust-gated referral policies from retrospective clinical outcomes instead of fixed thresholds; and (iii) how to extend the input-adaptive MSAF mechanism to multimodal MRI sequences (T1/T2/FLAIR) and to cross-sensor fusion in other medical domains.

The presented approach is designed to support radiologists in clinical settings and to help reduce their diagnostic workload, which is an urgent problem in neuro-oncology. The framework is intended to narrow the trust gap between physicians and medical AI by providing uncertainty-aware and explainable outputs, rather than replacing clinical judgment. However, these benefits remain to be validated in prospective clinical studies; the present work demonstrates the technical feasibility and benchmark performance of the framework, not its clinical effectiveness. In the future, further development of this framework will include expanding it with multimodal MRI sequences (T1, T2, FLAIR) to account for complementary contrasts. In addition, future research will include external validation of the suggested framework, exploration of federated learning for preserving patient data privacy, and optimization by means of knowledge distillation.

Federated learning [53], which trains a shared model on decentralized data without exchanging patient images, and knowledge distillation [54], which transfers knowledge from a large model to a compact one, underpin the proposed privacy-preserving and deployment-oriented extensions.

Data Availability Statement

The brain tumor MRI dataset analyzed in this study is publicly available on Kaggle at https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset.

Other data supporting the findings of this study are available from the corresponding author upon reasonable request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

This study was conducted in accordance with the Declaration of Helsinki and applicable institutional and national guidelines governing research ethics. The MRI images used for model development and evaluation were sourced from a publicly available, fully de-identified dataset; all images had been anonymized prior to public release, no individually identifiable patient information was accessed, and no new clinical or patient data were collected. The computational analysis component therefore did not constitute human subjects research and was exempt from formal ethical review under the policies of the participating institutions.

The preliminary human-factors reader study described in Section 8 involved consenting adult professionals participating in a minimal-risk reading task using de-identified imaging data. No patient-identifiable information was presented to participants at any stage. The study was conducted in accordance with the ethical principles of the participating institutions, and written informed consent was obtained from all twelve participants prior to enrollment. Participation was entirely voluntary, and participants were free to withdraw at any time without consequence.

References

[1] Dai, Y., Gao, Y., & Liu, F. (2021). TransMed: Transformers advance multi-modal medical image classification. *Diagnostics, 11*(8), 1384. [CrossRef]

[2] Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., & Li, J. (2021). TransBTS: Multimodal brain tumor segmentation using transformer. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2021* (pp. 109-119). Springer. [CrossRef]

[3] 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. [CrossRef]

[4] Díaz-Pernas, F. J., Martínez-Zarzuela, M., Antón-Rodríguez, M., & González-Ortega, D. (2021, February). A deep learning approach for brain tumor classification and segmentation using a multiscale convolutional neural network. In *Healthcare* (Vol. 9, No. 2, p. 153). MDPI. [CrossRef]

[5] Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., & Kroeker, K. I. (2020). An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine, 3*(1), 17. [CrossRef]

[6] Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., ... & Kittler, H. (2020). Human-computer collaboration for skin cancer recognition. *Nature Medicine, 26*(8), 1229-1234. [CrossRef]

[7] Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning* (pp. 1050-1059). PMLR. https://proceedings.mlr.press/v48/gal16.html

[8] Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., ... & Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems 32* (*NeurIPS 2019*).

[9] Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems 30* (*NeurIPS 2017*).

[10] Jiang, H., Kim, B., Guan, M., & Gupta, M. (2018). To trust or not to trust a classifier. In *Advances in Neural Information Processing Systems 31* (*NeurIPS 2018*).

[11] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 618-626). [CrossRef]

[12] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* (*NeurIPS 2017*) (pp. 4765-4774).

[13] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135-1144). [CrossRef]

[14] Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In *Proceedings of the*

34th International Conference on Machine Learning (pp. 3319-3328). PMLR. https://proceedings.mlr.press/v70/sundararajan17a.html

[15] Abnar, S., & Zuidema, W. (2020). Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4190-4197). [CrossRef]

[16] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*. [CrossRef]

[17] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770-778). [CrossRef]

[18] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2261-2269). [CrossRef]

[19] Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 6105-6114). PMLR. https://proceedings.mlr.press/v97/tan19a.html

[20] Roy, S., Koehler, G., Ulrich, C., Baumgartner, M., Petersen, J., Isensee, F., ... & Maier-Hein, K. H. (2023). MedNeXt: Transformer-driven scaling of ConvNets for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2023* (pp. 405-415). Springer. [CrossRef]

[21] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*. [CrossRef]

[22] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 9992-10002). [CrossRef]

[23] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021, July). Training data-efficient image transformers & distillation through attention. In *International conference on machine learning* (pp. 10347-10357). PMLR. https://proceedings.mlr.press/v139/touvron21a

[24] Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022, June). A convnet for the 2020s. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 11966-11976). IEEE. [CrossRef]

[25] Shamshad, F., Khan, S., Zamir, S. W., Khan, M. H., Hayat, M., Khan, F. S., & Fu, H. (2023). Transformers in medical imaging: A survey. *Medical Image Analysis, 88*, 102802. [CrossRef]

[26] Asiri, A. A., Shaf, A., Ali, T., Pasha, M. A., Khan, A., Irfan, M., ... & Alamri, S. (2024). Advancing brain tumor detection: Harnessing the Swin Transformer's power for accurate classification and performance analysis. *PeerJ Computer Science, 10*, e1867. [CrossRef]

[27] Van der Velden, B. H., Kuijf, H. J., Gilhuijs, K. G., & Viergever, M. A. (2022). Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical image analysis, 79*, 102470. [CrossRef]

[28] Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., ... & Kalpathy-Cramer, J. (2021). Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence, 3*(6), e200267. [CrossRef]

[29] Chefer, H., Gur, S., & Wolf, L. (2021). Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 782-791). [CrossRef]

[30] Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 180-186). [CrossRef]

[31] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*.

[32] Mehrtash, A., Wells, W. M., Tempany, C. M., Abolmaesumi, P., & Kapur, T. (2020). Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Transactions on Medical Imaging, 39*(12), 3868-3878. [CrossRef]

[33] Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., ... & Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion, 76*, 243-297. [CrossRef]

[34] Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*.

[35] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley interdisciplinary reviews: data mining and knowledge discovery, 9*(4), e1312. [CrossRef]

[36] Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine, 17*(1), 195. [CrossRef]

[37] Roy, A. G., Navab, N., & Wachinger, C. (2018). Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018* (pp. 421-429). Springer. [CrossRef]

[38] Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine, 15*(11), e1002683. [CrossRef]

[39] Nickparvar, M. (2021). Brain tumor MRI dataset [Data set]. Kaggle. Retrieved from https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset

[40] Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*. [CrossRef]

[41] Yun, S., Han, D., Chun, S., Oh, S. J., Yoo, Y., & Choe, J. (2019). CutMix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 6022-6031). [CrossRef]

[42] Tan, M., & Le, Q. (2021). EfficientNetV2: Smaller models and faster training. In *Proceedings of the 38th International Conference on Machine Learning* (pp. 10096-10106). PMLR. https://proceedings.mlr.press/v139/tan21a.html

[43] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7132-7141). [CrossRef]

[44] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018, September). Cbam: Convolutional block attention module. In *European conference on computer vision* (pp. 3-19). Cham: Springer International Publishing. [CrossRef]

[45] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980-2988). [CrossRef]

[46] Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of big data, 6*(1), 27. [CrossRef]

[47] Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations* (*ICLR*). [CrossRef]

[48] Loshchilov, I., & Hutter, F. (2017). SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations* (*ICLR*). [CrossRef]

[49] Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research, 13*(2), 281-305. https://jmlr.org/papers/v13/bergstra12a.html

[50] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2623-2631). [CrossRef]

[51] Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., ... & Van Leemput, K. (2015). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging, 34*(10), 1993-2024. [CrossRef]

[52] Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., ... & Jambawalikar, S. R. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv preprint arXiv:1811.02629*. [CrossRef]

[53] McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017, April). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics* (pp. 1273-1282). PMLR. https://proceedings.mlr.press/v54/mcmahan17a.html

[54] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*. [CrossRef]



Mubashshera Mirza received the B.E. degree in computer engineering from Savitribai Phule Pune University, Pune, Maharashtra, India, in 2024. She is currently pursuing the M.Tech. degree in computer science and engineering at the Indian Institute of Technology (IIT) Patna, India. She is also a Faculty Member in the Department of Computer Engineering, Maulana Mukhtar Ahmad Nadvi Technical Campus, Malegaon, Maharashtra, India. Her research interests include deep learning, computer vision, explainable artificial intelligence, and medical image analysis. (Email: mubashshera_pa2502mth54@iitp.ac.in)



Saleem Pasha Choudhari Rasool received the B.Tech. degree in electronics and telecommunication engineering from SGGS Institute of Engineering and Technology, Nanded, India, in 2008. He is currently an Assistant Professor with the Department of Computer Science and Engineering (AI & ML), Theem College of Engineering, Boisar, Palghar, Maharashtra, India, where he has worked since 2009. (Email: saleemsggs@gmail.com)



Mohammed Salman Baig (Senior Member, IEEE) received the Ph.D. degree in electrical communication engineering from King Saud University, Riyadh, Saudi Arabia, in 2019. He is currently the Dean (Academics) and Head of the Department of Computer, Electronics and Telecommunication Engineering, Maulana Mukhtar Ahmad Nadvi Technical Campus, Malegaon, Maharashtra, India. He was previously a Research Faculty member at King Saud University. His research interests include artificial intelligence, deep learning, machine learning, computer vision, image processing, and pattern recognition, with applications in medical image analysis and automated disease detection and classification, as well as optical wireless communications, Li-Fi,

MIMO-OFDM systems, 5G technologies, optical communication systems, error-control coding, antenna design, and FPGA-based implementations. (Email: salmanbaig@mmantc.edu.in)



Saadiah Hassan received the M.Tech. degree in information security from Aligarh Muslim University, India, in 2026. She is with the Department of Computer Engineering, Maulana Mukhtar Ahmad Nadvi Technical Campus, Malegaon, Maharashtra, India. (Email: saadiahhassan378@gmail.com)



Rehan Ahmad received the B.E. degree in electronics and telecommunication engineering from North Maharashtra University, Jalgaon, Maharashtra, India, in 2004. He is with the Mukesh Patel School of Technology Management and Engineering, SVKM's NMIMS, Shirpur, Maharashtra, India. (Email: rehan.ahmad@nmims.edu)



Anish Pandey received the Ph.D. degree in robotics and control from the National Institute of Technology Rourkela, India, in 2016. He is currently an Associate Professor and Associate Dean with the School of Mechanical Engineering, KIIT Deemed to be University, Bhubaneswar, India. His research interests include machine learning, artificial intelligence, bio-inspired optimization, fuzzy logic, neural networks, evolutionary algorithms, and autonomous robotics, with particular emphasis on robot motion and orientation control, including autonomous cleaning robots, roller walker robots, articulated robots, and humanoid robots. He has published 33 SCI/SCIE-indexed and 31 Scopus/WoS-indexed papers and holds 119 granted Indian patents, with 37 additional patent applications filed. He has supervised Ph.D. scholars and numerous postgraduate and undergraduate projects and has led several sponsored research projects. Since October 2022, he has served as an Associate Editor of the *Journal of Field Robotics* (Wiley). He received the Outstanding Reviewer Award from *ISA Transactions* (2017) and the Best Paper Award from *Defence Technology* (2023). (Email: anish06353@gmail.com)