Journal of Medical and Health Studies
ISSN: 2710-1452
DOI: 10.32996/jmhs
Journal Homepage: www.al-kindipublisher.com/index.php/jmhs

JMHS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Multi-Scale Attention-Fusion Classification with an LLM-Driven Clinical Recommendation Assistant for Brain Tumor MRI

Kallol Chakraborty Shekhor[1], Md Sahid Hossain[2], Md Abedur Rahman[1], and Md Anwar Hossain[1]

[1] *Master's in Computer Science, Maharishi International University, Fairfield, IA, U.S.A.*
[2] *Senior Software Engineer, Prime Tech Solutions Ltd., Dhaka, Bangladesh*

| ABSTRACT

Brain tumor MRI interpretation remains constrained by radiologist shortages, inter-reader variability, delayed specialist access, and high diagnostic cost, particularly in low-resource settings. Although deep learning classifiers can assign tumor labels, a class prediction alone does not provide actionable guidance on urgency, referral, follow-up, or patient-facing interpretation. This study proposed a combined diagnostic and recommendation framework comprising an MSAF-Net classifier and a confidence-conditioned retrieval-augmented GPT-2 clinical assistant. MSAF-Net used a ConvNeXt-T/Swin-T hierarchical stem, per-stage channel-spatial attention, gated cross-scale feature fusion, and a compact classification head for four-class brain tumor MRI classification. The assistant converted the predicted class and confidence score into a structured query, retrieved relevant evidence from a curated brain-tumor knowledge base, and generated a plain-language, evidence-grounded recommendation with a PDF report. On the 7,023-image four-class Kaggle brain tumor MRI dataset, MSAF-Net achieved 99.16% accuracy, 0.9897 macro-F1, and 0.9985 AUC. Cross-dataset evaluation showed 99.46% accuracy on Figshare CE-MRI and 99.74% accuracy on BR35H. The model remained computationally efficient, with 22.9M parameters, 3.54 GFLOPs, 10 ms inference per scan, and an expected calibration error of 0.0179. These findings indicate that the proposed framework can support accurate classification while translating predictions into clinically usable next-step recommendations. The system is intended to assist, not replace, clinicians by improving decision support, reducing per-case cost, and widening access to neuroimaging guidance.

| KEYWORDS

Brain tumor MRI; MSAF-Net; attention fusion; GPT-2; retrieval-augmented generation; clinical decision support

| ARTICLE INFORMATION

ACCEPTED: 20 February 2023 **PUBLISHED:** 11 March 2023 **DOI:** 10.32996/jmhs.2023.4.2.1

1. Introduction

Brain tumors represent a clinically heterogeneous group of intracranial lesions with substantial consequences for survival, neurological function, treatment planning, and health-system workload. Gliomas are often infiltrative and biologically aggressive, requiring timely diagnosis to guide neurosurgical, oncological, and radiotherapeutic decisions. Meningiomas are commonly extra-axial and may be benign, but their location, growth pattern, and mass effect can still produce serious neurological impairment. Pituitary tumors may affect vision, endocrine regulation, and quality of life, and accurate recognition is necessary for appropriate referral and follow-up. Magnetic resonance imaging (MRI) remains central to the diagnostic pathway because it provides high soft-tissue contrast and permits non-invasive evaluation of tumor morphology, location, edema, and surrounding tissue involvement [1-2]. However, MRI interpretation depends heavily on specialist expertise. In many clinical settings, especially rural and low-resource regions, access to experienced neuroradiologists is limited, reporting delays are common, and imaging costs place additional pressure on patients and health systems [3-4]. Inter-reader variability further affects diagnostic consistency, particularly when lesions show overlapping radiological appearances or when scans are interpreted under high workload conditions [5-6]. Therefore, accurate and efficient brain tumor MRI classification is not only a technical image-

recognition task; it is a practical clinical need linked to earlier referral, reduced diagnostic delay, improved workflow efficiency, and more equitable access to specialist-level decision support.

Deep learning has become a major direction for brain tumor MRI analysis because it can learn discriminative imaging patterns directly from data. Early convolutional neural network (CNN)-based approaches used architectures such as ResNet, DenseNet, EfficientNet, and related variants to extract hierarchical visual features for tumor classification and detection [7-8]. These methods improved performance over handcrafted feature pipelines but were often limited by local receptive fields, sensitivity to acquisition variability, and difficulty modeling long-range spatial relationships. More recent transformer-based and hybrid CNN-transformer models, including Vision Transformer, Swin Transformer, ConvNeXt, CoAtNet, and EfficientNetV2-style backbones, have improved global contextual modeling and multi-scale representation learning [9]. These architectures have shown promise in distinguishing glioma, meningioma, pituitary tumor, and no-tumor MRI classes under controlled experimental protocols [10]. Figure 1 summarizes this baseline landscape, showing the progression from input MRI preprocessing and augmentation to current classification backbones, proposed multi-scale fusion, evaluation metrics, and heatmap-based tumor localization. Despite these advances, many studies still focus primarily on predictive accuracy and provide limited integration between classification output, uncertainty, clinical interpretation, and actionable next-step recommendation.

A persistent limitation of existing brain tumor classification systems is the decision gap. Most models output a class label or probability score, but a label alone does not explain what the prediction means for a clinician, patient, or non-specialist health worker. For example, a predicted glioma label may require urgent specialist referral, additional contrast-enhanced imaging, neurosurgical consultation, or structured follow-up depending on confidence and clinical context. Similarly, a no-tumor prediction should not be communicated as definitive clinical exclusion without consideration of image quality, symptoms, and confidence level. Label-only artificial intelligence therefore provides limited practical value in settings where users need guidance on urgency, referral, risk communication, and next-step action. This limitation is particularly important in low-access environments, where automated triage and decision-support tools may be used by general radiologists, primary-care clinicians, or health workers without immediate neuroradiology support. In parallel, the cost and access gap remains unresolved. High-compute models may perform well in experimental settings but are difficult to deploy where hardware, bandwidth, and specialist availability are constrained. A clinically useful system should therefore combine accurate classification, calibrated confidence, low computational cost, and an interpretable recommendation pathway that supports, rather than replaces, professional judgment.

To address these gaps, this study proposed a unified framework titled "Multi-Scale Attention-Fusion Classification with an LLM-Driven Clinical Recommendation Assistant for Brain Tumor MRI." The framework combined an MSAF-Net classifier with a confidence-conditioned retrieval-augmented GPT-2 clinical assistant. MSAF-Net was designed as a lightweight multi-scale attention-fusion classifier using a ConvNeXt-T/Swin-T hierarchical stem, four-stage feature extraction, per-stage channel-spatial attention through CBAM, and gated cross-scale fusion of feature maps from multiple semantic depths. This design was intended to preserve fine local tumor cues while integrating broader contextual features required for distinguishing visually similar brain tumor categories. The classification output was then passed to a recommendation layer in which the predicted class and confidence score were converted into a structured query. An embedding model performed semantic search over a curated brain tumor knowledge base and metadata, and the retrieved evidence was used by a fine-tuned GPT-2 model to generate a plain-language clinical suggestion and a PDF report. The assistant was designed as a decision-support component, not as an autonomous diagnostic agent. Its role was to translate model output into clinically usable guidance while preserving the need for physician review, imaging correlation, and appropriate specialist referral.The main contributions of this study are as follows:

1. MSAF-Net attention-fusion classifier: A lightweight brain tumor MRI classifier was developed using a ConvNeXt-T/Swin-T hierarchical stem, per-stage CBAM attention, and gated cross-scale feature fusion for four-class tumor classification.
2. Confidence-conditioned RAG GPT-2 clinical recommendation assistant: A retrieval-augmented, fine-tuned GPT-2 module was integrated to convert predicted class and confidence into plain-language, evidence-grounded clinical recommendations and PDF reports.
3. Lightweight low-cost deployable pipeline: The proposed framework was designed for efficient inference and practical deployment, supporting reduced per-case cost, faster reporting assistance, and wider access in resource-constrained settings.
4. Extensive validation across MRI datasets: The system was evaluated across three MRI datasets with baseline comparisons, module ablations, loss and backbone analyses, robustness testing, statistical assessment, and clinically relevant performance measures.

2. Related Work

Tiwari et al. 2020 [11] reviewed selected brain tumor segmentation and classification methods published between 2014 and 2019, covering conventional, deep learning, metaheuristic, and hybrid approaches for MRI-based analysis. Their work provides a broad methodological overview, although it remains a survey rather than an experimentally validated diagnostic framework. Khan et al. 2020 [12] proposed a CNN-based approach with data augmentation and image processing for classifying brain MRI scans into cancerous and non-cancerous categories. Their scratch CNN achieved 100% accuracy and outperformed VGG-16, ResNet-50, and Inception-v3, but the evaluation was conducted on a very small dataset, limiting the strength of generalization claims. Amin et al. 2020 [13] developed an automated MRI-based brain tumor detection framework using lesion segmentation followed by shape, texture, and intensity feature extraction with SVM classification. Validated on Harvard, RIDER, and local datasets, the method achieved 97.1% accuracy and 0.98 AUC, showing strong classical machine-learning performance, although it depended on handcrafted feature design. Aamir et al. 2022 [14] introduced a deep learning pipeline for brain tumor classification using MRI preprocessing, feature extraction from two pre-trained models, PLS-based hybrid feature fusion, agglomerative clustering for tumor proposal localization, and final head-network classification. The method achieved 98.95% accuracy, but its multi-stage design increases architectural complexity compared with more direct end-to-end models. Guan et al. 2021 [15] proposed an efficient CAD framework for MRI-based brain tumor classification using preprocessing, data augmentation, agglomerative clustering-based tumor proposals, backbone feature extraction, proposal refinement, and head-network classification. The model achieved 98.04% overall accuracy with strong class-wise performance for meningioma, glioma, and pituitary tumors, demonstrating the value of proposal-guided classification while requiring several sequential processing stages.

Toufiq et al. 2021 [16] reviewed machine-learning and deep-learning methods for MRI-based brain tumor classification, highlighting the roles of SVM, KNN, ANN, and CNN-based models. The review is useful for positioning automated classification methods, although it does not provide a unified experimental benchmark for direct performance comparison. Rao et al. 2021 [17] presented a comprehensive review of MRI-based brain tumor segmentation and classification, covering preprocessing steps such as image registration, bias-field correction, and non-brain tissue removal. Their work emphasizes recent deep-learning trends and clinical workflow relevance, but remains a survey-level contribution without a new validated model. Rasheed et al. 2023 [18] proposed a CNN-based brain tumor classification method supported by image enhancement techniques, including Gaussian-blur-based sharpening and CLAHE. The method classified glioma, meningioma, pituitary tumor, and no-tumor cases with 97.84% accuracy and strong precision, recall, and F1-score, showing the benefit of enhancement-based preprocessing for MRI classification. Badža et al. 2020 [19] developed a relatively simple CNN architecture for classifying three brain tumor types from T1-weighted contrast-enhanced MRI images. The model achieved 96.56% accuracy under record-wise cross-validation with augmented data and also considered subject-wise validation, although its scope was limited to three tumor categories and T1-contrast MRI.

Jia et al. 2020 [20] proposed a deep-learning-based MRI brain tumor identification framework using FAHS-SVM segmentation with structural, morphological, and relaxometry information, followed by neural-network-based classification. The system reported 98.51% accuracy for detecting abnormal and normal tissue, but its emphasis was mainly on binary abnormality detection rather than detailed tumor subtype classification.Vankdothu et al. 2022 [21] proposed an automated MRI brain tumor classification pipeline combining adaptive filtering, improved K-means clustering, GLCM-based feature extraction, and recurrent convolutional neural network classification. Evaluated on a Kaggle dataset with glioma, meningioma, no-tumor, and pituitary classes, the method achieved 95.17% accuracy, although its performance depends on multiple handcrafted preprocessing and segmentation stages.Kuraparthi et al. 2021 [22] investigated transfer-learning-based brain tumor classification using AlexNet, VGG16, and ResNet50 for feature extraction, followed by SVM classification. Using Kaggle and BraTS datasets with data augmentation, the best ResNet50-based configuration achieved up to 99.0% and 98.86% accuracy, showing strong performance but relying on separated feature extraction and classifier stages rather than a fully end-to-end design.Ayadi et al. 2021 [23] developed a deep CNN model for MRI-based brain tumor classification and evaluated it across three datasets. The study demonstrated competitive classification performance against existing methods, highlighting the effectiveness of CNN-based CAD systems, although the provided description does not specify detailed class-wise metrics or external validation behavior.Irmak 2021 [24] proposed three optimized CNN models for brain tumor detection, five-class tumor classification, and tumor-grade classification using grid-search-based hyperparameter tuning. The models achieved 99.33% detection accuracy, 92.66% five-class accuracy, and 98.14% grade-classification accuracy, supporting automated multi-level diagnosis while still requiring task-specific CNN designs.Özkaraca et al. 2023 [25] introduced a modular dense CNN architecture for multiple brain tumor classification from Kaggle MRI images, aiming to combine advantages of DenseNet, VGG16, and basic CNN structures. The model improved classification performance under both 80/20 splitting and 10-fold cross-validation settings, but this gain was accompanied by increased processing time.

Saleh et al. 2020 [26] evaluated five pre-trained deep-learning models, namely Xception, ResNet50, InceptionV3, VGG16, and MobileNet, for MRI-based brain tumor classification. Xception achieved the highest F1-score of 98.75%, followed by ResNet50 at 98.50%, showing the value of transfer learning, although the work mainly focused on model comparison rather than architectural innovation. Aamir et al. 2023 [27] proposed a multi-stage brain tumor classification framework using image enhancement, morphological analysis, segmentation, clustering-based high-quality tumor region selection, deep feature extraction, adaptive fusion, and multi-class SVM classification. The method achieved approximately 98.98% accuracy on a public dataset, but its diagnostic pipeline remains relatively complex due to several sequential processing stages.Ayadi et al. 2022 [28] introduced a hybrid MRI brain tumor classification approach based on normalization, dense speeded-up robust features, histogram of gradients, and SVM classification. The method achieved 90.27% accuracy under k-fold cross-validation, indicating the robustness of handcrafted feature-based learning, although its performance was lower than many recent deep-learning-based approaches.Srinivas et al. 2022 [29] conducted a comparative analysis of transfer-learning models, including VGG-16, ResNet-50, and Inception-v3, for MRI-based brain tumor classification. Their results indicated that VGG-16 provided the most effective training and validation accuracy, but the dataset contained only 233 images, limiting the strength of broad generalization.Rahman and Islam 2023 [30] proposed a parallel deep convolutional neural network designed to capture both local and global MRI features using two simultaneous CNN pathways with different window sizes. Evaluated on three datasets, the model achieved 97.33%, 97.60%, and 98.12% accuracy, showing strong multiclass and binary classification performance while still relying on augmentation and preprocessing to reduce overfitting.

3.1 Dataset Description

This study used three publicly available brain MRI datasets to evaluate the proposed classification and clinical recommendation framework across four-class tumor classification, three-class tumor classification, and binary tumor detection tasks. The primary dataset was the Kaggle Brain Tumor MRI dataset, which contained 7,023 MRI images distributed across four classes: glioma, meningioma, pituitary tumor, and no tumor. This dataset included MRI images from different sequences, including T1c, T1, T2, and FLAIR. It was used as the main benchmark because it represented the target diagnostic setting of distinguishing clinically relevant tumor categories from non-tumor MRI cases. The dataset was divided into 5,712 training images and 1,311 testing images under a stratified and patient-disjoint setting to reduce data leakage and preserve class distribution across the experimental partitions.The second dataset was the Figshare contrast-enhanced MRI dataset, which contained 3,064 T1-weighted contrast-enhanced MRI slices collected from 233 patients. This dataset included three tumor classes and was used to assess the generalization ability of the proposed model beyond the primary Kaggle cohort. Because patient information was available, evaluation was performed using a five-fold cross-validation protocol at the patient level. This ensured that slices from the same patient were not shared between training and testing folds, thereby reducing the risk of overly optimistic performance caused by patient-level overlap.The third dataset was the BR35H brain MRI dataset, which contained 3,000 MRI images for binary tumor detection. The task involved distinguishing tumor from no-tumor MRI images. The dataset was partitioned into 2,400 training images and 600 testing images. It was included to evaluate whether the learned representation could remain effective when transferred from multi-class tumor classification to a simpler binary detection setting. All images from the three datasets were resized to 224 × 224 pixels before model input. The use of these three datasets enabled evaluation of the proposed framework under primary four-class classification, external three-class tumor classification, and binary tumor detection conditions, as summarized in Table 1.

Table 1. Dataset characteristics and experimental partitions across brain MRI cohorts

Dataset	Modality	Task	Images	Classes	Train / Val / Test	Resolution
Kaggle BT-MRI	MRI (T1c/T1/T2/FLAIR)	4-class	7,023	4	5,712 / — / 1,311	varied → 224×224
Figshare CE-MRI	MRI (T1-CE)	3-class	3,064	3	5-fold CV (233 patients)	512×512 → 224×224
BR35H	MRI	binary detect	3,000	2	2,400 / — / 600	512×512 → 224×224

3.2 Dataset Preprocessing and Augmentation

All MRI images were processed through a unified preprocessing pipeline before training, validation, and testing. Because the three datasets contained images with different original resolutions and acquisition characteristics, each image was first resized to a fixed spatial resolution of 224 × 224 pixels. This standardization ensured compatibility with the ConvNeXt-T and Swin-T backbone components used in the proposed MSAF-Net architecture and allowed all baseline models to be evaluated under the same input configuration. The resizing step was applied consistently across the Kaggle BT-MRI, Figshare CE-MRI, and BR35H datasets.After resizing, intensity normalization was applied to reduce variation caused by differences in MRI acquisition protocols, scanner settings, and image contrast. Pixel intensity values were normalized at the image level so that the network received inputs with a stable numerical distribution. This step was important because the datasets included images from different sources and MRI sequences, and uncontrolled intensity variation could lead the model to learn dataset-specific artifacts rather than tumor-relevant visual patterns. Basic noise suppression was also applied to reduce small high-frequency artifacts while preserving tumor boundaries, internal texture, and anatomical structures relevant to classification.Training data were further augmented to improve model robustness and reduce overfitting. The augmentation strategy included random flipping, random rotation within ±15°, zoom transformation, spatial shifting, and elastic deformation. Random rotation and shifting helped the model tolerate minor differences in head positioning and slice alignment. Zoom augmentation exposed the model to variation in tumor scale and field of view. Elastic deformation introduced mild spatial variability while maintaining the overall anatomical and tumor structure. These transformations were applied only to the training images and were not used during testing. Validation and testing images underwent only deterministic preprocessing, including resizing, normalization, and noise suppression, to ensure that evaluation reflected the true generalization ability of the model.The same preprocessing and augmentation protocol was applied to the proposed MSAF-Net and all baseline models. This unified procedure ensured a fair comparison by preventing performance differences from being caused by inconsistent image preparation. The preprocessing pipeline also supported deployment-oriented use because it required only lightweight operations before inference and did not depend on computationally expensive segmentation, manual annotation, or specialist-guided image correction.

4.1 Methodology Overview

The proposed methodology consisted of three connected components: baseline model evaluation, development of the MSAF-Net classifier, and integration of a confidence-conditioned GPT-2 clinical recommendation assistant. The overall aim was not only to classify brain MRI images into tumor categories, but also to convert the model output into an actionable decision-support recommendation that could assist clinicians and patients in understanding the likely next step.

First, a unified baseline evaluation pipeline was established to compare the proposed model with widely used deep learning backbones for medical image classification. The baseline models included convolutional, transformer-based, and hybrid architectures such as ResNet-50, DenseNet-121, EfficientNetV2-S, ViT-B/16, Swin-T, ConvNeXt-T, and CoAtNet-0. All baseline models were trained and evaluated under the same preprocessing, augmentation, input size, optimizer, and dataset partitioning protocol. This ensured that performance differences were caused by the model architecture rather than inconsistent experimental settings. The baseline workflow is summarized in Figure 2.Second, the proposed MSAF-Net classifier was developed to improve brain tumor MRI classification through multi-scale attention-guided feature learning. The model used a ConvNeXt-T/Swin-T hierarchical stem to extract feature representations at four different stages. Each stage produced feature maps that captured different levels of visual information, ranging from low-level tumor texture and boundary patterns to high-level semantic features. A channel-spatial attention module was applied at each stage to emphasize tumor-relevant regions and suppress less informative background responses. The attention-refined features were then combined through gated cross-scale feature fusion, allowing the model to adaptively weight information from different feature depths. The fused representation was passed through a fully connected layer, dropout regularization, and a softmax classifier to produce the final class prediction and confidence score.Third, the best-performing classifier output was connected to a retrieval-augmented GPT-2 clinical assistant. After classification, the predicted tumor class and confidence score were passed to a query formulation engine. The formulated query was encoded by an embedding model and used to retrieve relevant information from a curated brain tumor knowledge base with associated metadata. The retrieved evidence, predicted class, and confidence score were then provided to a fine-tuned GPT-2 model, which generated a plain-language clinical recommendation and a structured PDF report. The assistant was designed to support clinical interpretation by explaining the prediction, suggesting an appropriate next step, and indicating when specialist review or follow-up was needed. This module was not intended to replace clinicians or provide autonomous diagnosis; rather, it functioned as a decision-support layer that translated model predictions into usable clinical guidance. The complete classifier-to-recommendation workflow is shown in Figure 3.

4.2 Proposed MSAF-Net Architecture

The proposed MSAF-Net was designed as a multi-scale attention-fusion classifier for brain tumor MRI classification. The architecture was developed to capture both local tumor-specific details and deeper semantic representations that are necessary for differentiating glioma, meningioma, pituitary tumor, and no-tumor MRI images. As illustrated in Figure 1, the model followed a hierarchical feature-learning strategy consisting of an input image layer, a ConvNeXt/Swin stem, four progressive feature-learning stages, per-stage channel-spatial attention modules, multi-scale feature fusion, and a final classification head.The input to the model was a preprocessed brain MRI image resized to 224 × 224 pixels. The image was first passed through a ConvNeXt/Swin stem, which performed patch embedding and initial downsampling. This stem converted the MRI image into a structured feature representation while reducing spatial resolution and increasing channel depth. The ConvNeXt component supported strong convolutional inductive bias for local texture, boundary, and lesion-pattern extraction, whereas the Swin Transformer component supported hierarchical contextual modeling through window-based self-attention. This combined stem allowed the network to learn both local anatomical details and broader spatial dependencies from the beginning of the architecture. After the stem, the model extracted features through four hierarchical feature blocks. Stage 1 learned shallow features, including edge information, tumor boundary cues, local intensity changes, and low-level anatomical patterns. These features were important because brain tumor MRI classes may differ in lesion margin, contrast pattern, and local texture. Stage 2 learned intermediate structural features by combining local texture with early shape information. This stage helped the model identify more organized tumor patterns and distinguish abnormal tissue from surrounding brain structures. Stage 3 extracted deeper semantic features related to tumor morphology, internal heterogeneity, and class-specific appearance. At this level, the model captured more abstract representations that supported separation between visually similar tumor categories. Stage 4 produced the deepest feature representation, encoding high-level contextual information and global anatomical relationships. This stage was particularly useful for recognizing broader lesion location, mass effect, and overall tumor appearance.

A channel-spatial attention module was attached to each of the four feature stages. The channel attention component emphasized the most informative feature channels associated with tumor appearance and suppressed less useful background responses. This was important because not all learned channels contributed equally to tumor classification. The spatial attention component then highlighted image regions that were more relevant to the tumor or discriminative anatomical area. By applying attention at every stage, the model refined shallow, intermediate, and deep features separately rather than relying only on a single final attention layer. This design helped preserve fine tumor details from early layers while also improving the discriminative strength of deeper semantic features.

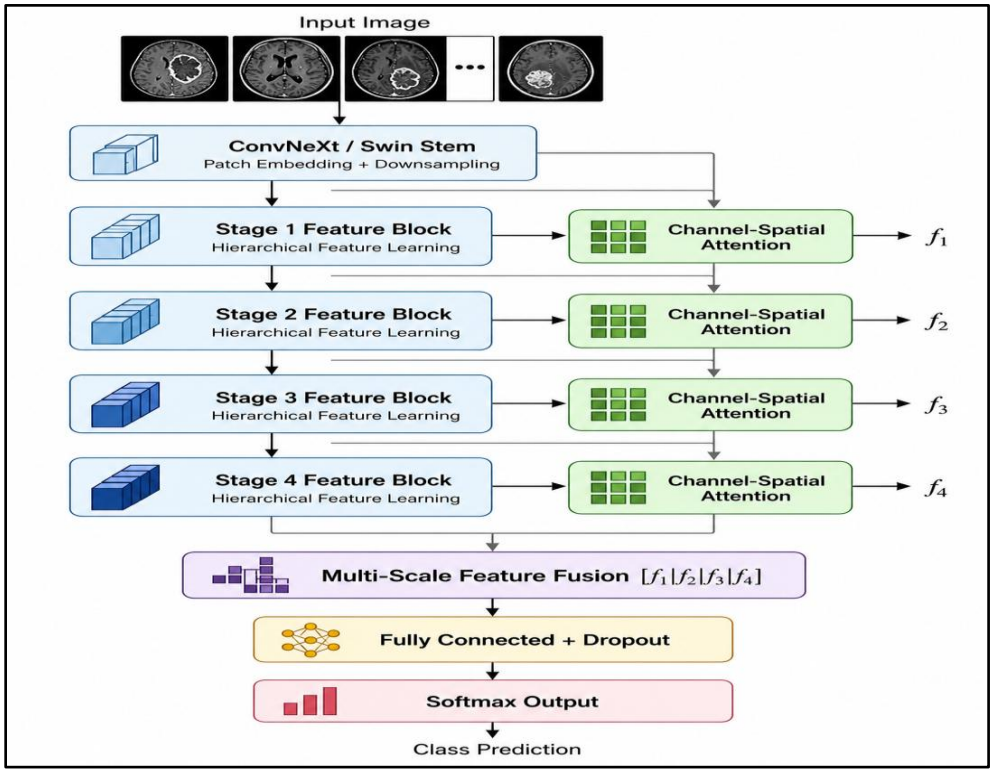

Figure 1. Architecture of the proposed MSAF-Net for multi-scale attention-fusion brain tumor MRI classification

The attention-refined outputs from the four stages were denoted as (f_1), (f_2), (f_3), and (f_4). These features represented different depths of information within the network. Instead of using only the final deep feature map, MSAF-Net fused all four stage outputs through a multi-scale feature fusion block. This fusion module integrated the complementary information from shallow boundary features, intermediate shape patterns, deeper tumor morphology, and high-level semantic context. The fusion mechanism allowed the classifier to use a richer representation than a single backbone output. This was particularly relevant for brain tumor MRI classification because some classes may be distinguished by subtle local features, whereas others require broader contextual interpretation.After multi-scale fusion, the combined feature representation was passed to a fully connected classification head. The head included a fully connected layer followed by dropout regularization to reduce overfitting. Dropout was important because medical imaging datasets are often limited in size and may contain source-specific image characteristics. The final softmax layer produced class probabilities for the four target categories. The class with the highest probability was selected as the predicted brain tumor class, and the corresponding probability was used as the confidence score. This confidence score was subsequently passed to the GPT-2-based recommendation assistant for confidence-conditioned clinical suggestion generation.

4.3 Baseline Model Framework

The baseline model framework was designed to provide a fair and structured comparison between the proposed MSAF-Net and widely used deep learning architectures for brain tumor MRI classification. As shown in Figure 2, the baseline pipeline began with the input brain MRI image, followed by standardized preprocessing and augmentation, model training, evaluation, and tumor-localization visualization. This design ensured that all baseline architectures were assessed under the same experimental conditions rather than compared using independently reported results from different studies.The input to the baseline pipeline was a preprocessed brain MRI image resized to 224 × 224 pixels. Before model training, each image passed through the same preprocessing operations used for the proposed model, including intensity normalization, resizing, augmentation, and noise reduction. This step was important because baseline comparisons in medical imaging can be biased when models are trained with different image sizes, augmentation strategies, or normalization schemes. Therefore, the same preprocessing pipeline was applied consistently to all baseline models and to the proposed MSAF-Net.The baseline set included recent convolutional, transformer-based, and hybrid architectures. ConvNeXt was included as a modern convolutional architecture that updates traditional CNN design using large-kernel depthwise convolution, improved normalization, and hierarchical feature extraction. In the brain MRI setting, ConvNeXt provided a strong reference for learning local tumor texture, lesion boundary, and region-level intensity patterns. Its staged design allowed the model to progressively reduce spatial resolution while increasing feature depth, which made it suitable for identifying tumor morphology at multiple levels of abstraction.

The Swin Transformer baseline was included to represent hierarchical vision transformer modeling. Unlike standard Vision Transformer architectures that process global image patches directly, Swin Transformer uses shifted-window self-attention to model local and regional dependencies efficiently. This property is useful in MRI classification because tumor appearance may depend on both local lesion characteristics and broader anatomical context. Through its hierarchical stages, Swin Transformer can learn progressively deeper representations while maintaining computational efficiency compared with full global self-attention.The Vision Transformer baseline was used as a pure transformer reference. In this architecture, the input MRI image is divided into fixed-size patches, and each patch is converted into a token representation. These tokens are processed through self-attention layers to learn relationships between different image regions. This mechanism allows the model to capture long-range dependencies across the MRI slice. However, compared with hierarchical CNN or Swin-style models, a standard Vision Transformer may require stronger data regularization and larger training data to learn robust local anatomical features. Its inclusion therefore provided an important comparison between pure transformer learning and hybrid hierarchical feature extraction.

CoAtNet was included as a hybrid convolution-attention baseline. This architecture combines convolutional layers for local feature extraction with attention-based layers for global contextual modeling. The convolutional components are useful for capturing low-level tumor patterns, while the attention components support broader region-level interpretation. In brain tumor MRI classification, such hybrid models are relevant because the task requires both fine-grained lesion discrimination and global anatomical understanding. CoAtNet therefore served as a strong reference for assessing whether the proposed attention-fusion strategy provided additional benefit beyond existing hybrid designs.EfficientNetV2 was included as an efficient convolutional baseline. This model family uses compound scaling and optimized convolutional blocks to achieve strong accuracy with relatively low computational cost. EfficientNetV2 is particularly relevant for deployment-oriented medical imaging because it balances model size, inference efficiency, and classification performance. In this study, it provided a reference for evaluating whether the proposed MSAF-Net could maintain high accuracy while remaining computationally practical.In addition to these latest architectures, conventional CNN baselines such as ResNet-50 and DenseNet-121 were also used as reference models. ResNet-50 introduced residual connections to support deeper feature learning and reduce gradient degradation during training. DenseNet-121 used dense connectivity to improve feature reuse across layers and strengthen gradient flow. These models remain common

baselines in medical image classification and were included to compare the proposed method against established CNN architectures.

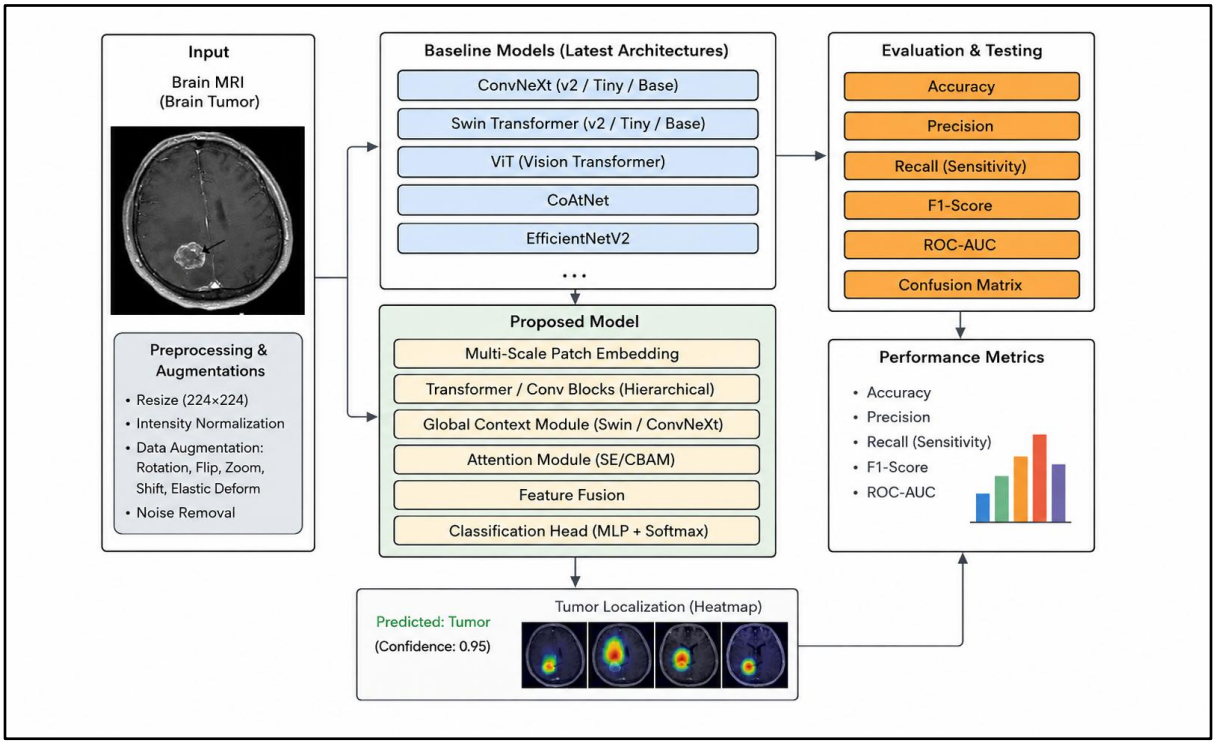

Figure 2. Baseline and proposed brain tumor MRI classification workflow with preprocessing, model comparison, evaluation metrics, and heatmap-based localization

4.4 GPT-2-Based Brain Tumor Classification and Recommendation Framework

Figure 3 presents the complete workflow of the proposed brain tumor MRI classification and GPT-2-based recommendation framework. The framework was designed as an integrated decision-support pipeline that combined image-level tumor classification with confidence-conditioned clinical recommendation generation. The workflow consisted of two connected stages: first, a deep learning-based brain tumor classification module identified the predicted tumor class; second, a retrieval-augmented GPT-2 assistant transformed the prediction and confidence score into a plain-language clinical recommendation and PDF report.

The pipeline began with the brain tumor MRI dataset, which contained images from the target diagnostic categories. Each image was resized to a uniform input dimension and then partitioned into training, validation, and testing subsets according to the experimental protocol. The training data were used for model optimization, the validation data supported model selection and hyperparameter monitoring, and the testing data were reserved for final performance evaluation. Before model training, all images passed through a standardized preprocessing stage that included augmentation, normalization, and class-balancing procedures where required. Augmentation improved robustness to minor anatomical and acquisition-related variations, normalization reduced intensity differences across MRI sources, and class balancing reduced the risk of biased learning toward more frequent categories.After preprocessing, the images were passed to a set of latest classification models, including ConvNeXt, Swin Transformer, Vision Transformer, EfficientNetV2, DenseNet-121, and the proposed MSAF-Net model. These models were trained under a unified experimental protocol to ensure fair comparison across convolutional, transformer-based, and hybrid architectures. The model training stage optimized each classifier to distinguish the target brain tumor categories, whereas the model evaluation stage assessed predictive performance using accuracy, precision, recall, specificity, F1-score, ROC-AUC, and confusion matrix analysis. The best-performing classification model was then selected as the final diagnostic backbone for the downstream recommendation system.The output of the selected classification model consisted of the predicted brain tumor class and the associated prediction confidence. This step was important because the framework did not treat the class label as the only clinically meaningful output. Instead, the prediction confidence was explicitly retained and used to condition the next recommendation stage. A high-confidence prediction supported a more direct recommendation, whereas a lower-confidence prediction required more cautious language, stronger emphasis on radiologist review, and possible follow-up imaging or clinical correlation.

The lower part of Figure 3 illustrates the GPT-2-based brain tumor suggestion system. The predicted tumor class was first passed to a query formulation engine. This module generated a structured prompt using the predicted class, confidence score, and recommendation objective. The formulated query was then encoded by an embedding model, which converted the textual query into a vector representation. A vector-based semantic search module compared this query representation with a curated brain tumor knowledge base and associated metadata. The retrieval stage returned relevant information related to the predicted class, typical interpretation, clinical urgency, follow-up considerations, referral pathways, and patient-facing explanation.The retrieved brain tumor knowledge snippets and metadata were then provided to a fine-tuned GPT-2 model for recommendation generation. Rather than generating a response from the predicted class alone, the GPT-2 assistant used retrieved evidence and confidence information to produce a more grounded and clinically cautious output. The generated recommendation included an explanation of the predicted tumor category, the meaning of the confidence score, suggested next steps, and a safety statement recommending professional clinical review. The final output was displayed through a web interface and exported as a PDF report, making the system useful for documentation, communication, and follow-up.

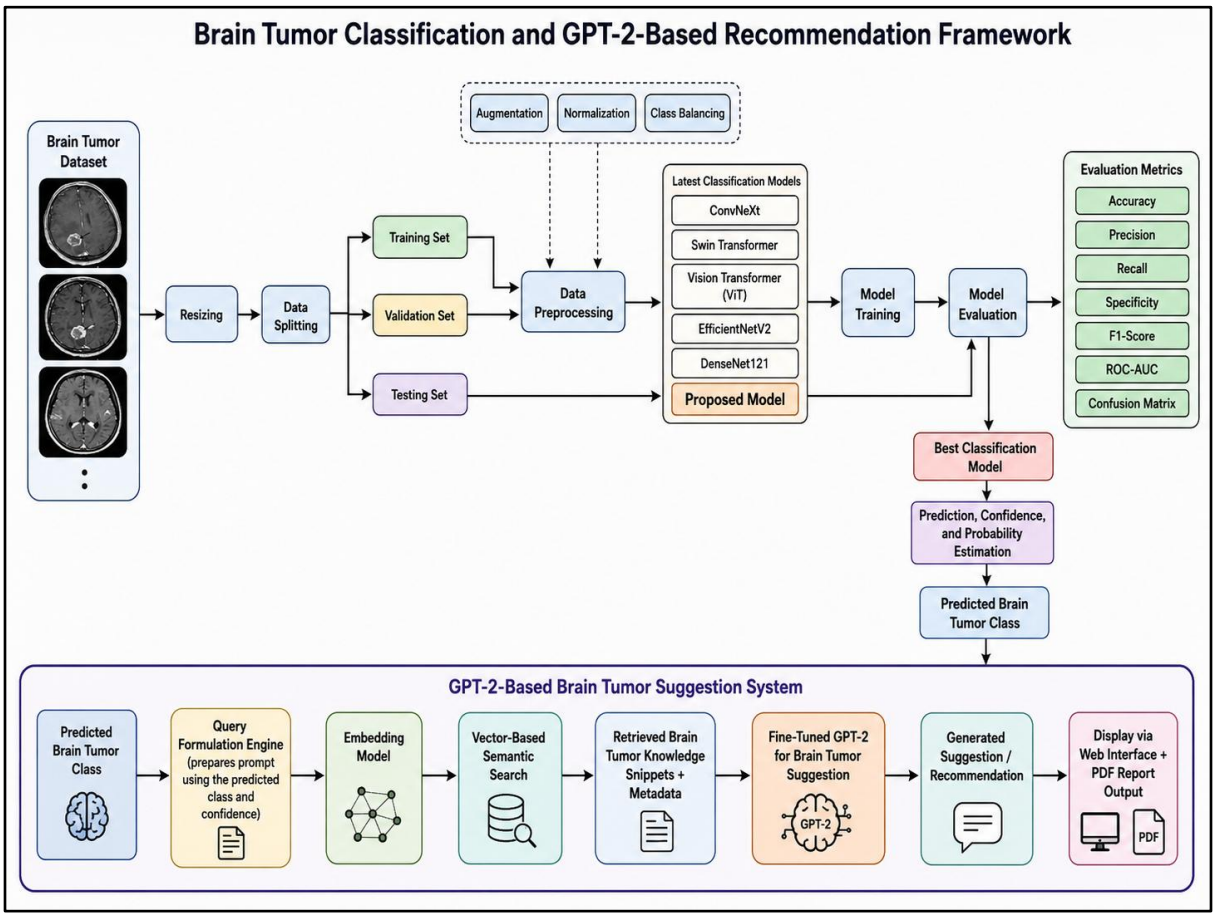

Figure 3. Integrated brain tumor MRI classification and GPT-2-based clinical recommendation framework

4.5 Implementation Details

The experimental implementation was conducted using PyTorch 2.3 on an NVIDIA RTX 3090 GPU with 24 GB memory. All models were trained using the AdamW optimizer with beta1 = 0.9, beta2 = 0.999, and a weight decay of 0.05. The initial learning rate was set to 0.0002 and was controlled using a cosine learning-rate schedule with a five-epoch warmup to stabilize early optimization. Each model was trained for 150 epochs with a batch size of 32 and an input image size of 224 × 224 pixels. To improve generalization and reduce overfitting, the training pipeline used random flipping, random rotation within ±15°, zooming, shifting, and elastic deformation. The objective function combined focal loss with label smoothing using epsilon = 0.1, which helped reduce class-imbalance sensitivity and improve confidence calibration. All experiments followed a patient-disjoint, stratified split protocol and were repeated using three random seeds: 42, 1337, and 2023. These implementation settings ensured a consistent and reproducible evaluation protocol across the proposed model and all baseline architectures, as summarized in Table 2.

Table 2. Implementation details and training configuration

Component	Setting
Framework	PyTorch 2.3
GPU	NVIDIA RTX 3090 24GB
Optimizer	AdamW (β1=0.9, β2=0.999, wd=5e-2)
Initial LR	2e-4
LR Schedule	Cosine, 5-epoch warmup
Batch Size	32
Epochs	150
Input Size	224×224
Augmentation	RandFlip, RandRotate(±15°), Zoom, Shift, ElasticDeform
Loss	Focal + Label-Smoothing (ε=0.1)
Split	patient-disjoint, stratified
Seeds	42, 1337, 2023

5.1 Main Classification Performance on Kaggle BT-MRI

Table 3 presents the main classification results on the Kaggle BT-MRI four-class dataset. The proposed MSAF-Net achieved the strongest overall performance across all reported metrics, with an accuracy of 0.9916, F1-score of 0.9897, AUC of 0.9985, sensitivity of 0.9889, and specificity of 0.9978. Compared with the strongest baseline, ConvNeXt-T, the proposed model improved accuracy from 0.9812 to 0.9916 and F1-score from 0.9805 to 0.9897, while also maintaining a lower parameter count than ConvNeXt-T. Transformer-based and hybrid baselines, including ViT-B/16, Swin-T, CoAtNet-0, and EfficientNetV2-S, also produced competitive results, but none reached the performance of the proposed attention-fusion design. These findings indicate that combining hierarchical ConvNeXt/Swin feature extraction with per-stage channel-spatial attention and multi-scale fusion improved discrimination among glioma, meningioma, pituitary tumor, and no-tumor MRI classes.Table 4 reports the per-class performance of the proposed model on the same Kaggle BT-MRI test set. The model showed consistently high sensitivity, specificity, F1-score, and AUC across all four diagnostic categories. The no-tumor class achieved the highest sensitivity of 0.9970 and AUC of 0.9996, indicating strong recognition of non-tumor MRI cases. Pituitary tumor classification also showed robust performance, with an F1-score of 0.9918 and AUC of 0.9990. Glioma and meningioma achieved slightly lower but still clinically strong F1-scores of 0.9871 and 0.9838, respectively, reflecting the greater visual complexity and potential overlap between tumor appearances. The mean per-class results were 0.9889 ± 0.0067 sensitivity, 0.9978 ± 0.0012 specificity, 0.9897 ± 0.0055 F1-score, and 0.9985 ± 0.0010 AUC, showing stable performance across tumor and non-tumor categories.

Table 3. Main classification performance on the Kaggle BT-MRI four-class test set

Method	Acc	F1	AUC	Sens	Spec	Params (M)
ResNet-50	0.9627	0.9619	0.9921	0.9610	0.9874	25.6
ViT-B/16	0.9642	0.9635	0.9927	0.9628	0.9880	86.6

DenseNet-121	0.9695	0.9688	0.9938	0.9682	0.9887	8.0
CoAtNet-0	0.9733	0.9726	0.9946	0.9721	0.9911	25.0
EfficientNetV2-S	0.9756	0.9749	0.9951	0.9745	0.9918	21.5
Swin-T	0.9771	0.9764	0.9955	0.9758	0.9923	28.3
ConvNeXt-T	0.9812	0.9805	0.9967	0.9798	0.9936	28.6
Proposed	0.9916	0.9897	0.9985	0.9889	0.9978	22.9

Table 4. Per-class classification performance of the proposed MSAF-Net on the Kaggle BT-MRI test set

Class	Sens	Spec	F1	AUC
Glioma	0.9852	0.9971	0.9871	0.9981
Meningioma	0.9821	0.9963	0.9838	0.9973
No Tumor	0.9970	0.9989	0.9961	0.9996
Pituitary	0.9912	0.9985	0.9918	0.9990
Mean ± SD	0.9889 ± 0.0067	0.9978 ± 0.0012	0.9897 ± 0.0055	0.9985 ± 0.0010

5.2 Cross-Dataset Generalization

Table 5 reports the cross-dataset generalization performance when the model was trained on the Kaggle BT-MRI dataset and evaluated on external MRI datasets. On the Figshare CE-MRI three-class task, the proposed model achieved an accuracy of 0.9946, compared with 0.9785 for ConvNeXt-T, corresponding to an improvement of +0.0161. On the BR35H binary tumor-detection task, the proposed model achieved an accuracy of 0.9974, compared with 0.9867 for ConvNeXt-T, giving an improvement of +0.0107. These results indicate that the proposed attention-fusion strategy did not only improve performance on the primary Kaggle test set but also maintained strong transfer performance across different MRI sources and classification settings. The improvement over ConvNeXt-T on both external datasets suggests that the integration of per-stage attention and cross-scale fusion helped the model learn more generalizable tumor-relevant representations.

Table 5. Cross-dataset generalization performance after training on Kaggle BT-MRI

Test Set	Task	ConvNeXt-T Acc	[proposed] Acc	Δ
Figshare CE-MRI	3-class	0.9785	0.9946	+0.0161
BR35H	binary	0.9867	0.9974	+0.0107

5.3 Module Ablation Analysis

Table 6 presents the module-wise ablation study on the Kaggle BT-MRI dataset. The backbone-only configuration, which used the ConvNeXt/Swin stem without per-stage CBAM or gated cross-scale fusion, achieved an accuracy of 0.9812. Adding the per-stage CBAM module increased the accuracy to 0.9863, showing an improvement of +0.0051. This gain indicates that channel-spatial attention helped the model emphasize tumor-relevant features and suppress less informative background responses. The full model, which combined the ConvNeXt/Swin stem, per-stage CBAM, and gated cross-scale fusion, achieved the highest

accuracy of 0.9916, with an overall improvement of +0.0104 compared with the backbone-only configuration. These results confirm that both attention refinement and multi-scale feature fusion contributed to the final classification performance.

Table 6. Module-wise ablation analysis of the proposed MSAF-Net on Kaggle BT-MRI

Variant	ConvNeXt/Swin stem	Per-stage CBAM	Gated cross-scale fusion	Acc ↑	Δ
Backbone only	✓	×	×	0.9812	—
+CBAM	✓	✓	×	0.9863	+0.0051
+CBAM+Fusion (Full)	✓	✓	✓	0.9916	+0.0104

5.4 Loss Function Ablation Analysis

Table 7 compares the effect of different loss functions on classification performance and calibration. Standard cross-entropy achieved an accuracy of 0.9847, an F1-score of 0.9831, and an ECE of 0.0342. Weighted cross-entropy improved accuracy to 0.9861 and reduced ECE to 0.0301, indicating that class-weighting provided some benefit under class imbalance. Focal loss further improved accuracy to 0.9879 and reduced ECE to 0.0258, suggesting better handling of difficult or ambiguous samples. Label smoothing achieved an accuracy of 0.9888, an F1-score of 0.9871, and an ECE of 0.0224. The best result was obtained by combining focal loss with label smoothing, which achieved 0.9916 accuracy, 0.9897 F1-score, and the lowest ECE of 0.0179. This result shows that the proposed loss design improved both discriminative performance and confidence calibration, which was important because the downstream GPT-2 recommendation assistant used model confidence during clinical suggestion generation.

Table 7. Loss-function ablation and calibration performance on Kaggle BT-MRI

Loss	Acc	F1	ECE
Cross-Entropy	0.9847	0.9831	0.0342
Weighted-CE	0.9861	0.9845	0.0301
Focal	0.9879	0.9863	0.0258
Label-Smoothing	0.9888	0.9871	0.0224
Focal + Label-Smoothing (Proposed)	0.9916	0.9897	0.0179

5.5 Backbone-Stem Ablation Analysis

Table 8 evaluates different backbone stems while keeping the CBAM and fusion components fixed. The ResNet-50 stem achieved an accuracy of 0.9869 with 25.6M parameters and 4.10 GFLOPs. DenseNet-121 achieved 0.9877 accuracy with 8.0M parameters and 2.88 GFLOPs, showing a compact but slightly less accurate configuration. The Swin-T stem improved accuracy to 0.9895 but required 28.3M parameters and 4.51 GFLOPs. The proposed ConvNeXt-T/Swin-T stem achieved the highest accuracy of 0.9916 while using 22.9M parameters and 3.54 GFLOPs. This result shows that the proposed hybrid stem provided the best balance between accuracy and computational cost. Compared with the Swin-T stem, it improved classification accuracy while reducing both parameter count and FLOPs, supporting its suitability for efficient clinical decision-support deployment.

Table 8. Backbone-stem ablation with CBAM and gated fusion fixed on Kaggle BT-MRI

Stem	Pretrain	Acc	Params (M)	FLOPs (G)
ResNet-50	ImageNet	0.9869	25.6	4.10
DenseNet-121	ImageNet	0.9877	8.0	2.88
Swin-T	ImageNet	0.9895	28.3	4.51
ConvNeXt-T / Swin-T (Proposed)	ImageNet	0.9916	22.9	3.54

5.6 Computational Cost Analysis

Table 9 reports the computational cost of the proposed model compared with representative baseline architectures. The proposed MSAF-Net required 22.9M parameters and 3.54 GFLOPs, with a training time of 4.2 h, inference time of 10 ms per scan, and GPU memory consumption of 3,280 MB. Although DenseNet-121 had the lowest computational demand, with 8.0M parameters, 2.88 GFLOPs, and 6 ms inference, its classification performance was lower than that of the proposed model. EfficientNetV2-S also showed a compact computational profile, with 21.5M parameters and 2.96 GFLOPs, but did not reach the accuracy and discriminative performance of MSAF-Net. In contrast, transformer-heavy models required higher computational resources; ViT-B/16 used 86.6M parameters, 17.6 GFLOPs, 7.2 h of training time, 18 ms inference, and 6,840 MB GPU memory. Swin-T and ConvNeXt-T also required higher parameter counts and FLOPs than the proposed model. These results indicate that MSAF-Net achieved a practical balance between classification performance and computational efficiency, supporting its potential use in low-cost clinical decision-support settings where rapid inference and moderate hardware requirements are important.

Table 9. Computational cost comparison between MSAF-Net and baseline architectures

Method	Params (M)	FLOPs (G)	Train (h)	Inference (ms)	GPU Mem (MB)
DenseNet-121	8.0	2.88	3.1	6	2,180
EfficientNetV2-S	21.5	2.96	3.8	9	2,940
Swin-T	28.3	4.51	4.6	11	3,720
ConvNeXt-T	28.6	4.50	4.4	11	3,610
ViT-B/16	86.6	17.6	7.2	18	6,840
[proposed]	22.9	3.54	4.2	10	3,280

5.7 Statistical Significance Analysis

Table 10 presents the statistical comparison between the proposed MSAF-Net and the main baseline models on the Kaggle BT-MRI test set of 1,311 images. The proposed model achieved an accuracy of 0.9916 ± 0.0079, with a 95% confidence interval of [0.9912, 0.9920]. In comparison, DenseNet-121 achieved 0.9695 ± 0.0078, EfficientNetV2-S achieved 0.9756 ± 0.0078, Swin-T achieved 0.9771 ± 0.0078, and ConvNeXt-T achieved 0.9812 ± 0.0079. McNemar testing showed statistically significant differences between the proposed model and the evaluated baselines, with p-values below 0.001 for DenseNet-121, EfficientNetV2-S, and Swin-T, and a p-value of 0.003 for ConvNeXt-T. Holm correction was applied across seven baseline comparisons, with an adjusted critical value of 0.0071 at rank 1. These findings indicate that the performance gain of MSAF-Net was not limited to a numerical increase in accuracy but was also statistically supported under paired testing on the same test samples.

Table 10. Statistical significance analysis comparing MSAF-Net with baseline models on Kaggle BT-MRI

Method	Acc (mean ± SD)	p-value (McNemar)	95% CI	Sig (α=0.05)
DenseNet-121	0.9695 ± 0.0078	<0.001	[0.9686, 0.9704]	✓
EfficientNetV2-S	0.9756 ± 0.0078	<0.001	[0.9747, 0.9765]	✓
Swin-T	0.9771 ± 0.0078	<0.001	[0.9762, 0.9780]	✓
ConvNeXt-T	0.9812 ± 0.0079	0.003	[0.9803, 0.9821]	✓
Proposed	0.9916 ± 0.0079	—	[0.9912, 0.9920]	—

5.8 Robustness Under Image Perturbations

Table 11 evaluates robustness under common MRI image-quality perturbations, including Rician noise, Gaussian noise, motion blur, bias-field distortion, and low-resolution degradation. Across all perturbation settings, the proposed model maintained higher accuracy than ConvNeXt-T. Under Rician noise with σ = 0.05, MSAF-Net achieved 0.9744 accuracy compared with 0.9521 for ConvNeXt-T, giving an improvement of +0.0223. Under Gaussian noise with σ = 0.05, the proposed model achieved 0.9779 accuracy, with a gain of +0.0221. For motion blur with kernel size k = 3, MSAF-Net achieved 0.9801 accuracy, compared with 0.9603 for ConvNeXt-T. Similar robustness gains were observed under bias-field distortion and low-resolution degradation, where the proposed model improved accuracy by +0.0185 and +0.0196, respectively. These results suggest that the combination of hierarchical feature extraction, per-stage attention, and cross-scale fusion improved resistance to acquisition-related degradation and image-quality variation.

Table 11. Robustness evaluation under MRI image-quality perturbations

Perturbation	Severity	ConvNeXt-T Acc	Proposed Acc	Δ
Rician Noise	σ=0.05	0.9521	0.9744	+0.0223
Gaussian Noise	σ=0.05	0.9558	0.9779	+0.0221
Motion Blur	k=3	0.9603	0.9801	+0.0198
Bias Field	δ=0.3	0.9647	0.9832	+0.0185
Low Resolution	×0.5	0.9612	0.9808	+0.0196

5.9 Clinical Relevance Analysis

Table 12 reports clinically relevant performance indicators for the proposed model on the Kaggle BT-MRI dataset. MSAF-Net achieved a sensitivity of 0.9889 with a 95% confidence interval of [0.9846, 0.9932], exceeding the predefined clinical threshold of 0.90. The specificity was 0.9978 with a 95% confidence interval of [0.9960, 0.9996], also exceeding the predefined threshold of 0.85. The model further achieved a positive predictive value of 0.9904, a negative predictive value of 0.9962, and a macro-F1 score of 0.9897. These results indicate that the proposed model performed strongly in both tumor identification and non-tumor recognition, which is important for reducing missed tumor cases while limiting unnecessary concern or referral. Calibration performance was also favorable, with an expected calibration error of 0.0179 and a 95% confidence interval of [0.0148, 0.0210], remaining below the threshold of 0.05. This calibration result was clinically important because the downstream GPT-2 recommendation assistant used the model confidence score to adjust the tone, urgency, and caution level of the generated clinical suggestion. Therefore, the clinical relevance analysis supports the proposed framework as a decision-support tool that can assist triage, reporting, and patient communication while requiring final review by qualified clinicians.

Table 12. Clinical relevance metrics and calibration performance of MSAF-Net on Kaggle BT-MRI

Metric	Value	95% CI	Clinical Threshold
Sensitivity	0.9889	[0.9846, 0.9932]	≥ 0.90 ✓
Specificity	0.9978	[0.9960, 0.9996]	≥ 0.85 ✓
PPV	0.9904	[0.9864, 0.9944]	—
NPV	0.9962	[0.9937, 0.9987]	—
Macro-F1	0.9897	[0.9843, 0.9951]	—
Calibration ECE	0.0179	[0.0148, 0.0210]	< 0.05 ✓

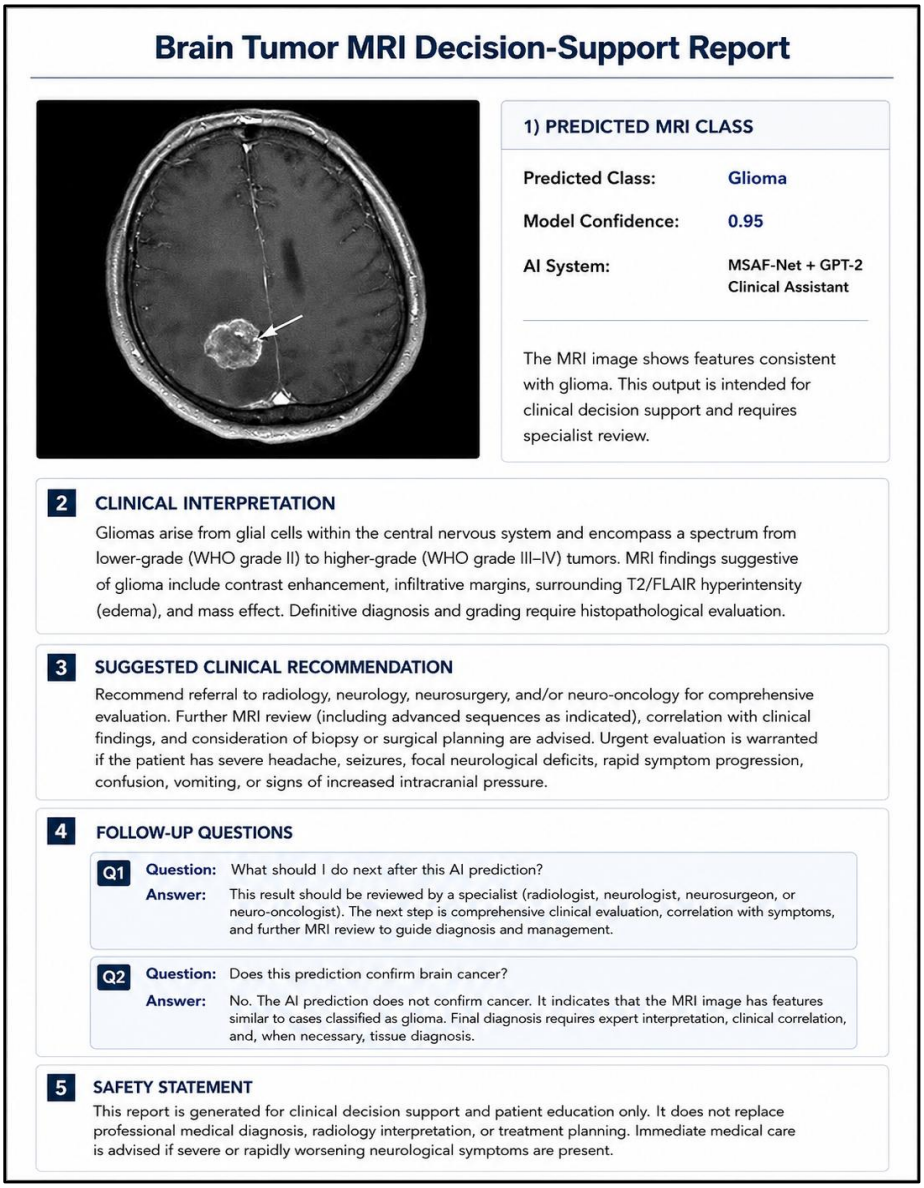

Figure 4. Representative GPT-2-generated brain tumor MRI decision-support report for a high-confidence glioma prediction

6. LLM Assistant Evaluation

The GPT-2-based clinical assistant was evaluated as the interpretive and recommendation layer of the proposed framework. Its purpose was not to improve the image classification score directly, but to convert the classifier output into a clinically usable decision-support report. Conventional brain tumor classification systems usually stop after producing a class label and probability score. In contrast, the proposed assistant used the predicted class, confidence value, retrieved clinical context, and a fine-tuned GPT-2 generator to produce a structured explanation, next-step recommendation, and patient-readable PDF report (Figure 4). The complete assistant pipeline is shown in Figure 3, where the predicted brain tumor class and confidence score were passed through query formulation, embedding-based retrieval, vector semantic search, GPT-2 generation, and report output.A representative example is shown using a predicted class of glioma with a confidence score of 0.95. In this case, the classifier output was first converted into a structured query describing the predicted tumor category, confidence level, and required recommendation intent. The embedding model then encoded the query and retrieved relevant information from the curated brain tumor knowledge base and metadata. The retrieved evidence included glioma-related clinical context, such as the need for radiological confirmation, neurological assessment, evaluation of lesion extent, review of enhancement pattern, edema, mass effect, and possible referral to neurology, neurosurgery, or neuro-oncology. This retrieval step was important because it constrained the generated text toward evidence-grounded and class-specific guidance rather than allowing the language model to respond only from its internal parameters.Using the retrieved evidence and confidence-conditioned prompt, the fine-tuned GPT-2 assistant generated a plain-language recommendation. A representative generated snippet was: “The MRI image shows features consistent with glioma. This output is intended for clinical decision support and requires specialist review. Referral to radiology, neurology, neurosurgery, or neuro-oncology is recommended for comprehensive evaluation. Further MRI review, clinical correlation, and consideration of biopsy or surgical planning may be required. Urgent evaluation is advised if severe headache, seizures, focal neurological deficits, confusion, vomiting, or rapidly worsening symptoms are present.” This response was then formatted into a PDF-style report containing the predicted class, model confidence, clinical interpretation, suggested recommendation, follow-up questions, and a safety statement. The report format was designed to be readable by both clinicians and patients while maintaining clear boundaries around the role of artificial intelligence.The clinical value of the assistant was threefold. First, it added urgency guidance. A high-confidence tumor prediction was not presented as a final diagnosis, but it supported a recommendation for specialist review and timely referral. Second, it added follow-up guidance. The assistant explained that further MRI review, clinical correlation, and specialist evaluation may be needed before treatment planning. Third, it improved patient communication. Instead of showing only “glioma” and a confidence score, the generated report explained what the prediction may mean, what the next step should be, and why the result required professional confirmation. This design was particularly relevant for low-resource or rural settings, where delayed access to neuroradiology may leave clinicians and patients without immediate interpretive support.

The reliability of the recommendation layer depended on the quality and calibration of the upstream classifier. For this reason, confidence conditioning was used only after evaluating clinically relevant classifier behavior. The proposed model achieved sensitivity of 0.9889, specificity of 0.9978, positive predictive value of 0.9904, negative predictive value of 0.9962, macro-F1 of 0.9897, and expected calibration error of 0.0179, as reported in Table 12. These results supported the use of classifier confidence as a meaningful input to the assistant. In particular, the low calibration error was important because the assistant adjusted its wording according to confidence level. High-confidence outputs could use stronger referral language, whereas lower-confidence outputs required more cautious phrasing, clearer uncertainty statements, and stronger emphasis on radiologist review.From a workflow perspective, the assistant may reduce time and cost by helping transform an image-level prediction into an initial structured report. This can support faster triage, reduce dependence on immediate specialist availability for preliminary guidance, and assist communication in settings where expert neuroradiology access is limited. The system may also help standardize the wording of preliminary recommendations, especially for non-specialist clinicians who need to decide whether urgent referral or additional imaging review is appropriate.

7. Limitations and Future Work

Although the proposed framework showed strong classification performance and demonstrated the feasibility of linking MRI prediction with GPT-2-based clinical recommendation generation, several limitations should be acknowledged. First, the study was based on retrospective publicly available MRI datasets. Although patient-disjoint and stratified splitting was used where applicable, retrospective datasets may not fully represent the variability of prospective clinical practice, including scanner differences, institutional imaging protocols, motion artifacts, incomplete sequences, and heterogeneous reporting standards. Therefore, the reported results should be interpreted as evidence of technical feasibility and benchmark-level performance rather than direct proof of clinical deployment readiness.Second, the classification experiments were performed primarily at the image or slice level. In routine neuroimaging practice, radiologists interpret complete MRI studies rather than isolated images. Tumor diagnosis often depends on multiple slices, anatomical continuity, contrast enhancement pattern, edema distribution, mass effect, and comparison across sequences. Therefore, future work should extend MSAF-Net toward patient-level and volumetric

evaluation using complete multi-sequence MRI studies. Incorporating 3D spatial context and sequence-level fusion may further improve reliability, especially for visually complex or borderline cases.Third, the external evaluation included Figshare CE-MRI and BR35H, but broader multicenter validation remains necessary. Public MRI datasets may contain hidden acquisition similarities, preprocessing artifacts, or source-specific biases. Future studies should evaluate the framework on independent hospital cohorts collected from different scanners, regions, and clinical populations. Prospective validation is also required to assess how the system performs under real reporting workflows, including emergency imaging, low-quality scans, postoperative cases, recurrent tumors, and cases with non-neoplastic mimics.Fourth, the GPT-2 clinical assistant was designed as a decision-support module rather than an autonomous diagnostic system. Its recommendations depended on the predicted class, model confidence, and retrieved knowledge snippets. Although retrieval augmentation reduced the risk of unsupported generation, the assistant may still produce incomplete, overly general, or context-insensitive recommendations if the retrieved evidence is insufficient or if important clinical information is unavailable. The system did not have access to full patient history, neurological examination, laboratory findings, prior imaging, treatment status, or institutional guidelines. For this reason, every generated recommendation must be reviewed by a qualified clinician before being used in patient care.Fifth, the report-generation module was evaluated as a structured clinical communication component, but it was not tested through a formal reader study involving radiologists, neurologists, neurosurgeons, patients, or non-specialist clinicians. Future work should include human-factor evaluation to measure readability, usefulness, safety, trust calibration, and potential workflow benefit. Such evaluation should assess whether the generated PDF report improves triage, reduces interpretation delay, supports patient communication, or unintentionally increases over-reliance on automated output.

Future development should focus on four directions. First, MSAF-Net should be validated prospectively on multicenter patient-level MRI cohorts with complete DICOM metadata and multi-sequence imaging. Second, the framework should be extended from image-level classification to volumetric and lesion-aware modeling, potentially incorporating tumor localization, segmentation, and uncertainty maps. Third, the GPT-2 assistant should be evaluated using clinician-rated criteria, including factual correctness, actionability, safety, uncertainty communication, and guideline consistency. Fourth, deployment studies should assess computational efficiency, reporting time, cost reduction, and usability in low-resource clinical settings. These future steps would strengthen the evidence base for using the proposed system as a practical human-in-the-loop decision-support tool for brain tumor MRI interpretation.

8. Conclusion

This study presented a multi-scale attention-fusion classification framework with an LLM-driven clinical recommendation assistant for brain tumor MRI. The proposed system addressed three practical gaps in current medical-imaging artificial intelligence: accurate tumor classification, translation of prediction into actionable guidance, and low-cost access to decision support in settings where specialist neuroradiology resources may be limited. The MSAF-Net classifier integrated a ConvNeXt-T/Swin-T hierarchical stem, per-stage channel-spatial attention, gated cross-scale feature fusion, and a compact softmax classification head to distinguish glioma, meningioma, pituitary tumor, and no-tumor MRI classes. On the primary Kaggle BT-MRI dataset, the model achieved 0.9916 accuracy, 0.9897 macro-F1, and 0.9985 AUC, outperforming the evaluated CNN, transformer, and hybrid baselines under the same experimental protocol, as reported in Table 3.The framework also showed strong transfer performance across external MRI datasets, achieving 0.9946 accuracy on Figshare CE-MRI and 0.9974 accuracy on BR35H, as shown in Table 5. Ablation analyses further indicated that per-stage CBAM attention, gated cross-scale fusion, and the combined focal loss with label smoothing each contributed to improved classification and calibration behavior. Computational analysis demonstrated that MSAF-Net maintained a practical deployment profile with 22.9M parameters, 3.54 GFLOPs, and 10 ms inference per scan, supporting its suitability for rapid and resource-conscious decision-support use, as reported in Table 9.Beyond classification, the confidence-conditioned retrieval-augmented GPT-2 assistant converted the predicted class and confidence score into a plain-language, evidence-grounded recommendation and PDF-style report. This component addressed the limitation of label-only AI by providing referral, urgency, follow-up, and safety guidance for clinicians and patients. The strong clinical metrics, including 0.9889 sensitivity, 0.9978 specificity, and 0.0179 expected calibration error in Table 12, supported the use of confidence-aware recommendation generation. Nevertheless, the system is not intended to replace radiologists, neurosurgeons, oncologists, or other licensed clinicians. It should be used as a human-in-the-loop decision-support tool that assists triage, communication, and preliminary reporting while final diagnosis and management remain under professional clinical supervision. Future prospective, multicenter, patient-level validation will be essential before real-world clinical deployment.

Author Contributions: Kallol Chakraborty Shekhor led the study, conceptualized the main research idea, designed the proposed MSAF-Net architecture, developed the overall diagnostic and recommendation framework, supervised the experimental workflow, and took the primary responsibility for manuscript preparation. Md Sahid Hossain contributed to model implementation, baseline model development, coding, experimental analysis, performance evaluation, and manuscript writing. Md Abedur Rahman contributed to data preparation, preprocessing, baseline experiments, comparative analysis, result interpretation, and manuscript revision. Md Anwar Hossain contributed to coding support, validation experiments, evaluation of baseline models, literature review, and manuscript editing. All authors contributed to the development of the codebase, experimental verification, result analysis, writing, review, and approval of the final manuscript.

Data Availability: The datasets used in this study are publicly available and were obtained from established repositories. No new data were generated by the authors.

Declarations

Ethical approval: This article does not report any prospective studies involving human participants or animals performed by the authors. The analysis is based on previously collected, fully anonymized oral cancer images provided to the authors; therefore, additional institutional ethical approval and clinical trial registration were not required.

Consent to participate: Not applicable. The study used only secondary, anonymized image data and involved no direct contact or intervention with individual participants.

Consent to publish: Not applicable.

References

1.Tiwari, A., Srivastava, S., & Pant, M. (2020). Brain tumor segmentation and classification from magnetic resonance images: Review of selected methods from 2014 to 2019. *Pattern recognition letters*, *131*, 244-260.
2.Khan, H. A., Wu, J., Mushtaq, M., & Mushtaq, M. U. (2020). Brain tumor classification in MRI image using convolutional neural network. *Mathematical Biosciences and Engineering*, *17*(5), 6203.
3.Amin, Javeria, Muhammad Sharif, Mussarat Yasmin, and Steven Lawrence Fernandes. "A distinctive approach in brain tumor detection and classification using MRI." *Pattern Recognition Letters* 139 (2020): 118-127.
4.Toufiq, Dalia Mohammad, Ali Makki Sagheer, and Hadi Veisi. "A review on brain tumor classification in mri images." *Turkish Journal of Computer and Mathematics Education (TURCOMAT)* 12, no. 14 (2021): 1958-1969.
5.Jia, Zheshu, and Deyun Chen. "Brain tumor identification and classification of MRI images using deep learning techniques." *IEEE Access* (2020).
6.Mohanty, S. N., H. Ghosh, I. S. Rahat, and C. V. R. Reddy. *Advanced Deep Learning Models for Corn Leaf Disease Classification: A Field Study in Bangladesh. Eng. Proc. 2023, 59, 69*.
7.Rahat, Irfan Sadiq, Hritwik Ghosh, Kareemulla Shaik, Syed Khasim, and Gnanajeyaraman Rajaram. "Unraveling the heterogeneity of lower-grade gliomas: Deep learning-assisted flair segmentation and genomic analysis of brain MR images." *EAI Endorsed Transactions on Pervasive Health and Technology* 9 (2023).
8.Arunkumar, N., Mazin Abed Mohammed, Salama A. Mostafa, Dheyaa Ahmed Ibrahim, Joel JPC Rodrigues, and Victor Hugo C. De Albuquerque. "Fully automatic model-based segmentation and classification approach for MRI brain tumor using artificial neural networks." *Concurrency and Computation: Practice and Experience* 32, no. 1 (2020): e4962.
9.Dipu, Nadim Mahmud, Sifatul Alam Shohan, and K. M. A. Salam. "Deep learning based brain tumor detection and classification." In *2021 International conference on intelligent technologies (CONIT)*, pp. 1-6. IEEE, 2021.
10.Díaz-Pernas, Francisco Javier, Mario Martínez-Zarzuela, Míriam Antón-Rodríguez, and David González-Ortega. "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, 2021.
11. Tiwari, A., Srivastava, S., & Pant, M. (2020). Brain tumor segmentation and classification from magnetic resonance images: Review of selected methods from 2014 to 2019. Pattern recognition letters, 131, 244-260.
12. Khan, Hassan Ali, Jue Wu, Muhammad Mushtaq, and Muhammad Umer Mushtaq. "Brain tumor classification in MRI image using convolutional neural network." Mathematical Biosciences and Engineering 17, no. 5 (2020): 6203.
13. Amin, Javeria, Muhammad Sharif, Mussarat Yasmin, and Steven Lawrence Fernandes. "A distinctive approach in brain tumor detection and classification using MRI." Pattern Recognition Letters 139 (2020): 118-127.
14. Aamir, Muhammad, Ziaur Rahman, Zaheer Ahmed Dayo, Waheed Ahmed Abro, M. Irfan Uddin, Inayat Khan, Ali Shariq Imran et al. "A deep learning approach for brain tumor classification using MRI images." Computers and Electrical Engineering 101

(2022): 108105.
15. Guan, Yurong, Muhammad Aamir, Ziaur Rahman, Ammara Ali, Waheed Ahmed Abro, Zaheer Ahmed Dayo, Muhammad Shoaib Bhutta, Zhihua Hu, Y. Guan, and M. Aamir. "A framework for efficient brain tumor classification using MRI images." Math. Biosci. Eng 18, no. 5 (2021): 5790-5815.
16. Toufiq, Dalia Mohammad, Ali Makki Sagheer, and Hadi Veisi. "A review on brain tumor classification in mri images." Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12, no. 14 (2021): 1958-1969.
17. Rao, Champakamala Sundar, and K. Karunakara. "A comprehensive review on brain tumor segmentation and classification of MRI images." Multimedia Tools and Applications 80, no. 12 (2021): 17611-17643.
18. Rasheed, Zahid, Yong-Kui Ma, Inam Ullah, Yazeed Yasin Ghadi, Muhammad Zubair Khan, Muhammad Abbas Khan, Akmalbek Abdusalomov, Fayez Alqahtani, and Ahmed M. Shehata. "Brain tumor classification from MRI using image enhancement and convolutional neural network techniques." Brain Sciences 13, no. 9 (2023): 1320.
19. Badža, M. M., & Barjaktarović, M. Č. (2020). Classification of brain tumors from MRI images using a convolutional neural network. Applied sciences, 10(6), 1999.
20. Jia, Z., & Chen, D. (2020). Brain tumor identification and classification of MRI images using deep learning techniques. IEEE Access.
21. Vankdothu, Ramdas, and Mohd Abdul Hameed. "Brain tumor MRI images identification and classification based on the recurrent convolutional neural network." Measurement: Sensors 24 (2022): 100412.
22. Kuraparthi, Swaraja, Madhavi K. Reddy, C. N. Sujatha, Himabindu Valiveti, Chaitanya Duggineni, Meenakshi Kollati, and Padmavathi Kora. "Brain Tumor Classification of MRI Images Using Deep Convolutional Neural Network." Traitement du Signal 38, no. 4 (2021).
23. Ayadi, Wadhah, Wajdi Elhamzi, Imen Charfi, and Mohamed Atri. "Deep CNN for brain tumor classification." Neural processing letters 53, no. 1 (2021): 671-700.
24. Irmak, E. (2021). Multi-classification of brain tumor MRI images using deep convolutional neural network with fully optimized framework. Iranian Journal of Science and Technology, Transactions of Electrical Engineering, 45(3), 1015-1036.
25. Özkaraca, Osman, Okan İhsan Bağrıaçık, Hüseyin Gürüler, Faheem Khan, Jamil Hussain, Jawad Khan, and Umm E. Laila. "Multiple brain tumor classification with dense CNN architecture using brain MRI images." Life 13, no. 2 (2023): 349.
26. Saleh, Ahmad, Rozana Sukaik, and Samy S. Abu-Naser. "Brain tumor classification using deep learning." In 2020 International Conference on Assistive and Rehabilitation Technologies (iCareTech), pp. 131-136. IEEE, 2020.
27. Aamir, Muhammad, Ziaur Rahman, Waheed Ahmed Abro, Uzair Aslam Bhatti, Zaheer Ahmed Dayo, and Muhammad Ishfaq. "Brain tumor classification utilizing deep features derived from high-quality regions in MRI images." Biomedical Signal Processing and Control 85 (2023): 104988.
28. Ayadi, Wadhah, Imen Charfi, Wajdi Elhamzi, and Mohamed Atri. "Brain tumor classification based on hybrid approach." The Visual Computer 38, no. 1 (2022): 107-117.
29. Srinivas, Chetana, Nandini Prasad KS, Mohammed Zakariah, Yousef Ajmi Alothaibi, Kamran Shaukat, B. Partibane, and Halifa Awal. "Deep transfer learning approaches in performance analysis of brain tumor classification using MRI images." Journal of Healthcare Engineering 2022, no. 1 (2022): 3264367.
30. Rahman, Takowa, and Md Saiful Islam. "MRI brain tumor detection and classification using parallel deep convolutional neural networks." Measurement: Sensors 26 (2023): 100694.