

Article

Glioma Grade Classification from Structural MRI: A Comparative Transfer Learning Study of Deep Learning Feature Extraction and Machine Learning Classifiers

Amir Khorasani ^{1,2,*}, Ghasem Azemi ³ and Antonio Di Ieva ³¹ Medical Image and Signal Processing Research Center, School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Isfahan 81746-73461, Iran² Department of Bioimaging, School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Isfahan 81746-73461, Iran³ Computational NeuroSurgery (CNS) Lab, Macquarie Medical School, Faculty of Medicine, Health and Human Sciences, Macquarie University, Sydney, NSW 2109, Australia; ghasem.azemi@mq.edu.au (G.A.); antonio.diieva@mq.edu.au (A.D.I.)

* Correspondence: a.khorasani@resident.mui.ac.ir or amir69k@yahoo.com; Tel.: +98-31-37923865

Abstract

Background: Accurate histological grading of gliomas, distinguishing low-grade (LGG) from high-grade (HGG) lesions, remains a critical determinant of treatment planning and patient prognosis. This study systematically investigates the optimal combination of structural MRI sequences, pretrained convolutional neural network (CNN) architectures, and supervised classifiers for automated glioma grade classification. **Methods:** Deep features were extracted from four MRI sequences (T1, contrast-enhanced T1 (T1Gd), T2, and FLAIR) from the BraTS 2023 dataset using five pretrained CNNs: VGG16, ResNet50, DenseNet121, EfficientNetB0, and InceptionV3. Features were refined via LASSO selection and classified using Random Forest, Support Vector Machine, XGBoost, Gradient Boosting, k-Nearest Neighbors (KNNs), and a shallow deep neural network. A patient-level 80/20 training–test partition was employed, with five-fold cross-validation used within the training set for feature selection and hyperparameter optimization. A total of 120 modality–extractor–classifier configurations were benchmarked on the held-out test set. **Results:** InceptionV3-derived features paired with KNN classifiers consistently yielded superior performance. Based on a TOPSIS multi-criteria ranking integrating accuracy, precision, recall, F1-score, and AUC, the best-performing configuration combined T1Gd features with KNN (closeness coefficient = 0.971; accuracy = 0.975, precision = 0.997, recall = 0.948, F1-score = 0.969, AUC = 0.996). The second-ranked configuration, using T1 features with KNN (closeness coefficient = 0.962), achieved significantly higher accuracy, recall, and F1-score, despite its slightly lower composite ranking. **Conclusion:** Deep feature extraction using InceptionV3 from T1-weighted MRI, coupled with KNN classification, represents a promising and practical approach for slice-level glioma grading and merits further validation in prospective clinical cohorts.

Keywords: artificial intelligence; magnetic resonance imaging; glioma grade classification; deep learning; transfer learning

Academic Editor: Krzysztof J Cios

Received: 11 June 2026

Revised: 25 July 2026

Accepted: 27 July 2026

Published: 3 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Gliomas are the most common primary brain tumors, and the World Health Organization (WHO) classifies them into grades 1–4 based on histopathological and molecular features [1]. Grades 1–2 are generally considered low-grade gliomas (LGGs), while grades 3–4 are high-grade gliomas (HGGs), with grade 4 (glioblastoma) associated with the poorest prognosis [2]. Accurate grading plays an important role in determining appropriate treatment strategies and supporting clinical decision-making. Clinically, glioma grade classification largely relies on visual interpretation of multimodal magnetic resonance imaging (MRI) image weights (T1, T1 post-contrast enhancement (T1Gd); T2, fluid-attenuated inversion recovery (FLAIR)) in combination with pathology diagnosis, a process that is time-consuming and susceptible to sampling errors and procedural complications [3,4].

In recent years, there has been growing interest in applying artificial intelligence (AI) to medical sciences, particularly in medical image analysis [5–7]. Prior work in radiomics and deep learning (DL) has demonstrated that machine learning (ML) and deep learning (DL) models applied to MRI sequences can noninvasively predict glioma grade [8–10]. A survey study conducted by Gao et al. [11] reported that radiomic features extracted from the peritumoral edema region in T2-weighted MRI images play an important role in glioma prognosis. However, these approaches still face several limitations, including poor reproducibility across imaging protocols, reliance on large annotated datasets, vulnerability to small sample sizes, and limited clinical interpretability [12–14]. Transfer learning (TL) offers a promising solution by leveraging pretrained models to reduce data requirements, improve generalizability across institutions, and enhance computational efficiency while maintaining diagnostic accuracy [15]. Furthermore, many DL-based glioma grading methods [16,17] rely on accurate tumor segmentation, which is time-consuming, prone to variability, and dependent on expert annotation. Although segmentation can help focus the model on tumor regions, it introduces an additional step that may affect workflow efficiency. Glioma grade classification approaches using MRI image slices eliminate this dependency by allowing models to learn tumor features along with relevant contextual information directly from full MRI scans. Recent studies [18–21] have shown that such methods can achieve reliable performance, making them a practical alternative for glioma grade classification. Therefore, in this study, we adopt an approach based on different MRI image weights combined with TL to develop an efficient and robust framework for LGG/HGG classification.

TL is especially useful in medical imaging tasks where labeled data are limited, as it allows models to reuse robust feature representations learned from large-scale datasets, reducing overfitting and training complexity. Consequently, TL has become a widely explored approach for glioma grade classification. Recent work [19] has established a two-stage TL paradigm where pre-trained convolutional neural networks (CNNs) (e.g., ResNet) serve as image feature extractors, with the extracted features then used to train different ML classifiers (e.g., Random Forest (RF) and Support Vector Machine (SVM)) for glioma grading. Many attempts have been made [22] to use different pretrained networks, such as VGG, ResNet, Inception, and others, as image feature extractors in glioma grade classification. On the other hand, an increasing number of studies have examined the impact of different ML and DL classifiers on glioma grade classification. Surveys conducted by previous studies [23,24] have shown the significant impact of MRI sequence selection on glioma grade classification.

Motivated by the variability of glioma imaging biomarkers across structural MRI sequences and the complementary strengths of deep learning and machine learning, this study introduces a comprehensive and systematic evaluation framework for glioma grade classification that addresses several critical yet underexplored gaps in the existing literature. Despite the widespread adoption of multimodal MRI in glioma grading, the

independent discriminative contribution of individual MRI weightings to classification performance remains largely uninvestigated, as the overwhelming majority of existing studies employ multiple sequences simultaneously as a combined input rather than evaluating each weighting independently [25–28]. This fundamental limitation obscures the specific informational value of each MRI contrast for glioma grade differentiation and hinders the development of streamlined, resource-efficient clinical workflows. Furthermore, the comparative performance of different pretrained CNN architectures used as fixed feature extractors, combined with various machine learning classifiers, has not been systematically evaluated within a unified pipeline that uses the entire MRI image, rather than segmented tumor boundaries, for feature extraction and classification. Consequently, the optimal configuration for this clinically important task remains unclear.

To address these gaps, the present study makes the following novel contributions. First, we systematically and independently evaluate four individual structural MRI weightings (T1, T2, FLAIR, and T1Gd) to quantify the unique discriminative capacity of each sequence for glioma grade classification. Second, we comprehensively benchmark five pretrained CNN architectures (VGG, ResNet, Inception, DenseNet, and EfficientNet) as fixed feature extractors across six machine learning classifiers (RF, SVM, XGBoost, Gradient Boosting, KNN, and DNN), yielding a total of 120 distinct pipeline configurations evaluated under a unified framework. Third, a dedicated ablation study is conducted using TOPSIS-based closeness coefficient analysis to quantitatively and transparently assess the independent contribution of each pipeline component to overall classification performance. To the best of our knowledge, this represents the first study to systematically investigate and benchmark such a comprehensive range of MRI weighting, CNN architecture, and classifier combinations within a single glioma grading framework, providing quantitative and clinically meaningful insights into the optimal pipeline configuration for this task.

2. Materials and Methods

This study proposes a comprehensive deep learning-based framework for the classification of LGG and HGG using the Brain Tumor Segmentation 2023 (BraTS 2023) MRI dataset. To enhance data diversity and mitigate overfitting, image augmentation techniques were applied during the preprocessing stage. Pretrained CNNs, including VGG16, ResNet50, InceptionV3, DenseNet121, and EfficientNet-B0, were employed as feature extractors, leveraging transfer learning to capture high-level discriminative representations from the MRI images. The extracted deep features were subsequently subjected to feature selection using the Least Absolute Shrinkage and Selection Operator (LASSO) to reduce dimensionality and retain the most informative features. The selected features were then used to train and evaluate multiple classification models, including tuned RF, SVM, XGBoost, Gradient Boosting, KNN, and a simple DNN. An overall workflow diagram, Figure 1, was designed to visually illustrate the sequential stages of the proposed methodology, from data loading and feature extraction to feature selection and final classification.

2.1. Magnetic Resonance Imaging (MRI) Data Set

The present study utilized the BraTS 2023 dataset (<https://www.synapse.org/Synapse:syn64952532>, <https://www.kaggle.com/datasets/bkb2024/brats-2023-training>, accessed on 10 January 2024), a widely recognized benchmark for the evaluation of brain tumor segmentation and classification algorithms. Ground-truth glioma grade labels (LGG/HGG) were not distributed as part of the BraTS 2023 release itself. As the BraTS 2023 dataset incorporates cases shared across earlier BraTS challenges, including BraTS 2021, in which glioma grade labels were provided, grade labels for the corresponding

subjects in this study were obtained by matching subject IDs between the BraTS 2023 glioma (GLI) cohort and the previously released BraTS 2021 labeled cases. Only subjects for which a validated LGG/HGG label could be retrieved through this matching procedure were included in the present analysis. This multi-institutional dataset comprises pre-operative MRI scans acquired using four imaging modalities: T1, T1Gd, T2, and FLAIR. The dataset includes a heterogeneous cohort of patients diagnosed with LGGs and HGGs, along with expert-provided annotations delineating the enhancing tumor, peritumoral edema, and necrotic or non-enhancing tumor core regions.

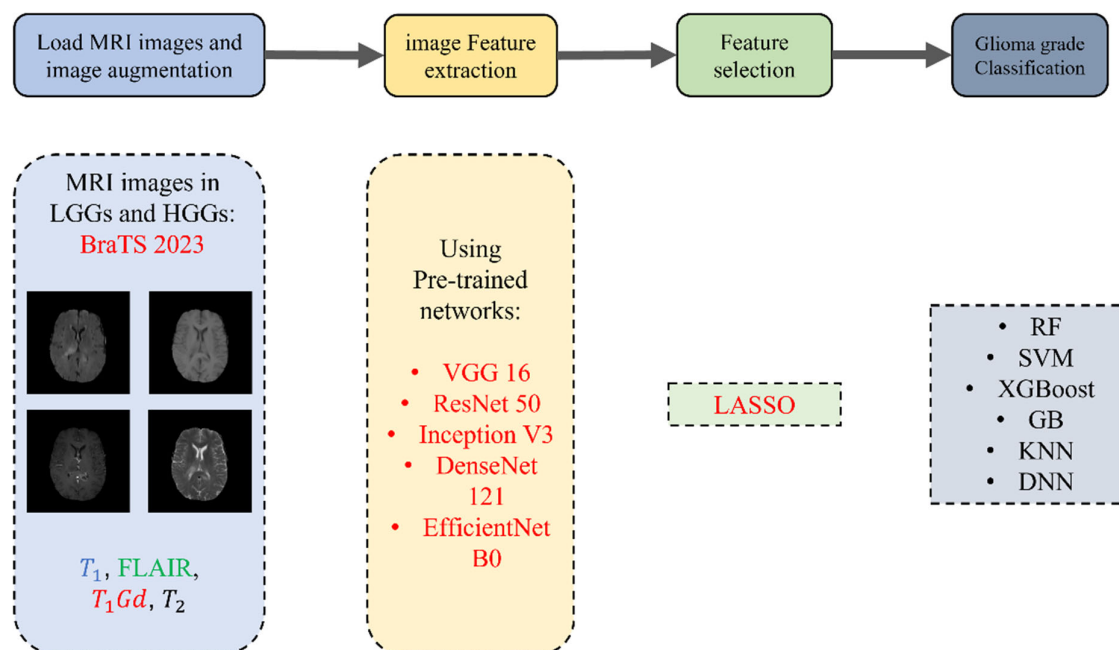


Figure 1. Overview of the proposed LGG/HGG classification framework. The workflow illustrates the sequential stages of the methodology, including input MRI images from the BraTS 2023 dataset, image augmentation, deep feature extraction using pretrained CNN models (VGG16, ResNet50, InceptionV3, DenseNet121, and EfficientNet-B0), feature selection via the Least Absolute Shrinkage and Selection Operator (LASSO), and final classification using tuned machine learning classifiers (Random Forest, Support Vector Machine, XGBoost, Gradient Boosting, k-Nearest Neighbors) and a deep neural network (DNN).

Prior to public release, the BraTS dataset underwent a standardized and rigorous preprocessing pipeline, including image co-registration, spatial resolution normalization, and skull stripping, to ensure data homogeneity and suitability for quantitative analysis. Following data acquisition, the MRI sequences (T1, T1Gd, T2, and FLAIR) were organized according to tumor grade, resulting in separate LGG and HGG groups for subsequent analysis.

Two cases from the BraTS 2023 dataset were excluded from the analysis due to the presence of imaging artifacts, such as Gibbs ringing and truncation effects, as identified by an experienced MRI expert. Consequently, a total of 400 glioma subjects were included in this study, consisting of 324 high-grade glioma (HGG) cases (81%) and 76 low-grade glioma (LGG) cases (19%). Following patient-level 80/20 splitting, the training set comprised 259 HGG patients and 60 LGG patients (319 patients total), while the held-out test set comprised 65 HGG patients and 16 LGG patients (81 patients total). Following data preparation and the application of image augmentation techniques, a total of 5145 two-dimensional images per MRI modality (T1, T2, FLAIR, and T1Gd) were used for training

across both glioma grades, with 41.59% corresponding to LGG and 58.41% to HGG cases. The held-out test set, augmented using the same class-specific strategy, comprised 1029 two-dimensional images per MRI modality (601 HGG and 428 LGG). This strategy ensured adequate representation of both tumor grades and all imaging sequences, supporting robust model training and evaluation.

2.2. Image Augmentation

In order to enhance the robustness and generalization capability of the proposed LGG/HGG classification framework, data augmentation was adopted. Spatial and intensity-based transformations were imposed on 2D MRI slices during training in order to increase data diversity and reduce overfitting. Spatial augmentations involved random rotations within $\pm 10^\circ$, while performing horizontal and vertical flipping with a probability of 0.5, in addition to random translations of up to 5 pixels in each direction. Furthermore, intensity-based augmentations—such as brightness and contrast variations within a $\pm 10\%$ range—were employed to account for the variability in MRI acquisition protocols. All augmentation operations were adopted in a controlled manner to maintain the underlying anatomical structures as well as tumor-related features. These data augmentation techniques were implemented using the ImageDataGenerator class from the Python library Keras (version 2.13.1).

2.3. Pre-Trained Networks

Five state-of-the-art pretrained networks, VGG, ResNet, Inception, DenseNet, and EfficientNet, were adopted as feature extractors in this study to evaluate their efficacy in capturing discriminative image features for glioma grade classification. Because these pretrained architectures require specific input dimensions, all MRI scans were resized to match the expected input size of each model. To enable task-independent deep feature extraction and to minimize overfitting, the fully connected layers (the last layers of the architecture) of all pretrained networks were excluded, while only the convolutional layers were preserved. The retained convolutional layers are responsible for learning hierarchical and spatially informative feature representations that are largely transferable across different tasks.

A pretrained VGG-16 model [29] with ImageNet was used as a feature extractor by removing the fully connected layers to generate $7 \times 7 \times 512$ deep feature maps for glioma grade classification. Details of the architecture are presented in Figure 2.

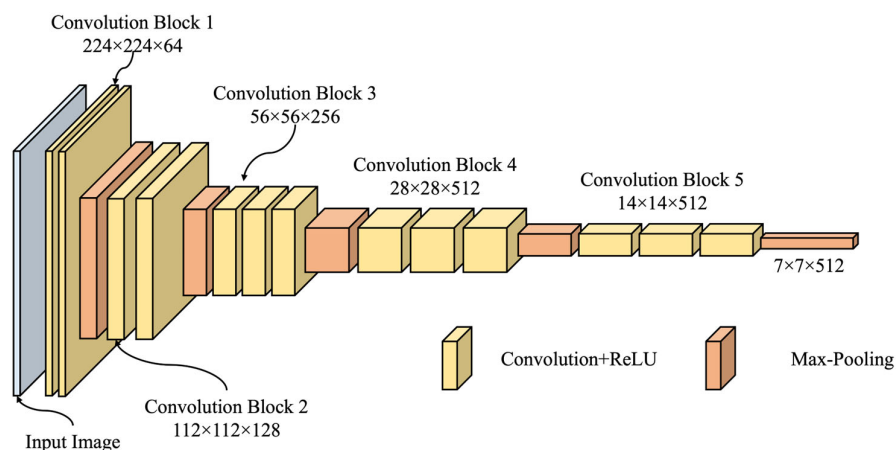


Figure 2. The VGG-16 convolutional neural network employed as a feature extractor. Features from the final convolutional layers were extracted and used as input for machine-learning classifiers.

ResNet-50 employs residual learning with skip connections and bottleneck blocks to enable deep feature extraction; in this study, a pretrained ImageNet-based ResNet-50 [30] was used as a feature extractor by removing the final fully connected layer, as illustrated in Figure 3.

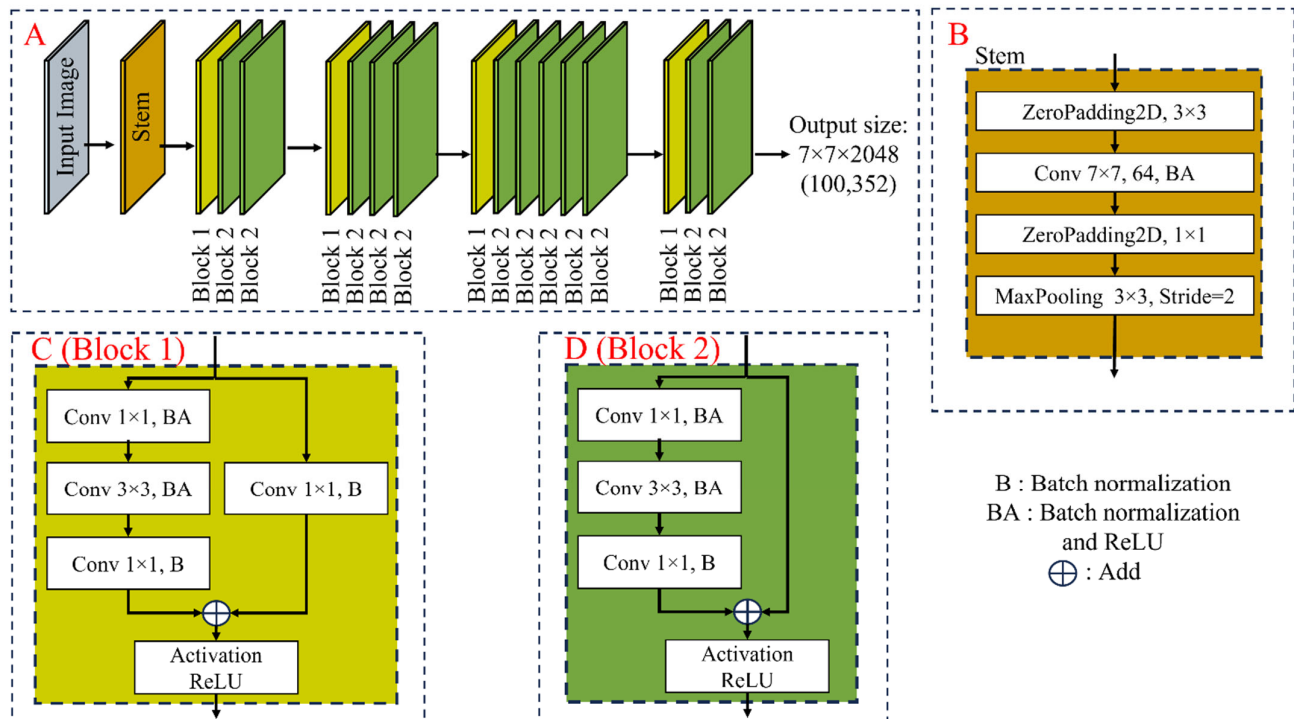


Figure 3. Architecture of the ResNet-50 model used for feature extraction in glioma grade classification. (A) Overview of the complete ResNet-50 architecture. (B) Detailed illustration of the stem block, consisting of the initial convolution and pooling layers. (C,D) Representative convolutional (residual) blocks employed in the ResNet-50 architecture, demonstrating the bottleneck structure and skip connections used for deep feature learning.

Inception-v3 [31] employs parallel multi-scale convolutions and factorized kernels to achieve computationally efficient deep feature extraction; in this study, a pretrained ImageNet-based Inception-v3 was used as a feature extractor by removing the final fully connected layer, as illustrated in Figure 4 (with detailed block structures provided in Supplementary Figures S1–S3).

DenseNet-121 [32] employs dense connectivity to enhance feature reuse and gradient flow across four dense blocks; in this study, a pretrained ImageNet-based DenseNet-121 was used as a feature extractor by removing the final fully connected layer for LGG/HGG classification, as shown in Figure 5.

EfficientNet-B0 [33] employs compound scaling with MBConv and squeeze-and-excitation blocks to achieve efficient deep feature extraction; in this study, a pretrained ImageNet-based EfficientNet-B0 was used as a feature extractor by removing the final fully connected layer for LGG/HGG classification, as illustrated in Figure 6.

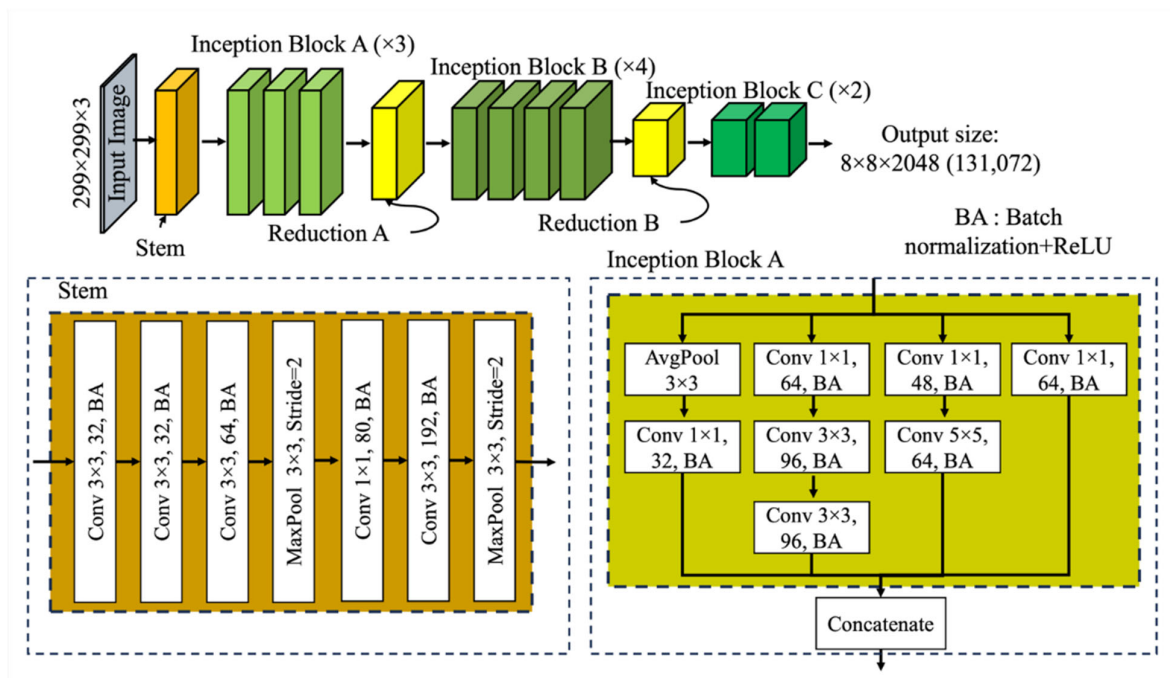


Figure 4. Inception-V3 architecture used for feature extraction in glioma grade classification. Overview of the complete Inception-V3 network architecture. Detailed representation of the stem block and Inception Block-A used within the network. The detailed structures of Inception Blocks-B and -C, as well as Reduction Blocks-A and -B, are provided in the Supplementary Material.

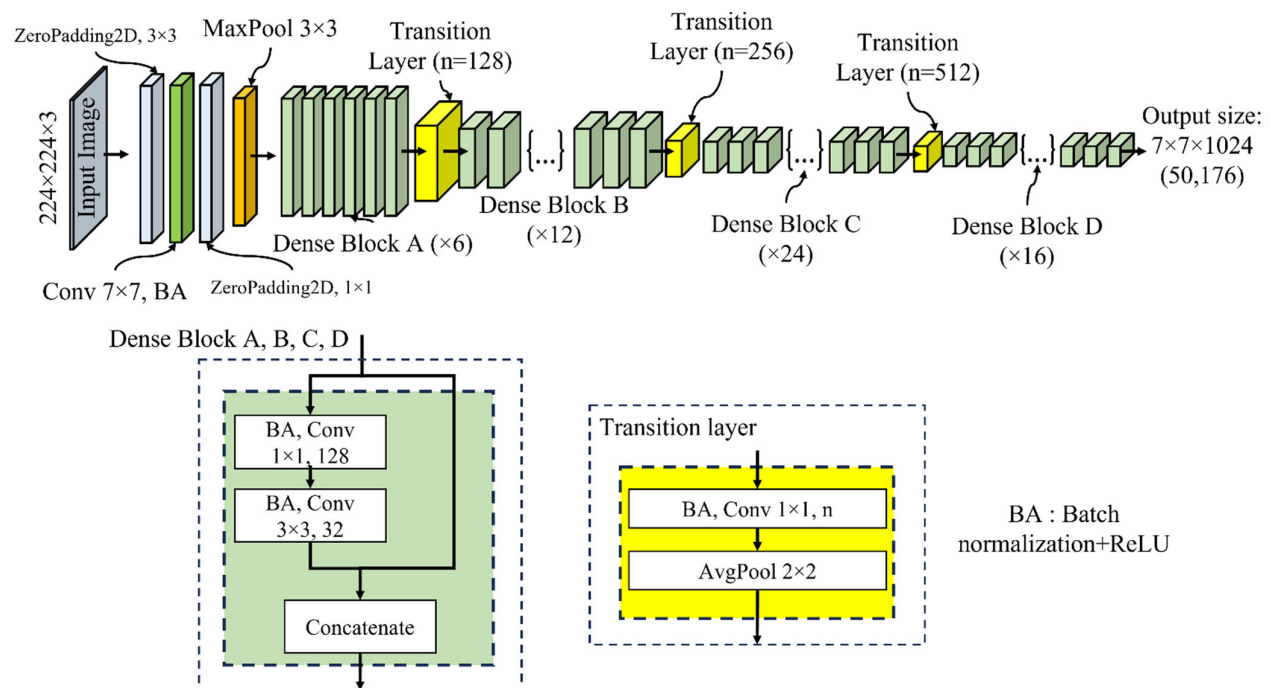


Figure 5. DenseNet-121 architecture used for feature extraction in glioma grade classification. Overview of the complete DenseNet-121 network architecture. Different Dense blocks illustrating the dense connectivity pattern among convolutional layers are presented. Transition layer showing the 1x1 convolution and pooling operations used to reduce feature map dimensions between dense blocks.

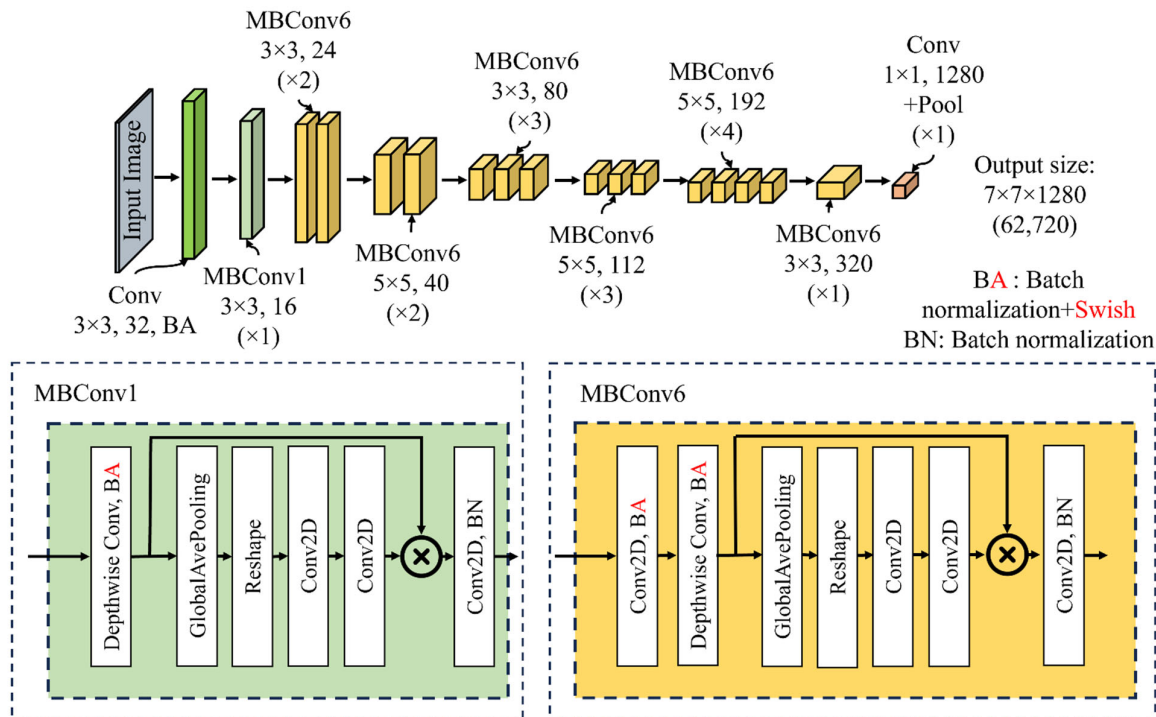


Figure 6. EfficientNet-B0 architecture used for feature extraction in glioma grade classification. The figure presents the overall network structure, including the convolutional layer and successive MBConv blocks with squeeze-and-excitation modules, illustrating how spatial resolution is progressively reduced while feature depth is increased for efficient and scalable feature learning.

2.4. Feature Selection

For feature selection, the Least Absolute Shrinkage and Selection Operator (LASSO) [34] was employed, as it is a well-established regularization technique capable of performing simultaneous coefficient shrinkage and feature selection. Prior to LASSO fitting, all extracted deep features were standardized using z-score normalization (StandardScaler), ensuring comparable feature scales and stable regularization behavior. Both the standardization and LASSO fitting procedures were performed exclusively on the training set; the fitted scaler and selected feature set were then applied to the held-out test set without refitting, preventing information leakage from the test data into the feature selection process.

By incorporating an L1 regularization penalty, LASSO promotes sparsity in the model coefficients, thereby enabling the identification of the most informative features while suppressing redundant and noisy variables. In this study, LASSO was applied with five-fold cross-validation ($CV = 5$) to determine the optimal regularization parameter, ensuring a robust and unbiased feature selection process. To ensure reproducibility and stable convergence, the model was configured with a fixed random seed (`random_state = 42`) and a maximum of 10,000 iterations (`max_iter = 10,000`). This feature selection strategy effectively reduced the dimensionality of the extracted deep feature sets while preserving the most discriminative features relevant for glioma LGG/HGG classification.

2.5. Classifiers

In this study, we evaluated multiple classifiers for glioma grade classification into LGG and HGG across different MRI weightings. These included conventional machine learning methods such as RF, SVM, XGBoost, GB, and KNN, alongside a custom DNN.

The DNN architecture comprised sequential dense layers (512, 256, 128, and 64 neuron units) with ReLU activation, interspersed with Dropout layers (0.3, 0.3, 0.2, 0.2) to mitigate overfitting. The final layer utilized a 2-neuron-unit dense layer with SoftMax activation to output probabilistic predictions for binary glioma grade classification. The DNN model was trained using the Adam optimizer (learning rate = 0.001) and the ‘categorical cross-entropy’ loss function over 500 epochs to optimize convergence while monitoring validation performance for early stopping. This comparative framework allowed us to assess the synergistic effects of deep feature extraction and diverse classification paradigms, balancing interpretability (traditional ML) and hierarchical learning (DNN) for optimized diagnostic performance.

In this study, we implemented a hyperparameter optimization strategy that combined Grid Search and Randomized Search cross-validation to improve the performance of the glioma grade classification models. Grid Search was applied to models with smaller and computationally manageable parameter spaces, including RF, SVM, KNN, and GB. For models with larger and more complex search spaces, such as XGBoost, we used Randomized Search with 50 iterations to efficiently sample from a broader set of possible configurations. All optimization procedures were performed using 5-fold stratified cross-validation with accuracy as the primary evaluation metric. The hyperparameter distributions for all classifiers were defined using Python’s SciPy (V 1.16.1), Stats (V 0.1.2a) and scikit-learn (V 1.7.1) libraries, as summarized in Table 1.

Table 1. Hyperparameter search and range settings for the glioma grade (LGG/HGG) classification models. The table summarizes the parameters and their corresponding ranges used for hyperparameter tuning via Grid Search and Randomized Search cross-validation across different classifiers, as implemented with 5-fold stratified cross-validation. Abbreviations: RF: Random Forest, SVM: Support Vector Machine, XGBoost: Extreme Gradient Boosting, GB: Gradient Boosting, and KNN: k-Nearest Neighbors.

Classifier	Search Method	Hyperparameters & Ranges
XGBoost	Randomized Search (50 iterations)	n_estimators [100, 200, 300]; max_depth [3, 5, 7]; learning_rate [0.01, 0.05, 0.1]; subsample [0.7, 0.9, 1]; colsample_bytree [0.7, 0.9, 1]
RF	Grid Search	n_estimators [100, 200, 300]; max_depth [none, 20, 40]; min_samples_split [2, 5]; min_samples_leaf [1, 2]; bootstrap [True, False]
SVM	Grid Search	C [0.01, 0.1, 1, 10, 100]; kernel (linear, RBF, polynomial); gamma [scale, auto, 0.001, 0.01]; degree [2, 3]
KNN	Grid Search	n_neighbors [3, 5, 7, 9, 11]; weights [uniform, distance]; metric [Euclidean, Manhattan]
GB	Grid Search	n_estimators [100, 200, 300]; learning_rate [0.01, 0.05, 0.1]; max_depth [3, 5, 7]; subsample [0.8, 1]

2.6. Grad-CAM

To visualize the image regions most influential to the model’s predictions, we employed Gradient-weighted Class Activation Mapping++ (Grad-CAM++) [35]. This method generates class-specific localization heatmaps by leveraging the gradients of the target class with respect to the final convolutional layer, thereby highlighting discriminative features that contribute most to the classification decision without altering the original model architecture.

2.7. Model Training and Evaluation

In this study, pretrained CNN architectures were utilized solely as fixed feature extractors, whereby their weights remained frozen throughout the process. No fine-tuning of the CNN models was performed. Instead, the extracted deep features served as input

to the machine learning classifiers, which were trained and optimized for LGG/HGG classification.

To prevent data leakage and ensure valid model evaluation, we implemented a strict patient-level data partitioning approach. The dataset was divided into training (80%) and testing (20%) sets at the patient level, ensuring that all 2D slices extracted from a given 3D volume remained exclusively within one subset. This approach maintained independence between training and testing data, as slices from the same patient are inherently correlated. Subsequently, 2D tumor region images were extracted from the 3D volumes based on segmentation masks. The BraTS segmentation masks were used only during preprocessing to identify and select slices containing tumor tissue. These masks were not provided to the model as input and were not used for segmentation during training or inference. The training set was used for model development with 5-fold cross-validation, maintaining patient-level assignments across folds. The test set was held out entirely during training and used only for final performance evaluation, ensuring unbiased assessment of the model's generalization capability for LGG/HGG classification. The BraTS 2023 dataset exhibits an inherent class imbalance between HGG and LGG cases, reflecting the true clinical prevalence disparity between these glioma grades. To address this imbalance while simultaneously increasing overall training sample size and improving model generalization, data augmentation was applied independently and separately within each class group, HGG and LGG, following strict patient-level splitting and tumor-containing slice selection. This class-specific augmentation strategy served two simultaneous and complementary purposes: (1) increasing the representational diversity of the minority LGG class to reduce the disparity between HGG and LGG sample counts within the training partition, thereby preventing systematic classifier bias toward the majority HGG class, and (2) increasing the overall number of training samples to improve classifier optimization and reduce the risk of overfitting. By applying augmentation independently within each class after patient-level splitting, the integrity of patient-level data separation and the balance between class representations were simultaneously preserved throughout the entire pipeline.

As patient-level partitioning was enforced prior to all subsequent processing steps, augmented slices within each dataset were derived exclusively from patients belonging to that dataset, preventing any form of data leakage between training and test sets. In this study, the performance of the proposed models was rigorously assessed using a comprehensive set of evaluation metrics for assessment of glioma grade classification performance. Standard quantitative measures, including accuracy, precision, recall, and F1-score, were employed to provide a balanced evaluation of classification performance. HGG was designated the positive class (LGG = 0, HGG = 1) for all binary classification metrics, consistent with its higher clinical priority for early and accurate detection. In addition, the area under the receiver operating characteristic curve (AUC-ROC) was used to quantify the models' discriminative capability across varying decision thresholds.

To systematically and objectively compare the 120 evaluated configurations after final testing, a post hoc multi-criteria decision-making approach based on the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) was employed [36]. This analysis was applied strictly for descriptive ranking of already-obtained test-set results and played no role in model training, tuning, or selection. TOPSIS ranks alternatives by simultaneously measuring their relative closeness to a positive ideal solution, representing the best attainable value across all criteria, and their distance from a negative ideal solution, representing the worst attainable value, providing a single composite closeness coefficient per alternative that enables transparent and reproducible multi-criteria ranking. All performance metrics, accuracy, precision, recall, F1-score, and AUC, were first normalized using vector (L2) normalization to ensure scale independence. The positive ideal

solution (A^+) and negative ideal solution (A^-) were determined as the maximum and minimum weighted normalized values for each criterion, respectively. Euclidean separation distances (S^+ and S^-) were computed for each combination, and a Closeness Coefficient ($CC_i = S^-/(S^+ + S^-)$) was derived, yielding values in the range $[0, 1]$ where 1 indicates perfect similarity to the ideal solution. All 120 combinations were ranked by CC_i in descending order.

It should be clarified that five-fold cross-validation was employed exclusively for LASSO feature selection and classifier hyperparameter optimization within the training partition and was not used to generate the final reported performance outcomes. All accuracy, precision, recall, F1-score, and AUC values, together with the corresponding TOPSIS rankings, were derived solely from the independent, patient-level held-out test set.

To systematically evaluate the independent contribution of each pipeline component to glioma grade classification performance, a comprehensive ablation study was conducted across all 120 configurations. The ablation analysis was performed by isolating each component (MRI image weighting, pretrained CNN architecture, and ML classifier), while averaging the TOPSIS closeness coefficient (CC) across all remaining components. This approach enables a fair and systematic comparison of each component's individual contribution to the overall classification performance, independent of the other pipeline elements. The average CC derived from the TOPSIS analysis was adopted as the unified ranking criterion, as it simultaneously incorporates all five performance metrics (accuracy, precision, recall, F1-score, and AUC) into a single composite score, providing a more balanced and comprehensive evaluation than any single metric alone.

To statistically evaluate whether the top TOPSIS-ranked configuration was meaningfully distinguishable from the remaining top-ranked alternatives, a comprehensive statistical comparison was conducted across the top 10 configurations. AUC values were compared using DeLong's test, applying the paired form for configurations sharing an identical test set and an independent-sample AUC comparison for configurations evaluated on different MRI sequences, together with 95% confidence intervals derived from stratified bootstrap resampling (2000 iterations). Classification error patterns at the standard 0.5 decision threshold were compared using McNemar's test. To additionally evaluate the remaining TOPSIS criteria, paired bootstrap resampling (2000 iterations) was used to compare accuracy, precision, recall, and F1-score between all pairs of the top 10 configurations, with 95% confidence intervals and two-sided p -values derived from the resampled difference distributions.

3. Results

Representative MRI examples from LGG and HGG patients across the four modalities are shown in Figure 7.

In this study, after image augmentation, we performed a comprehensive evaluation of glioma grade classification using five pretrained CNN architectures (Inception, VGG, ResNet, DenseNet, and EfficientNet) as feature extractors, combined with six machine learning classifiers (KNN, SVM, DNN, XGBoost, Gradient Boosting, and Random Forest) as well as a DNN classifier. In total, 30 unique combinations of feature extractors and classifiers (e.g., VGG + XGBoost, DenseNet + XGBoost, etc.) were systematically assessed across four MRI sequences. Because each pretrained CNN has its own architectural design, the dimensionality of the extracted deep features varied across models. The number of extracted features from each pretrained CNN is summarized in Table 2. These differences reflect the distinctive depth, width, and feature map structure inherent to each network. After extracting deep features from each pretrained model, we applied LASSO for feature selection, and the number of selected features in each MRI image weights is summarized in Table 2. An important observation from this table is that LASSO selected the smallest

subset of features from EfficientNet-B0 deep representations, suggesting that its features are more informative and less redundant compared to other pre-trained networks.

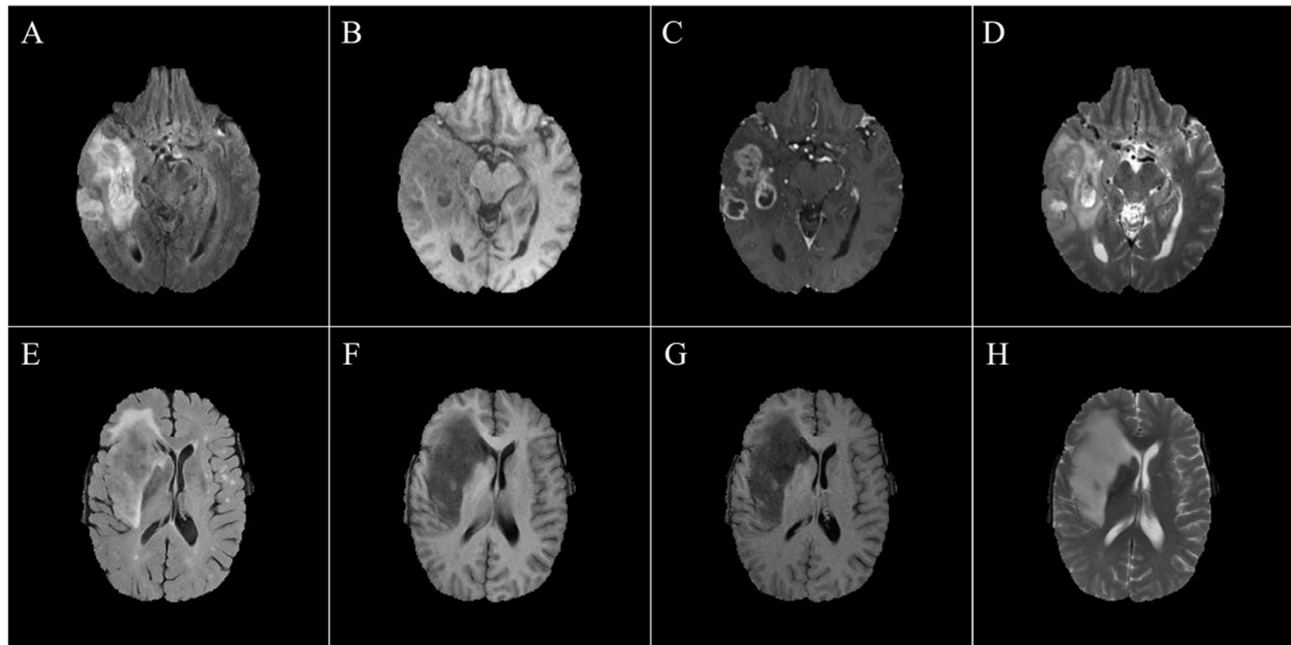


Figure 7. Representative MRI sequences of high-grade glioma (HGG, top row) and low-grade glioma (LGG, bottom row) from the BraTS dataset. (A) FLAIR, (B) T1-weighted, (C) T1-weighted post-contrast (T1Gd), and (D) T2-weighted images of an HGG case. (E–H) Corresponding sequences for an LGG case.

Table 2. Number of deep features extracted by each pretrained CNN and the corresponding number of features selected using LASSO for T1, T1Gd, T2, and FLAIR MRI images for LGG/HGG classification.

Pre-Trained Networks	Number of Extracted Features	MRI Sequences	Number of LASSO Selected Features
DenseNet 121	50,176	T1	482
		T1Gd	480
		FLAIR	489
		T2	568
EfficientNet B0	62,720	T1	132
		T1Gd	56
		FLAIR	29
		T2	50
Inception V3	131,072	T1	860
		T1Gd	587
		FLAIR	762
		T2	735
ResNet 50	100,352	T1	311
		T1Gd	157
		FLAIR	102
		T2	188
VGG 16	25,088	T1	518
		T1Gd	587

	FLAIR	680
	T2	433

The performance of all 120 evaluated combinations, including multiple pre-trained CNN deep feature extractors, machine learning classifiers, and MRI sequence inputs, is summarized as heatmaps in Figures 8–10, presenting Accuracy, Precision, Recall, F1-score, and AUC values across all combinations on the held-out test dataset. As a post hoc comparative step, TOPSIS-based multi-criteria ranking was applied to the held-out test-set results using all five performance metrics, Accuracy, Precision, Recall, F1-score, and AUC, as simultaneous decision criteria, with the complete ranked ordering of all 120 combinations presented in Supplementary Table S1. Clear performance variations can be observed among different feature extractor–classifier combinations. Notably, two low-ranking configurations, ResNet50+DNN and EfficientNetB0+SVM, both evaluated on T1Gd sequences, exhibited an identical precision of 1.00, recall of 0.50, and F1-score of 0.67 for the HGG class. Rather than reflecting majority-class prediction toward HGG, which would instead manifest as high recall and reduced precision given the HGG-majority test set (601 HGG vs. 428 LGG slices), this pattern indicates that both models adopted a conservative decision boundary, correctly avoiding false HGG predictions but failing to detect approximately half of true HGG cases. Consistent with their limited discriminative performance, both configurations ranked poorly in the overall TOPSIS analysis (rank 101 and 102 of 120, respectively; Supplementary Table S1), indicating that this behavior is confined to lower-performing pipeline combinations and does not affect the study’s primary findings regarding the top-performing InceptionV3-KNN configurations.

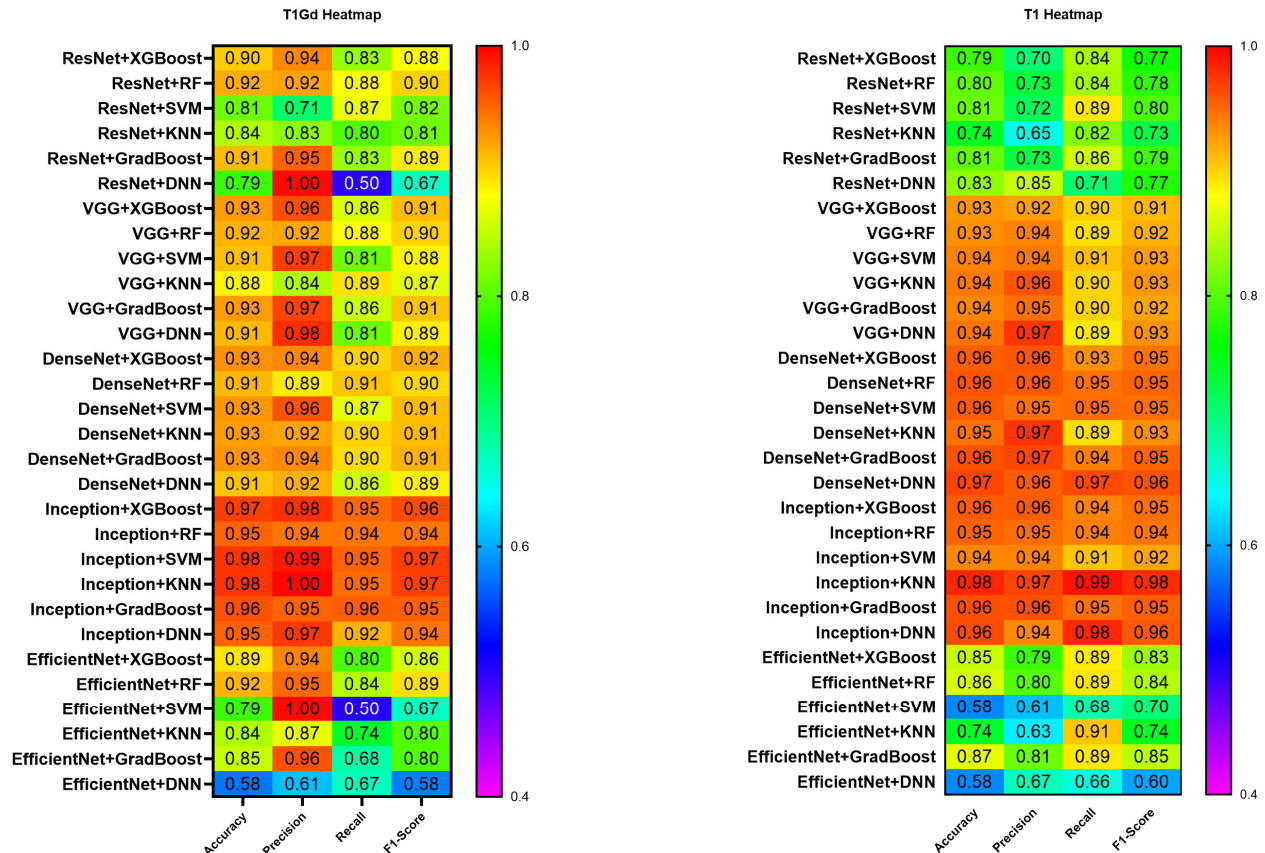


Figure 8. Heatmap of glioma classification performance metrics (Accuracy, Precision, Recall, F1-score) across 30 feature extractor–classifier combinations with T1 and T1Gd MRI image weights.

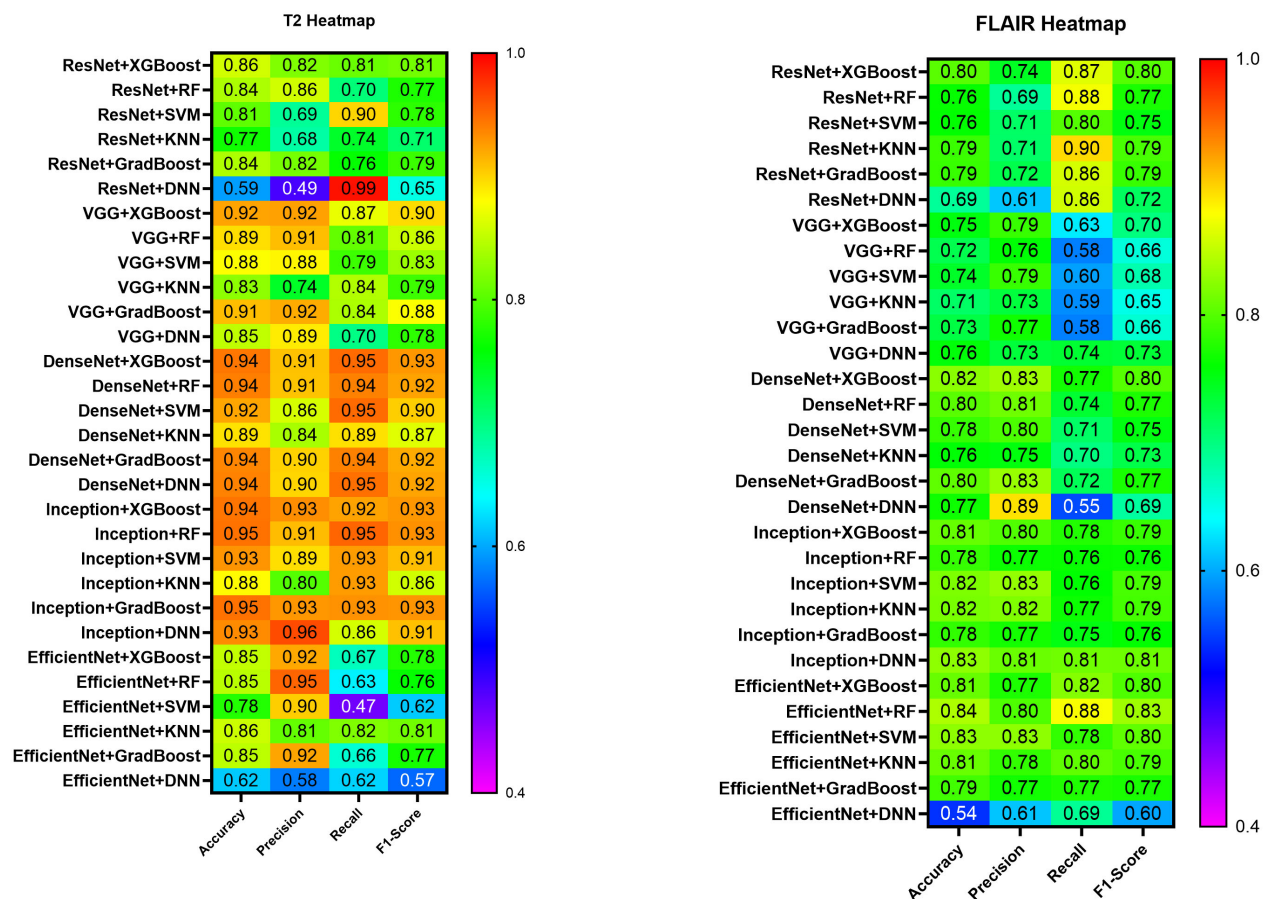


Figure 9. Heatmap of glioma classification performance metrics (Accuracy, Precision, Recall, F1-score) across 30 feature extractor–classifier combinations with T2 and FLAIR MRI image weights.

As illustrated in Figures 8 and 9 and the TOPSIS multi-criteria ranking (Table 3), the two top-ranked models among all 120 evaluated combinations were based on Inception-derived deep features coupled with KNN classifiers. According to the TOPSIS closeness coefficient, which integrates accuracy, precision, recall, F1-score, and AUC into a single composite ranking criterion, the T1Gd–InceptionV3–KNN configuration was identified as the overall best-performing model (CC = 0.971), followed closely by the T1–InceptionV3–KNN configuration (CC = 0.962). It should be noted that when individual metrics are considered separately, the T1-based configuration achieved marginally higher accuracy, recall, F1-score, and AUC than the T1Gd-based configuration; however, the T1Gd configuration attained the highest precision and the best overall balance across all five metrics simultaneously, resulting in its top TOPSIS ranking. The top-performing model utilized Inception features extracted from T1Gd images and a KNN classifier (configured with: metric = Manhattan, n_neighbors = 11, weights = uniform). This model achieved an accuracy of 0.9754, a precision of 0.997, a recall of 0.9489, and an F1-score of 0.9694. The second top-performing model employed Inception features extracted from T1 images in combination with a KNN classifier (metric = Manhattan, n_neighbors = 9, weights = uniform). This configuration yielded the performance with an accuracy of 0.9825, a precision of 0.968, a recall of 0.9907, and an F1-score of 0.9792.

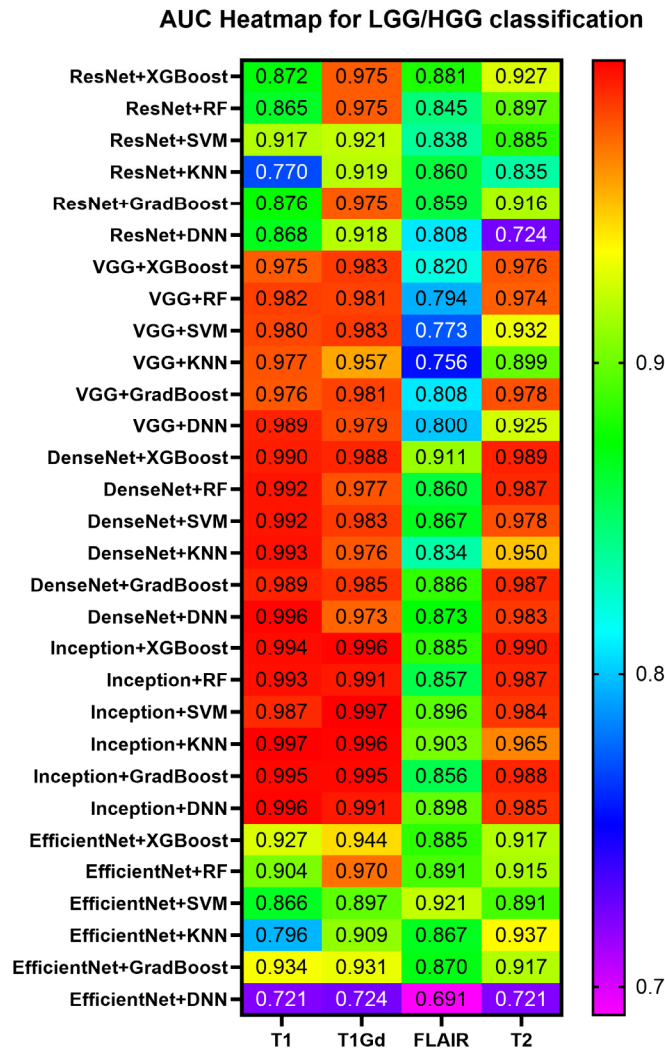


Figure 10. AUC heatmaps of glioma classification across 30 feature extractor–classifier combinations with T1, T1Gd, T2, and FLAIR MRI image weights.

Table 3. TOPSIS-based ranking of top 10 combinations of pre-trained CNN feature extractors, machine learning classifiers, and MRI modality inputs for glioma grade classification. ^a AUC and 95% confidence intervals estimated via 2000-iteration stratified bootstrap resampling on the respective held-out test set for each configuration. ^b Two-sided DeLong’s test comparing each configuration’s AUC against the Rank 1 configuration. ^c McNemar’s test on binarized predictions (threshold = 0.5). * p -value < 0.05.

Rank	CNN	Classifier	MRI Weight	Closeness_Co-efficient	AUC (95% CI) ^a	DeLong’s p -Value (vs. Rank 1) ^b	McNemar’s p -Value (vs. Rank 1) ^c
1	Inception	KNN	T1Gd	0.971	0.996 (0.993–0.999)	-	-
2	Inception	KNN	T1	0.962	0.997 (0.994–0.999)	0.021 *	0.0003 *
3	Inception	SVM	T1Gd	0.951	0.997 (0.995–0.999)	0.003 *	0.0023 *
4	Inception	XGBoost	T1Gd	0.949	0.996 (0.995–0.997)	0.004 *	0.0288 *
5	DenseNet	DNN	T1	0.945	0.996 (0.993–0.997)	0.177	0.2955
6	Inception	GradBoost	T1Gd	0.934	0.995 (0.993–0.997)	0.022 *	0.2531
7	Inception	GradBoost	T1	0.933	0.995 (0.993–0.997)	0.074	0.4338
8	Inception	DNN	T1	0.932	0.996 (0.994–0.998)	0.038 *	0.7463

9	DenseNet	RF	T1	0.929	0.992 (0.988–0.995)	0.662	0.5808
10	DenseNet	GradBoost	T1	0.928	0.989 (0.984–0.993)	0.592	0.5102

To assess whether the top TOPSIS-ranked configuration was statistically distinguishable from the remaining top 10 configurations, DeLong's test, McNemar's test, and paired bootstrap comparisons of accuracy, precision, recall, and F1-score were performed, with the results summarized in Table 3 and Supplementary Table S2. AUC values across the top 10 configurations were closely clustered (0.989–0.997), with substantially overlapping 95% confidence intervals. DeLong's test indicated that the top-ranked configuration (Inception+KNN+T1Gd) differed significantly in AUC in four of the nine remaining configurations ($p < 0.05$), while McNemar's test on binarized predictions indicated significant differences for three of the nine comparisons; no significant difference was found for the remaining majority of comparisons under either test. Paired bootstrap testing across accuracy, precision, recall, and F1-score similarly showed that only a minority of pairwise comparisons among the top 10 configurations reached statistical significance (44 of 180 comparisons, 24.4%). Notably, while the top-ranked configuration showed no significant difference in precision relative to the other InceptionV3-based classifiers evaluated on T1Gd (SVM, XGBoost, GradBoost; $p > 0.05$), it exhibited significantly lower recall than each of these alternatives ($p = 0.003$, 0.002 , and <0.001 , respectively), indicating that its top TOPSIS ranking reflects strong overall balance across the five composite criteria rather than uniform superiority on every individual metric. Taken together, these results indicate that several top-ranked pipelines, particularly those sharing the InceptionV3 feature extractor with T1 or T1Gd sequences, are not categorically distinguishable in discriminative performance, and that classifiers prioritizing higher recall, such as SVM or XGBoost on T1Gd, may represent preferable alternatives in clinical contexts where minimizing missed HGG detections is prioritized over minimizing false positives.

The results of the ablation study are summarized in Figure 11 according to three pipeline components: CNN architecture, ML classifier, and MRI image weighting.

Effect of CNN Architecture: Among the five evaluated pretrained CNN architectures, Inception achieved the highest average TOPSIS CC of 0.823, followed by DenseNet (CC = 0.784), VGG (CC = 0.693), ResNet (CC = 0.589), and EfficientNet (CC = 0.546). These findings indicate that Inception, with its multi-scale parallel feature extraction capability, provides the most discriminative deep feature representations for glioma grade classification across all MRI weightings and classifiers. The substantial performance gap between Inception and EfficientNet suggests that architectural design plays a critical role in determining the quality of extracted features within a segmentation-independent pipeline.

Effect of ML Classifier: Among the six evaluated classifiers, XGBoost achieved the highest average CC of 0.741, followed closely by RF (CC = 0.727) and GradBoost (CC = 0.722), indicating that ensemble-based tree methods are the most compatible classifiers for deep features extracted by pretrained CNNs. KNN and SVM achieved equal average CC scores of 0.667, while DNN demonstrated the lowest performance (CC = 0.600), likely attributable to its higher sensitivity to hyperparameter configuration and the relatively limited size of the feature input space.

Effect of MRI Weighting: Among the four evaluated MRI image weightings, T1Gd achieved the highest average CC of 0.772, followed by T1 (CC = 0.761), T2 (CC = 0.694), and FLAIR (CC = 0.521). The superior performance of T1Gd is consistent with its well-established clinical relevance in glioma characterization, as contrast enhancement patterns in T1Gd reflect blood–brain barrier disruption and tumor vascularity, which are key biological features differentiating HGG from LGG. The comparable performance of T1 further supports the utility of non-contrast sequences in glioma grading, while the notably

lower performance of FLAIR suggests that its diffuse signal characteristics provide less discriminative information for grade classification when used independently.

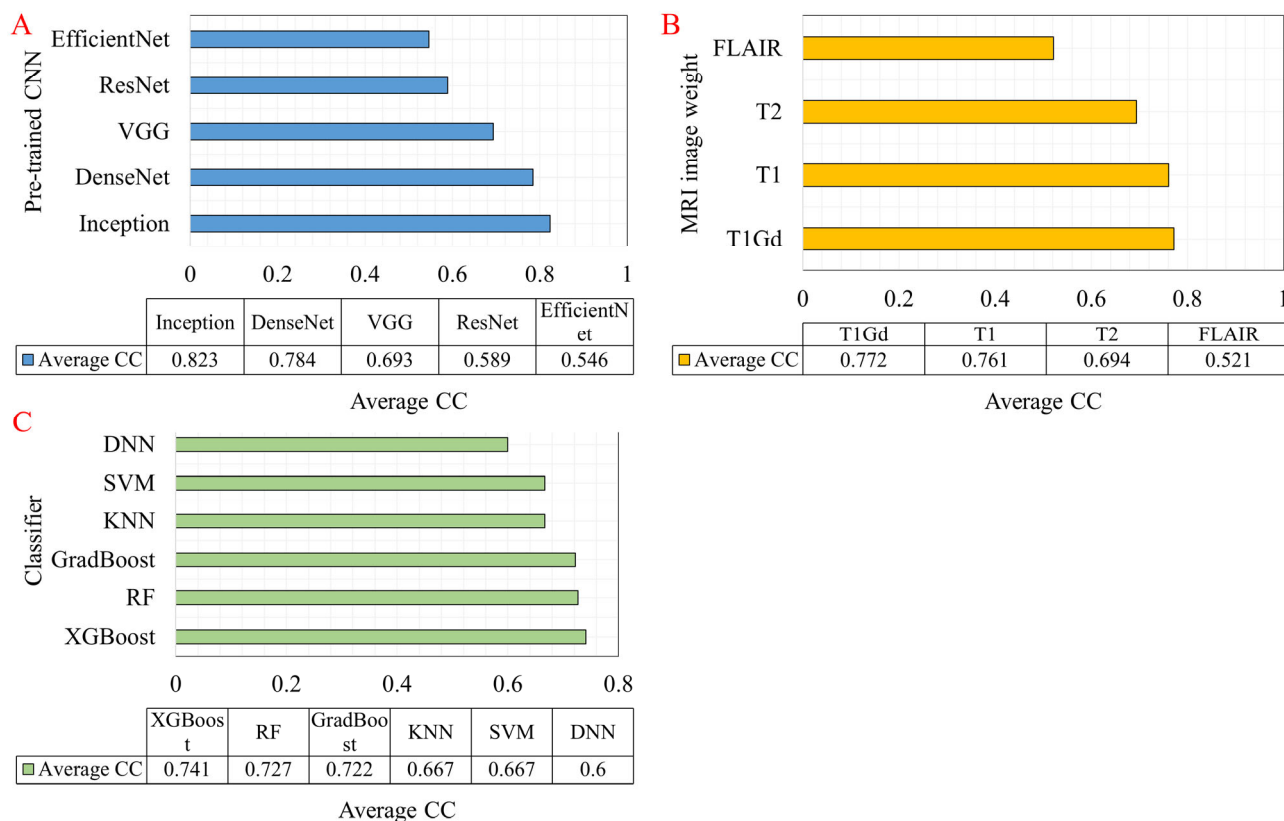


Figure 11. Ablation study results illustrating the independent contribution of each pipeline component to glioma grade classification performance, quantified by the average TOPSIS closeness coefficient (CC). **(A)** Effect of CNN architecture, comparing five pretrained feature extractors (Inception, DenseNet, VGG, ResNet, and EfficientNet) averaged across all MRI weightings and classifiers. **(B)** Effect of MRI image weighting, comparing four structural MRI sequences (T1Gd, T1, T2, and FLAIR) averaged across all CNN architectures and classifiers. **(C)** Effect of ML classifier, comparing six classifiers (XGBoost, RF, GradBoost, KNN, SVM, and DNN) averaged across all CNN architectures and MRI weightings. Higher average CC values indicate greater overall contribution to classification performance across all evaluated configurations.

The AUC heatmaps for all feature extractor–classifier combinations across the four MRI sequences are shown in Figure 10. As shown in Figures 8–10, heatmaps, and TOPSIS analysis, our results demonstrate that models trained on T1 and T1Gd MRI sequences consistently achieved higher classification performance compared to those trained with FLAIR and T2 MRI weighted images. Notably, the DNN classifier consistently underperformed across all feature extractors, achieving the lowest performance in glioma grading compared to ensemble methods like XGBoost, Gradient Boosting, RF, SVM, and KNN.

Figure 12 presents Grad-CAM++ visualizations for representative cases with a T1Gd MRI image weight for the LGG case, highlighting the image regions that most strongly influenced the classification decisions of the evaluated pretrained CNN models. Notably, the Inception-V3 pretrained CNN models consistently attended to both the tumor and brain regions, indicating an effective capture of discriminative features relevant to glioma grading. This spatial attention pattern is consistent with the strong quantitative

performance achieved by the InceptionV3 models and suggests their strong capacity to extract discriminative deep features for glioma grade classification.

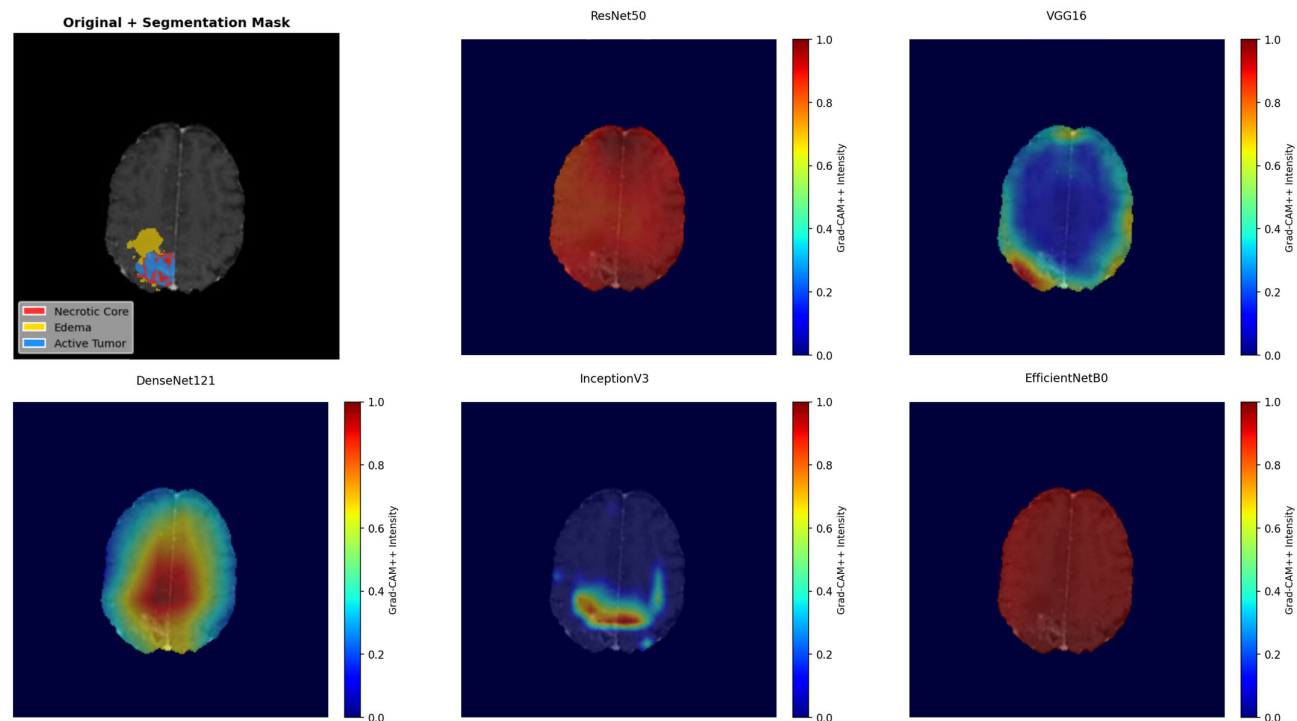


Figure 12. Grad-CAM++ visualizations for glioma grade classification using different pretrained CNN architectures. The figure presents the input contrast-enhanced T1-weighted MRI (T1Gd), followed by Grad-CAM++ heatmaps generated from ResNet-50, InceptionV3, DenseNet-121, VGG-16, and EfficientNet-B0, highlighting regions that most strongly influenced each model's prediction. The color bar represents the Grad-CAM++ activation intensity. Corresponding tumor masks are overlaid on the original T1Gd images, where red denotes edema regions, white indicates active tumor regions, and purple represents necrotic core regions.

Figure 13 illustrates the ROC curves and the corresponding confusion matrices for the two highest-performing models in glioma grade classification. Both models are based on KNN classifiers trained on deep features extracted using the Inception-V3 architecture. The first model utilizes deep features derived from T1Gd MRI image weights, whereas the second model employs deep features extracted from T1 MRI image weights.

As indicated by the ROC curves, both models demonstrate strong discriminative capability, achieving high sensitivity and specificity in differentiating between LGG and HGG cases. The accompanying confusion matrices further corroborate these findings, showing a high proportion of correctly classified samples with minimal misclassification across both glioma grades. Collectively, these results underscore the effectiveness of integrating Inception-based deep feature representations with KNN classifiers, particularly when applied to T1Gd and T1 MRI sequences.

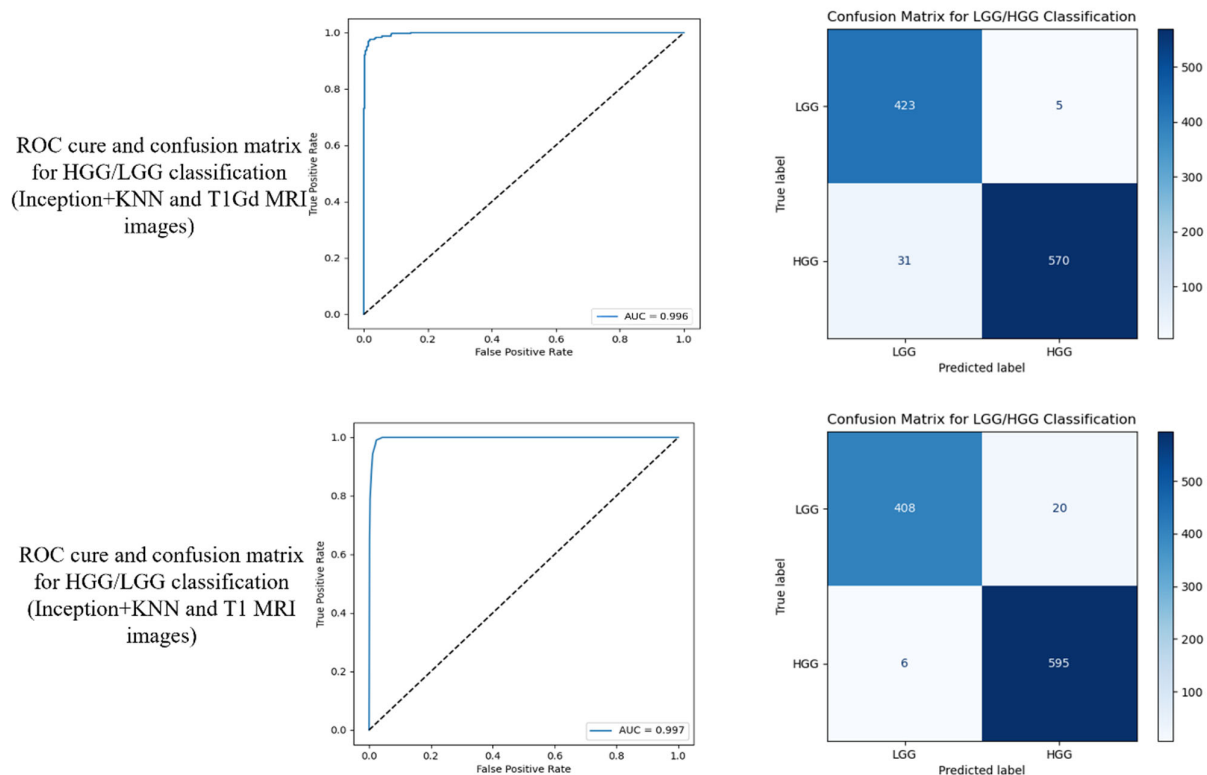


Figure 13. ROC curves and confusion matrices for the two best-performing glioma grade classification models. (**Top row**) ROC curve and confusion matrix for the KNN classifier trained on Inception deep features extracted from T1Gd MRI images. (**Bottom row**) ROC curve and confusion matrix for the KNN classifier trained on Inception deep features extracted from T1 MRI images.

4. Discussion

Accurate glioma grading is of critical clinical importance, as it directly influences treatment planning, prognostic assessment, and eligibility for clinical trials, thereby underscoring the need for automated, accurate, and efficient classification systems to support standardized clinical decision-making. Numerous studies [28,37–40] have demonstrated that ML, Radiomics, and DL approaches can substantially enhance the accuracy and efficiency of glioma grade and tumor classification compared with conventional diagnostic methods. In this context, the present study was designed to systematically examine the combined effects of different pretrained deep learning architectures employed for transfer learning as feature extractors, various machine learning classification algorithms, and structural MRI image weights on glioma grade classification performance, thereby addressing a critical gap in the existing literature.

In this study, we employed 2D MRI slices as model inputs and extracted deep features from the entire image without performing any tumor segmentation. In contrast, many previous studies [11,15,16] have relied on features extracted from manually or automatically segmented tumor regions, which require an additional segmentation step. By eliminating the need for tumor segmentation, our approach simplifies the processing pipeline and reduces dependence on expert annotations or complex segmentation algorithms. Based on our results, the image deep feature extraction strategy demonstrates strong classification performance while enabling a fully automatic and computationally efficient framework. This streamlined design contributes to faster processing and improved practicality for potential clinical usage.

Based on the accuracy, precision, recall, and F1-score values shown in Figures 8–10, ablation results in Figure 11, and TOPSIS analysis, features extracted from T1 and T1Gd MRI images consistently outperformed those derived from T2 and FLAIR images for glioma grade classification. This performance difference was observed across multiple feature extractor–classifier combinations, indicating that T1-weighted and contrast-enhanced T1Gd images provide more discriminative information for distinguishing between LGG and HGG. These findings highlight the critical role of MRI image weighting in glioma grade classification and suggest that the choice of MRI modality can substantially influence model performance. Previous radiomics-based studies on glioma grade classification using structural MRI [10,41,42] have reported that feature selection methods tend to retain a greater number of informative features from T1Gd images compared with other MRI sequences due to the high and discriminative information contained within these images for glioma grade classification. These findings are consistent with our results. The superior model performance trained with T1Gd and T1 MRI sequences may be attributed to the distinct biological information they capture. T1Gd enhancement reflects contrast agent extravasation across a disrupted blood–brain barrier, a phenomenon more pronounced in higher-grade gliomas due to their heightened angiogenesis and structural vascular abnormality, making it a robust correlate of malignancy-associated neovascularity and necrosis [43]. Even without contrast, T1 retains discriminative capacity through its sensitivity to intrinsic tumor features, including necrosis, hemorrhage, and elevated cellularity, alongside grade-dependent T1 relaxation-time alterations, collectively rendering it informative for tumor grade classification [24,44].

One of the main aims of this study was to compare the performance of different pre-trained CNN architectures as image feature extractors for glioma grade classification. According to our results, the InceptionV3 network consistently outperformed the other pre-trained models evaluated in this study. This superior performance was further supported by Grad-CAM++ visualizations, which demonstrated that InceptionV3 focused more precisely on both tumor regions and relevant surrounding brain structures, indicating more effective extraction of discriminative features for glioma grading. A previous study [22] employed various pre-trained CNN architectures combined with two to three fully connected layers for glioma grade classification, reporting that InceptionV3 achieved superior performance compared to other pre-trained networks. These findings are in line with the results of the present study. The higher performance of InceptionV3 in this study can be attributed to its ability to capture multi-scale features through parallel convolutional filters, allowing simultaneous extraction of fine tumor details and broader contextual information from brain MRI images. Compared with VGG16's sequential architecture and the deeper but more redundant representations of ResNet50 and DenseNet121, InceptionV3 provides a balanced combination of feature diversity and computational efficiency. While EfficientNetB0 emphasizes parameter efficiency, its compact design may limit task-specific feature richness. These architectural advantages are supported by Grad-CAM++ visualizations, which show that InceptionV3 focuses more effectively on tumor and relevant brain regions, contributing to improved glioma grade classification.

The most significant finding of this study is that across 120 different combinations of MRI sequences, deep feature extractors, and machine learning classifiers used in this study, deep features extracted using the InceptionV3 architecture from T1 and T1Gd MRI image weights, when coupled with a KNN classifier, yielded the highest classification performance for glioma grading. Moreover, features derived from T1Gd images demonstrated a marginal performance advantage over those obtained from T1 images, although this improvement comes at the cost of requiring contrast agent administration during MRI acquisition. Statistical comparison further supported this observation: the T1-based InceptionV3+KNN configuration achieved significantly higher accuracy, recall, and F1-score

than the T1Gd-based configuration ranked first by the composite TOPSIS score (McNemar's test, $p < 0.001$; paired bootstrap test, $p < 0.001$), while precision did not differ significantly between the two. Given this statistically supported performance advantage, combined with the practical benefit of not requiring contrast agent administration, our findings suggest that InceptionV3-extracted deep features from T1 images combined with a KNN classifier represent a statistically well-supported and practical candidate approach for glioma grade classification, although further validation on larger, independent, and molecularly characterized cohorts remains warranted before any clinical translation can be considered.

In evaluating the clinical utility of the two top-performing classification models, the direction of misclassification, as shown in Figure 13's confusion matrix, emerged as a critical consideration. The Inception-KNN model based on T1Gd features predominantly misclassified HGGs as LGGs, with 31 HGG cases incorrectly labeled as LGGs. This misclassification pattern is particularly concerning in a clinical setting, as underestimation of tumor grade may lead to insufficient treatment, rapid disease progression, and severely compromised patient outcomes. In contrast, the Inception-KNN model using T1-derived features misclassified 20 LGG cases as HGG. Although this type of misclassification is also clinically undesirable, it is more likely to result in overtreatment rather than failure to detect aggressive disease. While overtreatment may increase the risk of unnecessary toxicity and negatively affect quality of life, its consequences are generally less severe than those associated with missed HGG diagnosis. Therefore, despite comparable overall error rates, the T1Gd-based Inception-KNN model presents a greater potential risk to patient safety, underscoring the importance of prioritizing HGG sensitivity in future model optimization and clinical validation.

Previous studies [17] have applied transfer learning to structural MRI images for glioma grade classification and reported top-performing models with performance comparable to that of our best results. However, these approaches relied on features extracted from segmented tumor regions. This distinction highlights the advantage of our approach, which eliminates the need for tumor segmentation while maintaining competitive classification performance. In another study [45], researchers utilized T1Gd images from the BraTS 2021 dataset for LGG/HGG classification using various pre-trained CNN architectures. The average performance reported for networks such as DenseNet121 and EfficientNetB0 was comparable to the results obtained with these architectures in our study. However, the top-performing model in our work achieved higher classification performance than the best and proposed BoT-YOLOv6 model reported in their study. This performance gain can be attributed to our use of fine-tuned machine learning classifiers following deep feature extraction and feature selection, which enhanced the discriminative capability of the learned representations.

The classification performance achieved by our proposed pipeline was comparable to previously reported segmentation-based and radiomics-driven glioma grade classification approaches [10,46–49]. The aim of this comparison is to highlight the potential advantages of our method relative to segmentation-based methods. Radiomics-based methods, while demonstrating competitive diagnostic accuracy, inherently require precise tumor segmentation as a mandatory prerequisite for feature extraction, either through time-consuming manual delineation by expert neuroradiologists or through computationally demanding automated segmentation models that necessitate high-performance hardware infrastructure, substantial GPU memory, and considerably longer training times. Furthermore, manual segmentation is subject to inter-observer variability, which may compromise the reproducibility and consistency of extracted features across different clinical settings. In contrast, the present study adopts a mask-assisted slice-selection pipeline with segmentation-free model inputs. BraTS masks were used only during preprocessing to

identify which slices contained tumor tissue. Deep features were then extracted directly from the full MRI slice, with no tumor delineation involved in the classification step itself. This design substantially reduces computational demands, hardware requirements, and overall processing time while maintaining competitive classification performance. Notably, Grad-CAM++ visualization of the top-performing models revealed that the most discriminative features were predominantly extracted from tumor regions, providing post-hoc evidence that the models implicitly focused on clinically relevant areas despite the absence of explicit segmentation guidance. These findings collectively suggest that our approach offers a more practical and streamlined framework with comparable spatial specificity to segmentation-based methods, thereby enhancing its potential for clinical applicability, particularly in resource-limited settings where high-performance computing infrastructure may not be readily available.

An important observation in this study is that the DNN classifier exhibited comparatively lower performance than the other evaluated classifiers. This outcome is likely attributable to the fact that no hyperparameter optimization or architectural fine-tuning was applied to the DNN, including adjustments to the number of fully connected layers, the number of neurons per layer, or the optimization strategy, while the other machine learning classifiers in this study were extensively fine-tuned. As a result, the DNN was unable to fully exploit its representational capacity. Future studies may therefore benefit from systematic optimization of DNN architectures to achieve improved classification performance.

This study has several limitations that should be acknowledged. First, the number of LGG cases was relatively limited, which may have introduced class imbalance. Despite the application of data augmentation and class imbalance weighting, the substantial disparity between HGG and LGG case counts represents an inherent limitation of the current study. Future work should investigate more advanced imbalance handling strategies, including synthetic minority oversampling and cost-sensitive learning frameworks, to further improve classification performance on the minority LGG class. Performance metrics reported in this study are derived from slice-level evaluation of tumor-containing 2D slices, selected based on BraTS 2023 segmentation masks. While patient-level splitting was strictly enforced throughout all stages to prevent data leakage, a formal patient-level prediction aggregation strategy, such as majority voting or averaged-probability fusion across slices, was not implemented in the current study. As correlated slices originating from the same patient may inflate apparent slice-level performance relative to true patient-level diagnostic accuracy, the reported metrics should be interpreted as slice-level classification performance within a segmentation-independent 2D pipeline, rather than as a direct estimate of per-patient diagnostic accuracy. This study was intentionally designed as a preliminary, slice-level benchmarking investigation to establish the relative value of different MRI sequences, CNN feature extractors, and classifiers prior to more clinically oriented development. Future work will extend this framework to three-dimensional volumetric analysis and formal patient-level decision fusion, which together will more accurately reflect real-world clinical diagnostic workflows and enable rigorous estimation of per-patient diagnostic performance. Additionally, the present study relied on two-dimensional MRI slices, which may not fully capture the three-dimensional structural characteristics of gliomas. Future work should therefore investigate three-dimensional imaging-based approaches and systematically compare their performance with 2D-based methods. The LGG and HGG labels used in this study were obtained by matching subject identifiers between the BraTS 2023 glioma cohort and the previously labeled BraTS 2021 dataset, as BraTS 2023 itself does not distribute grade labels directly (see Section 2.1). These labels follow the WHO 2016 CNS tumor classification framework and are primarily histology-based and do not incorporate the integrated molecular-genetic criteria introduced in the

WHO Classification of Tumors of the Central Nervous System, 5th edition [50], including IDH1/IDH2 mutation status, 1p/19q co-deletion, and other molecular markers now used to determine tumor grade independent of histological appearance. This distinction is clinically significant, as a subset of histologically lower-grade-appearing tumors may be reclassified as high grade under WHO 2021 criteria based on molecular features alone, and vice versa. Given that our labeling procedure itself relied on legacy, histology-based grade annotations rather than direct molecular characterization, readers should interpret the grading labels accordingly, and future work should aim to validate findings using molecularly characterized cohorts, with transparently documented and verifiable labeling procedures, aligned with current WHO 2021 standards [50]. Finally, more comprehensive hyperparameter optimization and architectural tuning of DNN classifiers are recommended in future studies to potentially enhance classification performance. Another limitation of this study is that TOPSIS ranking was performed on the held-out test set across all 120 combinations, which may introduce optimistic bias. Future work should employ stricter protocols such as nested cross-validation to ensure unbiased model selection.

It is important to emphasize that the results reported in this study were obtained exclusively within the standardized and preprocessed BraTS 2023 experimental framework and should not be interpreted as evidence of inter-center robustness or guaranteed clinical transferability. Real-world deployment across heterogeneous clinical environments, involving variable MRI acquisition protocols, scanner manufacturers, field strengths, and patient demographics, would likely introduce additional sources of variability that could meaningfully affect classification performance. External validation on independent, multi-institutional, and molecularly characterized cohorts acquired under diverse real-world clinical conditions is therefore a critical and necessary step before any clinical translation of the proposed pipeline can be responsibly considered. Although a parameter-level ablation analysis was conducted by computing the average TOPSIS Closeness Coefficient independently as a function of pre-trained CNN architecture and MRI modality input, a more comprehensive component-wise ablation study, systematically evaluating the isolated contribution of individual pipeline components such as LASSO feature selection, data augmentation, and class imbalance weighting on overall classification performance, was not performed in the current study and represents an important direction for future investigation.

5. Conclusions

The present study was designed to determine the effect of different pre-trained CNNs, ML classifiers, and structural MRI image weights on glioma grade classification. In this investigation, the aim was to find the highest performing image feature extractor, classifier, and MRI image weight for LGG/HGG classification. One of the more significant findings to emerge from this study is that combining deep features extracted using the InceptionV3 architecture with a KNN classifier yields highly effective slice-level classification performance within the standardized BraTS 2023 dataset, particularly when applied to T1-weighted MRI images. Statistical comparison confirmed that this T1-based configuration significantly outperformed the composite-ranked top configuration in accuracy, recall, and F1-score, while offering the practical advantage of requiring no contrast agent. These findings should be interpreted as a promising benchmark result rather than a demonstration of clinical robustness, given the slice-level evaluation and single-dataset setting of this study. Taken together, these results suggest that careful selection of MRI image weight, deep feature extractor, and classifier plays a crucial role in optimizing glioma grade classification performance. Future studies will assess the effect of extending the proposed framework to three-dimensional MRI volumes and external multi-center datasets to further evaluate its robustness, generalizability, and clinical applicability.

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/make8080228/s1>. Figure S1: Detailed components of the Inception Block-B of Inception-V3 architecture used for feature extraction; Figure S2: Detailed components of the Inception Block-C of Inception-V3 architecture used for feature extraction; Figure S3: Detailed components of the Reduction Block-A and B of Inception-V3 architecture used for feature extraction. Table S1: TOPSIS-based ranking of all 120 combinations of pre-trained CNN feature extractors, machine learning classifiers, and MRI modality inputs for glioma grade classification. Table S1. Pairwise statistical comparison of accuracy, precision, recall, and F1-score across all combinations of the top 10 TOPSIS-ranked pipeline configurations (Table 3), derived from 2000-iteration paired bootstrap resampling.

Author Contributions: Conceptualization, A.K., G.A. and A.D.I.; Methodology, A.K. and G.A.; Software, A.K.; Validation, A.K. and A.D.I.; Formal analysis, A.K.; Investigation, A.K.; Resources, A.K.; Data curation, A.K.; Writing—original draft, A.K.; Writing—review and editing, A.K., G.A. and A.D.I.; Visualization, A.K.; Supervision, A.K.; Project administration, A.K.; Funding acquisition, A.K. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by Isfahan University of Medical Sciences, Isfahan, I.R. Iran (grant number 2404197).

Institutional Review Board Statement: This study was conducted in accordance with the principles of the Declaration of Helsinki, and ethical approval for this study was granted by the research ethics committee of Isfahan University of Medical Sciences, Isfahan, Iran, with the ID: IR.MUI.DHMT.REC.1404.147 (Approval Date: 28 October 2025).

Informed Consent Statement: Not applicable. The study used de-identified, publicly available imaging data from the Multimodal Brain Tumor Segmentation (BraTS) dataset, for which informed consent was obtained or waived by the original contributing institutions as part of the challenge’s ethical approvals.

Data Availability Statement: The data presented in this study are available in the BraTS 2023 Training dataset repository on Kaggle at <https://www.kaggle.com/datasets/bkb2024/brats-2023-training> (Accessed on 10 January 2024). These data were derived from the Brain Tumor Segmentation (BraTS) 2023 Challenge dataset, originally distributed by the Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania (contact: brats@cbica.upenn.edu). The trained models and additional analyzed data generated in this study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors express their gratitude to Isfahan University of Medical Sciences for providing the financial support for this work. During the preparation of this manuscript, the authors used Claude (Anthropic) [Claude Sonnet 5] for English language and grammar editing purposes only. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Conflicts of Interest: The authors declare that they have no conflicts of interest.

References

1. Fernandes, R.T.; Teixeira, G.R.; Mamere, E.C.; Bandeira, G.A.; Mamere, A.E. The 2021 World Health Organization classification of gliomas: An imaging approach. *Radiol. Bras.* **2023**, *56*, 157–161. <https://doi.org/10.1590/0100-3984.2022.0089-en>.
2. Segura, P.P.; Quintela, N.V.; García, M.M.; Berrón, S.d.B.; Sarrió, R.G.; Gómez, J.G.; Castaño, A.G.; Martín, L.M.N.; Rubio, O.G.; Losada, E.P. SEOM-GEINO clinical guidelines for high-grade gliomas of adulthood (2022). *Clin. Transl. Oncol.* **2023**, *25*, 2634–2646. <https://doi.org/10.1007/s12094-023-03245-y>.
3. Wu, M.; Luan, J.; Zhang, D.; Fan, H.; Qiao, L.; Zhang, C. Development and validation of a clinical prediction model for glioma grade using machine learning. *Technol. Health Care* **2024**, *32*, 1977–1990. <https://doi.org/10.3233/thc-231645>.

4. Pandey, S.; Trivedi, M.C.; Bharadwaj, B.; Singh, V.K. Brain Tumor Classification into High-Grade and Low-Grade Gliomas Using Adaboost. In *Advances in Computational Intelligence and Communication Technology: Proceedings of CICT 2021*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 433–440.
5. Kaur, A.; Kukreja, V.; Ati, M.; Bansal, A.; Hariharan, S. BrainDx: A dual-transformer framework using PVT and SegFormer for tumor diagnosis. *Biomed. Signal Process. Control* **2026**, *113*, 108917. <https://doi.org/10.1016/j.bspc.2025.108917>.
6. Attallah, O.; Pacal, I. Comparative Evaluation of Lightweight Convolutional Neural Network and Vision Transformer Models for Multi-Class Brain Tumor Classification Using Merged Large MRI Datasets. *Chemom. Intell. Lab. Syst.* **2025**, *269*, 105609. <https://doi.org/10.1016/j.chemolab.2025.105609>.
7. Gökmen, N. AI techniques for brain tumor segmentation in MRI: A review (2019–2024). *Netw. Model. Anal. Health Inform. Bioinform.* **2025**, *14*, 168. <https://doi.org/10.1007/s13721-025-00650-x>.
8. Tas, I.C. Brain Tumor Classification of Mri Images via Deep Learning Approximations: A Review. *Biomed. Eng. Appl. Basis Commun.* **2025**, *38*, 2530005. <https://doi.org/10.4015/s1016237225300056>.
9. Alongi, P.; Arnone, A.; Vultaggio, V.; Fraternali, A.; Versari, A.; Casali, C.; Arnone, G.; DiMeco, F.; Vetrano, I.G. Artificial intelligence analysis using MRI and PET imaging in gliomas: A narrative review. *Cancers* **2024**, *16*, 407. <https://doi.org/10.3390/cancers16020407>.
10. Khorasani, A. Glioma Grading by Integrating Radiomic Features from Peritumoral Edema in Fused MRI Images and Automated Machine Learning. *J. Imaging* **2025**, *11*, 336. <https://doi.org/10.3390/jimaging11100336>.
11. Gao, M.; Cheng, J.; Qiu, A.; Zhao, D.; Wang, J.; Liu, J. Magnetic resonance imaging (MRI)-based intratumoral and peritumoral radiomics for prognosis prediction in glioma patients. *Clin. Radiol.* **2024**, *79*, e1383–e1393. <https://doi.org/10.1016/j.crad.2024.08.005>.
12. Xu, Y.; Li, Y.; Wang, F.; Zhang, Y.; Huang, D. Addressing the current challenges in the clinical application of AI-based Radiomics for cancer imaging. *Front. Med.* **2025**, *12*, 1674397. <https://doi.org/10.3389/fmed.2025.1674397>.
13. Namdar, K.; Wagner, M.W.; Ertl-Wagner, B.B.; Khalvati, F. Open-radiomics: A collection of standardized datasets and a technical protocol for reproducible radiomics machine learning pipelines. *BMC Med. Imaging* **2025**, *25*, 312. <https://doi.org/10.1186/s12880-025-01855-2>.
14. Wan, B.; Hu, B.; Zhao, M.; Li, K.; Ye, X. Deep learning-based magnetic resonance image segmentation technique for application to glioma. *Front. Med.* **2023**, *10*, 1172767. <https://doi.org/10.3389/fmed.2023.1172767>.
15. Srivastava, G.R.; Gera, P.; Rani, R.; Jaiswal, G.; Sharma, A. A novel method for glioma segmentation and classification on pre-operative MRI scans using 3D U-Nets and transfer learning. *Multimed. Tools Appl.* **2025**, *84*, 3569–3609. <https://doi.org/10.1007/s11042-024-19261-1>.
16. Tandel, G.S.; Tiwari, A.; Kakde, O. Performance enhancement of MRI-based brain tumor classification using suitable segmentation method and deep learning-based ensemble algorithm. *Biomed. Signal Process. Control* **2022**, *78*, 104018. <https://doi.org/10.1016/j.bspc.2022.104018>.
17. Yang, J.; Siddique, M.A.; Ullah, H.; Gilanie, G.; Por, L.Y.; Alshathri, S.; El-Shafai, W.; Aldossary, H.; Gadekallu, T.R. BrainCNN: Automated brain tumor grading from magnetic resonance images using a convolutional neural network-based customized model. *SLAS Technol.* **2025**, *34*, 100334. <https://doi.org/10.1016/j.slant.2025.100334>.
18. Shirehjini, O.F.; Mofrad, F.B.; Shahmohammadi, M.; Karami, F. Grading of gliomas using transfer learning on MRI images. *Magn. Reson. Mater. Phys. Biol. Med.* **2023**, *36*, 43–53. <https://doi.org/10.1007/s10334-022-01046-y>.
19. Aziz, A.; Attique, M.; Tariq, U.; Nam, Y.; Nazir, M.; Jeong, C.-W.; Mostafa, R.R.; Sakr, R.H. An Ensemble of Optimal Deep Learning Features for Brain Tumor Classification. *Comput. Mater. Contin.* **2021**, *69*, 2653–2670. <https://doi.org/10.32604/cmc.2021.018606>.
20. Tripathi, P.C.; Bag, S. A computer-aided grading of glioma tumor using deep residual networks fusion. *Comput. Methods Programs Biomed.* **2022**, *215*, 106597. <https://doi.org/10.1016/j.cmpb.2021.106597>.
21. Khazaei, Z.; Langarizadeh, M.; Ahmadabadi, M.E.S. Developing an artificial intelligence model for tumor grading and classification, based on MRI sequences of human brain gliomas. *Int. J. Cancer Manag.* **2022**, *15*, e120638. <https://doi.org/10.5812/ijcm.120638>.
22. Rubio, S.V.; García-Ordás, M.T.; Olivera, O.G.-O.; Alaiz-Moretón, H.; González-Alonso, M.-I.; Benítez-Andrades, J.A. Survival and grade of the glioma prediction using transfer learning. *PeerJ Comput. Sci.* **2023**, *9*, e1723. <https://doi.org/10.7717/peerj-cs.1723>.
23. Tavakoli, M.B.; Khorasani, A.; Jalilian, M. Improvement grading brain glioma using T2 relaxation times and susceptibility-weighted images in MRI. *Inform. Med. Unlocked* **2023**, *37*, 101201. <https://doi.org/10.1016/j.imu.2023.101201>.

24. Khorasani, A.; Tavakoli, M.B. Multiparametric study for glioma grading with FLAIR, ADC map, eADC map, T1 map, and SWI images. *Magn. Reson. Imaging* **2023**, *96*, 93–101. <https://doi.org/10.1016/j.mri.2022.12.004>.
25. Irmak, E. Multi-Classification of Brain Tumor MRI Images Using Deep Convolutional Neural Network with Fully Optimized Framework. *Iran. J. Sci. Technol. Trans. Electr. Eng.* **2021**, *45*, 1015–1036. <https://doi.org/10.1007/s40998-021-00426-9>.
26. Nazir, M.; Shakil, S.; Khurshid, K. End-to-End Multi-task Learning Architecture for Brain Tumor Analysis with Uncertainty Estimation in MRI Images. *J. Imaging Inform. Med.* **2024**, *37*, 2149–2172. <https://doi.org/10.1007/s10278-024-01009-w>.
27. Safy, M.; Abd-Ellah, M.K.; Bayoumi, E.S.; Salama, G.M. Adaptive multi-feature fusion architecture with optimized learning for high-fidelity brain tumor classification in MRI. *Sci. Rep.* **2026**, *16*, 8498. <https://doi.org/10.1038/s41598-026-38672-8>.
28. Yathirajam, S.S.; Gutta, S. Efficient glioma grade prediction using learned features extracted from convolutional neural networks. *J. Med. Artif. Intell.* **2024**, *7*, 2. <https://doi.org/10.21037/jmai-23-161>.
29. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
30. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2016; pp. 770–778.
31. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2016; pp. 2818–2826.
32. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2017; pp. 4700–4708.
33. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning*; PMLR: London, UK, 2019; pp. 6105–6114.
34. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1996**, *58*, 267–288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>.
35. Chattopadhyay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*; IEEE: New York, NY, USA, 2018; pp. 839–847.
36. Hwang, C.-L.; Yoon, K. *Multiple Attribute Decision Making: Methods and Applications a State-of-the-Art Survey*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2012.
37. Remzan, N.; Tahiry, K.; Farchi, A. Advancing brain tumor classification accuracy through deep learning: Harnessing radimagenet pre-trained convolutional neural networks, ensemble learning, and machine learning classifiers on MRI brain images. *Multimed. Tools Appl.* **2024**, *83*, 82719–82747. <https://doi.org/10.1007/s11042-024-18780-1>.
38. Lysdahlgaard, S.; Jørgensen, M.D. Artificial intelligence and advanced MRI techniques: A comprehensive analysis of diffuse gliomas. *J. Med. Imaging Radiat. Sci.* **2024**, *55*, 101736. <https://doi.org/10.1016/j.jmir.2024.101736>.
39. Tan, R.; Sui, C.; Wang, C.; Zhu, T. MRI-based intratumoral and peritumoral radiomics for preoperative prediction of glioma grade: A multicenter study. *Front. Oncol.* **2024**, *14*, 1401977. <https://doi.org/10.3389/fonc.2024.1401977>.
40. Aouadi, S.; Torfeh, T.; Arunachalam, Y.; Paloor, S.; Riyas, M.; Hammoud, R.; Al-Hammadi, N. Investigation of radiomics and deep convolutional neural networks approaches for glioma grading. *Biomed. Phys. Eng. Express* **2023**, *9*, 35020. <https://doi.org/10.1088/2057-1976/acc33a>.
41. Zachariah, R.M.; Priya, P.S.; Pendem, S. Classification of low- and high-grade gliomas using radiomic analysis of multiple sequences of MRI brain. *J. Cancer Res. Ther.* **2023**, *19*, 435–446. https://doi.org/10.4103/jcrt.jcrt_1581_22.
42. Ding, J.; Zhao, R.; Qiu, Q.; Chen, J.; Duan, J.; Cao, X.; Yin, Y. Developing and validating a deep learning and radiomic model for glioma grading using multiplanar reconstructed magnetic resonance contrast-enhanced T1-weighted imaging: A robust, multi-institutional study. *Quant. Imaging Med. Surg.* **2022**, *12*, 1517–1528. <https://doi.org/10.21037/qims-21-722>.
43. Goryawala, M.; Roy, B.; Gupta, R.K.; Maudsley, A.A. T1-weighted and T2-weighted Subtraction MR Images for Glioma Visualization and Grading. *J. Neuroimaging* **2021**, *31*, 124–131. <https://doi.org/10.1111/jon.12800>.
44. Sanada, T.; Yamamoto, S.; Sakai, M.; Umehara, T.; Sato, H.; Saito, M.; Mitsui, N.; Hiroshima, S.; Anei, R.; Kanemura, Y.; et al. Correlation of T1- to T2-weighted signal intensity ratio with T1- and T2-relaxation time and IDH mutation status in glioma. *Sci. Rep.* **2022**, *12*, 18801. <https://doi.org/10.1038/s41598-022-23527-9>.
45. Zhang, H.; Huang, S.-L.; Kuruoglu, E.E. Classification and recognition of brain tumors from MRI images based on improved YOLOv6. *Biomed. Signal Process. Control* **2026**, *113*, 109048. <https://doi.org/10.1016/j.bspc.2025.109048>.
46. Feng, L.; Hao, G.; Jian, W. Efficient hybrid fuzzy weighted 3D FCNN with TSO PSO optimization for accurate multi modal MRI brain tumor classification. *Sci. Rep.* **2025**, *15*, 36037. <https://doi.org/10.1038/s41598-025-19889-5>.

47. Scientific, L.L. Brain tumour segmentation and grade classification using modified VGG19 with support vector machine. *J. Theor. Appl. Inf. Technol.* **2025**, *103*, 1095–1116.
48. Sultan, M.N.; Yehia, S.; Khalil, M.M. Potential of MR-based radiomics and optimized statistical machine learning in grading patients with glioma. *Egypt. J. Radiol. Nucl. Med.* **2025**, *56*, 30. <https://doi.org/10.1186/s43055-025-01443-y>.
49. Shi, Y.; Jiang, W.; Zhou, L.; Shi, Y.; Ji, X.; Zheng, W.; Zhao, H.; Yao, N. Radiomics-based machine learning for glioma grade classification: A multicenter study with SHapley Additive exPlanations interpretability analysis. *Transl. Cancer Res.* **2025**, *14*, 7329–7346.
50. Louis, D.N.; Perry, A.; Wesseling, P.; Brat, D.J.; Cree, I.A.; Figarella-Branger, D.; Hawkins, C.; Ng, H.K.; Pfister, S.M.; Reifenberger, G.; et al. The 2021 WHO classification of tumors of the central nervous system: A summary. *Neuro-Oncology* **2021**, *23*, 1231–1251. <https://doi.org/10.1093/neuonc/noab106>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.