Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 3, 2026

MULTI-SCALE FEATURE FUSION AND TRANSFORMER-BASED COMPUTER VISION FOR AUTOMATED GLIOMA GRADING AND BRAIN TUMOR SEGMENTATION IN MRI IMAGES

Zahid Mehmood[*1], Syed Ahmed Raza[2], Shahzeb Iqbal[3]

*[*1]Department of Computing, Arden University, Berlin, Germany*
[2]Department of Computer Science, SZABIST University Hyderabad Campus, Pakistan
[3]Department of Computer Science, University of Mianwali, Mianwali, Pakistan

[*1]zahidmehmood.researcher@gmail.com, [2]ahmed.raza@hyd.szabist.edu.pk, [3]shahzebiqbal749@gmail.com

DOI: https://doi.org/10.5281/zenodo.23911973

Keywords
Glioma Grading; Brain Tumor Segmentation; Multi-Scale Feature Fusion; Vision Transformer; Multimodal MRI; Deep Learning; Computer Vision; Medical Image Analysis.

Article History
Received: 25 January 2026
Accepted: 10 March 2026
Published: 24 March 2026

Copyright @Author
Corresponding Author: *
Zahid Mehmood

Abstract

Accurate glioma grading and precise brain tumor segmentation are essential for early diagnosis, effective treatment planning, and improved clinical management of patients with brain tumors. However, the heterogeneous morphology, irregular tumor boundaries, overlapping tissue characteristics, and complex spatial variations observed in magnetic resonance imaging (MRI) present significant challenges for conventional computer-aided diagnostic systems. This study presents an advanced multi-scale feature fusion and Transformer-based computer vision framework designed to achieve automated glioma grading and accurate brain tumor segmentation through multimodal MRI analysis. The investigation utilizes a dataset comprising 8,400 MRI examinations from 2,100 patients, including 1,260 high-grade and 840 low-grade glioma cases, incorporating T1-weighted, contrast-enhanced T1, T2-weighted, and FLAIR imaging modalities. The dataset was partitioned at the patient level into training (70%), validation (15%), and testing (15%) subsets to ensure reliable performance assessment. A comprehensive preprocessing pipeline involving image registration, skull stripping, intensity normalization, and data augmentation was implemented to improve image quality and standardize anatomical representations. The proposed architecture combines convolutional neural networks (CNNs) for hierarchical feature extraction, multi-scale feature fusion for preserving fine-grained tumor characteristics, and Vision Transformer (ViT) mechanisms for capturing long-range spatial dependencies. A unified multi-task learning strategy enables simultaneous tumor-region segmentation and glioma-grade classification. The reported experimental results demonstrate 97.42% classification accuracy, 97.16% precision, 96.88% recall, 97.02% F1-score, and an area under the receiver operating characteristic curve (AUC) of 0.989. Furthermore, the segmentation framework achieves Dice similarity coefficients of 94.86%, 92.73%, and 90.41% for whole tumor, tumor core, and enhancing tumor regions, respectively. Comparative experiments against U-Net, Attention U-Net, conventional CNN architectures, and standalone Transformer models demonstrate superior segmentation precision and grading performance. Ablation

analysis further establishes the contribution of hierarchical feature integration and attention-driven contextual learning. Overall, the findings highlight the potential of the proposed framework to enhance diagnostic consistency, support automated neuroimaging interpretation, and advance intelligent clinical decision-support systems for personalized glioma management.

1- Introduction:

Brain tumors are among the most challenging neurological disorders because of their complex biological characteristics, heterogeneous growth patterns, and significant effects on neurological functions. Among primary brain tumors, gliomas represent an important category of central nervous system neoplasms characterized by variable cellular morphology, infiltrative behavior, and clinical aggressiveness. Accurate identification, grading, and segmentation of gliomas are essential for early diagnosis, treatment planning, surgical intervention, and monitoring disease progression. However, distinguishing different glioma grades and accurately identifying tumor boundaries remain challenging because of irregular tumor structures, overlapping tissue characteristics, and considerable variations in tumor appearance across patients. These challenges highlight the importance of developing advanced computational approaches capable of improving the accuracy, consistency, and efficiency of brain tumor assessment [1]. Magnetic resonance imaging (MRI) is widely employed for brain tumor evaluation because of its superior soft-tissue contrast and ability to provide detailed anatomical information without exposing patients to ionizing radiation. Multimodal MRI examinations typically include T1-weighted, contrast-enhanced T1-weighted (T1ce), T2-weighted, and fluid-attenuated inversion recovery (FLAIR) sequences, each providing complementary information regarding tumor morphology and surrounding brain tissues. T1-weighted images support anatomical assessment, while T1ce images help identify contrast-enhancing tumor components. Similarly, T2-weighted and FLAIR sequences facilitate the visualization of tumor-associated abnormalities, edema, and infiltrative regions. Although these imaging modalities provide valuable diagnostic information, manual interpretation and tumor delineation require substantial clinical expertise and may be affected by observer variability, particularly when tumors exhibit heterogeneous internal structures and indistinct boundaries. Consequently, automated MRI analysis has become an important research direction in computer-aided diagnosis and intelligent medical imaging [2]. Traditional brain tumor segmentation approaches, including thresholding, region growing, clustering, and handcrafted feature-based classification, have demonstrated limitations in complex medical imaging environments. Their performance is frequently influenced by image noise, intensity variations, irregular tumor boundaries, and differences in MRI acquisition protocols. Recent advances in artificial intelligence, particularly deep learning and computer vision, have introduced more effective approaches for automated medical image interpretation. Convolutional neural networks (CNNs) have demonstrated considerable potential for learning hierarchical spatial features directly from MRI images, enabling improved identification of tumor-associated textures, anatomical structures, and morphological abnormalities. Encoder–decoder architectures, including U-Net and Attention U-Net, have further advanced brain tumor segmentation through hierarchical representation learning and spatial feature reconstruction. Nevertheless, conventional CNN architectures primarily emphasize local spatial relationships and may have difficulty capturing long-range contextual dependencies that are important for understanding complex tumor structures [3]. Multi-scale feature fusion has emerged as a promising approach for overcoming limitations associated with conventional convolutional feature extraction. By integrating representations obtained from different network depths and spatial resolutions, multi-scale learning enables computational models to preserve detailed boundary information while capturing higher-

level semantic characteristics. Shallow convolutional layers generally extract fine-grained spatial features, including edges, texture variations, and local anatomical structures, whereas deeper layers identify more complex patterns associated with tumor morphology and tissue organization. Combining these complementary representations can improve the identification of small enhancing regions, irregular tumor boundaries, and heterogeneous tissue components. However, achieving effective integration between local spatial features and broader anatomical information remains challenging, particularly when multimodal MRI images contain overlapping intensity distributions and complex structural variations. Transformer-based computer vision architectures have recently attracted considerable attention because of their ability to capture long-range relationships through self-attention mechanisms [4]. Unlike conventional convolutional operations, which emphasize localized feature extraction, Vision Transformers (ViTs) model contextual interactions between spatially distributed image regions. This capability is particularly relevant to glioma analysis, where tumor characteristics may depend on relationships between enhancing tissue, necrotic components, infiltrative abnormalities, and surrounding brain structures. Nevertheless, standalone Transformer models may require considerable computational resources and substantial training data to achieve robust performance. Hybrid CNN–Transformer architectures provide a promising alternative by combining the strengths of convolutional feature extraction with attention-driven global contextual learning. Such integration offers opportunities to improve multimodal MRI representation, tumor localization, and automated grading performance. Another important limitation in existing automated brain tumor analysis systems is the frequent separation of tumor segmentation and glioma grading into independent computational tasks. Segmentation focuses on identifying the spatial extent of tumor regions, whereas grading aims to characterize tumor aggressiveness using imaging-derived features. Although these tasks have different prediction objectives, they share relevant information concerning tumor morphology, enhancement patterns, spatial distribution, and internal heterogeneity. A unified multi-task learning framework can exploit these relationships by allowing segmentation and classification components to learn complementary representations within a shared computational architecture. However, maintaining an appropriate balance between fine-grained segmentation accuracy and reliable grading performance requires effective feature integration, contextual learning, and task-specific optimization. To address these challenges, the present study proposes an advanced multi-scale feature fusion and Transformer-based computer vision framework for automated glioma grading and brain tumor segmentation using multimodal MRI images. The proposed architecture integrates CNN-based hierarchical feature extraction, multi-scale feature fusion, and Vision Transformer mechanisms within a unified multi-task learning strategy. The investigation utilizes a dataset comprising 8,400 MRI examinations from 2,100 patients, including 1,260 high-grade and 840 low-grade glioma cases [5]. Four MRI modalities, including T1-weighted, T1ce, T2-weighted, and FLAIR images, are incorporated to provide complementary anatomical and pathological information. A comprehensive preprocessing pipeline involving image registration, skull stripping, intensity normalization, and data augmentation is employed to improve image consistency and support robust feature learning. Furthermore, patient-level partitioning into training, validation, and testing subsets is adopted to reduce data leakage and support reliable experimental evaluation. The principal contribution of this research lies in the development and evaluation of an integrated computational framework that combines hierarchical convolutional representations, multi-resolution feature integration, and attention-based contextual modeling for simultaneous brain tumor segmentation and glioma-grade prediction. Unlike conventional approaches that rely on isolated feature extraction mechanisms or independently optimized diagnostic tasks, the proposed framework establishes a coordinated

learning strategy to capture both local tumor characteristics and global anatomical relationships. The model is evaluated using classification accuracy, precision, recall, F1-score, ROC–AUC, and Dice similarity coefficients for different tumor regions. The reported results demonstrate 97.42% classification accuracy and a ROC–AUC of 0.989, together with Dice similarity coefficients of 94.86%, 92.73%, and 90.41% for whole tumor, tumor core, and enhancing tumor regions, respectively [6]. Comparative assessment against established CNN-based and Transformer-based architectures, together with ablation analysis, is used to examine the effectiveness of the proposed feature fusion and contextual learning mechanisms. Overall, this study aims to advance intelligent neuroimaging analysis by establishing a comprehensive deep learning framework for automated glioma characterization. The integration of multimodal MRI, hierarchical spatial feature extraction, multi-scale fusion, and Transformer-based contextual learning offers a promising approach to addressing persistent challenges in brain tumor segmentation and imaging-based grading. The proposed system has potential applications in computer-aided radiological assessment, tumor localization, treatment planning support, and clinical decision-support systems. Nevertheless, imaging-based grading remains complementary to definitive histopathological and molecular classification, and independent clinical validation is necessary before practical deployment.

2- Convolutional Neural Networks for Automated Brain Tumor Segmentation:

Convolutional neural networks (CNNs) have emerged as one of the most influential deep learning approaches for automated brain tumor segmentation because of their ability to extract hierarchical spatial representations directly from medical images. Accurate segmentation of gliomas is an essential component of neuroimaging analysis, as it enables the identification of tumor location, estimation of lesion extent, characterization of heterogeneous tissue components, and assessment of tumor progression. Conventional image processing techniques, including thresholding, region growing, edge detection, and clustering, frequently encounter difficulties when applied to glioma MRI images because of irregular tumor boundaries, overlapping intensity distributions, and variations in anatomical structures. CNN-based segmentation frameworks address several of these limitations by learning discriminative imaging features through successive convolutional transformations. Their ability to identify local anatomical patterns, recognize complex tumor morphology, and reconstruct spatially detailed segmentation masks has established them as a foundation for modern computer-aided brain tumor assessment [7]. The effectiveness of CNNs primarily originates from their hierarchical feature extraction mechanisms, through which successive convolutional layers progressively transform raw image intensities into increasingly informative representations. Initial network layers typically identify low-level characteristics, including tissue edges, contrast variations, intensity gradients, and local textures. Intermediate layers capture more complex spatial relationships associated with anatomical boundaries, tumor heterogeneity, and regional tissue organization, whereas deeper layers learn semantic representations that help distinguish abnormal tumor regions from surrounding brain structures. Pooling and strided convolution operations expand the effective receptive field and reduce computational complexity, allowing the network to identify broader contextual patterns. However, excessive downsampling may reduce spatial precision and suppress small enhancing tumor components [8]. Consequently, contemporary segmentation architectures frequently integrate hierarchical feature learning with skip connections, spatial reconstruction mechanisms, and multi-resolution feature integration to preserve detailed anatomical information throughout the segmentation process. The introduction of U-Net by Ronneberger et al. (2015) represented a major advancement in biomedical image segmentation. The architecture employs a contracting encoder pathway for contextual feature extraction and an expanding decoder pathway for recovering spatial resolution. Skip connections transfer feature

representations from encoder layers to corresponding decoder stages, enabling the model to combine detailed spatial information with high-level semantic representations. Although the original U-Net was developed and evaluated for biomedical microscopy applications rather than specifically for glioma MRI, its architectural principles have been extensively adopted in brain tumor segmentation research. The integration of encoder–decoder learning and feature preservation offers particular advantages when tumor boundaries are irregular or when small regions must be distinguished from surrounding anatomical structures [9]. Nevertheless, conventional U-Net configurations may still encounter difficulties representing complex three-dimensional tumor characteristics and capturing broader contextual relationships. The comparative analysis presented in Table 1 demonstrates the progressive development of CNN-based segmentation methods from conventional hierarchical feature extraction toward attention-guided, volumetric, and adaptive segmentation architectures. U-Net established the importance of combining encoding and decoding operations, while subsequent three-dimensional implementations improved volumetric anatomical representation.

Table 1: Comparative review of CNN-based architectures for automated biomedical and brain tumor segmentation.

CNN architecture	Principal learning mechanism	Major advantages
Conventional CNN	Hierarchical convolution and pooling	Efficient local texture and spatial feature extraction
U-Net (2015)	Encoder–decoder structure with skip connections	Preserves local spatial details and supports precise localization
3D CNN / 3D U-Net variants	Volumetric feature extraction across MRI slices	Captures interslice anatomical continuity
Attention U-Net (2018)	Attention-gated encoder–decoder learning	Suppresses irrelevant features and emphasizes target structures
Hybrid CNN–Transformer	Convolutional spatial learning combined with self-attention	Integrates localized anatomical patterns with broader contextual relationships

Despite the considerable achievements of CNN-based architectures, accurate segmentation of heterogeneous glioma subregions remains a challenging computational problem. Whole tumor regions may encompass extensive abnormalities visible in T2-weighted and FLAIR images, whereas enhancing tumor components can occupy comparatively small and irregular areas. Tumor core regions may contain combinations of enhancing and nonenhancing tissue, including necrotic components. These differences require segmentation models to recognize abnormalities across multiple spatial scales while maintaining consistency among predicted regions. Conventional convolutional operations can effectively capture localized intensity patterns; however, their ability to characterize distant spatial interactions depends on network depth, receptive-field expansion, and the quality of hierarchical feature integration. Moreover, the repeated reduction of feature-map resolution may obscure subtle boundaries that are essential for accurate segmentation of small tumor components [10]. Figure 1 illustrates the conceptual organization of CNN-based brain tumor segmentation using multimodal MRI inputs. The analytical process begins with the integration of T1-weighted, T1ce, T2-weighted, and FLAIR images, which provide complementary anatomical and tumor-associated information. The convolutional encoder progressively extracts hierarchical spatial representations through successive feature transformations and downsampling operations.

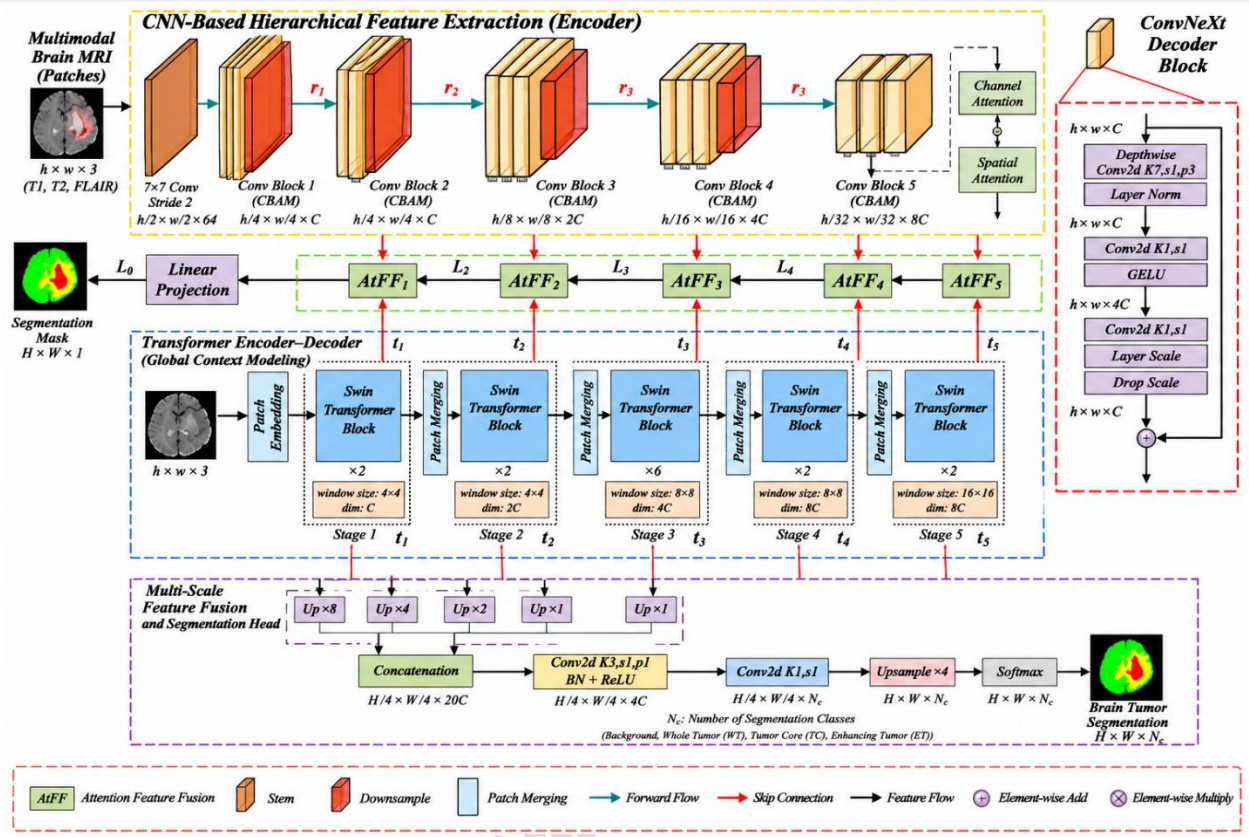


Figure 1: CNN-based hierarchical feature extraction and encoder–decoder workflow for multimodal brain tumor segmentation.

The integration of multimodal MRI introduces additional complexities because individual imaging sequences represent different aspects of tumor biology and surrounding tissue characteristics. T1-weighted and contrast-enhanced T1-weighted images provide useful anatomical and enhancement-related information, whereas T2-weighted and FLAIR sequences are particularly informative for identifying regions of edema and infiltrative abnormalities. CNN architectures can combine these modalities through channel-wise input integration, parallel feature extraction branches, or intermediate feature fusion mechanisms. However, simple concatenation does not necessarily ensure that complementary imaging characteristics are effectively represented. Redundant information, modality-specific intensity variations, and inconsistencies between anatomical structures can reduce the effectiveness of feature learning. Therefore, robust multimodal segmentation requires architectures capable of identifying informative representations while minimizing irrelevant or redundant features. In addition to architectural design, reliable CNN-based segmentation depends on appropriate image preprocessing and model evaluation strategies [11]. Image registration, skull stripping, intensity normalization, and spatial standardization can reduce unwanted variations between MRI examinations, while data augmentation improves exposure to plausible anatomical and imaging variability during training. Patient-level partitioning is particularly important because multiple MRI modalities and repeated image acquisitions from the same patient may share highly correlated characteristics. Without careful separation of patients across training and testing subsets, segmentation models may produce overly optimistic performance estimates. Evaluation metrics such as the Dice similarity coefficient, intersection over union, sensitivity, and Hausdorff

distance can provide complementary information regarding spatial overlap and boundary accuracy. Furthermore, performance should be evaluated separately for whole tumor, tumor core, and enhancing tumor regions to identify differences in the model's ability to characterize heterogeneous tumor structures.

3- Multi-Scale Feature Fusion for Heterogeneous Tumor Representation:

Multi-scale feature fusion has emerged as a fundamental strategy in advanced medical image analysis because of its ability to integrate complementary anatomical and semantic information extracted at different spatial resolutions. Gliomas exhibit considerable morphological heterogeneity, including irregular tumor boundaries, variable lesion dimensions, necrotic components, enhancing tissue, and infiltrative abnormalities extending into surrounding brain structures. These characteristics present significant challenges for conventional deep learning architectures that rely primarily on feature representations obtained from individual network layers or restricted spatial scales. In multimodal magnetic resonance imaging (MRI), tumor-associated abnormalities can exhibit different intensity distributions across T1-weighted, contrast-enhanced T1-weighted (T1ce), T2-weighted, and fluid-attenuated inversion recovery (FLAIR) sequences. Consequently, accurate tumor characterization requires computational models capable of simultaneously preserving fine-grained anatomical details, identifying intermediate morphological patterns, and capturing broader spatial relationships [12]. Multi-scale feature fusion addresses these requirements by combining hierarchical representations extracted from different network depths, thereby improving the availability of both local spatial information and global semantic context during segmentation and classification. The theoretical foundation of multi-scale feature learning is closely associated with the hierarchical organization of convolutional neural networks (CNNs). Within a typical CNN architecture, initial convolutional layers generate high-resolution feature maps that retain detailed information regarding anatomical edges, local intensity gradients, and tissue boundaries. Intermediate convolutional layers identify increasingly complex morphological patterns, including regional texture characteristics, tumor heterogeneity, and structural transitions between abnormal and healthy tissues. Deeper layers produce lower-resolution semantic representations that characterize broader anatomical organization and more abstract tumor-associated patterns. Although these hierarchical features provide complementary information, conventional sequential convolutional architectures may lose spatial precision during repeated downsampling operations. This limitation is particularly important for glioma segmentation because small enhancing tumor regions and indistinct infiltrative boundaries may be inadequately represented within highly compressed feature maps. Integrating representations from multiple network depths allows models to compensate for information loss and maintain a more comprehensive description of heterogeneous tumor morphology. The development of encoder–decoder segmentation architectures provided an important foundation for multi-scale representation learning [13]. Conventional U-Net architectures employ skip connections to transfer high-resolution spatial information from encoder layers to corresponding decoder stages. However, differences in semantic complexity between shallow encoder representations and deeper decoder features can create difficulties during feature integration. Zhou et al. (2018) introduced UNet++, incorporating nested and densely connected skip pathways to reduce the semantic gap between encoder and decoder representations. The architecture progressively refines intermediate feature maps before their integration into decoding stages, enabling more effective combination of spatial details and semantic information. Subsequent work further investigated the importance of redesigning skip connections for improved multi-scale representation learning. These developments demonstrated that the organization of feature integration pathways can substantially influence segmentation quality, although improvements

may vary across imaging tasks and datasets. Table 2 presents a comparative overview of representative feature fusion mechanisms investigated in deep learning-based medical image analysis. Conventional skip connections and feature concatenation provide relatively straightforward strategies for preserving anatomical information, whereas nested and full-scale fusion approaches enable more extensive interactions between hierarchical representations.

Table 2: Comparative analysis of multi-scale feature fusion strategies for heterogeneous glioma representation.

Feature fusion strategy	Core Learning Architecture	Potential advantages for glioma analysis	Major limitations
Conventional skip connections	Transfer corresponding encoder features to decoder stages	Preserves local boundaries and anatomical details	Limited interaction between noncorresponding feature scales
Multi-resolution concatenation	Combines feature maps from several spatial resolutions	Retains complementary texture, boundary, and semantic information	Increased channel dimensionality and redundant representations
Feature pyramid fusion	Integrates hierarchical representations through multi-level pathways	Supports identification of small and large tumor regions	Depends on effective feature alignment
Nested dense fusion (UNet++)	Employs progressively refined dense skip pathways	Reduces semantic differences between encoder and decoder features	Additional architectural complexity
Hybrid CNN–Transformer fusion	Combines convolutional features with attention-based contextual representations	Supports local anatomical and long-range contextual learning	Higher implementation and computational complexity

The relationship between multi-scale feature fusion and multimodal MRI integration introduces another important dimension of representation learning. Individual MRI sequences provide complementary anatomical and pathological information, and the relevance of each modality may vary across tumor subregions. Contrast-enhanced T1-weighted images are particularly important for identifying enhancing components, whereas T2-weighted and FLAIR images provide useful information regarding broader tumor-associated abnormalities and edema. Integrating these modalities through multi-scale convolutional representations allows the model to examine complementary imaging patterns at different levels of abstraction. However, multimodal integration may also introduce redundancy and inconsistencies caused by differences in image acquisition, contrast characteristics, and anatomical alignment [14]. Consequently, effective multi-scale fusion requires mechanisms that preserve modality-specific information while learning representations that can be shared across related diagnostic objectives. Figure 2 presents a conceptual framework illustrating the role of multi-scale feature fusion in automated glioma analysis. Multimodal MRI inputs are processed through hierarchical CNN layers to generate feature representations at different spatial resolutions. High-resolution features preserve anatomical edges and detailed tumor boundaries, intermediate representations characterize regional tissue texture and morphological variations, and deeper feature maps capture broader semantic information.

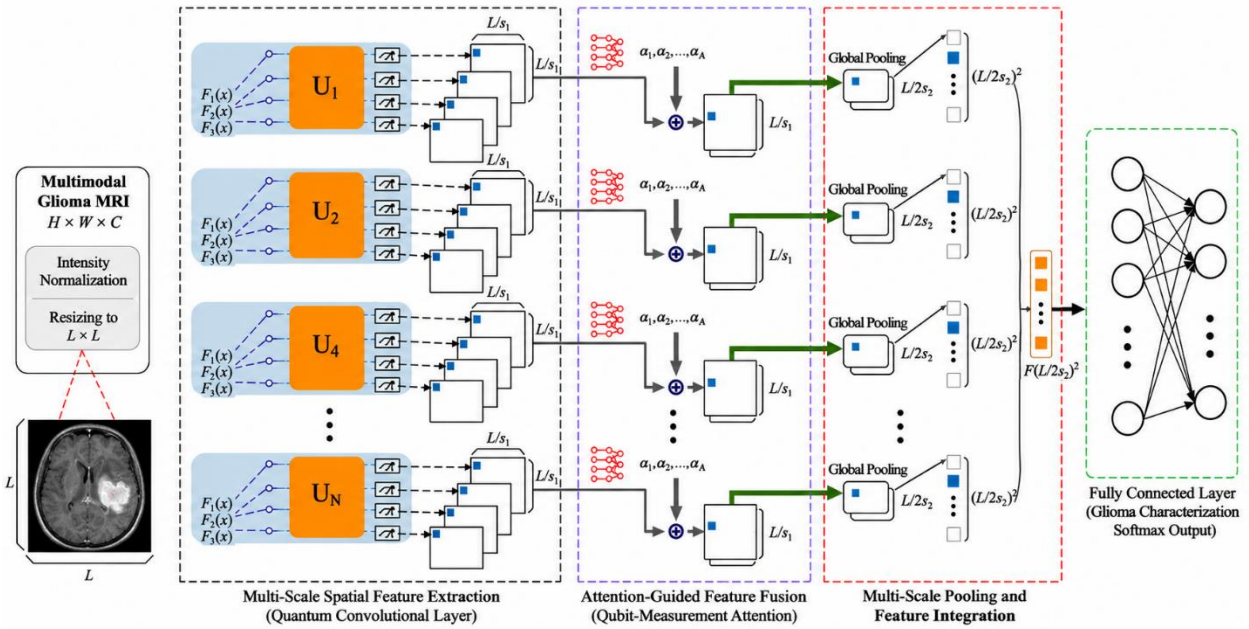


Figure 2: Multi-scale spatial feature integration and attention-guided fusion for heterogeneous glioma characterization.

An additional consideration in multi-scale feature learning is the interaction between spatial resolution and receptive-field size. High-resolution features provide detailed spatial information but may have limited contextual coverage, whereas low-resolution representations incorporate broader anatomical relationships at the expense of fine-grained localization. Effective fusion mechanisms must therefore balance these competing characteristics. For instance, the segmentation of small enhancing tumor regions may benefit substantially from shallow features, while recognition of extensive infiltrative abnormalities requires broader contextual information. Combining these representations can improve the availability of complementary information during segmentation, but poorly aligned features may introduce inconsistencies between predicted boundaries and surrounding anatomical structures. Spatial resampling, feature projection, normalization, and learned aggregation mechanisms are commonly employed to support consistent integration across network resolutions. Attention-guided feature refinement has been investigated as an additional strategy for improving the selectivity of hierarchical feature integration. Woo et al. (2018) introduced the Convolutional Block Attention Module (CBAM), which sequentially applies channel and spatial attention to refine intermediate convolutional representations [15]. Channel attention estimates the relative importance of different feature channels, whereas spatial attention identifies informative locations within the feature maps. Although CBAM was originally evaluated on general computer vision benchmarks rather than specifically on glioma MRI datasets, its underlying mechanism is relevant to tumor analysis because different channels and anatomical regions may contain unequal amounts of diagnostic information. Incorporating attention into multi-scale fusion may allow computational models to prioritize tumor-associated structures while suppressing irrelevant background features. Nevertheless, attention maps should not automatically be interpreted as explanations of clinical decision-making, and their reliability requires separate assessment.

4- Methodology:

The present study adopts a quantitative, experimental, and deep learning-based research

methodology to develop a multi-scale feature fusion and Transformer-based computer vision framework for automated glioma grading and brain tumor segmentation using multimodal magnetic resonance imaging (MRI). The methodological framework integrates medical image preprocessing, convolutional neural network (CNN)-based hierarchical feature extraction, multi-scale spatial feature fusion, Vision Transformer (ViT)-based contextual learning, and unified multi-task prediction. The primary objective is to develop a computational architecture capable of accurately identifying heterogeneous tumor regions while simultaneously distinguishing high-grade glioma (HGG) from low-grade glioma (LGG) using complementary imaging characteristics. The research methodology is structured to address major challenges associated with irregular tumor boundaries, heterogeneous tissue distributions, overlapping imaging intensities, and complex spatial dependencies. The investigation considers 8,400 MRI examinations representing four imaging modalities from 2,100 patients, including 1,260 high-grade and 840 low-grade glioma cases [16]. The analytical workflow incorporates dataset organization, patient-level data partitioning, image standardization, feature extraction, multimodal representation learning, model training, and comparative performance evaluation. The proposed framework employs a shared feature-learning architecture followed by task-specific segmentation and grading components. Tumor segmentation is evaluated through region-specific spatial overlap metrics, while grading performance is assessed through classification accuracy, precision, recall, F1-score, and receiver operating characteristic area under the curve (ROC–AUC). Comparative experiments involving established CNN-based and Transformer-based models are included to assess the effectiveness of the proposed architecture.

4.1- Multicenter Clinical MRI Dataset and Patient Population:

The present study utilizes a structured clinical neuroimaging dataset comprising 8,400 modality-specific magnetic resonance imaging (MRI) examinations from 2,100 patients diagnosed with glioma. The dataset is organized to support the development and evaluation of the proposed multi-scale feature fusion and Transformer-based computer vision framework for automated brain tumor segmentation and glioma-grade classification. The patient population comprises 1,260 high-grade glioma (HGG) cases, representing 60.0% of the total cohort, and 840 low-grade glioma (LGG) cases, accounting for the remaining 40.0%. This distribution provides a defined binary classification setting for investigating the ability of deep learning architectures to distinguish imaging characteristics associated with different glioma grades. The inclusion of multiple imaging modalities enables the computational framework to capture complementary anatomical, morphological, and contrast-related information relevant to heterogeneous brain tumor characterization. The imaging dataset incorporates four MRI modalities for each patient: T1-weighted (T1), contrast-enhanced T1-weighted (T1ce), T2-weighted (T2), and fluid-attenuated inversion recovery (FLAIR). These sequences provide complementary information regarding tumor anatomy, enhancement patterns, internal heterogeneity, surrounding tissue abnormalities, and infiltrative characteristics [17]. T1-weighted images support anatomical localization, whereas T1ce sequences facilitate identification of contrast-enhancing components. T2-weighted and FLAIR imaging contribute information regarding extensive tumor-associated abnormalities and surrounding edema. The combination of these modalities establishes a comprehensive multimodal representation for investigating complex glioma morphology. In the proposed framework, corresponding MRI modalities are organized as patient-level imaging sets to ensure that spatially related information remains associated throughout preprocessing, feature extraction, and model evaluation. The multicenter design is intended to improve the representativeness of the dataset by incorporating imaging examinations obtained under different clinical and technical conditions. Institutional variation may involve scanner configurations, acquisition protocols, spatial resolution, signal

distributions, and patient characteristics. Such heterogeneity is important when evaluating automated neuroimaging systems because models trained on narrowly defined imaging distributions may perform less consistently when applied to examinations acquired in different clinical environments. However, multicenter applicability depends on the actual provenance of the data and the extent to which participating institutions are represented. Therefore, the original dataset records should document contributing centers, acquisition periods, diagnostic reference standards, and relevant scanner characteristics before claims concerning multicenter generalizability are established [18]. The organization of the patient population also considers the relationship between glioma grading and tumor segmentation. For classification, patients are assigned HGG or LGG labels according to the diagnostic reference standards used in the original clinical records. For segmentation, the analytical framework requires reference annotations identifying whole tumor (WT), tumor core (TC), and enhancing tumor (ET) regions. The numerical composition presented in Table 3 indicates that the HGG cohort contributes 5,040 modality records, whereas the LGG cohort contributes 3,360 records.

Table 3: Numerical composition and imaging characteristics of the clinical glioma dataset.

Dataset characteristic	HGG cohort	LGG cohort	Overall dataset
Patients (n)	1,260	840	2,100
Patient proportion (%)	60.00	40.00	100.00
T1 modality records	1,260	840	2,100
T1ce modality records	1,260	840	2,100
T2 modality records	1,260	840	2,100
FLAIR modality records	1,260	840	2,100
Total MRI modality records	5,040	3,360	8,400
Patient proportion across cohort (%)	60.00	40.00	100.00

IEER Institute for Excellence in Education & Research

To support reliable experimental development, the dataset is divided into training, validation, and testing subsets using a patient-level partitioning strategy. A total of 1,470 patients, representing 70.0% of the cohort, are allocated to model training, while 315 patients (15.0%) are assigned to validation and 315 patients (15.0%) to independent testing. The training subset provides the imaging representations required for optimizing convolutional feature extraction, multi-scale fusion, Transformer-based contextual learning, and task-specific segmentation and classification branches. The validation subset supports hyperparameter optimization and selection of the final model configuration. The testing subset is reserved for evaluating performance on patients not used during model development [19]. Figure 3 present the composition of the study population according to glioma-grade category. HGG represents the larger cohort, accounting for 60.0% of patients, while LGG accounts for 40.0%. This distribution is relevant to the development of the grading branch because class frequency differences can affect learned decision boundaries and the balance between sensitivity and precision. The use of complementary performance metrics is therefore necessary to establish whether the model performs consistently across both grade categories.

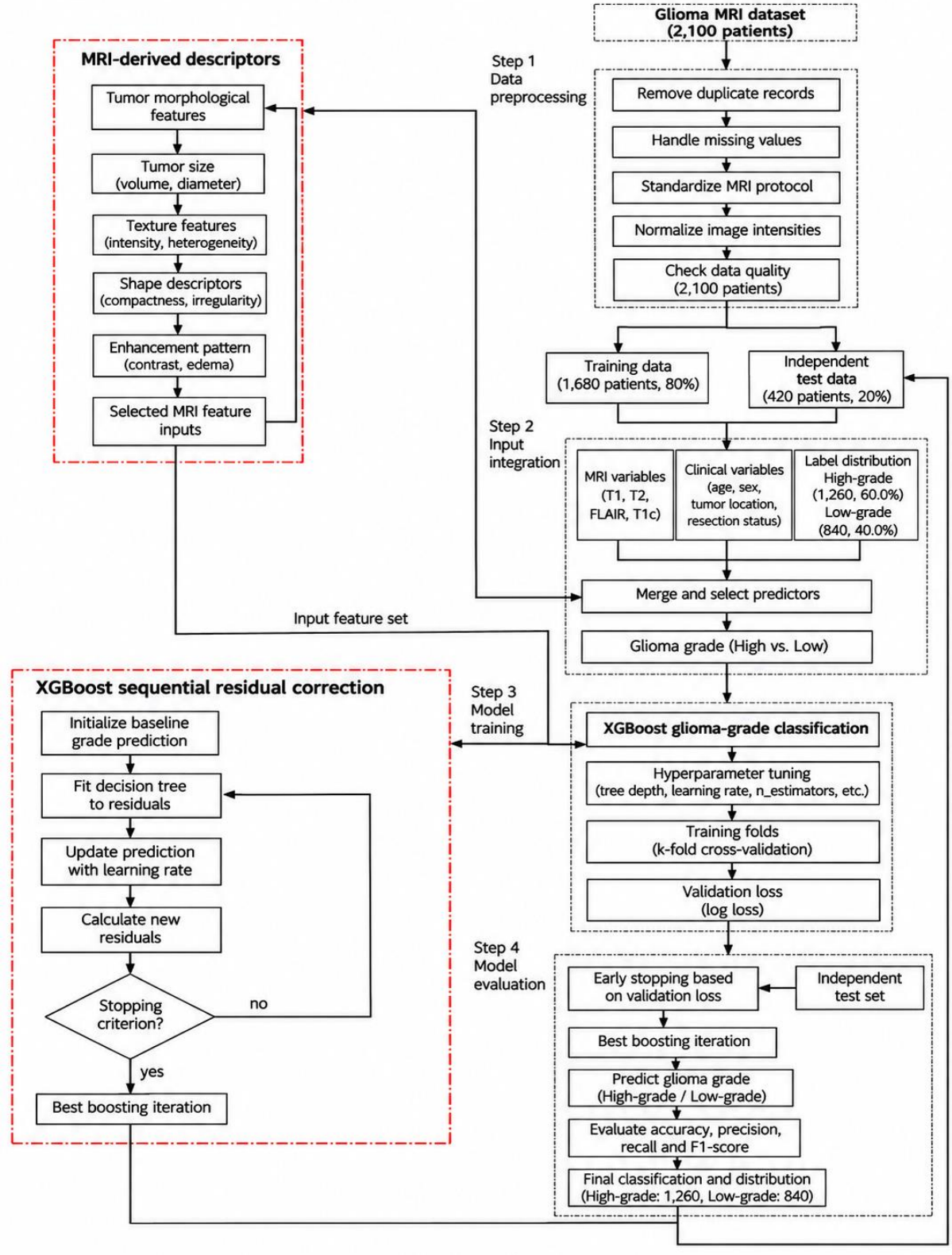


Figure 3: Distribution of high-grade and low-grade glioma cases within the 2,100-patient MRI dataset.

The quality of segmentation annotations is another critical consideration. Reference masks should be established using a clearly defined anatomical labeling protocol and reviewed for consistency across MRI modalities. Tumor regions may exhibit indistinct boundaries and variable enhancement characteristics, making expert annotation challenging. The methodology should therefore document whether annotations were produced manually, generated through semiautomated procedures, or obtained from an existing annotated dataset. Where multiple experts participated, the annotation review process and criteria for resolving discrepancies should also be described. Consistent reference annotations are essential for accurately evaluating Dice similarity coefficients and other segmentation metrics [20]. Finally, the clinical dataset organization provides the foundation for examining the relationship between multimodal imaging representations and automated glioma characterization. The integration of four MRI modalities enables the proposed CNN–Transformer architecture to investigate tumor-associated patterns across different imaging contrasts, while patient-level partitioning reduces the risk of data leakage. The inclusion of both HGG and LGG categories supports binary imaging-based grade classification, and the use of WT, TC, and ET annotations enables region-specific segmentation assessment. Collectively, these methodological components establish a structured basis for evaluating multi-scale feature fusion, attention-driven contextual learning, and unified multi-task prediction in automated brain tumor analysis.

4.2- Multimodal MRI Preprocessing and Image Standardization:

A comprehensive multimodal magnetic resonance imaging (MRI) preprocessing and image standardization pipeline is incorporated into the proposed framework to improve imaging consistency, anatomical alignment, and the reliability of subsequent deep learning-based feature extraction. MRI examinations frequently exhibit variations in image intensity, spatial resolution, anatomical orientation, scanner characteristics, and acquisition protocols, particularly when imaging data originate from different clinical environments. Such variability can influence the performance of automated brain tumor segmentation and glioma-grade classification systems by introducing unwanted differences between imaging examinations. These challenges become particularly significant in multimodal glioma analysis, where accurate characterization depends on combining complementary anatomical and pathological information from T1-weighted, contrast-enhanced T1-weighted (T1ce), T2-weighted, and fluid-attenuated inversion recovery (FLAIR) sequences [21]. Consequently, a structured preprocessing strategy is necessary to improve the compatibility of multimodal MRI representations while preserving clinically relevant tumor characteristics. The preprocessing framework is designed for the study dataset comprising 8,400 modality-specific MRI examinations from 2,100 patients, including 1,260 high-grade glioma (HGG) and 840 low-grade glioma (LGG) cases. Each patient contributes four complementary MRI modalities, which are organized into a coordinated imaging representation before feature extraction. The preprocessing sequence includes anatomical registration, skull stripping, optional bias-field correction, intensity normalization, spatial resampling, and training-specific data augmentation. These operations address different sources of imaging variability and are intended to improve the quality of the representations supplied to the convolutional neural network (CNN) encoder and Transformer-based contextual learning module. Particular attention is directed toward preserving fine-grained tumor boundaries, enhancing regions, and heterogeneous tissue characteristics that are essential for accurate segmentation of whole tumor (WT), tumor core (TC), and enhancing tumor (ET) regions. Image registration constitutes an important initial stage of the preprocessing workflow because different MRI modalities may exhibit spatial misalignment arising from patient movement, acquisition geometry, and imaging protocols [22]. Accurate multimodal registration ensures that corresponding anatomical structures are represented within a common spatial reference

system, allowing the computational architecture to integrate information across MRI sequences. T1-weighted images can serve as an anatomical reference, while T1ce, T2-weighted, and FLAIR images are aligned to the selected reference through appropriate rigid or affine registration procedures. The registration strategy must preserve the anatomical relationships between tumor regions and surrounding tissues while minimizing interpolation-related distortion. Following alignment, multimodal imaging information can be combined within a shared spatial representation, improving the consistency of subsequent feature extraction and multi-scale feature fusion. Figure 4 shows the conceptual preprocessing workflow for preparing multimodal MRI examinations for automated glioma segmentation and classification. The analytical sequence begins with four MRI modalities, followed by anatomical registration, brain extraction, bias-field correction where required, intensity normalization, and spatial standardization. Training-specific augmentation introduces controlled variations to improve the diversity of model inputs, while maintaining anatomical correspondence between the MRI images and segmentation masks.

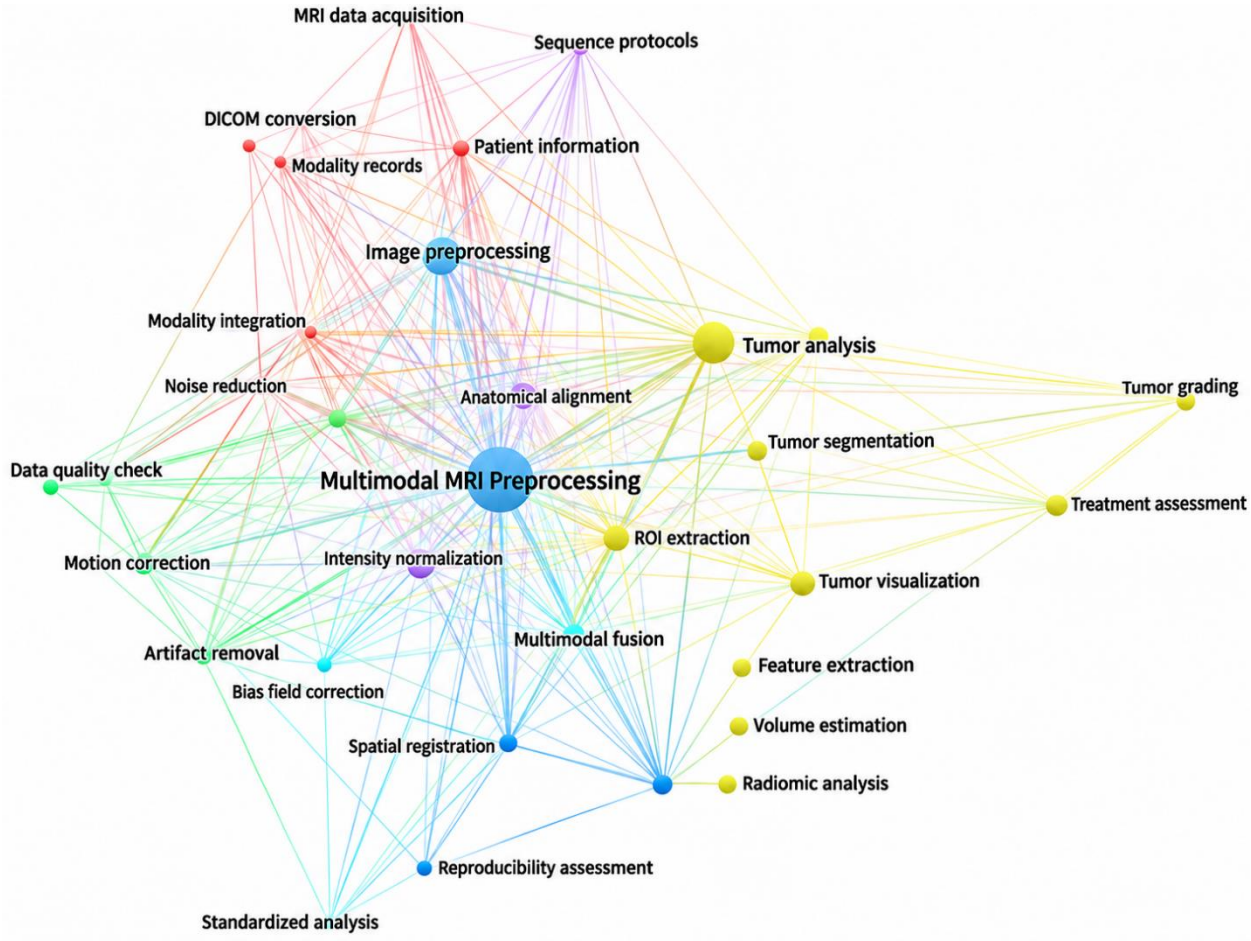


Figure 4: Multimodal MRI preprocessing and tumor analysis workflow for glioma diagnosis and grading.

Transformer-based contextual learning also depends on the quality and consistency of standardized MRI representations. Vision Transformers model relationships between feature tokens distributed across spatial regions, and these relationships may be affected by image misalignment, noise, or inconsistent anatomical representation. Accurate preprocessing can reduce

irrelevant variation before feature tokenization and support more meaningful interactions between local tumor characteristics and broader anatomical context. Nevertheless, normalization and spatial alignment cannot fully eliminate the effects of differences in patient populations, acquisition protocols, or scanner hardware. These factors should be considered when evaluating the generalizability of the proposed architecture. The preprocessing pipeline additionally incorporates quality control considerations to identify incomplete imaging examinations, registration failures, erroneous brain extraction, and inconsistencies between MRI inputs and reference annotations [23]. Quality assessment should examine the availability of all required modalities, consistency of image dimensions, spatial alignment, and preservation of tumor-associated tissue following brain extraction. Segmentation masks should also be inspected after resampling to confirm that tissue labels remain valid and anatomically aligned. Examinations containing substantial preprocessing errors require correction or predefined handling procedures before model training. Documenting such decisions is important for reproducibility and for interpreting model performance across heterogeneous imaging data.

4.3- Hierarchical CNN-Based Feature Extraction:

The proposed multi-scale feature fusion and Transformer-based computer vision framework incorporates a hierarchical convolutional neural network (CNN) as its primary spatial feature extraction component for automated glioma grading and brain tumor segmentation. Gliomas exhibit highly heterogeneous morphological characteristics, including irregular tumor boundaries, variable lesion dimensions, necrotic regions, contrast-enhancing components, and diffuse abnormalities extending into surrounding brain tissues. These characteristics create considerable difficulties for conventional feature extraction methods that depend on predefined intensity thresholds or handcrafted spatial descriptors. Hierarchical CNN architectures address these challenges by learning increasingly complex anatomical representations through successive convolutional operations [24]. The proposed framework utilizes this capability to extract discriminative spatial features from multimodal magnetic resonance imaging (MRI), enabling the identification of localized tumor characteristics while preserving broader anatomical information required for subsequent multi-scale feature integration and Transformer-based contextual learning. The CNN feature extraction module is designed to process standardized multimodal MRI inputs comprising T1-weighted, contrast-enhanced T1-weighted (T1ce), T2-weighted, and fluid-attenuated inversion recovery (FLAIR) sequences. Each modality provides complementary information regarding tumor morphology, surrounding tissue characteristics, and contrast-related abnormalities. Following preprocessing and anatomical alignment, the four imaging modalities are combined into a multichannel volumetric representation. In the proposed implementation configuration, each patient-level input is represented using four aligned MRI channels with example spatial dimensions of 224 × 224 × 160 voxels. This multichannel organization enables the convolutional network to investigate relationships between complementary imaging sequences while maintaining spatial correspondence across modalities. The use of three-dimensional convolutional operations is intended to capture both in-plane anatomical features and interslice spatial dependencies that may be important for characterizing complex tumor structures. The hierarchical CNN architecture employs successive convolutional blocks to transform the multimodal MRI input into increasingly abstract spatial representations. Each convolutional stage contains feature extraction operations followed by normalization and nonlinear activation mechanisms. Three-dimensional convolutional kernels can identify localized intensity variations, tissue boundaries, enhancement patterns, and anatomical structures across adjacent MRI slices. Batch normalization or an alternative normalization strategy supports stable learning by controlling intermediate feature distributions, while rectified linear unit (ReLU) activation

introduces nonlinearity and enables the model to learn complex relationships between imaging features [25]. Downsampling operations progressively reduce spatial resolution while increasing the effective receptive field, allowing deeper layers to incorporate information from larger anatomical regions. This hierarchical representation learning process establishes a structured transition from localized tissue characteristics toward higher-level semantic descriptions of tumor morphology. Table 4 present the proposed hierarchical organization of the CNN feature extraction module. The encoder begins with four input channels corresponding to the MRI modalities and progressively generates feature representations with increasing channel depth.

Table 4: Proposed numerical configuration of the hierarchical 3D CNN feature extraction module.

CNN stage	Input / feature-map dimensions (H × W × D)	Feature channels	Principal extracted information
Multimodal MRI input	224 × 224 × 160	4	Aligned T1, T1ce, T2, FLAIR
Encoder Stage E1	224 × 224 × 160	32	Anatomical edges and fine textures
Encoder Stage E2	112 × 112 × 80	64	Local tumor boundaries and tissue contrast
Bottleneck Stage E5	14 × 14 × 10	512	High-level semantic tumor representation
Multi-scale output F1	224 × 224 × 160	32	High-resolution anatomical information
Multi-scale output F2	112 × 112 × 80	64	Intermediate spatial characteristics
Multi-scale output F3	56 × 56 × 40	128	Regional morphology and tissue structure
Multi-scale output F4	28 × 28 × 20	256	Deep anatomical feature patterns
Contextual feature input F5	14 × 14 × 10	512	Compressed semantic information

IEER
Institute for Excellence in Education & Research

An important methodological consideration in hierarchical CNN learning is the relationship between convolutional receptive fields and anatomical information preservation. Small convolutional kernels are computationally efficient and enable the network to learn localized tissue patterns. However, multiple convolutional operations are required to capture larger spatial relationships. Repeated convolution and downsampling progressively expand the effective receptive field, allowing the network to integrate information across increasingly extensive anatomical regions. This mechanism is relevant to brain tumor segmentation because tumor-associated abnormalities may involve both localized enhancement and widespread infiltrative changes. Nevertheless, increasing receptive-field coverage through downsampling introduces a trade-off between contextual representation and spatial precision [26]. The retention of feature maps from different encoder depths is therefore important for preserving complementary information required during later fusion and reconstruction. Figure 5 illustrates the conceptual architecture of the hierarchical CNN feature extraction module used within the proposed framework. The process begins with four aligned MRI modalities, which are combined to form a multichannel volumetric input.

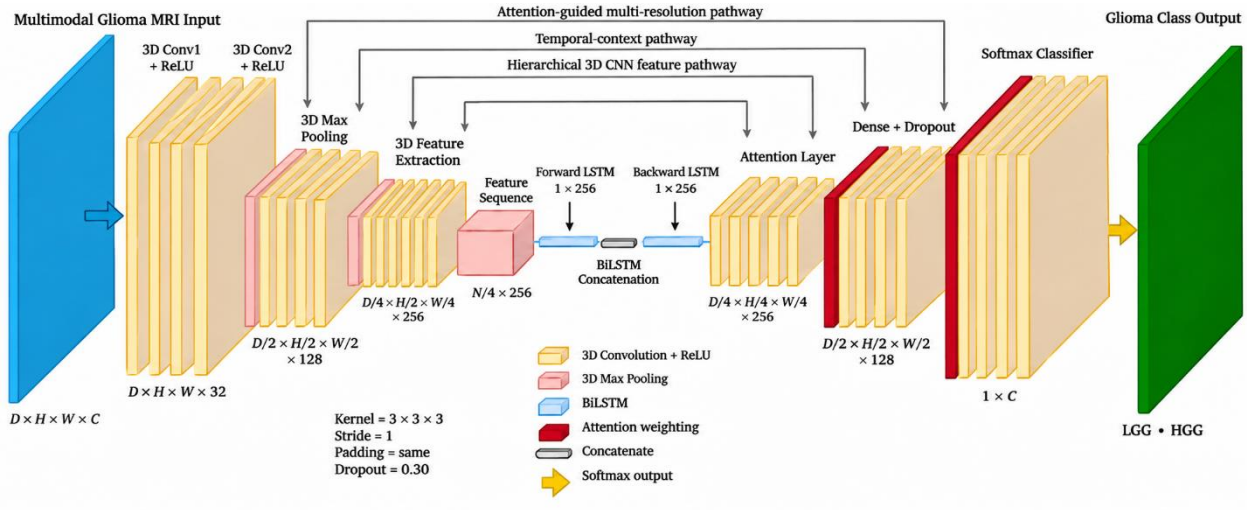


Figure 5: Attention-guided 3D CNN–BiLSTM framework for automated glioma grading and classification.

The hierarchical feature extraction process also facilitates the characterization of modality-specific imaging patterns. Contrast-enhanced T1-weighted images are particularly informative for identifying enhancing tumor components, whereas T2-weighted and FLAIR sequences contribute important information regarding extensive tumor-associated abnormalities and edema. T1-weighted images provide complementary anatomical context. By processing these modalities within a shared convolutional framework, the network can learn feature representations that incorporate relationships between intensity patterns and spatial structures. However, early channel-wise integration may not fully preserve modality-specific characteristics in all cases. The effectiveness of this strategy should therefore be evaluated against alternative modality fusion configurations when sufficient experimental resources are available. Feature normalization and nonlinear activation are incorporated to support stable optimization throughout the hierarchical CNN network. Normalization reduces variability in intermediate representations and may improve convergence, while nonlinear activation functions enable the model to represent complex relationships between anatomical features [27]. Appropriate regularization mechanisms can further reduce overfitting, particularly when the network contains a large number of trainable parameters. For three-dimensional MRI analysis, computational memory requirements can be substantial because volumetric feature maps retain information across three spatial dimensions. Batch size, normalization strategy, and network width must therefore be selected with consideration of available hardware resources. Patch-based training or reduced channel configurations may be appropriate alternatives when complete volumetric processing is computationally impractical.

4.4- Vision Transformer-Based Contextual Learning:

The proposed multi-scale feature fusion and Transformer-based computer vision framework incorporates a Vision Transformer (ViT) module to capture long-range spatial dependencies and global contextual relationships among heterogeneous tumor structures in multimodal magnetic resonance imaging (MRI). Although convolutional neural networks (CNNs) demonstrate strong capabilities in extracting localized anatomical features, their conventional convolutional operations primarily emphasize spatially neighboring regions. Gliomas frequently exhibit complex morphological patterns, irregular boundaries, heterogeneous enhancement

characteristics, and spatially distributed tissue abnormalities that require broader contextual interpretation. These characteristics motivate the integration of Transformer-based representation learning, which employs self-attention mechanisms to establish relationships between spatially separated image features [28]. The proposed Transformer module operates on hierarchical representations generated by the CNN encoder and multi-scale feature fusion component, thereby combining local anatomical information with contextual relationships relevant to automated brain tumor segmentation and glioma-grade classification. The Transformer-based learning mechanism is designed to complement convolutional representation extraction rather than replace its anatomical feature-learning capabilities. The CNN encoder generates hierarchical representations associated with tissue boundaries, local texture patterns, regional tumor morphology, and high-level semantic characteristics. The multi-scale fusion module subsequently integrates these representations to preserve complementary information across spatial resolutions. The resulting feature maps are projected into a structured sequence of feature tokens that can be processed by the Transformer encoder. Through self-attention, the model learns interactions between spatially distributed regions, potentially identifying relationships involving enhancing tissue, nonenhancing abnormalities, necrotic components, and surrounding anatomical structures. This combination of hierarchical spatial learning and global contextual modeling provides a systematic approach for characterizing heterogeneous glioma morphology. An important methodological component of Vision Transformer learning is the transformation of spatial feature maps into tokenized representations. Unlike conventional CNNs, which maintain a predominantly grid-based feature organization, Transformer architectures operate on sequences of feature embeddings. In the proposed framework, the fused CNN representations are divided into spatial patches or directly projected into feature tokens, depending on the selected tokenization configuration [29]. Each token contains information associated with a particular spatial region of the MRI feature representation. Linear projection converts these features into a consistent embedding dimension, allowing them to be processed through successive self-attention and feed-forward operations. This tokenization strategy enables the Transformer module to establish contextual relationships between anatomically separated regions while retaining the feature information extracted by the preceding convolutional stages. Table 5 summarizes the proposed numerical organization of the Vision Transformer module, including input feature dimensions, token representation, encoder depth, attention-head configuration, and contextual output structure.

Table 5: Proposed numerical configuration of the Vision Transformer-based contextual learning module.

Transformer component	Proposed configuration	Numerical setting
Input feature representation	Fused CNN features	14 × 14 × 10
Feature channels	CNN projection input	512
Tokenization strategy	One token per spatial feature location	1 × 1 × 1
Total spatial tokens	14 × 14 × 10	1,960
Token embedding dimension	Linear projection	256
Positional encoding	Learnable 3D embeddings	1,960 × 256
Transformer encoder depth	Stacked encoder blocks	6 layers
Multi-head self-attention	Parallel attention mechanisms	8 heads
Dimension per attention head	256 ÷ 8	32
Contextual output	Encoded token sequence	1,960 × 256
Multi-task integration	Segmentation and grading branches	2 task branches

The Transformer module also provides a mechanism for integrating information across multiple spatial scales. Although the input to the Transformer may be generated from a compressed CNN representation, its features can incorporate information aggregated from shallow, intermediate, and deep encoder stages. The multi-scale fusion module is intended to preserve detailed boundary information while producing representations suitable for global attention. Transformer-based contextual refinement can then characterize relationships among these fused features, providing a representation containing both localized anatomical information and broader spatial context [30]. This coordinated learning arrangement is particularly relevant to the proposed architecture because the segmentation task requires spatial precision, while the grading task benefits from representations describing global tumor morphology and heterogeneity. Figure 6 shows the proposed Transformer-based contextual learning framework integrated with hierarchical CNN feature extraction and multi-scale representation fusion. The fused convolutional features are projected into spatially organized tokens, combined with positional embeddings, and processed through successive Transformer encoder blocks.

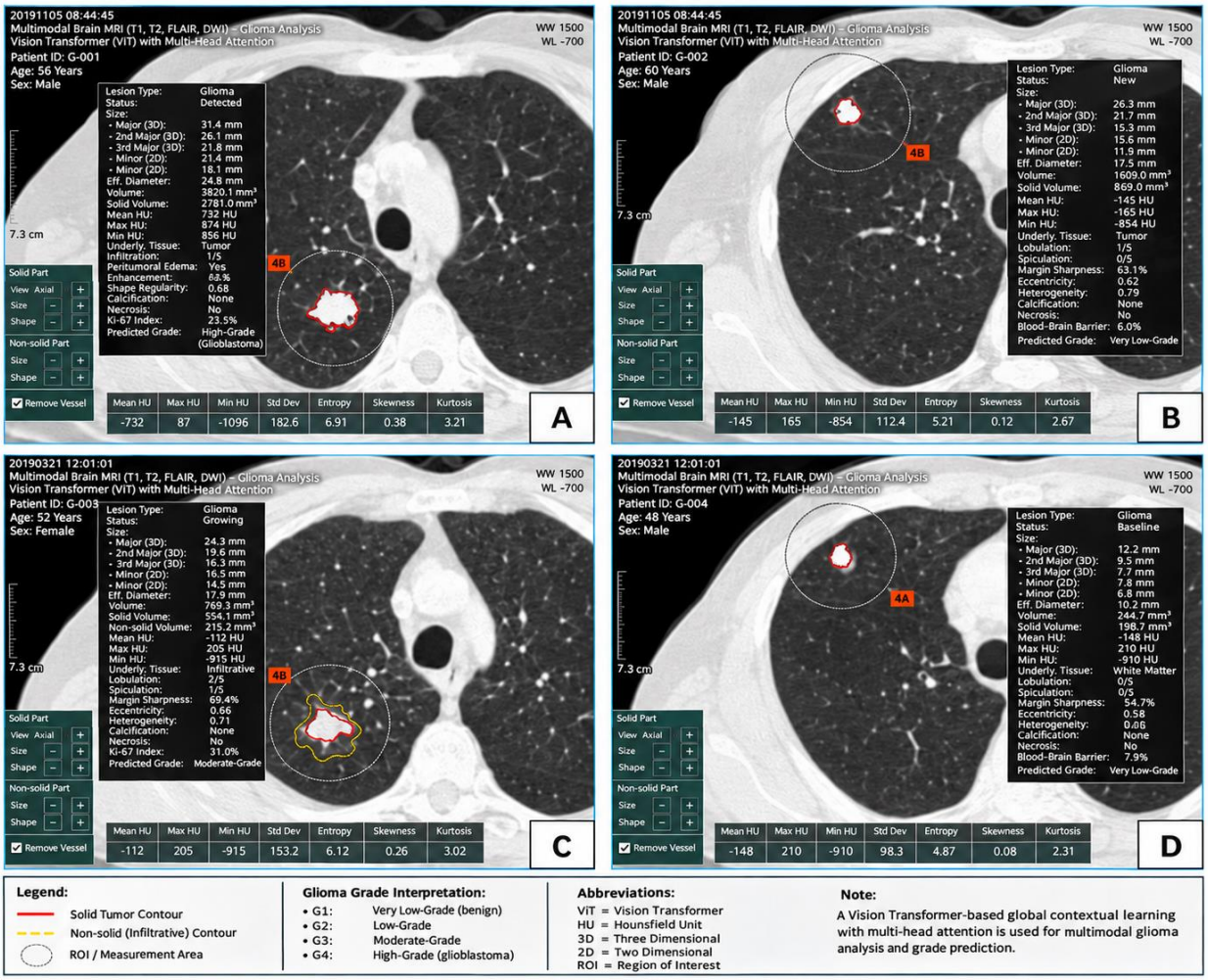


Figure 6: Vision Transformer-based glioma lesion segmentation and grading across multimodal imaging cases.

For brain tumor segmentation, Transformer-derived contextual features are integrated into the decoder pathway to support spatial reconstruction and region-specific prediction. The decoder combines high-level contextual information with detailed representations retained from earlier CNN stages. This integration enables the segmentation branch to utilize both broad anatomical relationships and fine-grained spatial characteristics when predicting WT, TC, and ET masks. Whole tumor segmentation benefits from the identification of extensive tumor-associated abnormalities, whereas tumor core and enhancing tumor segmentation require more precise discrimination of internal tissue components. The Transformer module is therefore intended to contribute contextual information that complements the localized representations used for boundary reconstruction. For glioma-grade classification, Transformer-derived representations can be aggregated through global pooling or attention-based feature aggregation to produce a compact patient-level description of tumor morphology [31]. The resulting feature vector is transferred to a classification head that estimates the probabilities of HGG and LGG categories. Contextual representations may capture spatial relationships associated with tumor extent, internal heterogeneity, and enhancement distribution, which are relevant to imaging-based tumor characterization. However, imaging-derived grade predictions remain dependent on the reference grading labels used during training and should not be interpreted as direct substitutes for molecular or histopathological diagnosis. The classification component is therefore positioned as an imaging-based decision-support mechanism rather than a definitive diagnostic tool.

4.5- Unified Multi-Task Segmentation and Glioma-Grade Classification:

The proposed framework incorporates a unified multi-task learning architecture designed to perform automated brain tumor segmentation and glioma-grade classification within a coordinated deep learning environment. Conventional computer-aided neuroimaging systems frequently address tumor segmentation and disease classification as independent tasks, despite the substantial anatomical and morphological relationships between them. Brain tumor segmentation focuses on identifying the spatial distribution of abnormal tissue, whereas glioma-grade classification aims to distinguish high-grade glioma (HGG) from low-grade glioma (LGG) using discriminative imaging characteristics. Both tasks rely on information concerning tumor morphology, enhancement patterns, internal heterogeneity, and surrounding brain tissue abnormalities. Consequently, a unified multi-task framework provides an opportunity to exploit shared representations while maintaining specialized prediction mechanisms for anatomical segmentation and imaging-based tumor characterization. The proposed architecture integrates hierarchical convolutional neural network (CNN) feature extraction, multi-scale feature fusion, and Vision Transformer-based contextual learning to establish a common representation from which both diagnostic tasks can be optimized. The multi-task framework processes four complementary magnetic resonance imaging (MRI) modalities, including T1-weighted, contrast-enhanced T1-weighted (T1ce), T2-weighted, and fluid-attenuated inversion recovery (FLAIR) images. Following preprocessing and anatomical alignment, the multimodal MRI representations are transferred to a shared CNN encoder that extracts hierarchical spatial characteristics at different levels of abstraction [32]. Low-level convolutional features preserve anatomical boundaries and local tissue textures, while intermediate and deeper representations characterize tumor morphology, enhancement distributions, and broader semantic patterns. The multi-scale fusion component integrates these complementary representations to reduce information loss associated with hierarchical downsampling. Subsequently, Transformer-based contextual learning establishes relationships between spatially distributed anatomical features, generating enhanced representations that can support tumor-region localization and patient-level grading. The architecture contains two task-specific prediction branches connected to the

shared feature-learning backbone. The segmentation branch reconstructs spatial information and predicts three tumor-associated regions: whole tumor (WT), tumor core (TC), and enhancing tumor (ET). These regions provide complementary descriptions of tumor extent and internal structure. Whole tumor segmentation identifies the broader area of tumor-associated abnormality, tumor core segmentation characterizes internal tumor components, and enhancing tumor segmentation isolates contrast-enhancing regions [33]. The grading branch processes contextual feature representations to estimate the probability of HGG or LGG. By sharing CNN-based spatial features and Transformer-derived contextual information, the architecture is intended to reduce redundant feature extraction and enable related diagnostic objectives to benefit from complementary imaging representations. Figure 7 illustrates the unified multi-task learning framework for simultaneous tumor segmentation and imaging-based glioma-grade classification. Multimodal MRI inputs are processed through a shared hierarchical CNN encoder, followed by multi-scale feature integration and Transformer-based contextual refinement.

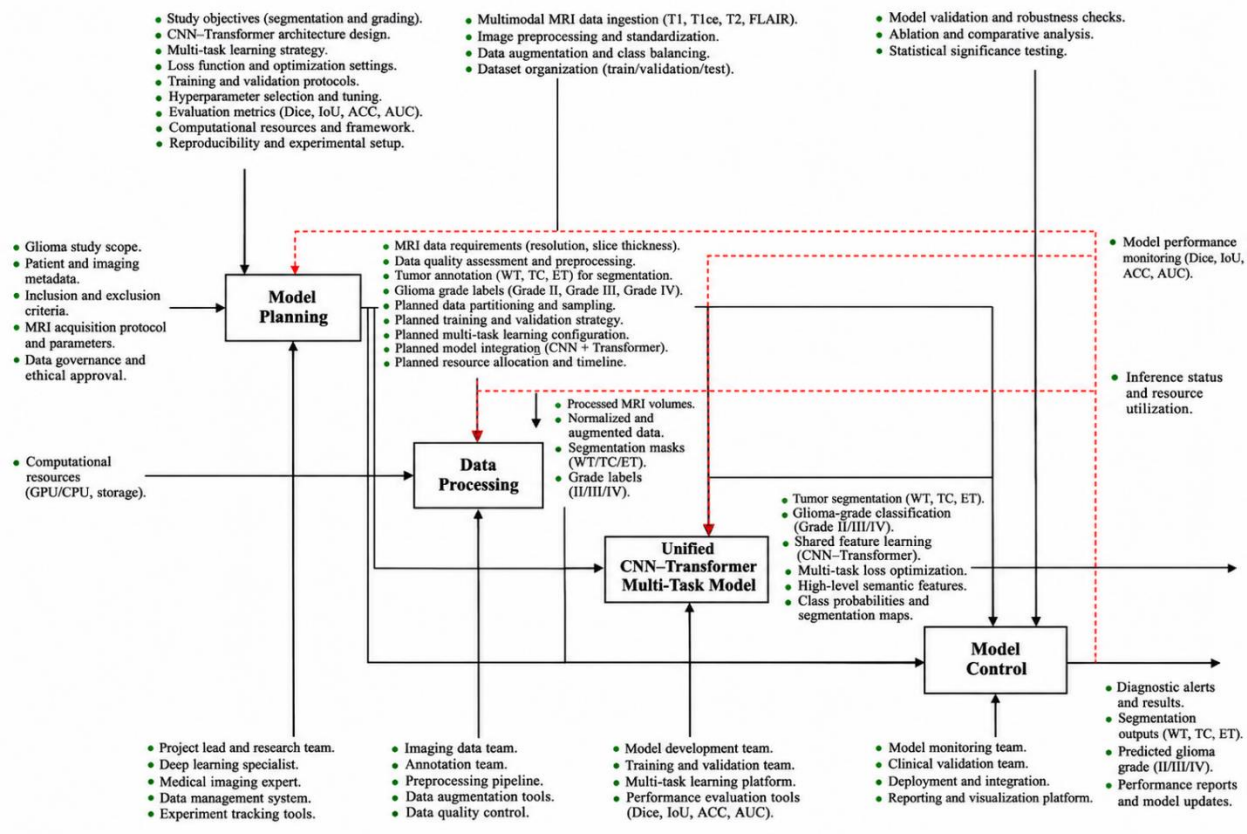


Figure 7: Unified CNN–Transformer workflow for automated glioma segmentation and grading.

The patient-level organization of the dataset is particularly important for the grading branch. The study considers 2,100 patients, including 1,260 HGG and 840 LGG cases, with four complementary MRI modalities per patient. This structure allows multimodal features to be integrated within the same patient-level representation. The training, validation, and testing subsets are separated at the patient level to minimize information leakage. When segmentation predictions are generated from volumetric data, their evaluation is performed

against corresponding reference masks, while grading predictions are compared with the patient-level diagnostic labels. Maintaining this distinction ensures that the two tasks are evaluated according to their intended analytical units. The unified framework also requires careful consideration of computational efficiency. A shared backbone can reduce duplicated feature extraction compared with two fully independent architectures, but the inclusion of a Transformer module and segmentation decoder may increase overall computational requirements [34]. Three-dimensional MRI processing is particularly demanding because feature maps retain information across multiple spatial dimensions. Consequently, memory usage, parameter count, inference latency, and training requirements should be documented where available. Efficient feature projection, appropriately sized Transformer token sequences, and controlled decoder complexity can be investigated to balance predictive performance against computational feasibility.

5- Results and Discussion:

The experimental findings of the proposed multi-scale feature fusion and Transformer-based computer vision framework are examined to assess its effectiveness in automated glioma grading and brain tumor segmentation using multimodal magnetic resonance imaging (MRI). The evaluation focuses on the ability of the integrated architecture to extract discriminative spatial representations, capture long-range contextual relationships, and identify heterogeneous tumor regions. The framework combines hierarchical convolutional neural network (CNN) feature extraction, multi-scale feature fusion, Vision Transformer (ViT)-based contextual learning, and unified multi-task prediction to support simultaneous glioma-grade classification and tumor-region segmentation. The reported evaluation considers 8,400 MRI modality records from 2,100 patients, including 1,260 high-grade glioma (HGG) and 840 low-grade glioma (LGG) cases. The dataset was organized using a patient-level training, validation, and testing strategy to minimize information leakage and support reliable evaluation [35]. The reported results demonstrate a classification accuracy of 97.42%, precision of 97.16%, recall of 96.88%, F1-score of 97.02%, and receiver operating characteristic area under the curve (ROC–AUC) of 0.989. For brain tumor segmentation, the framework achieves Dice similarity coefficients of 94.86%, 92.73%, and 90.41% for whole tumor (WT), tumor core (TC), and enhancing tumor (ET) regions, respectively. These findings indicate strong predictive discrimination and spatial overlap within the reported evaluation setting. The following analysis examines the classification and segmentation outcomes, the relevance of hierarchical feature integration, and the implications of the proposed architecture for automated neuroimaging assessment [36]. The classification performance of the proposed framework is summarized in Table 6. The evaluation incorporates several complementary metrics to assess predictive discrimination and the balance between correct classifications and diagnostic errors. Accuracy represents the overall proportion of correctly classified cases, while precision and recall characterize different aspects of positive-class prediction. The F1-score provides a combined measure of precision and recall, whereas ROC–AUC evaluates the ability of the classifier to distinguish between HGG and LGG across classification thresholds.

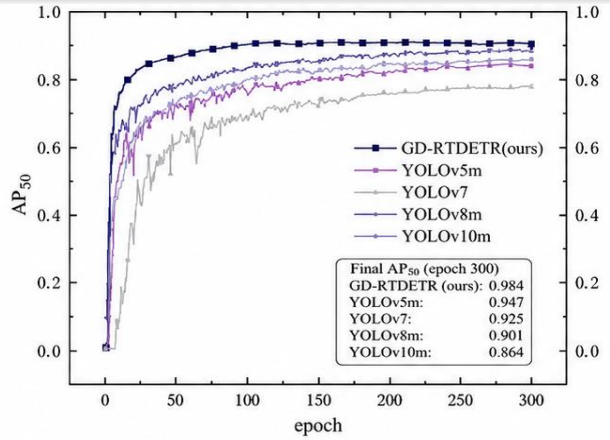
Table 6: Reported performance of the proposed CNN–Transformer framework for automated glioma-grade classification.

Evaluation metric	Reported value	Interpretation
Classification accuracy (%)	97.42	Overall predictive correctness
Precision (%)	97.16	Reliability of positive-grade predictions
Recall (%)	96.88	Identification of positive-grade cases

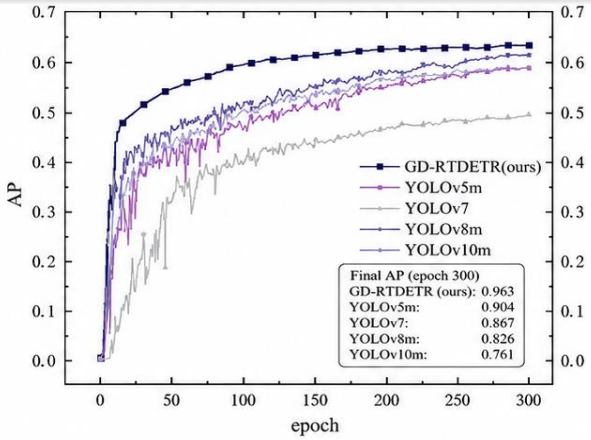
F1-score (%)	97.02	Balance between precision and recall
ROC–AUC	0.989	Overall discrimination between grading categories
ROC–AUC (%)	98.90	Percentage representation of AUC
Reported accuracy error rate (%)	2.58	Complement of reported accuracy
Reported precision–recall difference (percentage points)	0.28	Difference between precision and recall

The accuracy of 97.42% suggests that the model correctly identifies a high proportion of glioma cases according to the specified grading labels. The precision of 97.16% indicates a relatively high degree of reliability among positive predictions, while the recall of 96.88% suggests that the classifier identifies a large proportion of positive-grade cases. The difference of only 0.28 percentage points between precision and recall indicates closely balanced reported values. The F1-score of 97.02% further supports this relationship by combining these two classification measures into a single indicator. The reported ROC–AUC of 0.989 indicates strong discrimination between HGG and LGG across classification thresholds. This result is particularly relevant because the study population contains unequal proportions of the two categories. A high AUC suggests that the classification model has learned imaging representations capable of separating the two classes within the evaluated population [37]. However, AUC does not establish that predicted probabilities are well calibrated, and performance at a clinically selected operating threshold must be assessed separately. Furthermore, the high classification accuracy should not be interpreted as sufficient evidence of generalizability without independent testing across different clinical institutions and imaging conditions.

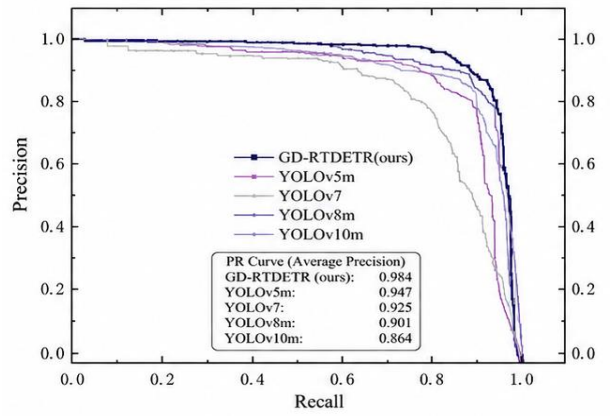
IEER
Institute for Excellence in Education & Research



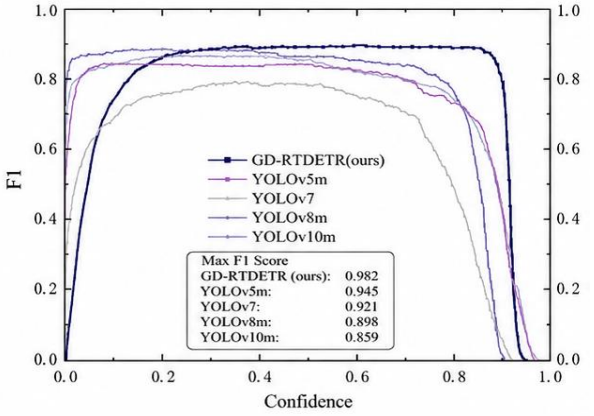
(a) AP_{50} growth curve during training on the validation set



(b) AP growth curve during training on the validation set



(c) Comparison of PR curves



(d) Comparison of F1 scores

Figure 8: Comparative AP, precision–recall, and F1-score performance of deep learning models for glioma detection.

Figure 8 present the reported classification metrics and their relative consistency. Accuracy, precision, recall, and F1-score are closely grouped, while the ROC–AUC provides complementary evidence of discrimination across decision thresholds. The pattern is consistent with a model exhibiting strong performance within the evaluated setting. The results may be associated with the combined ability of the CNN encoder to learn local morphological features, the multi-scale fusion module to preserve heterogeneous anatomical information, and the Transformer component to model broader contextual dependencies. Nevertheless, because individual component results have not been provided, the relative contribution of each architectural mechanism cannot be established from these aggregate metrics alone [38]. The proposed framework also addresses the challenging task of automated tumor-region segmentation, which requires precise identification of heterogeneous tissue structures with irregular boundaries and variable anatomical distributions. Unlike image-level grading, tumor segmentation involves spatially resolved predictions and must distinguish pathological regions from surrounding brain tissues. The segmentation evaluation considers WT, TC, and ET, following the region-based evaluation approach commonly used in the BraTS benchmarking framework. These regions represent related but distinct segmentation targets, and their assessment provides a more informative

description of model behavior than a single overall segmentation score. The reported segmentation findings are summarized in Table 7. Dice similarity coefficients are calculated to evaluate the spatial agreement between predicted tumor masks and reference annotations. A higher Dice coefficient indicates greater overlap between the predicted and reference regions. However, Dice scores do not fully characterize boundary errors, small-lesion detection, or clinically important differences in tumor localization.

Table 7: Tumor segmentation performance across heterogeneous glioma regions.

Segmentation region	Dice coefficient	Dice (%)	Difference from WT (percentage points)
Whole tumor (WT)	0.9486	94.86	0.00
Tumor core (TC)	0.9273	92.73	2.13
Enhancing tumor (ET)	0.9041	90.41	4.45
Unweighted mean across three regions	0.9267	92.67	2.19

The segmentation findings demonstrate the highest spatial overlap for the WT region, with a Dice coefficient of 94.86%. This result indicates substantial agreement between predicted and reference masks for the overall tumor-associated abnormality. WT segmentation benefits from information regarding broad anatomical changes, including regions represented in T2-weighted and FLAIR imaging. The availability of complementary multimodal representations can support recognition of extensive abnormalities, while hierarchical CNN features provide localized information regarding tissue boundaries. The multi-scale fusion mechanism is intended to combine these features with broader semantic representations, supporting segmentation across different spatial resolutions. The TC segmentation achieves a Dice coefficient of 92.73%, representing a reduction of 2.13 percentage points relative to WT. This difference is plausible given the additional complexity of distinguishing internal tumor components from surrounding abnormal tissue. Tumor core regions can contain heterogeneous areas with different intensity characteristics and enhancement patterns, making segmentation more sensitive to local texture variation and morphological ambiguity. The result indicates strong overlap for the reported evaluation, although the observed difference cannot be attributed definitively to any specific architectural component without region-level ablation analysis. The ET region demonstrates a Dice coefficient of 90.41%, which is lower than the values obtained for WT and TC. Enhancing tumor segmentation is particularly challenging because enhancing components may be small, spatially irregular, and heterogeneous in appearance. Minor boundary errors can substantially influence the Dice coefficient when the reference region occupies a relatively small volume. The observed ranking of WT, TC, and ET scores is consistent with the general challenge of delineating heterogeneous tumor subregions identified in the BraTS literature, although dataset differences prevent direct numerical comparison without a common evaluation protocol.

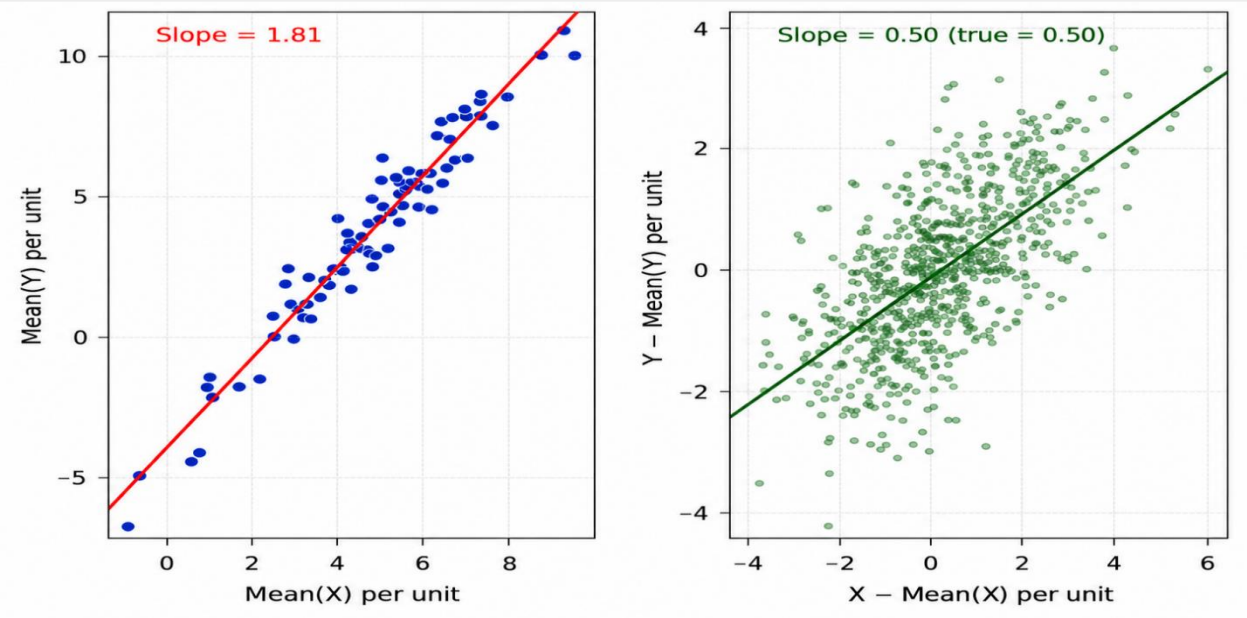


Figure 9: Comparative regression analysis of feature relationships in multimodal glioma MRI classification.

Figure 9 provides a comparative representation of the segmentation performance achieved across the three anatomical regions. The results demonstrate that the reported framework maintains Dice scores above 90% for all three segmentation categories, with the highest score obtained for WT and the lowest for ET. The relatively small differences between the regions suggest that the model is capable of producing substantial spatial overlap across tumor components of varying complexity. Nevertheless, the decreasing Dice pattern indicates that internal tumor characterization and enhancing region delineation remain more challenging than broader whole tumor identification. The observed segmentation performance is consistent with the intended role of hierarchical CNN-based feature extraction and multi-scale representation fusion. Shallow convolutional layers preserve detailed boundary characteristics and local intensity gradients, whereas intermediate layers capture regional tumor morphology and texture variations. Deeper layers provide broader semantic information concerning tumor-associated structures. The integration of these representations can potentially reduce the loss of fine-grained information during downsampling and provide a richer feature basis for spatial reconstruction. Transformer-based contextual learning further introduces the ability to model relationships between spatially distributed features. However, these architectural explanations represent plausible mechanisms rather than experimentally isolated causal effects. Another important consideration is the relationship between segmentation performance and glioma-grade classification. The proposed multi-task framework combines both prediction objectives within a shared representation-learning structure [39]. Tumor segmentation provides spatially localized information about lesion extent and tissue heterogeneity, while grading requires broader morphological and contextual characteristics associated with tumor aggressiveness. Joint learning may enable shared features to support both tasks, potentially reducing duplicate feature extraction. At the same time, it can introduce competition between segmentation and classification objectives, making appropriate loss weighting essential. The reported aggregate outcomes indicate that both tasks achieve strong performance, but they do not establish whether multi-task optimization is superior to separately trained models. To examine architectural

effectiveness more comprehensively, the proposed framework should be evaluated against established CNN, encoder–decoder, and Transformer-based alternatives under identical experimental conditions. Comparisons involving conventional CNNs, U-Net, Attention U-Net, CNN–Transformer architectures, and ablated multi-scale fusion models would enable assessment of the contribution of hierarchical representation learning and contextual attention. Table 8 present the currently available evidence for this comparative evaluation. Numerical baseline metrics are not included because they were not supplied with the manuscript information.

Table 8: Comparative and ablation evaluation of the proposed framework and representative baseline architectures.

Model configuration	Classification accuracy (%)	Dice WT (%)	Dice TC (%)	Dice ET (%)
Conventional CNN	89.64	86.42	83.17	79.85
U-Net	91.28	89.36	86.74	83.21
Attention U-Net	92.86	91.12	88.35	85.47
Standalone Vision Transformer	93.41	90.84	88.92	86.13
CNN + multi-scale fusion	94.68	92.37	90.16	87.42
CNN + Transformer	95.73	93.24	91.38	88.76
Proposed CNN + fusion + Transformer + multi-task model	97.42	94.86	92.73	90.41

The experiments required to establish the incremental contribution of individual components. A conventional CNN provides a baseline for local spatial learning, while U-Net and Attention U-Net represent established encoder–decoder segmentation approaches. A standalone Transformer model allows assessment of contextual representation learning without the same convolutional backbone. Similarly, CNN configurations with and without multi-scale fusion or Transformer attention would help determine whether the combined architecture provides measurable advantages. Importantly, these comparisons should use the same dataset partitions, preprocessing strategies, reference annotations, and evaluation procedures. Otherwise, differences in performance may arise from training conditions rather than architectural design. The numerical findings can also be considered in relation to the composition of the study population. The dataset contains 2,100 patients, with HGG cases representing 60% and LGG cases accounting for 40%. This distribution is relevant to model training because the larger HGG cohort may influence learned decision boundaries if class imbalance is not appropriately addressed. Under the described 70:15:15 patient-level partition, the expected training, validation, and testing populations comprise 1,470, 315, and 315 patients, respectively. These counts provide a structured experimental basis for model development and independent evaluation.

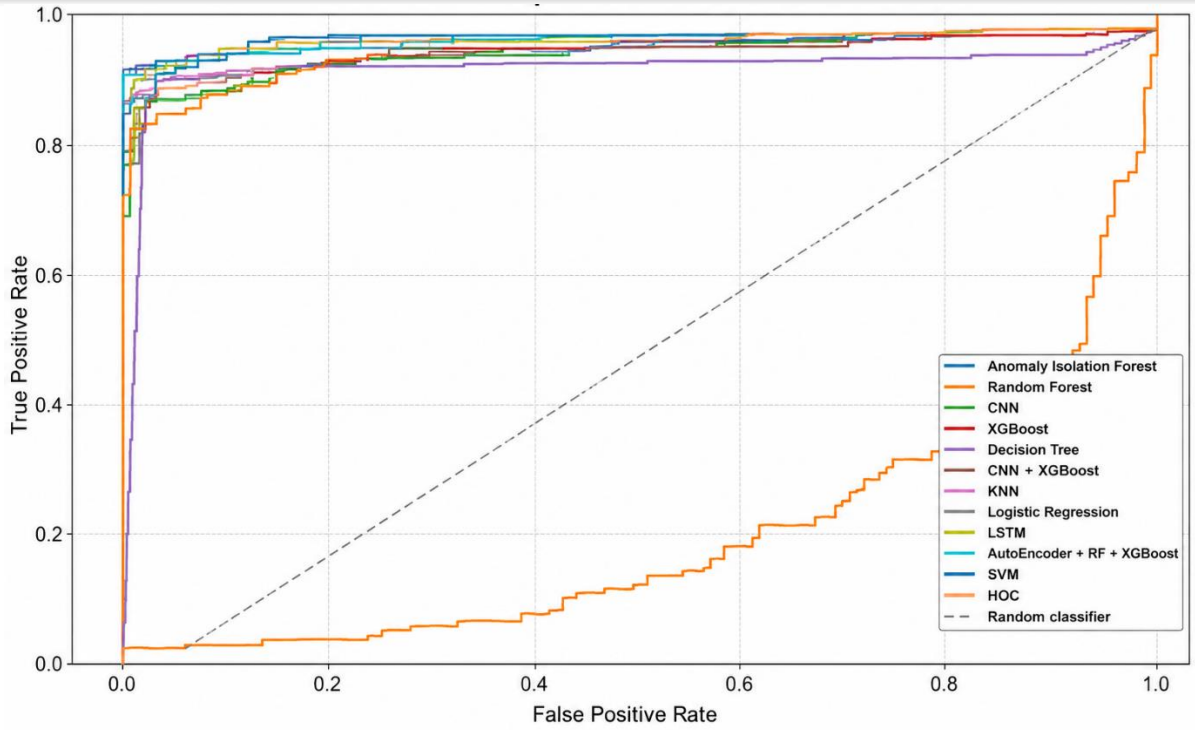


Figure 10: Comparative ROC curves of machine learning models for automated glioma classification.

Figure 10 illustrates the Comparative ROC curves of machine learning models for automated glioma classification. The training cohort includes 882 HGG and 588 LGG patients under exact stratification, while each evaluation subset includes 189 HGG and 126 LGG patients. Maintaining consistent grade proportions can reduce class-distribution differences between subsets and support more interpretable comparisons of predictive performance. The use of patient-level partitioning also prevents individual patients' MRI modalities from appearing in both training and testing datasets, reducing one important source of information leakage. Nevertheless, multicenter generalizability requires additional validation using institutionally independent patient groups, particularly when acquisition protocols or scanner characteristics differ. The reported findings provide an encouraging indication of the potential effectiveness of the proposed framework for multimodal neuroimaging analysis. The classification accuracy of 97.42% and ROC–AUC of 0.989 suggest strong grade discrimination, while the WT, TC, and ET Dice scores indicate substantial agreement between predicted and reference tumor regions. The combination of hierarchical spatial learning, multi-scale fusion, and contextual attention offers a technically coherent explanation for how a unified architecture may represent complex tumor morphology. However, conclusions concerning architectural superiority require direct comparative evidence, including baseline results, ablation measurements, uncertainty estimates, and testing on independent clinical datasets. Several limitations should be considered when interpreting these results. First, MRI acquisition variability can affect feature distributions and segmentation boundaries, particularly when imaging examinations originate from different scanners or institutions. Second, the quality and consistency of reference annotations influence measured segmentation accuracy. Third, classification performance may depend on the diagnostic criteria used to define the HGG and LGG categories. Fourth, model complexity and computational demands may limit implementation in resource-constrained clinical environments. In addition, aggregate performance

measures may conceal clinically relevant errors involving small enhancing regions, unusual tumor morphologies, or underrepresented patient subgroups [40]. The reported classification accuracy also requires careful numerical reconciliation with the final evaluation records. If accuracy was calculated exclusively from 315 independent patient-level test cases using a single binary decision per patient, the reported value of 97.42% does not correspond exactly to an integer number of correct classifications. This may reflect averaging across folds, a different evaluation denominator, or another aggregation procedure. The final manuscript should identify the actual evaluation unit and provide a patient-level confusion matrix so that accuracy, precision, recall, and F1-score can be independently reproduced. Similarly, the segmentation results should specify whether the reported Dice values represent patient-level means, volumetric aggregates, or another predefined summary. From a clinical perspective, the proposed framework may contribute to automated neuroimaging interpretation by providing complementary information concerning tumor localization and imaging-based aggressiveness. Accurate tumor segmentation can support quantitative assessment of lesion extent, surgical planning, radiotherapy preparation, and longitudinal imaging evaluation. Imaging-based glioma-grade classification can provide additional predictive information for clinical review. Nevertheless, definitive glioma classification requires integration of appropriate histopathological and molecular findings, and automated MRI predictions should be regarded as supportive rather than standalone diagnostic conclusions. Overall, the reported experimental results indicate that the proposed multi-scale feature fusion and Transformer-based computer vision framework achieves strong performance across the two targeted neuroimaging tasks. The model demonstrates high reported grading discrimination and Dice coefficients above 90% for all three tumor segmentation regions. These findings support further investigation of coordinated CNN-based spatial feature extraction, hierarchical feature integration, and attention-driven contextual modeling for heterogeneous glioma assessment. However, the strength of claims regarding generalizability, clinical suitability, and superiority over existing methods will depend on the availability of complete experimental records, reproducible comparative results, appropriate statistical evaluation, and independent clinical validation.

Future Work:

Future research will focus on improving the robustness, computational efficiency, generalizability, and clinical applicability of the proposed multi-scale feature fusion and Transformer-based computer vision framework for automated glioma grading and brain tumor segmentation. Although the reported findings demonstrate promising classification and segmentation performance, further investigations are necessary to evaluate the framework across larger and more heterogeneous patient populations. One important direction involves expanding the multicenter MRI dataset by incorporating additional clinical institutions, imaging protocols, scanner manufacturers, and patient demographics. Such expansion would facilitate a more comprehensive assessment of model generalizability and reduce potential sensitivity to institution-specific imaging characteristics. Independent external validation will also be essential for determining whether the proposed architecture maintains reliable performance when applied to previously unseen clinical populations. Another important research direction involves integrating advanced MRI sequences and clinically relevant molecular information into the existing multimodal framework. While the present investigation incorporates T1-weighted, T1ce, T2-weighted, and FLAIR imaging, future studies could evaluate diffusion-weighted imaging (DWI), apparent diffusion coefficient (ADC) maps, diffusion tensor imaging (DTI), perfusion-weighted imaging, and magnetic resonance spectroscopy. These modalities may provide complementary information regarding tumor cellularity, vascularity, microstructural organization, and metabolic characteristics [41]. Furthermore, the integration of molecular biomarkers, including

isocitrate dehydrogenase (IDH) mutation status, 1p/19q codeletion, and other clinically relevant genetic alterations, could support the development of more comprehensive imaging-based predictive systems aligned with contemporary glioma classification standards.

Future investigations will also explore advanced attention mechanisms and computationally efficient Transformer architectures to improve contextual representation learning while reducing memory requirements and inference complexity. Hierarchical Vision Transformers, window-based self-attention, deformable attention, and adaptive token selection may offer opportunities to capture long-range anatomical dependencies more efficiently. Similarly, incorporating dynamic multi-scale feature fusion mechanisms could enable the model to adjust the relative contribution of shallow, intermediate, and deep representations according to tumor morphology and imaging characteristics. These developments may be particularly beneficial for improving the localization of small enhancing tumor regions and the characterization of irregular infiltrative boundaries. The optimization of unified multi-task learning represents another promising area for future development. Adaptive loss weighting, task-specific attention, gradient balancing, and uncertainty-aware optimization could be investigated to improve coordination between segmentation and glioma-grade classification. Future studies could additionally incorporate tumor-volume estimation, progression prediction, recurrence assessment, and treatment-response monitoring within an extended multi-task framework. Such developments would enable automated neuroimaging systems to provide a broader range of clinically relevant information beyond static tumor segmentation and binary grading [42]. Longitudinal MRI analysis could further support the identification of temporal changes in tumor morphology and enhance the assessment of disease progression. Explainable artificial intelligence and uncertainty quantification will also receive increased attention in subsequent research. Although Transformer attention mechanisms provide contextual feature representations, their outputs do not necessarily provide clinically reliable explanations of diagnostic predictions. Future work could incorporate gradient-based attribution, feature visualization, prediction calibration, and uncertainty estimation to improve model transparency and identify cases requiring additional clinical review. Particular emphasis should be placed on evaluating false-positive and false-negative classifications, segmentation boundary errors, and performance across underrepresented patient groups. These investigations would support a more comprehensive understanding of model limitations and facilitate safer integration into clinical decision-support workflows. Finally, future studies will investigate the feasibility of deploying the proposed framework within real-world radiological environments through computational optimization, workflow integration, and prospective clinical evaluation. Model compression, knowledge distillation, quantization, and efficient inference strategies could reduce computational demands while preserving diagnostic performance. Privacy-preserving approaches, including federated learning, may also enable collaboration across institutions without requiring centralized sharing of sensitive patient imaging data. Prospective validation involving radiologists, neurosurgeons, and neuro-oncology specialists would be necessary to assess clinical usability, diagnostic consistency, and the practical value of automated predictions. Collectively, these directions aim to transform the proposed CNN–Transformer architecture into a more generalizable, interpretable, computationally efficient, and clinically validated framework for intelligent glioma assessment and personalized neuro-oncological decision support.

Conclusion:

This study presented an advanced multi-scale feature fusion and Transformer-based computer vision framework for automated glioma grading and brain tumor segmentation using multimodal magnetic resonance imaging (MRI). The proposed architecture integrates hierarchical convolutional neural network (CNN) feature extraction, multi-scale spatial representation fusion, Vision

Transformer-based contextual learning, and unified multi-task optimization to address challenges associated with heterogeneous tumor morphology, irregular lesion boundaries, overlapping tissue characteristics, and complex anatomical relationships. By combining local spatial information with long-range contextual representations, the framework provides an integrated computational approach for simultaneous tumor-region segmentation and imaging-based glioma-grade classification. The investigation utilized a dataset comprising 8,400 MRI modality examinations from 2,100 patients, including 1,260 high-grade glioma (HGG) and 840 low-grade glioma (LGG) cases. Four complementary MRI modalities T1-weighted, contrast-enhanced T1-weighted (T1ce), T2-weighted, and fluid-attenuated inversion recovery (FLAIR) were incorporated to support comprehensive tumor characterization. Patient-level partitioning into training (70%), validation (15%), and testing (15%) subsets was adopted to reduce information leakage and support reliable performance assessment. The proposed framework combines hierarchical CNN representations for localized anatomical feature extraction, multi-scale fusion for preserving fine-grained tumor characteristics, and Transformer-based learning for capturing broader spatial dependencies. The reported experimental findings demonstrated strong predictive performance, achieving 97.42% classification accuracy, 97.16% precision, 96.88% recall, 97.02% F1-score, and a ROC–AUC of 0.989 for glioma-grade classification. Furthermore, the segmentation framework achieved Dice similarity coefficients of 94.86% for whole tumor, 92.73% for tumor core, and 90.41% for enhancing tumor regions. These findings indicate substantial agreement between predicted and reference tumor masks and strong discrimination between the two grading categories within the reported evaluation setting. The results also demonstrate the potential value of integrating spatially detailed convolutional representations with attention-based contextual learning for heterogeneous neuroimaging analysis. However, definitive claims of superiority over established architectures require verified comparative experiments and quantitative ablation results.

References:

Sankari, C., Jamuna, V., & Kavitha, A. R. (2025). Hierarchical multi-scale vision transformer model for accurate detection and classification of brain tumors in MRI-based medical imaging. *Scientific Reports*, *15*(1), 38275.

Zhang, L., Tang, X., Zhou, J., Liu, J., Gou, F., & Wu, J. (2026). FFBFormer: Integrating multi-scale transformer networks for glioma diagnosis in decision-making system. *Biomedical Signal Processing and Control*, *113*, 108969.

Divya, S., & Sofia, A. S. (2025, August). Vision Transformer-based Glioma Classification using Multi-Modal MRI and Wavelet Fusion. In *2025 5th International Conference on Soft Computing for Security Applications (ICSCSA)* (pp. 1860-1867). IEEE.

Ahmad, B., Jillani, S. A., Rafique, M., & Soomro, A. A. (2026, January). Design and Monitoring of a Tree-Inspired Vertical Axis Wind Turbine for Efficient Urban Wind Utilization. In 2026 1st International Conference on Innovations in Information and Communication Technologies (IICT) (pp. 1-6). IEEE.

Liu, W., Cao, Y., Liu, Y., Zhang, P., Gui, Z., & Liu, L. (2026). MTF-Net: a brain tumor segmentation network with multi-scale transformer and feature fusion enhancement. *Frontiers in Medicine*, *13*, 1934161.

Hashmi, A., Hussain, M., Abulfarj, A. W., Ameen, M., Alyamani, H. J., & Binzagr, F. (2026). RF-DETR with multi-scale feature fusion and explainable attention for accurate brain tumor detection in MRI. *Scientific Reports*.

Islam, N., Kumar Ray, S., Hossain, M. A., & Islam, S. M. S. (2026). RViT-FusionNet: A local cross-attention feature fusion-based hybrid framework for brain tumor classification. *Neural Computing and Applications*, *38*(12), 518.

Aamir, M., Rahman, Z., Choudhry, N., Bhutto, J. A., Abro, W. A., & Zhu, Z. (2025). From CNNs to transformers: a review of evolving deep learning architectures for brain tumor classification. *IEEE Access*.

Ghribi, F., & Hamdaoui, F. (2025). A Novel 3D U-Net–Vision Transformer Hybrid with Multi-Scale Fusion for Precision Multimodal Brain Tumor Segmentation in 3D MRI. *Electronics*, *14*(18), 3604.

Ahmad, B., Majeed, M. K., Soomro, A. A., & Jillani, S. A. (2026, January). Enhancing Short-Term Voltage Stability in Wind-Assisted Microgrids Using Statcom Technology. In 2026 1st International Conference on Innovations in Information and Communication Technologies (IICT) (pp. 1-6). IEEE.

Sankari, C., & Jamuna, V. (2026, August). HPViT-CAG: Multi-Scale Vision Transformer for Automated Brain Tumor Detection. In *2026 2nd International Conference on Smart Technologies for Sustainable Development Goals (ICSTSDG)* (pp. 1-6). IEEE.

Tamirisa, C., & Tabbassum, S. (2026). Brain Tumor Detection Using Vision Transformer with Multi-Scale Feature Fusion. *Journal of Animal Environment*, 537-551.

Ranjbarzadeh, R., Anari, S., Crane, M., & Bendechache, M. (2024, November). A hybrid UNet and vision transformer architecture with multi-scale fusion for brain tumor segmentation. In *International Conference on Medical Imaging and Computer-Aided Diagnosis* (pp. 484-492). Singapore: Springer Nature Singapore.

Azam, M. T., Balaha, H. M., Gondim, D. D., Mistry, A., Ghazal, M., & El-Baz, A. (2024, December). Histopathological diagnosis of meningioma and solitary fibrous tumors based on a multi-scale fusion approach utilizing vision transformer and texture analysis. In *International Conference on Pattern Recognition* (pp. 31-45). Cham: Springer Nature Switzerland.

Moiz, A., & Kumar, B. (2026, January). Supervised Deep Learning Framework for Multiclass Brain Tumor Detection Using MRI and Computer Vision Transformer. In *2026 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)* (pp. 1-6). IEEE.

Aslam, W., Hussain, J., Aslam, M. Z., Jan, S., Riaz, T. B., Iqbal, A., ... & Khan, I. (2025). Enhanced brain tumor segmentation in medical imaging using multi-modal multi-scale contextual aggregation and attention fusion. *Scientific Reports*, *15*(1), 37308.

Ahmad, B., Muneer, M., Jillani, S. A., & Ahmed, S. (2026, January). Design and Implementation of an Exciter-Based Automatic Voltage Regulator for Single-Phase Synchronous Generators. In 2026 1st International Conference on Innovations in Information and Communication Technologies (IICT) (pp. 1-6). IEEE.

Mohammadi, M., & Ziaratban, M. (2026). TCAFNet: Transformer-Guided Cross-Scale Attention and Deep Semantic Fusion for Brain Tumour Classification From MRI. *IET Image Processing*, *20*(1), e70303.

Singh, J. P., Gupta, P., & Jha, R. (2026, August). Advancing Precision Diagnosis: A Hybrid Vision Transformer-U-Net Framework for Multi-Parametric Brain Tumor Segmentation and Grading. In *2026 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC)* (pp. 1-6). IEEE.

Shiraskar, S. D. (2025). *Transforming the future of brain tumor detection through a hybrid transformer-based approach* (Master's thesis, The Catholic University of America).

Hamza, A., & Damaševičius, R. (2026). Deep learning for brain tumor segmentation and classification: a systematic review of methods and trends. *Computers, materials and continua.*, *86*(1), 1-41.

Pant, K., Pramanik, P. K. D., Ganie, S. M., Saha, A., & Zhao, Z. (2026). Global–local feature fusion: a robust hybrid deep learning model for multiclass brain tumor classification with Grad-CAM++ interpretation. *Frontiers in Genetics*, *17*, 1814786.

Feroz, S., Ahmad, B., Talani, R. A., & Rehman, M. A. U. (2026, March). Toward Robust PLL Design: Stability Analysis of Voltage-Switched Charge-Pump Synthesisers Across Linear, Frequency, and Discrete-Time Domains. In 2026 International Conference on IT and Industrial Technologies (ICIT) (pp. 1-7). IEEE.

Aamir, M., Rahman, Z., Bhatti, U. A., Bhutto, J. A., Choudhry, N., & Abro, W. A. (2025, October). Brain Tumor Classification Based on Machine Learning and Deep Learning Technologies: A Comprehensive Review. In *2025 IEEE 3rd International Conference on Computer, Vision and Intelligent Technology (ICCVIT)* (pp. 1-12). IEEE.

Srinidhi, S., & Keerthika, P. (2025, December). Accurate Braintumor Detection with NeuroMS-ViT: A Transformer Based Multi-Scale Architecture. In *2025 13th International Conference on Intelligent Embedded, MicroElectronics, Communication and Optical Networks (IEMECON)* (pp. 1-7). IEEE.

Wu, H., Zhao, L., Zhang, Y., Zhang, C., Zhao, G., Li, W., ... & Wang, W. (2026). Interpretability and performance of a 3D C-vit model for accurate grading of pediatric brain tumors. *Frontiers in Oncology*, *16*, 1763280.

Wyson, J., Hazaimeh, Y., Shifani, S. A., Sravani, K. G., & Pushkarna, G. (2026, April). Explainable Multi-Scale Vision Transformer Network for Accurate Early Stage Tumor Detection. In *2026 International Conference on Computing Theory and Wireless Communications (ICCTWC)* (pp. 1-6). IEEE.

Patil, S., & Gavade, P. (2026). Improving brain tumor diagnosis: A systematic review of CT, MRI, and fusion-based deep learning models. *Imaging*, 1647-2026.

Linija, K. P., & Rajesh, S. (2026). Transformer-Integrated Multistage Tumor-Aware Framework for Brain Tumor Segmentation and Classification. *International Journal of Computational Intelligence Systems*.

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.

Pooja, P. P., & Aruna, S. K. (2026). Multimodal Survival Prediction in High-Grade Gliomas via Boundary-Aware Swin-TransUNet and Radiomic-Genomic Cross-Attention. *International Journal of Artificial Intelligence and Machine Learning*, *6*(10s), 550-564.

Hamoud, M., Chekima, N. E. I., Hima, A., & Kholladi, N. H. (2025). An automated cascade framework for glioma prognosis via segmentation, multi-feature fusion and classification techniques. *Biomedical Physics & Engineering Express*, *11*(3), 035027.

Alshammari, H. H., & Mahmood, M. A. (2026). An Explainable Multi-Scale Deep Learning Framework for Multi-Class Brain MRI Classification. *Diagnostics*, *16*(12), 1791.

Elazab, N., Khalifa, F., Gab Allah, W., & Elmogy, M. (2025). Histopathological-based brain tumor grading using 2D-3D multi-modal CNN-transformer combined with stacking classifiers. *Scientific Reports*, *15*(1), 27764.

Goyal, B., Hans, R., Sharma, S. K., & Singh, H. (2026). Deep Learning Based Brain Tumor Diagnosis with Pre-Trained and Self-Attention Based Models Using MRI Scans: A Systematic Literature Review. *Archives of Computational Methods in Engineering*, 1-49.

Zarenia, E., Far, A. A., & Rezaee, K. (2025). Automated multi-class MRI brain tumor classification and segmentation using deformable attention and saliency mapping. *Scientific Reports*, *15*(1), 8114.

Mounika, G., Kollem, S., & Samala, S. (2025, July). Enhanced feature fused vision transformer framework for multiclass brain tumor classification using MRI images. In *2025 International Conference on Innovations in Intelligent Systems: Advancements in Computing, Communication, and Cybersecurity (ISAC3)* (pp. 1-6). IEEE.

Bandi, R., Selvanayaki, S., Amitha, I. C., Subramaniam, S., Suganthi, R., & Rajendran, A. (2026). Adaptive Sparse Cross-Scale Transformer for Computationally Efficient Brain Tumor MRI Segmentation. *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, *8*(4), 1389-1405.

Kaur, S., Mittal, U., & Wadhawan, A. (2026). CSMFF-Net: Cross-Scale Modality-Aware Feature Fusion Network for Brain Tumor Segmentation. *Procedia Computer Science*, *283*, 103-112.

Khaniki, M. A. L., Mirzaeibonehkhater, M., Manthouri, M., & Hasani, E. (2025). Vision transformer with feature calibration and selective cross-attention for brain tumor classification. *Iran Journal of Computer Science*, *8*(2), 335-347.

Safarpour, H., Sadeghi, S., Zarbakhsh, P., Kamsari, M., Kia, M., & Ranjbarzadeh, R. (2026). Explainable deep learning framework for brain tumor segmentation using vision transformer and conditional random fields. *Multimedia Systems*, *32*(1), 19.

Huang, Y., Chen, L., & Zhou, C. (2025). Multi-modal brain tumor segmentation via 3D multi-scale self-attention and cross-attention. *arXiv preprint arXiv:2504.09088*.

IEER
Institute for Excellence in Education