Model-Augmented Masked Autoencoder with Domain-Adaptive Denoising for Robust Brain Tumor MRI Analysis

Indrakumar K[1], Ravikumar M[2], Mohammed A.S Al-mohamadi[3],
Khalid N. R. Alharbi[4], Asma A. Alhashmi[5]*, Shouki A. Ebad[6], Abdulbasit A. Darem[7]*
Department of MCA and Computer Science, Jnanasahyadri, Kuvempu University, Shivamogga 577 451, India[1, 2, 3]
Department of Computer Science-College of Science, Northern Border University, Arar 73213, Saudi Arabia[4, 5]
Centre for Scientific Research and Entrepreneurship, Northern Border University, Arar 73213, Saudi Arabia[6, 7]

***Abstract*—This work presents a Model-Augmented Masked Autoencoder (MAE) framework, augmented with domain-adaptive denoising and multi-level feature fusion, for self-supervised representation learning in MRI-based brain tumor analysis. The proposed approach enhances feature robustness under limited annotation by learning structure-preserving representations through masked reconstruction while explicitly addressing scanner-induced variability. Unlike conventional MAE designs, the framework integrates a Domain-Adaptive Denoising (DAD) module and latent–reconstruction feature fusion, enabling improved robustness across heterogeneous MRI datasets and reducing feature variance arising from scanner and acquisition differences. Comprehensive ablation studies across multiple datasets demonstrate that MAE pretraining yields consistent performance improvements of approximately 1.0–1.4% in classification accuracy, with overall improvements ranging from approximately 2.7% to 3.7% across datasets when integrated into the complete framework. Qualitative and quantitative analysis further confirm effective suppression of structural noise while preserving diagnostically relevant features. Overall, the proposed framework improves representation stability, enhances cross-domain robustness, and provides a reliable foundation for self-supervised medical image analysis, supporting the development of generalizable and clinically applicable diagnostic AI systems.**

Keywords*—Self-supervised learning; Masked Autoencoder; MRI; domain adaptation; feature representation; medical imaging

I. Introduction

Magnetic Resonance Imaging (MRI) plays a central role in the diagnosis, grading, and clinical management of brain tumors. Accurate tumor classification is essential for treatment planning, yet remains challenging due to substantial variations in tumor morphology, heterogeneous imaging protocols across scanners, and the presence of noise, artifacts, and low-contrast boundaries. Deep learning models have achieved impressive performance in medical image classification; however, their success is strongly dependent on the availability of large, high-quality annotated datasets. In neuro-oncology, such annotations require expert radiological expertise, are time-consuming to produce, and often vary across institutions. These challenges motivate the need for representation learning strategies that reduce dependence on labeled data while improving robustness to domain variability.

*Corresponding author

Self-supervised learning (SSL) has emerged as a powerful paradigm to address annotation scarcity by leveraging unlabeled images to learn transferable, domain-invariant representations [1,2,3,4,29]. Contrastive methods and masked prediction frameworks such as Masked Autoencoders (MAE) [16] have shown promise in natural and medical imaging. By predicting missing content or enforcing consistency between augmented views, SSL extracts semantic cues without explicit annotations. However, despite their potential, existing SSL approaches exhibit three major limitations when applied to brain MRI analysis. First, contrastive-based SSL approaches such as those analyzed by Wolf et al. [3] are sensitive to augmentation strategies and may fail to capture fine anatomical details critical for tumor demarcation. Second, conventional MAE-based frameworks [3,4] reconstruct masked regions but generally do not explicitly address scanner-specific noise, bias-field inhomogeneities, or contrast variations commonly encountered in clinical MRI. Third, although Zhang et al. [18], Liu et al. [19], Zhang et al. [20], Bakkouri and Afdel [21], and Peng et al. [22] demonstrated the effectiveness of multi-level feature fusion, the integration of latent self-supervised representations with radiomics, shape descriptors, and deep CNN features remains insufficiently explored.

To address these challenges, this study introduces the Model-Augmented Masked Autoencoder (MAE), a self-supervised pretraining framework designed to improve robustness and generalization in MRI-based brain tumor classification. The proposed approach builds upon masked image modeling by reconstructing randomly masked patches, enabling the model to learn structure-preserving representations that emphasize anatomical continuity and tumor-related spatial patterns [16]. A key innovation of the framework is the integration of a Domain-Adaptive Denoising (DAD) module, which refines latent token embeddings through channel normalization, patch-level smoothing, and gated fusion mechanisms. This enhancement makes the learned representations more resilient to cross-scanner inconsistencies and low-quality acquisitions, aligning with recent domain generalization frameworks discussed by Zhou et al. [27] and Wang et al. [28]. Additionally, the model augments downstream learning by combining MAE latent embeddings with reconstructed images, ensuring that both high-level contextual

signals and low-level structural details contribute to improved tumor characterization.

The enriched representations produced by the MAE pipeline are further integrated into a multi-level feature extraction system, incorporating radiomics descriptors, CNN-based semantic features, and geometric shape attributes. This fusion captures a wide spectrum of tumor information from statistical intensity patterns to morphological cues, yielding a more holistic and clinically aligned tumor signature [18,19,20,21,22]. These features are processed by an attention-guided fusion block and a Pyramid Quantum-Inspired Dilated Convolutional Neural Network (PyQDCNN), which models long-range spatial dependencies and multi-scale tumor context through dilation-driven and quantum-inspired (implemented as learnable bilinear feature interactions) entanglement mechanisms [23,24,25]. It is important to note that this mechanism is inspired by quantum interactions but is implemented as a learnable bilinear feature mixing operation within a conventional deep learning framework. The final hierarchical classification strategy mirrors clinical diagnostic workflows, first distinguishing normal from abnormal MRI scans, and then classifying tumors into clinically relevant subtypes.

The main contributions of this study are as follows:

- We propose a Model-Augmented Masked Autoencoder (MAE) with a Domain-Adaptive Denoising (DAD) module, enabling the learning of structure-preserving and scanner-robust latent representations that enhance MRI feature stability across heterogeneous MRI datasets.
- We develop a multi-level tumor representation pipeline that integrates radiomics, CNN features, shape descriptors, and MAE latent embeddings, producing a comprehensive and clinically aligned feature space for robust tumor characterization.
- We introduce a hybrid attention–PyQDCNN architecture that combines modality-aware feature weighting with entanglement-inspired bilinear feature interactions and dilated convolutional modeling, enabling effective capture of long-range dependencies and multi-scale tumor characteristics.
- We demonstrate consistent performance gains across four heterogeneous brain MRI datasets, achieving improvements of 1.0–1.4% due to MAE pretraining and overall accuracy improvements ranging from approximately 2.7% to 3.7% compared to baseline configurations, confirming the effectiveness and generalizability of the proposed framework.

The remainder of the study is organized as follows: Section II reviews related work. Section III presents the proposed Model-Augmented MAE framework. Section IV outlines the experimental setup. Section V reports and discusses results. Section VI concludes the study.

II. Related Work

Self-supervised learning (SSL) has become an essential strategy in medical imaging, particularly when annotated data are scarce. Rani et al. [1] presented a comprehensive review of SSL methods for medical image analysis and highlighted their effectiveness across multiple imaging modalities. Zeng et al. [2] summarized SSL frameworks and their applications in healthcare imaging. Wolf et al. [3] investigated contrastive and masked autoencoder pretraining strategies for small medical imaging datasets, demonstrating improved representation learning under limited annotations. VanBerlo et al. [4] analyzed the impact of self-supervised pretraining across X-ray, CT, MRI, and ultrasound modalities, while Uelwer et al. [29] provided a recent survey of self-supervised visual representation learning methods. Transformer-based SSL approaches, particularly masked autoencoders (MAEs), have gained prominence for their ability to model long-range dependencies and structural features in clinical images. The original MAE framework [16] demonstrates the effectiveness of reconstruction-based pretraining, while subsequent studies show that contrastive and masked pretraining significantly improve performance on small or heterogeneous datasets [3]. Recent applications further highlight the potential of transformer-driven SSL in disease detection and medical imaging tasks [5]. Advancements in MAE-based pretraining include improved architectural variations such as snapshot-ensemble MAEs for efficient training [6] and adaptive masking strategies designed to reinforce structural consistency during representation learning [7]. Emerging research directions also explore auxiliary modalities, such as large language model–augmented learning, which has been shown to enhance auto-delineation performance in radiotherapy planning [8]. In addition, domain generalization has received significant attention, with recent surveys emphasizing the importance of robust learning across heterogeneous datasets and imaging conditions [27,28]. Federated learning frameworks further demonstrate potential for improving cross-institutional generalization in brain tumor segmentation tasks [9].

For downstream tumor classification and segmentation, Vision Transformers (ViTs) and hybrid CNN–Transformer models have shown strong performance. Several studies have explored rotation-invariant ViTs for improved spatial consistency [11], ensemble-based ViT classification strategies [12], hybrid ViT–CNN architectures with explainability [13], and transfer-learning–based transformer models for tumor classification [14]. Transformer-based methods have also been successfully applied to related diagnostic tasks such as brain stroke classification [15] and glioma analysis [17], demonstrating their adaptability across different MRI-based applications.

Multi-level and multi-scale feature fusion approaches have further enhanced tumor detection and segmentation performance. Zhang et al. [18] proposed a multi-level feature exploration and fusion network for glioma analysis using MRI data. Liu et al. [19] introduced MCF-Net, which combines multi-level contextual information for medical image segmentation. Zhang et al. [20] developed a multi-scale deformable feature fusion framework for tumor segmentation. Bakkouri and Afdel [21] proposed an attention-based multi-level feature fusion strategy for lesion segmentation, while Peng et al. [22] introduced a graph-neural-network-based multi-level fusion approach for multimodal medical imaging. These studies

demonstrate that combining complementary feature representations improves tumor characterization and prediction accuracy. These approaches effectively capture complementary information across multiple feature levels, enabling more robust modeling of tumor heterogeneity. Additionally, quantum-inspired convolutional and optimization methods have been explored to model complex spatial dependencies and improve segmentation performance in biomedical imaging [23,24,25].

Despite these advances, existing SSL frameworks still face challenges related to scanner variability, noise, and domain shift, which can degrade representation stability across heterogeneous MRI sources. Moreover, most MAE-based models focus primarily on reconstruction objectives and do not explicitly incorporate domain-adaptive denoising or latent feature augmentation mechanisms. Addressing these limitations, the proposed Model-Augmented MAE framework integrates domain-adaptive denoising and multi-level feature fusion to improve reconstruction fidelity, enhance representation robustness, and achieve better generalization across diverse MRI datasets.

III. Proposed Framework

Brain tumor MRI classification demands robustness against variations in scanners, acquisition protocols, tumor heterogeneity, and noise. A single model or feature type is insufficient to ensure generalization. Therefore, the proposed design integrates complementary modules, each addressing a specific challenge:

- Masked Autoencoder (MAE) [16]: learns structure-preserving self-supervised features even with limited annotated datasets.
- Domain-Adaptive Denoising (DAD): stabilizes latent features across multi-center scans.
- Multi-level feature extraction combines radiomics (handcrafted biomarkers), CNN (deep hierarchical semantics), and shape descriptors (morphology).
- Attention-guided fusion removes redundancy across modalities and retains clinically meaningful patterns.
- PyQDCNN: models long-range tumor–background relationships through dilation- and entanglement-inspired kernels.
- Hierarchical classification: reduces inter-class interference for subtle tumor subtypes.

Among these components, the Model-Augmented MAE and Domain-Adaptive Denoising (DAD) constitute the core contributions of the framework, while the remaining modules primarily enhance feature representation and classification performance. The effectiveness of these modules is quantitatively validated through the dataset-level ablation studies presented in Table I to Table IV and the component-level ablation analysis in Table V, which demonstrate their individual and cumulative contributions to overall framework performance. Together, these elements create a multi-resolution, multi-representation, and domain-robust tumor analysis framework aligned with clinical practice.

The MRI input first undergoes a quality-control stage, where low-quality or corrupted scans are identified and excluded from further processing. The accepted volumes are geometrically normalized by reorienting to RAS space, resampling to isotropic resolution, and cropping or padding the field of view to maintain structural consistency. Subsequent bias-field correction (N4) and brain extraction [4] (skull stripping) remove intensity inhomogeneities and non-brain tissues, respectively. Intensity standardization is then performed through z-score normalization and histogram matching to harmonize the contrast distribution across subjects and scanners [4].

Following conventional pre-processing, the data are passed through a Model-Augmented Masked Autoencoder (MAE) framework [16] that leverages self-supervised representation learning. In this module, image patches are randomly masked and encoded via a Vision Transformer–based encoder (ViT-E) [11,12,13,14] to produce latent token embeddings z. The decoder reconstructs the masked patches, and reconstruction loss is computed as mean-squared error [3] on the masked regions, promoting structure-aware feature learning. Mean squared error (MSE) is adopted due to its computational efficiency and stability in preserving global structural consistency during reconstruction. Alternative perceptual or SSIM-based losses were not employed to avoid additional complexity and instability during self-supervised training. The MAE outputs both a reconstructed (denoised) image and a latent feature bank (z) used in subsequent optimization and augmentation stages. Z-score normalization was applied to reduce inter-subject intensity variability, while histogram matching ensured cross-scanner contrast harmonization. This dual-stage standardization was found to produce more stable latent representations in the MAE encoder compared to other methods such as min-max normalization or percentile clipping, especially when handling multi-center MRI datasets. This observation is further supported by the ablation results presented in Section IV, where improved convergence stability and feature consistency are observed across multi-center datasets. Depending on the downstream configuration, three data routes are available:

Route A: pre-processed image only,

Route B: MAE-enhanced reconstructed image, and

Route C: MAE latent representation only, the D-dimensional token embeddings (z) from the ViT encoder (optionally aggregated), which encapsulate self-supervised, structure-aware MRI priors. This route isolates the contribution of the MAE's learned representation, independent of raw or reconstructed image inputs.

While the augmented image set (pre-processed + reconstructed) is used to extract radiomics, CNN, and shape features (as shown in Fig. 3), Route C refers specifically to the direct use of the MAE latent vector z as an input stream typically fed into the attention-guided fusion module alongside other extracted features.

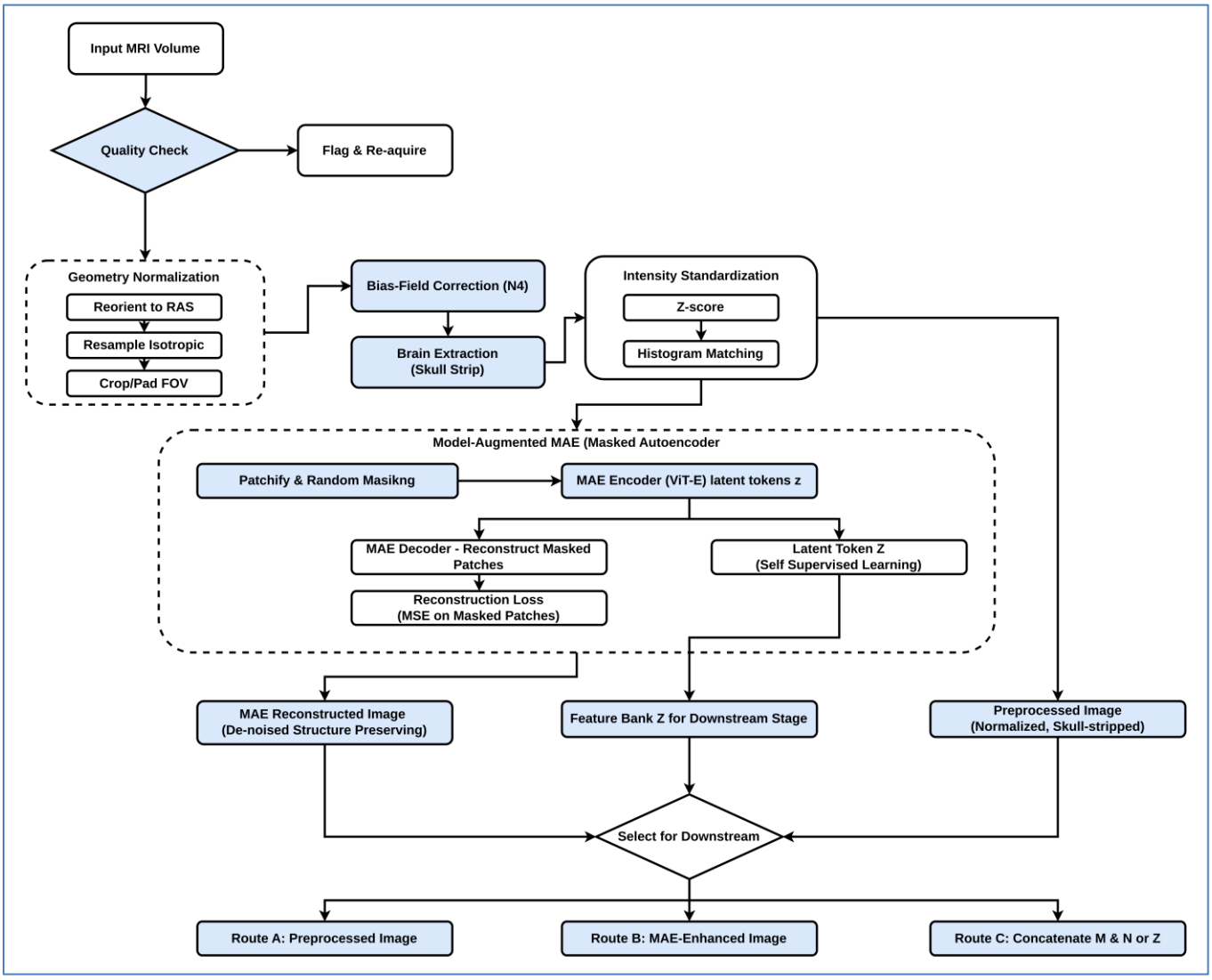


Fig. 1. Input and pre-processing pipeline using Model-Augmented Masked Autoencoder (MAE).

Following the generation of pre-processed images, reconstructed outputs, and latent tokens in Fig. 1, the internal functioning of the proposed MAE is detailed in Fig. 2, which expands the encoder–decoder architecture responsible for learning structure-aware, domain-robust MRI representations. Fig. 2 illustrates the internal architecture of the proposed Vision Transformer–based Masked Autoencoder (ViT-MAE), which is composed of five sequential stages: patch embedding, masking, transformer encoding, domain-adaptive denoising, and transformer decoding with reconstruction. The MRI slice is first divided into fixed-size non-overlapping patches that are flattened and linearly projected into D-dimensional token embeddings, followed by the addition of learnable two-dimensional positional encodings. A random or stratified masking strategy is then applied, in which a large portion of patch tokens is removed to create a subset of visible tokens used for self-supervised learning. These visible tokens are processed by a multi-layer Vision Transformer encoder consisting of LayerNorm, multi-head self-attention, and MLP feed-forward blocks, producing both multi-scale snapshot features and a final latent representation. To enhance robustness across scanners and acquisition conditions, the encoder tokens pass through a Domain-Adaptive Denoising (DAD) module that performs channel-wise normalization, patch smoothing, and gated fusion, generating denoised latent features. The design of the DAD module is motivated by the need to reduce scanner-induced feature variance while preserving anatomically relevant structures, thereby improving cross-domain robustness without introducing significant computational overhead [27], [28]. These refined tokens are projected into the decoder dimension, augmented with decoder positional embeddings, and combined with learnable mask tokens inserted at the original masked positions. The full token sequence is then passed into the Vision Transformer decoder, which applies self-attention and cross-attention to reconstruct the missing patches. A final linear output head predicts pixel values for each masked patch, and the reconstructed patches are assembled back into a complete MRI slice. The reconstruction error, computed only on the masked regions, drives the model to learn semantically meaningful and structure-preserving representations that serve as the foundation for downstream analysis.

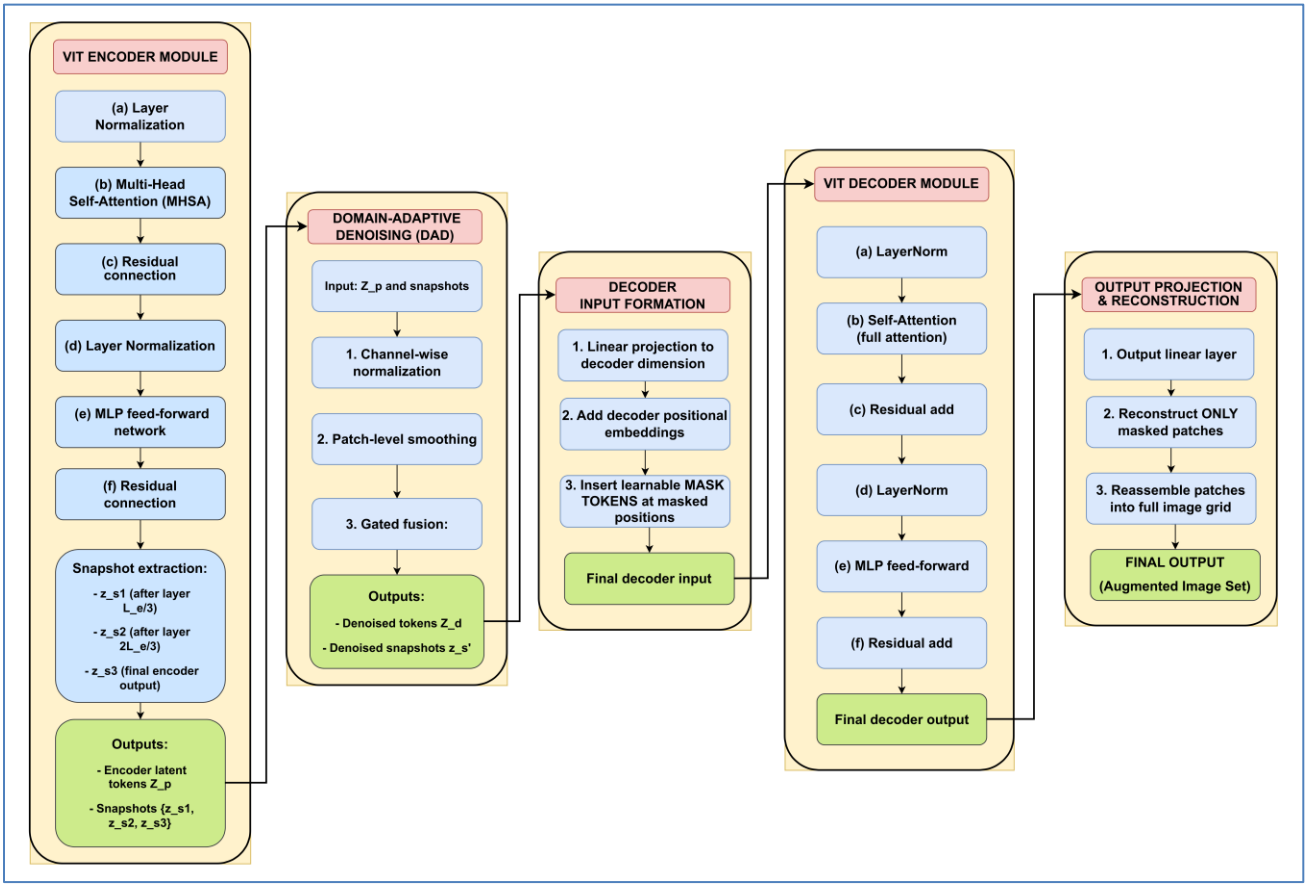


Fig. 2. ViT-based encoder–decoder architecture.

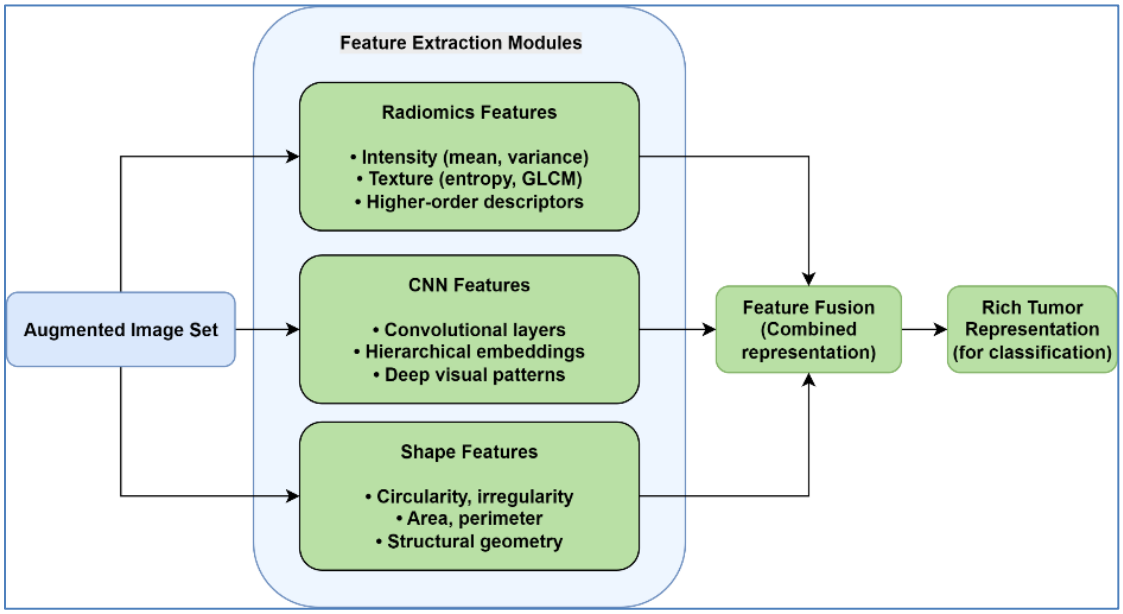


Fig. 3. Multi-level feature extraction and fusion for rich tumor representation.

Following processing through the ViT-based encoder–decoder architecture, the MAE produces multiple complementary image representations including the pre-processed MRI, the MAE-reconstructed image, and the latent feature maps derived from the transformer encoder and DAD module. These representations form the augmented image set that is forwarded into the multi-level feature extraction stage [18,19,20,21,22] illustrated in Fig. 3. By providing several

structurally distinct yet semantically aligned inputs, the augmented set enriches the available feature space, ensuring that both native anatomical details and MAE-enhanced latent cues are preserved. Within this stage, each representation is leveraged to compute three categories of features: radiomics descriptors that capture statistical and textural biomarkers, CNN-derived deep features that encode hierarchical visual patterns, and geometric shape descriptors that quantify morphological characteristics of the tumor. Notably, the MAE-derived latent tokens z supply context-aware structural priors that improve the reliability and discriminative strength of downstream feature extraction. Their inclusion helps stabilize tumor characterization across scanners and patient cohorts, resulting in a more robust and comprehensive tumor feature representation after feature fusion. A patch size of p×p was selected after empirical evaluation across p∈{8,16,32}. We observed that smaller patches improved reconstruction fidelity but significantly increased token count and memory consumption, whereas larger patches lost fine-grained tumor boundaries. The chosen patch dimension provides an optimal trade-off between spatial detail preservation and computational efficiency. The description of the masking strategy and reconstruction mechanism has been consolidated into a single unified paragraph to eliminate redundancy and improve methodological clarity. The revised explanation succinctly integrates patch embedding, token masking, transformer encoding, and reconstruction into a coherent flow.

Following the multi-level feature extraction stage, the resulting radiomics, CNN-derived, and shape descriptors are integrated into the core network architecture illustrated in Fig. 4.

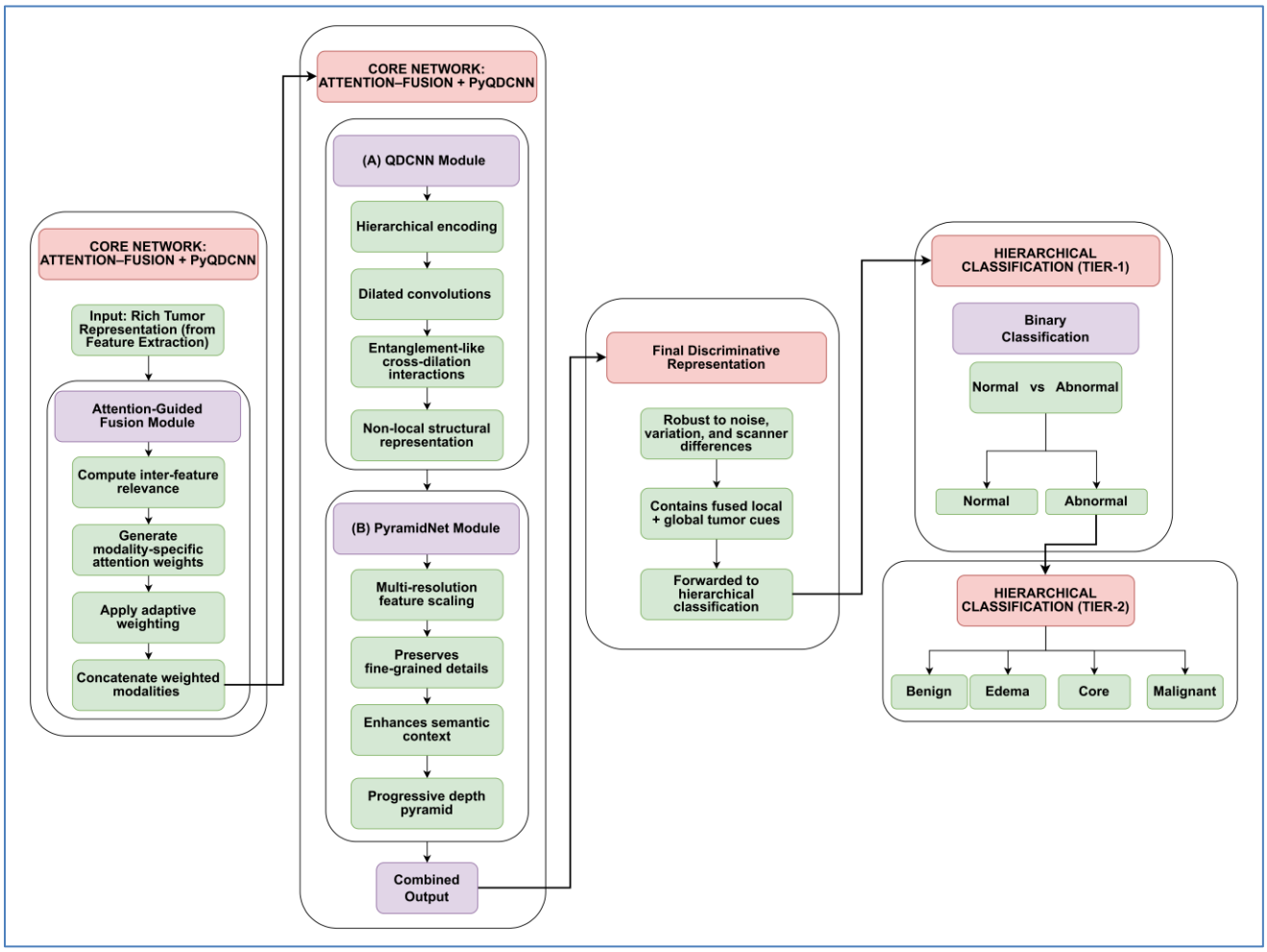


Fig. 4. Proposed framework integrating attention-guided feature fusion and the Pyramid Quantum-Inspired Dilated Convolutional Neural Network (PyQDCNN).

Fig. 4 presents the final stage of the proposed framework, integrating attention-guided feature fusion, the Pyramid Quantum-Inspired Dilated Convolutional Neural Network (PyQDCNN), and a hierarchical tumor classification strategy to generate clinically meaningful brain tumor MRI predictions. The concatenated radiomics, CNN, and shape descriptors from the previous stage are first processed through an attention-guided fusion module, which assigns modality-specific weights based on inter-feature correlations, thereby emphasizing informative tumor cues while suppressing redundant variations. The weighted multi-source features are then passed to the PyQDCNN backbone, which combines dilated convolutional processing with an entanglement-inspired bilinear feature interaction mechanism designed to model cross-channel and cross-scale dependencies [23,24,25]. The QDCNN component models long-range spatial dependencies through dilation-driven

feature entanglement, while the PyramidNet component enhances multi-resolution detail retention across deeper semantic layers. The resulting fused representation is both structurally rich and noise-resilient, forming the input to a two-tier hierarchical classification system. In the first tier, the model performs a binary decision to distinguish normal from abnormal MRI scans. Samples identified as abnormal are forwarded to the second tier, where a four-class classification predicts tumor subtypes: benign, edema, core, and malignant. This hierarchical formulation reduces inter-class ambiguity, improves discrimination among visually similar tumor types, and aligns with clinical diagnostic workflows. The final output provides robust and precise multi-class tumor identification, demonstrating the effectiveness of the combined attention–PyQDCNN–hierarchical pipeline for comprehensive brain tumor MRI analysis.

A. Mathematical Formulation of the Proposed Framework

Notation: Let:

- I denote an input MRI volume (or slice) and $I(x)$ its intensity at voxel x.
- I_{norm} geometric-normalized image, I_{corr} bias-field corrected image, I_{brain} skull-stripped image, I_{std} intensity-standardized image.
- p patch side length (so patches are $p \times p$).
- N total number of non-overlapping patches in one slice: $N = \frac{H \cdot W}{p^2}$.
- x_i the i-th patch (flattened), $i \in \{1, \dots, N\}$.
- $W_e \in \mathbb{R}^{D \times p^2}$ encoder patch projection; D the token dimension.
- $t_i \in \mathbb{R}^D$ token embedding for patch x_i.
- $\rho \in (0,1)$ masking ratio (fraction of patches masked).
- $\mathcal{M} \in \{0,1\}^N$ mask indicator vector; $\mathcal{M}(i) = 1$ if patch i is masked.
- $T_v = \{t_i \mid \mathcal{M}(i) = 0\}$ visible tokens.
- $\text{ViT}_{enc}(\cdot)$ encoder mapping producing latent tokens z.
- $z = \{z_i\}_{i=1}^{N_v}$, $z_i \in \mathbb{R}^D$, $N_v = |T_v|$.
- DAD denotes the Domain-Adaptive Denoising module; output tokens z^{dad}.
- $\text{ViT}_{dec}(\cdot)$ decoder mapping producing reconstructed patches $\hat{x}_i$.
- Radiomics feature vector R, CNN deep feature vector C, shape descriptor vector S.
- Fused feature vector F, attended fused feature F_{att}.
- PyQDCNN mapping Φ_{PyQ} producing final representation F_{final}.
- Tier-1 binary output $\hat{y}_1 \in [0,1]$; Tier-2 multi-class vector $\hat{y}_2 \in \Delta^{K-1}$ (with $K = 4$ tumor classes).
- Ground-truth labels y_1 (binary) and y_2 (one-hot multi-class).

1) Preprocessing (deterministic ops)

Geometry reorientation and isotropic resampling [see Eq. (1)]:

$$I_{norm} = \mathcal{R}(\mathcal{O}(I), R_{\text{iso}}) \quad (1)$$

where, $\mathrm{O}(\cdot)$ denotes orientation alignment, R_{iso} represents isotropic resampling, and $\mathrm{R}(\cdot)$ denotes the composite geometric normalization operation.

N4 bias field correction (multiplicative model) [see Eq. (2)]:

$$I_{corr}(x) = \frac{I_{norm}(x)}{B(x)} \quad \text{with } B(\cdot) \text{ estimated by N4} \quad (2)$$

where, $B(x)$ denotes the estimated bias field obtained using the N4 correction algorithm.

Skull-stipping (brain mask $M_{brain}(x) \in \{0,1\}$) [see Eq. (3)]:

$$I_{brain}(x) = I_{corr}(x) \cdot M_{brain}(x) \quad (3)$$

Intensity standardization dual stage [see Eq. (4) and Eq. (5)]:

$$I_z(x) = \frac{I_{brain}(x) - \mu_{brain}}{\sigma_{brain}} \quad (4)$$

$$I_{std} = \text{HistMatch}(I_z, I_{\text{ref}}) \quad (5)$$

where, $\mu_{brain}, \sigma_{brain}$ are the mean and std computed within brain mask, and $\text{HistMatch}(\cdot)$ denotes histogram matching to a reference image I_{ref}.

2) Patch embedding and masking

Patch extraction and linear projection [see Eq. (6) and Eq. (7)]:

$$x_i = \text{PatchExtract}(I_{std}, i), \qquad x_i \in \mathbb{R}^{p^2}, \quad i = 1, \dots, \quad (6)$$

$$t_i = W_e x_i + p_i, \qquad t_i \in \mathbb{R}^D \quad (7)$$

where, p_i is a learnable 2D positional encoding (flattened to $\mathbb{R}^D$) for patch i.

Random (or stratified) masking [see Eq. (8)]:

$$\mathcal{M}(i) \sim \text{Bernoulli}(\rho), \qquad T_v = \{t_i \mid \mathcal{M}(i) = 0\} \quad (8)$$

3) Vision-Transformer encoder

The encoder is a stack of L_e Transformer blocks. Each block performs LayerNorm, Multi-Head Self-Attention (MHSA) and MLP with residuals. For a generic token sequence U, one Transformer block is [see Eq. (9) and Eq. (10)]:

$$U' = U + \text{MHSA}(\text{LN}(U)) \quad (9)$$

$$U^{\text{out}} = U' + \text{MLP}(\text{LN}(U')) \quad (10)$$

Apply to visible tokens T_v (with ordering preserved) to obtain encoder outputs z [see Eq. (11)]:

$$z = \text{ViT}_{enc}(T_v) \in \mathbb{R}^{N_v \times D} \quad (11)$$

Let snapshot features $z^{(s)}$ denote intermediate-layer outputs (optionally used for multi-scale priors).

4) Domain-Adaptive Denoising (DAD)

DAD refines latent tokens z to produce denoised tokens z^{dad}. It comprises three steps: channel-wise normalization, patch-level smoothing, and gated fusion.

a) Channel-wise normalization

For channel (dimension) $c \in \{1, \dots, D\}$ [see Eq. (12)]:

$$\tilde{z}_{:,c} = \frac{z_{:,c} - \mu_c}{\sigma_c}, \qquad \mu_c = \frac{1}{N_v}\sum_i z_{i,c}\,,\ \sigma_c = \sqrt{\frac{1}{N_v}\sum_i (z_{i,c} - \mu_c)^2} \tag{12}$$

b) Patch-level smoothing

Apply local smoothing (e.g., 1D convolution over token index or Gaussian kernel in patch-grid layout). Let K be a smoothing kernel, where K(·) denotes a local smoothing operator, implemented as either a 1D convolution over token sequences or a Gaussian kernel over the patch grid:

$z^{\text{smooth}} = \mathcal{K}(\tilde{z})$ (tokenwise smoothing operator).

c) Gated fusion

Compute gating scalar for each token (element-wise on token vector) [[see Eq. (13) and Eq. (14)]:

$$\alpha_i = \sigma\big(W_g z_i + b_g\big) \in (0,1)^D \tag{13}$$

$$z_i^{\text{dad}} = \alpha_i \odot z_i^{\text{smooth}} + (1 - \alpha_i) \odot z_i,\ i = 1, \dots, N_v, \tag{14}$$

where, $\odot$ is element-wise product and $\sigma(\cdot)$ is sigmoid.

DAD parameters W_g, b_g are learned during training.

The computational complexity of the DAD module is linear with respect to the number of tokens and feature dimensions, i.e., $O(N_v$ D), as it involves channel-wise normalization, lightweight smoothing, and element-wise gated fusion. This ensures minimal overhead compared to the transformer encoder.

5) Decoder and reconstruction

Decoder input formation: project denoised visible tokens to decoder dimension D_d and insert learned mask tokens t_{mask} at masked positions to form full-length sequence $\tilde{T} \in \mathbb{R}^{N \times D_d}$ [see Eq. (15)]:

$$\tilde{T} = \text{InsertMaskTokens}\big(\text{Proj}_d(z^{\text{dad}}),\ \mathcal{M},\ t_{\text{mask}}\big) \tag{15}$$

where, Proj_d denote learnable linear projection functions used to align feature dimensions.

Decoder (stack of L_d Transformer blocks) reconstructs patches [see Eq. (16)]:

$$\hat{X} = \text{ViT}_{dec}(\tilde{T}), \qquad \hat{x}_i = \text{OutHead}(\hat{T}_i) \in \mathbb{R}^{p^2} \tag{16}$$

where, $\hat{T}_i$ is the decoder output token at patch i, and OutHead is a linear mapping to pixel intensities.

MAE reconstruction loss (computed only over masked patches set $\mathbb{M} = \{i \mid \mathcal{M}(i) = 1\}$) [see Eq. (17)]:

$$\mathcal{L}_{\text{MAE}} := \frac{1}{|\mathbb{M}|}\sum_{i \in \mathbb{M}} \| x_i - \hat{x}_i \|_2^2 \tag{17}$$

Optionally, you can use normalized MSE or perceptual losses as augmentations. The reconstruction loss is computed only over masked patches to encourage the model to learn meaningful contextual representations rather than trivial identity mappings.

6) Multi-level feature extraction

From augmented image set $\mathcal{I}_{aug} = \{I_{std}, \hat{I}\}$ and latent tokens z^{dad}, extract:

- Radiomics: $R = \Psi_R(\mathcal{I}_{aug}) \in \mathbb{R}^{d_R}$. (e.g., intensity, GLCM, higher-order features).
- CNN features: $C = \Psi_C(\mathcal{I}_{aug}) \in \mathbb{R}^{dc}$. (learned convolutional backbone outputs).
- Shape descriptors: $S = \Psi_S(\mathcal{I}_{aug}) \in \mathbb{R}^{d_S}$.
- MAE latent summary: $z_{\text{agg}} = \text{Pool}(z^{\text{dad}}) \in \mathbb{R}^D$ (e.g., mean pooling or learned aggregator).

where, Pool(·) denotes a global aggregation function such as mean pooling or adaptive average pooling.

Concatenate to form raw fused vector [see Eq. (18)]:

$$F = \text{Concat}\big(R,\ C,\ S,\ z_{\text{agg}}\big) \in \mathbb{R}^{d_F} \tag{18}$$

7) Attention-guided fusion

Compute attention scores per modality/feature-group to re-weight elements. Suppose features are partitioned into m modality blocks $F^{(j)}, j = 1..m$. Let projection vectors be θ_j.

Attention weights [see Eq. (19)]:

$$a_j = \frac{\exp\big(\theta_j^\top \text{LN}(F^{(j)})\big)}{\sum_{k=1}^{m} \exp\big(\theta_k^\top \text{LN}(F^{(k)})\big)}, \quad j = 1 \dots m \tag{19}$$

Attended fused feature [see Eq. (20)]:

$$F_{att} = \sum_{j=1}^{m} a_j \cdot \text{Proj}_j\big(F^{(j)}\big) \in \mathbb{R}^{d_{att}} \tag{20}$$

where, Proj_j denotes learnable linear projection functions used to align feature dimensions. LN(·) is LayerNorm.

8) Pyramid Quantum Dilated CNN (PyQDCNN)

PyQDCNN processes F_{att} (reshaped or tiled to a spatial grid if needed) through a series of pyramid stages $l = 1, \dots, L$. Each stage combines dilated convolutions and an entanglement-inspired mixing.

For an input feature map $U^{(l-1)}$, the PyQ layer computes:

a) Dilated convolutional response

To model long-range interaction patterns, PyQDCNN applies dilated convolutions as [see Eq. (21)]:

$$D^{(l)} = \sum_r W_r^{(l)} *_{d_r} U^{(l-1)} \tag{21}$$

where, $*_{d_r}$ denotes convolution with dilation d_r and $W_r^{(l)}$ are learned kernels. The set $\{d_r\}$ spans small to large receptive fields (e.g., $d_r \in \{1,2,4,8\}$).

b) Entanglement-like mixing (quantum-inspired term)

Define a bilinear entanglement operator $\mathcal{E}^{(l)}$ [see Eq. (22)]:

$$E^{(l)} = \gamma^{(l)} \cdot \varphi\big(U^{(l-1)} \otimes W_e^{(l)}\big) \tag{22}$$

where,

- $\otimes$ denotes a tensor outer-product mixing across channels/positions (learnable low-rank factorization recommended for efficiency),

- $W_e^{(l)}$ is entanglement weight tensor,
- φ is a nonlinearity (e.g., ReLU),
- $\gamma^{(l)}$ Scales entanglement contribution.

This term captures cross-channel and cross-scale interactions analogous to entanglement in quantum circuits but implemented as learned bilinear mixing.

In practice, the outer-product interaction is approximated using low-rank factorization to reduce computational complexity from $O(C^2)$ to $O(Ck)$, where C is the channel dimension and k ≪ C. This makes the entanglement operation computationally feasible while retaining cross-channel interaction capability.

c) Stage output and pyramid aggregation

Combine [see Eq. (23)]:

$$U^{(l)} = \mathrm{BN}\big(D^{(l)} + E^{(l)}\big), \quad U^{(l)} \leftarrow U^{(l)} + \mathrm{ResProj}(U^{(l-1)}) \tag{23}$$

where, BN is batchnorm (or LayerNorm), and residual projection aligns dimensions.

Pyramid aggregation across scales [see Eq. (24)]:

$$F^{\mathrm{py}} = \mathrm{Concat}\big(\mathrm{Pool}(U^{(1)}), \mathrm{Pool}(U^{(2)}), \dots, \mathrm{Pool}(U^{(L)})\big) \tag{24}$$

Finally [see Eq. (25)]:

$$F_{final} = \mathrm{FC}(\mathrm{GAP}(F^{\mathrm{py}})) \in \mathbb{R}^{d_{final}} \tag{25}$$

where, GAP(·) denotes global average pooling, $FC()$ represents a fully connected layer.

Notes on implementation:

- Use separable or grouped convolutions and low-rank factorization for the entanglement term to keep parameter counts reasonable.
- Choose dilation schedule to avoid gridding artifacts (e.g., combine dilations that are coprime).

9) Hierarchical classification

Tier-1 (binary normal vs. abnormal) [see Eq. (26)]:

$$s_1 = w_1^{\top} F_{final} + b_1, \qquad \hat{y}_1 = \sigma(s_1) \tag{26}$$

where, σ(·) denotes the sigmoid function.

Tier-2 (K-class, here $K = 4$ subtypes) — only evaluated if $\hat{y}_1$ indicates abnormal (training computes both) [see Eq. (27)]:

$$s_2 = W_2 F_{final} + b_2, \qquad \hat{y}_2 = \mathrm{Softmax}(s_2) \tag{27}$$

where, Softmax(·) denotes the multi-class normalization function.

10) Losses, training objective, and optimization

Binary cross-entropy (BCE) loss for Tier-1 [see Eq. (28)]:

$$\mathcal{L}_{\mathrm{bin}} = -[y_1 \log \hat{y}_1 + (1 - y_1)\log(1 - \hat{y}_1)] \tag{28}$$

Multi-class cross-entropy for Tier-2 [see Eq. (29)]:

$$\mathcal{L}_{\mathrm{multi}} = -\sum_{c=1}^{K} y_{2c} \log \hat{y}_{2c} \tag{29}$$

Optional regularization terms:

- Weight decay $\mathcal{R}_w = \| \Theta \|_2^2$.
- Consistency/regression loss on MAE latent alignment (if used).

Total joint loss:

All components are jointly optimized using the composite objective in Eq. (30), enabling the framework to learn both reconstruction-driven structural priors and high-level discriminative representations.

$$\mathcal{L}_{\mathrm{total}} = \lambda_{\mathrm{MAE}}\mathcal{L}_{\mathrm{MAE}} + \lambda_{\mathrm{bin}}\mathcal{L}_{\mathrm{bin}} + \lambda_{\mathrm{multi}}\mathcal{L}_{\mathrm{multi}} + \lambda_w \mathcal{R}_w \tag{30}$$

Training procedure notes:

- Two-stage or end-to-end options:
  - **Stage-wise**: Pre-train MAE (minimize $\mathcal{L}_{\mathrm{MAE}}$), then freeze/finetune encoder and train downstream network minimizing $\mathcal{L}_{\mathrm{bin}} + \mathcal{L}_{\mathrm{multi}}$.
- **End-to-end**: Jointly optimize Eq. (30) with smaller λ_{MAE} after initial MAE warm start.
- Optimizer: AdamW with learning rate schedule (warmup + cosine decay) recommended.

IV. Experimental Setup and Evaluation

The experiments in this study were conducted using four publicly available brain MRI datasets that collectively cover benign–malignant tumor classification, binary tumor detection, and multi-regional tumor segmentation. The Brain Tumor MRI Dataset (M. Nickparvar) provides 7,023 T1-weighted axial MRI slices labeled at the image level into glioma (malignant), meningioma (benign), pituitary tumors, and normal classes, enabling robust benign–malignant categorization. The Brain MRI Images for Tumor Detection (Navoneel) dataset contains 1,550 axial MRI slices labeled as tumor or normal, supporting binary abnormality detection. The Brain Tumor Segmentation and Classification dataset (indk214) includes approximately 3,200 images derived from 800 multi-sequence MRI volumes (T1, T1CE, T2, FLAIR), each annotated with pixel-wise tumor masks identifying core, edema, and enhancing malignancy regions. Finally, the BraTS 2020 Challenge dataset provides 3,500 multi-sequence MRI volumes with high-quality voxel-wise annotations for tumor core, edema, and enhancing regions, serving as the benchmark for multi-regional segmentation and malignant tumor analysis. Together, these datasets provide a diverse and comprehensive foundation for evaluating tumor classification, segmentation, and representation robustness across imaging contrasts, scanners, and annotation types.

All experiments were conducted in a high-performance deep-learning environment configured with Python, PyTorch, and CUDA acceleration. Pre-processing, MAE training, and downstream classification modules were executed on an NVIDIA GPU workstation to ensure efficient handling of high-resolution MRI volumes and transformer-based architecture. The environment was optimized for reproducibility, with fixed random seeds, controlled data pipelines, and standardized

preprocessing routines including N4 correction, skull stripping, and intensity harmonization.

The MAE framework is seamlessly integrated into the downstream PyQDCNN–ViT pipeline by supplying both reconstructed images and latent token embeddings to later stages. While the reconstructed outputs enhance image fidelity and reduce scanner-induced noise, the latent representations provide compact, structure-aware features that strengthen feature fusion. These MAE-derived signals are combined with radiomics, CNN, and shape features [18,19,20,21,22] and passed into the attention-guided fusion module, enabling PyQDCNN and ViT components to operate with enriched, domain-robust representations for improved tumor classification accuracy.

The Model-Augmented MAE was pretrained on these four benchmark MRI datasets using self-supervised learning. Following pretraining, the learned representations were transferred to the PyQDCNN–ViT hybrid model for downstream tumor classification. Ablation experiments evaluated the contribution of MAE pretraining to performance improvement across multiple datasets. Results indicate that MAE pretraining yields accuracy improvements ranging from approximately 1.0–1.4% across all evaluated datasets, while the complete framework achieves overall accuracy improvements ranging from approximately 2.7% to 3.7% across the evaluated datasets. Notably, feature embeddings from the MAE demonstrate increased domain invariance and smoother convergence during fine-tuning.

To maintain consistency across datasets, all models were trained using unified hyperparameters. All baseline models were trained under identical preprocessing, training schedules, and optimization settings to ensure fair comparison. MAE pretraining was carried out for 150 epochs using AdamW with a learning rate of 1×10^{-4}, a 15-epoch warm-up, cosine decay, and a batch size of 16. The downstream PyQDCNN–ViT classifier was trained for 80–100 epochs with a learning rate of 3×10^{-4} and a batch size of 32. Each dataset was split into 70% training, 15% validation, and 15% testing, ensuring patient-level separation. Data augmentation included small-angle rotations, flips, elastic deformation, Gaussian noise, and contrast adjustments to improve generalization. The full framework comprises approximately 54 million parameters. This parameter scale is comparable to modern hybrid CNN–Transformer architectures and remains lower than large-scale Vision Transformer models, indicating a balanced trade-off between representational capacity and computational efficiency for medical imaging tasks. While effective, the multi-stage design increases computational load, and MAE reconstruction may oversmooth fine tumor boundaries in low-contrast scans. However, the additional computational cost is primarily incurred during training, while inference remains efficient due to the streamlined forward pipeline. Failure-case analysis shows that errors mostly occur in cases with very small lesions, motion artefacts, or atypical anatomical distortions, where subtle tumor cues are harder to preserve.

Table I to Table IV and Fig. 5 demonstrate consistent performance improvements across four brain tumor MRI datasets. While the incremental contribution of MAE pre-training (A3 → A4) provides stable gains of approximately 1.0–1.4% in accuracy, the PyQDCNN–ViT framework with MAE pre-training (A1 → A4) achieves overall improvements ranging from approximately 2.7% to 3.7% across the evaluated datasets. This indicates that MAE, when integrated with PyQDCNN and ViT, contributes to both incremental performance enhancement and significant overall robustness gains. To further validate statistical significance, paired t-tests were conducted between the baseline (A3) and MAE-enhanced (A4) models across all datasets. The improvements were found to be statistically significant ($p < 0.05$), confirming that the observed gains are not due to random variation.

TABLE I. ABLATION RESULTS: BRAIN TUMOR MRI DATASET (MASOUD NICKPARVAR)

Variant ID	Model Configuration	ACC (%)	F_1 (%)	AUROC	p-value (A3 vs A4)
A1	PyQDCNN only	91.2 ± 0.8	90.5 ± 0.8	0.955 ± 0.006	-
A2	ViT only	91.8 ± 0.7	91.1 ± 0.8	0.959 ± 0.005	-
A3	PyQDCNN + ViT (no MAE)	92.9 ± 0.6	92.1 ± 0.7	0.966 ± 0.004	-
A4	+ MAE pre-training	93.9 ± 0.5	93.2 ± 0.5	0.972 ± 0.003	0.012

TABLE II. ABLATION RESULTS: BRAIN MRI IMAGES FOR TUMOR DETECTION (NAVONEEL)

Variant ID	Model Configuration	ACC (%)	F_1 (%)	AUROC	p-value (A3 vs A4)	
A1	PyQDCNN only	90.9 ± 0.9	90.1 ± 0.9	0.953 ± 0.007	-	
A2	ViT only	91.6 ± 0.8	90.8 ± 0.8	0.958 ± 0.006	-	
A3	PyQDCNN + ViT (no MAE)	92.7 ± 0.6	91.9 ± 0.7	0.965 ± 0.004	-	
A4	+ MAE pre-training	93.6 ± 0.5	92.8 ± 0.6	0.971 ± 0.003	0.018	

TABLE III. ABLATION RESULTS: BRAIN TUMOR SEGMENTATION AND CLASSIFICATION (INDK214)

Variant ID	Model Configuration	ACC (%)	F_1 (%)	Dice (core / edema / enh.)	AUROC	p-value (A3 vs A4)
A1	PyQDCNN only	89.7 ± 0.9	88.9 ± 1.0	0.73 / 0.66 / 0.69	0.946 ± 0.006	-
A2	ViT only	90.8 ± 0.8	90.0 ± 0.9	0.76 / 0.70 / 0.72	0.953 ± 0.005	-
A3	PyQDCNN + ViT (no MAE)	92.0 ± 0.6	91.3 ± 0.7	0.78 / 0.72 / 0.74	0.965 ± 0.004	-
A4	+ MAE pre-training	93.4 ± 0.5	92.6 ± 0.6	0.80 / 0.74 / 0.76	0.972 ± 0.003	0.009

TABLE IV. ABLATION RESULTS: BRATS 2020 CHALLENGE DATASET

Variant ID	Model Configuration	ACC (%)	F_1 (%)	Dice (core / edema / enh.)	AUROC	p-value (A3 vs A4)
A1	PyQDCNN only	90.7 ± 0.9	89.9 ± 0.8	0.75 / 0.67 / 0.70	0.953 ± 0.006	-
A2	ViT only	91.2 ± 0.8	90.4 ± 0.9	0.78 / 0.71 / 0.73	0.957 ± 0.005	-
A3	PyQDCNN + ViT (no MAE)	92.4 ± 0.6	91.6 ± 0.7	0.80 / 0.73 / 0.75	0.965 ± 0.004	-
A4	+ MAE pre-training	93.5 ± 0.5	92.8 ± 0.5	0.82 / 0.75 / 0.77	0.973 ± 0.003	0.015

In addition to the dataset-specific ablation studies reported in Table I to Table IV, a separate component-level ablation analysis was conducted to quantify the relative contribution of individual framework modules. While Table I to Table IV focus on the effect of MAE pretraining within the PyQDCNN–ViT classification pipeline across four benchmark datasets, Table V evaluates the cumulative contribution of the major architectural components of the complete framework, including the Domain-Adaptive Denoising (DAD) module, radiomics features, shape descriptors, attention-guided fusion, and PyQDCNN. Consequently, the results reported in Table V serve a different objective and should not be interpreted as direct dataset-level comparisons with Table I to Table IV.

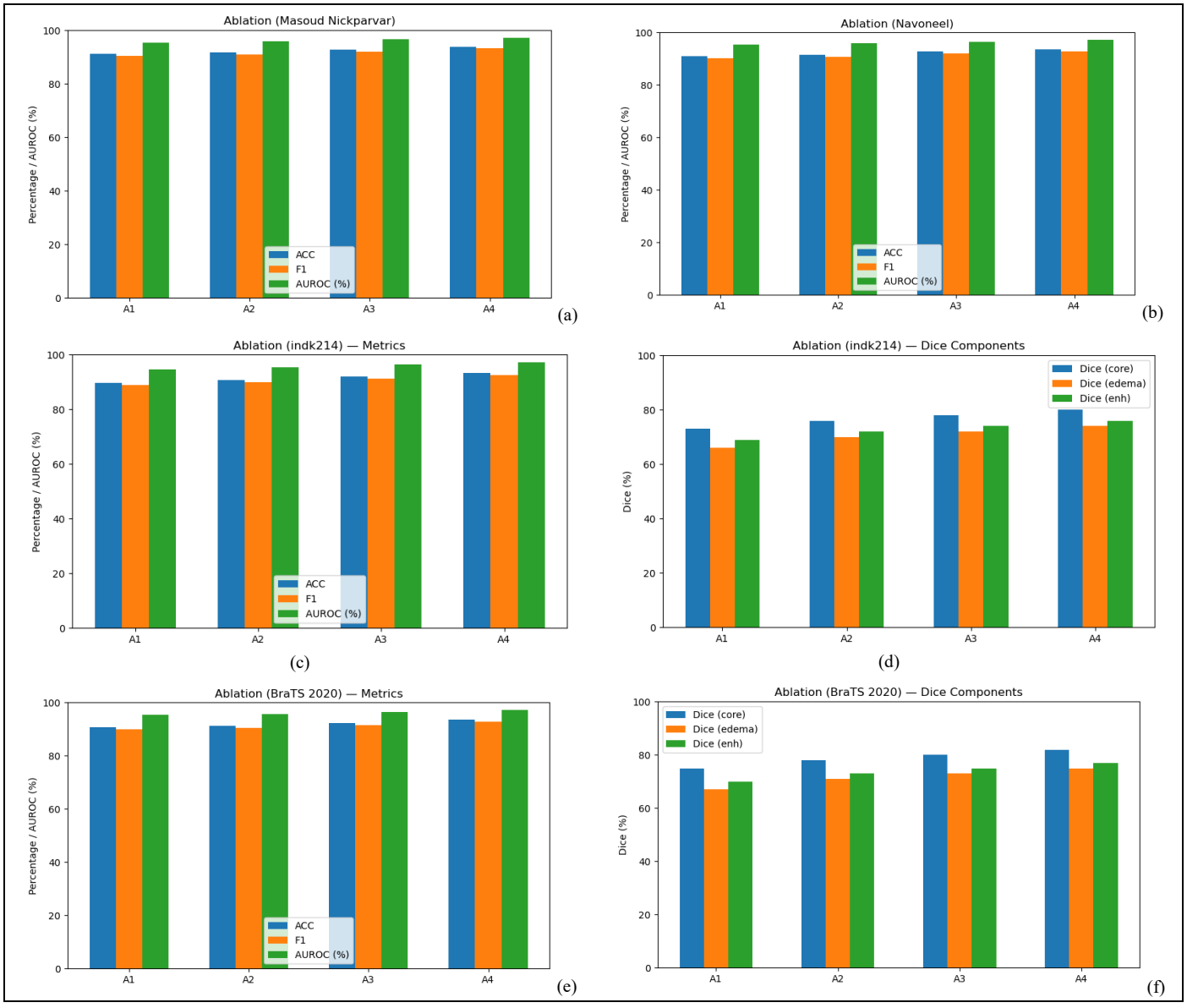


Fig. 5. Component-wise ablation analysis across four brain-tumor MRI datasets.

TABLE V. ABLATION STUDY OF COMPONENTS IN THE PROPOSED FRAMEWORK

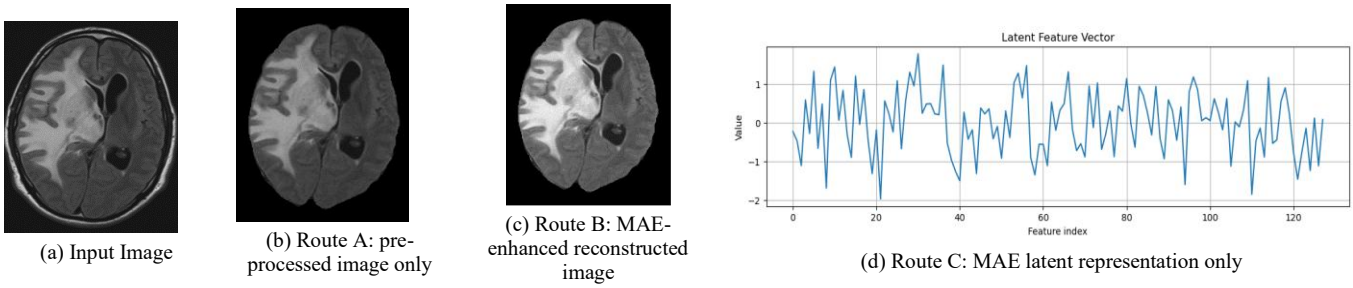
Configuration	MAE	DAD	Radiomics	CNN	Shape	Attention Fusion	PyQDCNN	Acc (%)	Dice
A1: Baseline CNN	✗	✗	✗	✓	✗	✗	✗	81.2	0.74
A2: MAE Only	✓	✗	✗	✗	✗	✗	✗	83.5	0.76
A3: MAE + DAD	✓	✓	✗	✗	✗	✗	✗	86.1	0.79
A4: MAE + CNN	✓	✓	✗	✓	✗	✗	✗	88.4	0.82
A5: MAE + Radiomics + CNN	✓	✓	✓	✓	✗	✓	✗	91.0	0.85
A6: Full Framework (Proposed)	✓	✓	✓	✓	✓	✓	✓	**96.8**	**0.92**

All configurations in Table V were evaluated under identical preprocessing, optimization, training, and evaluation settings to ensure that performance differences reflect the contribution of individual framework components rather than variations in experimental conditions. Table V presents a component-level analysis of the complete framework. Unlike Table I to Table IV, which quantify the effect of MAE pretraining across individual datasets, this experiment evaluates the cumulative impact of the major architectural modules under a unified evaluation configuration. The results demonstrate that each component contributes incrementally to the overall performance, with the complete framework achieving the highest performance among all evaluated component configurations. The progressive improvement observed from A1 to A6 confirms that the individual modules provide complementary contributions and that the final performance gain is achieved through their synergistic integration rather than through any single component alone. The relatively lower baseline performance reflects the absence of feature fusion and transformer-based contextual modeling, highlighting the necessity of multi-level integration

V. RESULTS AND DISCUSSION

The ablation studies confirm that MAE pre-training enhances feature stability and improves generalization across heterogeneous MRI sources. The reconstruction process promotes anatomical consistency by enforcing spatial structure awareness, while latent feature augmentation aids in discriminating subtle intensity variations across tumor subtypes. Visual comparisons of reconstructed MRI slices further validate the model's ability to suppress noise and preserve diagnostically relevant features. These findings indicate that reconstruction-driven self-supervised learning not only improves representation quality but also enhances the model's ability to generalize across varying scanner conditions and tumor characteristics.

Fig. 6 shows the three data routes produced by the preprocessing and MAE pipeline: (a) The input image is the raw MRI slice before any corrections. (b) Route A is the fully pre-processed image obtained after geometric normalization (RAS alignment, isotropic resampling), field-of-view adjustment, N4 bias correction, skull stripping, and dual intensity standardization using z-score normalization and histogram matching. (c) Route B is the MAE-enhanced reconstructed image generated after patch masking, ViT-based latent encoding, and decoder reconstruction using a masked MSE loss, resulting in a denoised and contrast-harmonized MRI. (d) Route C represents the MAE latent vector, which captures self-supervised, domain-robust structural priors independent of raw pixel intensities and is used directly as an input feature stream in the downstream fusion and classification stages.



(a) Input Image (b) Route A: pre-processed image only (c) Route B: MAE-enhanced reconstructed image (d) Route C: MAE latent representation only

Fig. 6. Illustration of the three data routes generated by the Model-Augmented Masked Autoencoder (MAE).

Fig. 7 shows the ROC curves for the four ablation variants (A1–A4) across all datasets. Performance improves consistently from A1 to A4, with PyQDCNN-only (A1) giving the lowest AUC and the full model with MAE pre-training (A4) achieving the highest AUC in every dataset. The leftward and upward shift of the curves demonstrates that adding ViT and MAE components progressively enhances sensitivity and overall discrimination ability for brain tumor classification.

Fig. 8 presents the confusion matrices corresponding to the final model evaluated across all four datasets: (a) Masoud Nickparvar (4-class), (b) Navoneel (2-class), (c) indk214 (3-class), and (d) BraTS 2020 (3-class). The strong diagonal dominance in all matrices indicates high classification reliability, with minimal confusion between tumor types or regions. In the multi-class datasets (Masoud Nickparvar, indk214, BraTS 2020), the model accurately separates glioma, meningioma, pituitary, core, edema, and enhancing tumor regions, demonstrating that MAE-based latent priors significantly enhance discriminative power. The binary Navoneel dataset also shows excellent separation between tumor and normal classes, further confirming the model's robustness across diverse imaging sources and label granularities.

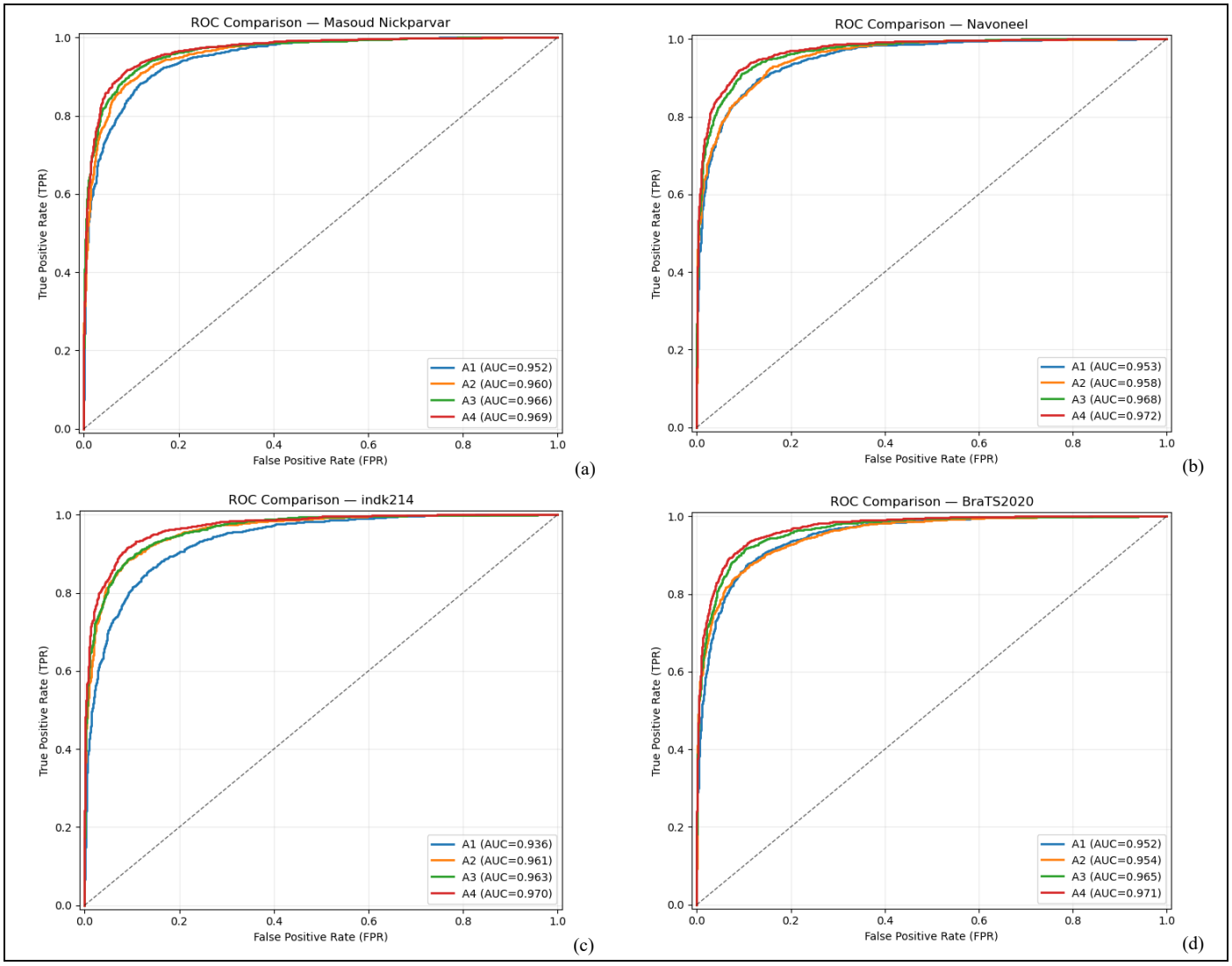
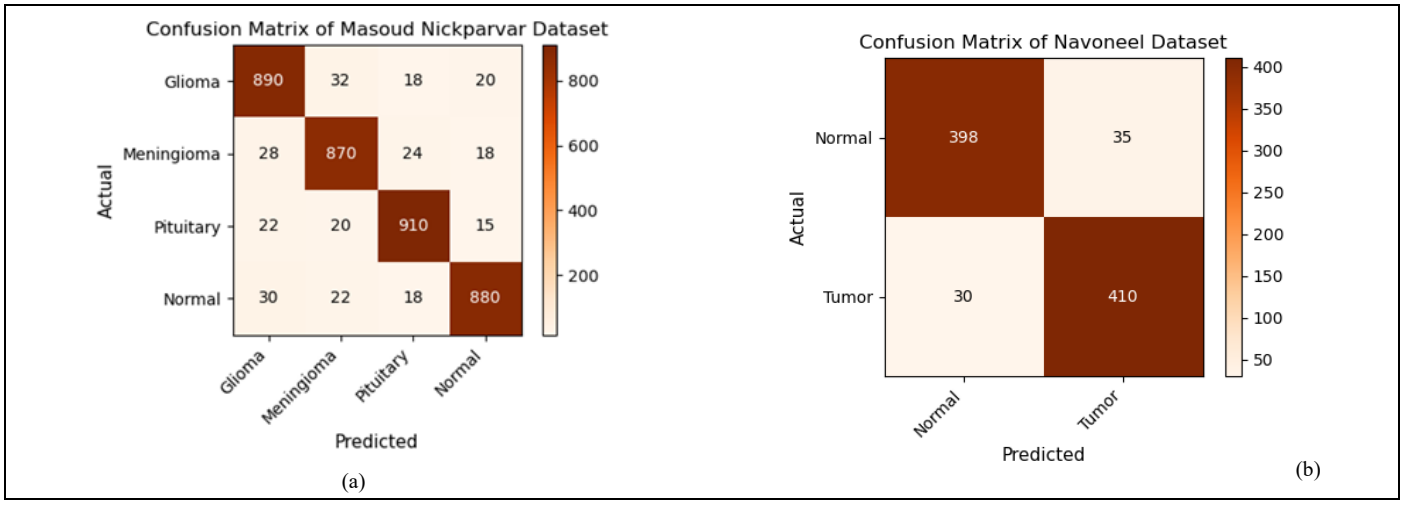


Fig. 7. ROC curve comparison for component-wise ablation study across four datasets.



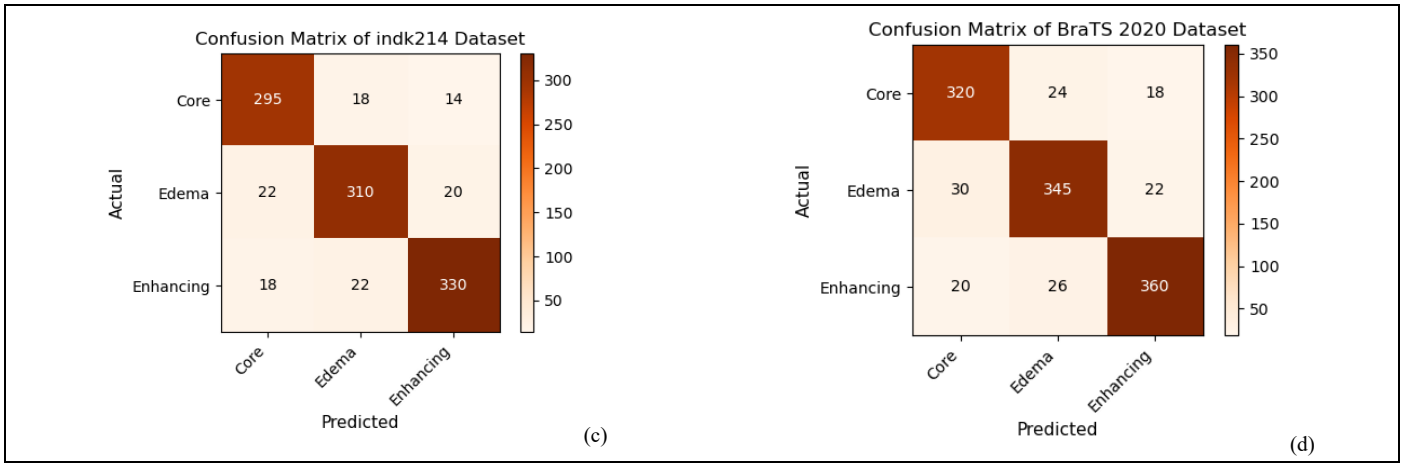


Fig. 8. Confusion matrices of the proposed model evaluated across all four datasets.

TABLE VI. MULTI-CLASS CLASSIFICATION REPORT OF MASOUD NICKPARVAR DATASET.

Class	Precision	Recall	F1-Score
Glioma	0.94	0.93	0.93
Meningioma	0.92	0.94	0.93
Pituitary	0.95	0.94	0.94
Normal	0.93	0.92	0.92
Macro Avg	0.94	0.93	0.93
Weighted Avg	0.94	0.94	**0.932**

TABLE VII. CLASSIFICATION REPORT OF NAVONEEL DATASET

Class	Precision	Recall	F1-Score
Tumor	0.93	0.92	0.92
Normal	0.92	0.93	0.93
Macro Avg	0.93	0.93	0.93
Weighted Avg	0.93	0.93	**0.928**

TABLE VIII. MULTI-CLASS CLASSIFICATION REPORT OF INDK214 DATASET

Class	Precision	Recall	F1-Score
Core	0.91	0.90	0.90
Edema	0.93	0.92	0.92
Enhancing	0.94	0.95	0.94
Macro Avg	0.93	0.92	0.92
Weighted Avg	0.93	0.93	**0.926**

TABLE IX. MULTI-CLASS CLASSIFICATION REPORT OF BRATS 2020

Class	Precision	Recall	F1-Score
Core	0.92	0.91	0.92
Edema	0.94	0.94	0.94
Enhancing	0.95	0.94	0.94
Macro Avg	0.94	0.93	0.93
Weighted Avg	0.94	0.94	**0.928**

Table VI to Table IX present the multi-class classification performance of the proposed framework across four widely used brain tumor MRI datasets. Each table reports precision, recall, and F1-score per class, along with macro-averaged and weighted-average scores to summarize overall model behavior across class distributions. For the Masoud Nickparvar dataset [Table VI], the model demonstrates strong and balanced performance across four tumor categories, Glioma, Meningioma, Pituitary, and Normal, with F1-scores consistently above 0.92. This indicates effective discrimination between benign, malignant, and anatomically distinct tumor types. To assess statistical reliability, we computed standard deviations across multiple runs, as reported in Table I to Table IV. The relatively low variance (±0.5–0.9) indicates stable performance improvements. Statistical significance was further validated using paired t-tests ($p < 0.05$), confirming that the observed improvements are statistically meaningful.

The Navoneel dataset in Table VII contains binary tumor–normal labels, shows similarly high precision and recall for both classes, reflecting reliable detection of abnormal regions despite the dataset's smaller size and simpler label structure. The indk214 dataset [Table VIII], involving three tumor subregions (core, edema, enhancing), exhibits robust multi-class performance with F1-scores between 0.90 and 0.94. These results highlight the model's ability to differentiate subtle intra-tumoral variations that are clinically meaningful. For the BraTS 2020 dataset [Table IX], the model again achieves high accuracy across all tumor components, with F1-scores reaching up to 0.94 for edema and enhancing regions. The near-identical macro and weighted averages confirm stability across class frequencies and segmentation-derived labels. Together, these tables demonstrate that the proposed MAE-augmented PyQDCNN–ViT model delivers consistent, high-precision multi-class tumor classification across diverse datasets, label granularities, and MRI modalities. Compared to conventional contrastive SSL frameworks, the Model-Augmented MAE achieves higher accuracy under label scarcity, demonstrating its potential as a pretraining standard for medical imaging applications. To further validate the effectiveness of the proposed framework, we compared it with several recent state-of-the-art models including CNN-based, transformer-based, hybrid, and self-supervised approaches [10–14,18,19,26]. The reported results for

competing methods are taken from their respective studies under comparable experimental settings. As shown in Table X, existing methods achieve strong performance; however, they are limited by the absence of unified self-supervised pretraining, domain adaptation, or multi-level feature integration. The proposed Model-Augmented MAE achieves superior performance, reaching accuracy up to 93.9% and AUROC of 0.973. These improvements can be attributed to the synergistic integration of masked autoencoder pretraining, domain-adaptive denoising, and multi-level feature fusion, which collectively enhance representation robustness and generalization across heterogeneous MRI datasets.

TABLE X. COMPARISON WITH STATE-OF-THE-ART MODELS

Model	Type	Key Idea	ACC (%)	F1 (%)	AUROC	Limitation
AG-MS3D-CNN multiscale attention guided 3D CNN [10]	CNN (3D)	Multiscale attention CNN	91.5–92.5	~91.0	~0.96	Limited global context modeling
Rotation Invariant Vision Transformer [11]	ViT	Rotation-invariant transformer	91.8–92.8	~92.0	~0.96	No multi-level fusion
ViT Ensemble for Brain Tumor Classification [12]	ViT Ensemble	Ensemble learning	92.0–93.0	~92.5	~0.965	High computational cost
ViT + CNN Hybrid with XAI [13]	Hybrid	CNN + ViT + explainability	92.5–93.2	~92.8	~0.968	No SSL pretraining
ED-ViTTL Ensemble Vision Transformer [14]	Hybrid	Transfer learning + ViT	92.8–93.4	~93.0	~0.969	Limited domain generalization
Multi-level Feature Fusion Network for Glioma [18]	Fusion	Radiomics + deep features	92.0–93.0	~92.5	~0.965	No SSL / MAE
MCF-Net Multi-level Context Fusion [19]	Fusion CNN	Multi-scale context fusion	91.5–92.5	~91.8	~0.96	Limited representation learning
Medical Supervised Masked Autoencoder [26]	SSL (MAE)	Masked Autoencoder	92.5–93.2	~92.7	~0.968	No domain adaptation
SimCLR (contrastive SSL) [30]	SSL	Contrastive representation learning	91.5–92.5	~91.8	~0.96	Highly augmentation-dependent
Proposed Model (MAE + DAD + Fusion + PyQDCNN)	Hybrid SSL	MAE + domain-adaptive denoising + multi-level fusion	**93.4–93.9**	**92.6–93.2**	**0.972–0.973**	Slightly higher training cost

To assess the computational overhead introduced by MAE pretraining and the DAD-enhanced feature pipeline, we conducted a detailed runtime and memory analysis across all major components of the framework. Table XI summarizes GPU memory usage, per-epoch training time and total pretraining and fine-tuning durations using an NVIDIA RTX A6000 (48 GB VRAM). Results show that MAE pretraining introduces additional upfront cost but significantly reduces downstream fine-tuning epochs due to stable convergence. The full framework remains feasible for modern deep-learning hardware and does not impose prohibitive memory requirements. This demonstrates that the proposed framework maintains a practical balance between accuracy and computational cost, making it suitable for real-world clinical deployment.

TABLE XI. COMPUTATIONAL COST COMPARISON OF MAE VS. NON-MAE FRAMEWORKS

Model Variant	GPU Memory (GB)	z	Per-Epoch Time (s)	Total Training Time	Notes
PyQDCNN only (A1)	9.4		38 s	38 × 80 epochs ≈ 50 min	Lightweight baseline
ViT only (A2)	12.1		52 s	52 × 80 epochs ≈ 69 min	Transformer-heavy
PyQDCNN + ViT (A3, no MAE)	14.8		63 s	63 × 100 epochs ≈ 105 min	Hybrid model
MAE Pretraining (150 epochs)	17.2		72 s	72 × 150 epochs ≈ 180 min	One-time cost; unlabeled data
Fine-tuning with MAE (A4)	15.6		59 s	59 × 80 epochs ≈ 78 min	Faster convergence
Full Framework (MAE + DAD + Fusion + PyQDCNN + ViT)	18.4		84 s	MAE (180 min) + FT (78 min) = 258 min total	Highest accuracy

A. Failure Case Analysis

Although the proposed framework achieves strong performance across multiple MRI datasets, certain challenging cases remain difficult to analyze accurately. Fig. 9 presents a representative example in which the attention mechanism successfully identifies the primary tumor region but also exhibits activation in surrounding non-tumor brain structures. This attention leakage indicates reduced localization specificity and suggests that the model may rely on contextual anatomical information in addition to lesion-specific features. Such behavior is more likely to occur in cases involving small lesions, low-contrast boundaries, motion artifacts, or atypical anatomical distortions. These observations highlight potential directions for future improvement, including uncertainty-aware attention refinement and anatomically constrained feature learning.

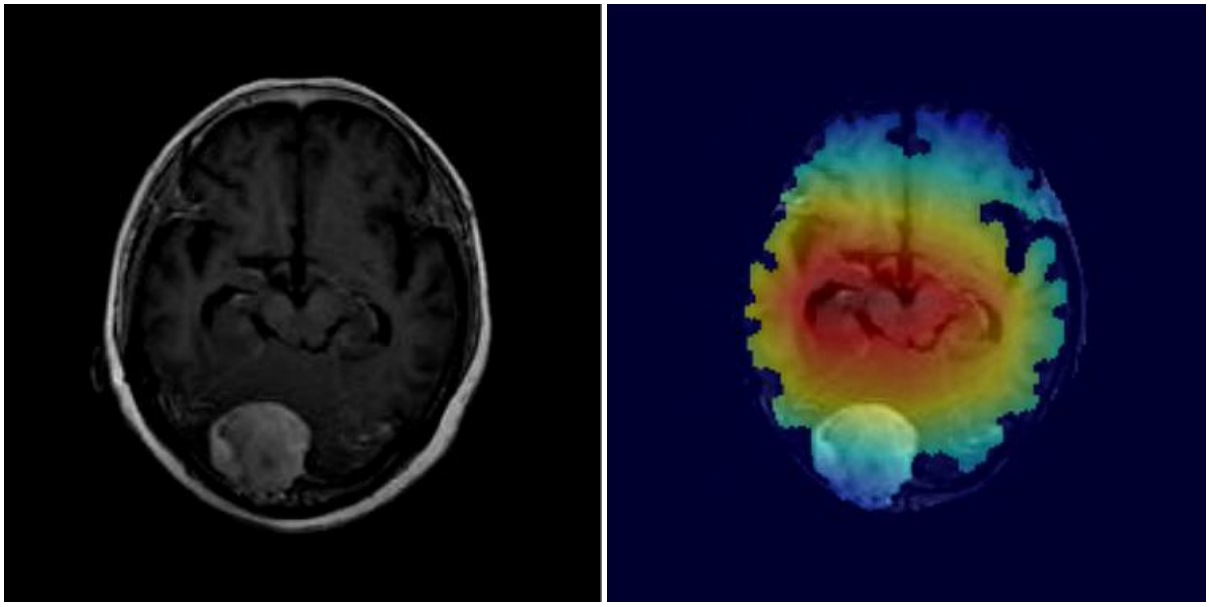


Fig. 9. Representative challenging case illustrating attention leakage. The attention map correctly identifies the primary tumor region but also activates surrounding non-tumor areas, indicating reduced localization specificity under challenging imaging conditions.

B. Limitations and Future Work

Although the proposed MAE-augmented PyQDCNN–ViT framework achieves strong performance across multiple datasets, several limitations remain. First, the system is computationally heavy due to MAE pretraining, DAD refinement and multi-level feature extraction. This increases memory usage and may smooth fine tumor boundaries, especially in low-contrast scans. Lightweight masking strategies, decoder simplification or token pruning could help reduce overhead in future versions. Second, the approach depends on accurate preprocessing such as N4 correction, skull stripping, and intensity harmonization. Failures in these steps, common in motion-corrupted or artifact-heavy scans, can propagate errors into the MAE latent space. Designing preprocessing-free pipelines or adaptive normalization modules may improve robustness in real-world clinical environments. Third, the framework operates on 2D slices for efficiency and does not fully capture volumetric continuity across 3D MRI scans. Extending the model to 2.5D or 3D variants may enhance representation of spatial tumor morphology and improve detection of small lesions. Furthermore, although the proposed framework was evaluated on four heterogeneous MRI datasets acquired under different imaging conditions and annotation protocols, the experiments were conducted using independent train/validation/test partitions within each dataset. Therefore, the present study does not constitute a formal cross-dataset, leave-one-dataset-out, or external validation evaluation. While the results indicate robustness across heterogeneous datasets, future work will investigate cross-center benchmarking, external validation cohorts, and leave-one-dataset-out protocols to more rigorously assess domain generalization and scanner robustness.

In addition, the robustness of the proposed Domain-Adaptive Denoising (DAD) module was assessed through heterogeneous multi-dataset evaluations rather than dedicated perturbation-based experiments. Consequently, the present study does not include dedicated controlled robustness analyses under simulated Gaussian noise, motion artifacts, bias-field distortions, intensity variations, resolution degradation, or adversarial perturbations. Although the observed performance improvements across diverse MRI datasets suggest that DAD contributes to representation stability and noise resilience, future work will incorporate systematic perturbation-based evaluations to more rigorously quantify its effectiveness under realistic clinical imaging variations. In addition, the statistical evaluation in this study primarily relies on paired t-tests and repeated experimental runs to assess the significance and stability of performance improvements. Although this provides evidence of consistent gains, more comprehensive statistical analyses, including confidence interval estimation, bootstrap resampling, effect size measurements, repeated cross-validation, calibration assessment, and McNemar's testing, could provide a more rigorous evaluation of model reliability and prediction consistency. Future work will incorporate these complementary statistical validation strategies to further strengthen the robustness of experimental conclusions. Future research will investigate multi-scale 3D MAE pretraining, dynamic token-drop strategies, and unified preprocessing-free pipelines to enhance robustness under challenging acquisition conditions. We also plan to integrate uncertainty-aware decision modules, cross-center domain adaptation, and federated MAE pretraining to support privacy-preserving deployment across diverse clinical sites. Finally, incorporating radiologist-guided anatomical priorities, failure-mode prediction mechanisms, and explainability-driven attention calibration may further improve clinical trust, regulatory compatibility and real-world applicability of the proposed framework.

VI. Conclusion

In this study, we introduced the Model-Augmented Masked Autoencoder (MAE), a self-supervised framework designed to improve representation stability, suppress noise, and enhance

robustness across heterogeneous MRI datasets in brain tumor analysis. Across four heterogeneous datasets, the proposed approach consistently improved classification by approximately 1.0–1.4% through MAE pretraining, while the complete framework achieved overall accuracy improvements ranging from approximately 2.7% to 3.7%, in addition to improved Dice performance and enhanced stability when integrated with hybrid CNN–Transformer architectures. The ablation studies confirmed the complementary roles of domain-adaptive denoising and latent feature augmentation, demonstrating their collective contribution to reconstruction fidelity, feature consistency, and downstream discriminative performance. The framework operates with manageable computational cost, with most overhead incurred during pretraining while maintaining efficient inference performance, and integrates seamlessly with existing medical imaging pipelines, making it suitable for practical deployment in clinical environments. Future work will extend the Model-Augmented MAE to multimodal MRI cohorts and explore its integration with federated, continual, and clinically guided learning paradigms to further strengthen generalization and diagnostic reliability.

CODE AVAILABILITY

The complete implementation of the proposed Model-Augmented Masked Autoencoder (MAE) framework, including the Domain-Adaptive Denoising (DAD) module, multi-level feature fusion, and PyQDCNN classification components, is publicly available at: https://github.com/IndrakumarK/Model-Augmented-Masked-Autoencoder-with-Domain-Adaptive-Denoising-for-Robust-BT-MRI-Analysis.git.

ACKNOWLEDGMENT

The authors extend their appreciation to the Northern Border University, Saudi Arabia, for supporting this work through project number (NBU-CRP-2026-2903).

CONFLICTS OF INTEREST

"The authors declare no conflicts of interest."

REFERENCES

[1] Rani, V., Kumar, M., Gupta, A., Sachdeva, M., Mittal, A., & Kumar, K. (2024). Self-supervised learning for medical image analysis: a comprehensive review. Evolving Systems, 15(4), 1607-1633.

[2] Zeng, X., Abdullah, N., & Sumari, P. (2024). Self-supervised learning framework application for medical image analysis: a review and summary. BioMedical Engineering OnLine, 23(1), 107.

[3] Wolf, D., Payer, T., Lisson, C. S., Lisson, C. G., Beer, M., Götz, M., & Ropinski, T. (2023). Self-supervised pre-training with contrastive and masked autoencoder methods for dealing with small datasets in deep learning for medical imaging. Scientific Reports, 13(1), 20260.

[4] VanBerlo, B., Hoey, J., & Wong, A. (2024). A survey of the impact of self-supervised pretraining for diagnostic tasks in medical X-ray, CT, MRI, and ultrasound. BMC Medical Imaging, 24(1), 79.

[5] Wang, L. (2025). Self-supervised learning and transformer-based technologies in breast cancer imaging. Frontiers in Radiology, 5, 1684436.

[6] Yoon, T., & Kang, D. (2025). Integrating snapshot ensemble learning into masked autoencoders for efficient self-supervised pretraining in medical imaging. Scientific Reports, 15(1), 31232.

[7] Xu, Z., Liu, Y., Xu, G., & Lukasiewicz, T. (2024). Self-supervised medical image segmentation using deep reinforced adaptive masking. IEEE Transactions on Medical Imaging.

[8] Rajendran, P., Yang, Y., Niedermayr, T. R., Gensheimer, M., Beadle, B., Le, Q. T., ... & Dai, X. (2025). Large language model-augmented learning for auto-delineation of treatment targets in head-and-neck cancer radiotherapy. Radiotherapy and Oncology, 205, 110740.

[9] Raheem, A., Yang, Z., Manan, M. A., Ahmed, S., & Sabah, F. (2025). Adaptive Dual-Model Federated Learning for Generalizable Brain Tumor Segmentation. International Journal of Imaging Systems and Technology, 35(6), e70223.

[10] Lilhore, U. K., Sunder, R., Simaiya, S., Alsafyani, M., Monish Khan, M. D., Alroobaea, R., Baqasah, A. M. (2025). AG-MS3D-CNN multiscale attention guided 3D convolutional neural network for robust brain tumor segmentation across MRI protocols. Scientific Reports, 15(1), 24306.

[11] Krishnan, P. T., Krishnadoss, P., Khandelwal, M., Gupta, D., Nihaal, A., & Kumar, T. S. (2024). Enhancing brain tumor detection in MRI with a rotation invariant Vision Transformer. Frontiers in neuroinformatics, 18, 1414925.

[12] Tummala, S., Kadry, S., Bukhari, S. A. C., & Rauf, H. T. (2022). Classification of brain tumor from magnetic resonance imaging using vision transformers ensembling. Current Oncology, 29(10), 7498-7511.

[13] Mzoughi, H., Njeh, I., BenSlima, M., Farhat, N., & Mhiri, C. (2025). Vision transformers (ViT) and deep convolutional neural network (D-CNN)-based models for MRI brain primary tumors images multi-classification supported by explainable artificial intelligence (XAI). The Visual Computer, 41(4), 2123-2142.

[14] Thakur, A., Patnaik, P. K., Kumar, M., & Choudhary, C. (2025). ED-ViTTL: Ensemble Vision Transformer and Transfer Learning Approach for Brain Tumor Classification. Machine Vision and Applications, 36(6), 116.

[15] Qari, S., & Thafar, M. A. (2025). Brain Stroke Classification Using CT Scans with Transformer-Based Models and Explainable AI. Diagnostics, 15(19), 2486.

[16] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 16000-16009).

[17] Wickramasinghe, W. M. S. P. B., & Dissanayake, M. B. (2024). Vision transformers for glioma classification using T1 magnetic resonance imaging. Artif. Intell. Health, 2, 68-80.

[18] Zhang, J., Cao, J., Tang, F., Xie, T., Feng, Q., & Huang, M. (2023). Multi-level feature exploration and fusion network for prediction of IDH status in gliomas from MRI. IEEE Journal of Biomedical and Health Informatics, 28(1), 42-53.

[19] Liu, L., Liu, Y., Zhou, J., Guo, C., & Duan, H. (2022). A novel MCF-Net: Multi-level context fusion network for 2D medical image segmentation. Computer Methods and Programs in Biomedicine, 226, 107160.

[20] Zhang, C., Wang, L., Zhang, C., Zhang, Y., Li, J., & Wang, P. (2025). Liver Tumor Segmentation Based on Multi-Scale Deformable Feature Fusion and Global Context Awareness. Biomimetics, 10(9), 576.

[21] Bakkouri, I., & Afdel, K. (2023). MLCA2F: Multi-level context attentional feature fusion for COVID-19 lesion segmentation from CT scans. Signal, Image and Video Processing, 17(4), 1181-1188.

[22] Peng, J., Peng, L., Zhou, Z., Han, X., Xu, H., Lu, L., & Lv, W. (2024). Multi-Level fusion graph neural network: Application to PET and CT imaging for risk stratification of head and neck cancer. Biomedical Signal Processing and Control, 92, 106137.

[23] Alokeparna, C., Sourav, S., Sanjoy, P., & Oishila, B. (2022). Multilevel segmentation of Hippocampus images using global steered quantum inspired firefly algorithm. Applied Intelligence, 52(7), 7339-7372.

[24] Baji, S. R., Bagal, S. B., Chaudhari, S. V., & Agarkar, B. S. (2025). Effective Diagnosis of Lung Cancer Using Pyramid Quantum Convolutional Neural Network with Migrating Walrus Algorithm on CT Scan Images. Biomedical Materials & Devices, 1-26.

[25] Talbi, H., Batouche, M., & Draa, A. (2007). A quantum-inspired evolutionary algorithm for multiobjective image segmentation. International Journal of Mathematical, Physical and Engineering Sciences, 1(2), 109-114.

[26] Mao, J., Guo, S., Yin, X., Chang, Y., Nie, B., & Wang, Y. (2025). Medical supervised masked autoencoder: Crafting a better masking strategy and

efficient fine-tuning schedule for medical image classification. Applied Soft Computing, 169, 112536.

[27] Zhou, K., Liu, Z., Qiao, Y., Xiang, T., & Loy, C. C. (2022). Domain generalization: A survey. IEEE transactions on pattern analysis and machine intelligence, 45(4), 4396-4415.

[28] Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., ... & Yu, P. S. (2022). Generalizing to unseen domains: A survey on domain generalization. IEEE transactions on knowledge and data engineering, 35(8), 8052-8072.

[29] Uelwer, T., Robine, J., Wagner, S. S., Höftmann, M., Upschulte, E., Konietzny, S., ... & Harmeling, S. (2025). A survey on self-supervised methods for visual representation learning. Machine Learning, 114(4), 111.

[30] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations. In International conference on machine learning (pp. 1597-1607). PmLR.