Dual-Center Study: Noninvasive Prediction of Ki-67 Status in Intracranial Tumors from Contrast-Enhanced MRI with Explainable XGBoost

Yang Yang[1,2,4*], Yang Lv[3*], Zhongying Li[5*], Dasheng Wang[1,2], Xianchao Hu[1,2], Longfei Hu[4], Yong Guan[4], Fei Wang[1,2#], Zheng Jiang[4#], Yongfei Dong[1,2#]

[1]Department of Neurosurgery, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui, China

[2] Department of Neurosurgery, Centre for Leading Medicine and Advanced Technologies of IHM, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui, China

[3]Department of Radiology, the First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui, China.

[4]National Synchrotron Radiation Laboratory, University of Science and Technology of China, Hefei, Anhui, China

[5]Department of Neurosurgery, the First Affiliated Hospital of Anhui Medical University, Hefei, Anhui, China

*are co-first authors

[#]**Correspondence:** Yongfei Dong (dyf.w@163.com)

Zheng Jiang (jiangz@ustc.edu.cn)

Fei Wang (neurosurgeonahwf@126.com)

Keywords:

Magnetic resonance imaging, Intracranial Tumors, Ki-67 Prediction, Machine Learning, XGBoost, Deep Learning Features, Explainable Artificial Intelligence, Shapley Additive Explanations

Ethics and trial registration

This study was approved by the Research Ethics Committees of the First Affiliated Hospital of USTC (2026-BE(H)-035), and the requirement for written informed consent was waived.

ABSTRACT

Background and Objective: The Ki-67 proliferation index is a key biomarker for assessing the malignancy and prognosis of intracranial tumors, but its acquisition relies on invasive surgery. This study aimed to develop an explainable machine-learning model that non-invasively predicts Ki-67 status in intracranial tumors using preoperative contrast-enhanced T1 (T1CE)-weighted magnetic resonance imaging (MRI). Methods: We retrospectively analyzed a final cohort of 573 pathologically confirmed cases, including 394 from the First Affiliated Hospital of USTC and 179 from the First Affiliated Hospital of Anhui Medical University, across five tumor types: glioma (n=369), meningioma (n=53), acoustic neuroma (n=52), metastatic tumor (n=50), and sellar lesion (n=49). Patients were dichotomized into high- (Ki-67 >=10%) and low-expression groups. After standardized preprocessing and nnU-Net-based tumor segmentation, we extracted clinical (age, gender), hand-crafted features (n=23), and deep-learning (DL) features (n=256) from T1CE. A selection strategy identified 46 features (2 clinical, 14 hand-crafted, 30 DL) for model development. Results: Using 5-fold stratified cross-validation, an XGBoost classifier achieved an out-of-fold AUC of 0.804 (95% CI: 0.769-0.838), accuracy 0.743, sensitivity 0.785, and specificity 0.691, with acceptable probability calibration. SHapley Additive exPlanations (SHAP) provided global and local interpretability, highlighting location, enhancement, and ring-related descriptors (e.g., skullbase_distance_mm, rim_core_entropy_diff, ring_continuity_pct) alongside influential DL features. Conclusion: These results indicate that a T1CE-based, explainable XGBoost model shows promise for non-invasively predicting Ki-67 status across diverse intracranial tumor types, particularly for gliomas, offering imaging-derived evidence to support clinical decision-making.

1. Introduction

Intracranial tumors are among the most prevalent malignant lesions in the central nervous system, displaying notable heterogeneity in terms of malignancy, invasiveness, and prognosis. Accurately assessing the biological behavior of tumors is the fundamental prerequisite for optimizing treatment strategies[1, 2]. The Ki-67 proliferation index, as the

gold-standard biomarker for reflecting the proliferative activity of tumor cells, directly correlates with tumor malignancy and clinical prognosis by labeling proliferating cells in the G1, S, G2, and M phases[3, 4]. In common intracranial tumors such as gliomas, meningiomas, and metastatic tumors, high Ki-67 expression (usually with a 10% threshold) not only serves as a crucial basis for tumor grading (for example, high - grade gliomas often have a Ki-67 index >10%), but is also closely associated with resistance to radiotherapy and chemotherapy, an increased risk of recurrence, and a shortened survival time for patients[3-5]. Moreover, the Ki-67 index can guide treatment selection. For example, in meningiomas with high Ki-67 expression, postoperative adjuvant radiotherapy significantly reduces the recurrence rate, while those with low expression may only need surgical resection[5, 6].

Currently, the acquisition of the Ki-67 index relies on histopathological examination of surgical or biopsy specimens, a process that has inherent limitations. First, invasive procedures may lead to complications such as bleeding and infection, particularly for tumors located in critical functional areas like the brainstem or thalamus, where surgical biopsy carries extremely high risks. Second, tumors exhibit intratumoral heterogeneity; biopsy sampling may result in "sampling bias," failing to comprehensively reflect the proliferative activity of the entire tumor. Third, pathological results typically require several days post-surgery to obtain, thus failing to provide timely support for preoperative treatment decisions (such as surgical approach planning or the need for preoperative neoadjuvant radiotherapy or chemotherapy)[7-9]. Therefore, developing a non-invasive, accurate, and preoperatively applicable method for assessing the Ki-67 index holds significant clinical importance for advancing the precision diagnosis and treatment of intracranial tumors.

Magnetic resonance imaging (MRI) has become the preferred imaging modality for t he diagnosis and evaluation of intracranial tumors due to its high soft tissue resolution, m ulti-sequence imaging capability, and non-invasiveness[10-12]. It provides rich anatomical a nd functional information. In recent years, with the advancement of radiomics and deep l earning technologies, the extraction of quantitative features from MRI to predict the mole cular pathological status of tumors has become a research hotspot. Previous studies have primarily focused on Ki-67 prediction for single tumor types (e.g., gliomas) and relied on

manually delineated radiomics features, which have limited generalization and clinical applicability[3, 7]. This study aimed to address this challenge by constructing a unified machine learning model to achieve non-invasive prediction of Ki-67 status for five common intracranial tumors, including gliomas, meningiomas, acoustic neuromas, metastatic tumors, and sellar lesions (primarily pituitary adenomas)[13-16]. We integrated clinical information, interpretable hand-crafted features, and deep learning features. These features were extracted from supervised learning tasks to comprehensively capture tumor proliferation-related imaging phenotypes. We hypothesized that this multimodal feature fusion strategy can enhance the model's predictive performance and generalization ability, and reveal the model's decision-making mechanisms through advanced interpretable tools (SHAP), thereby improving its credibility in clinical applications[17-19].

This study had three primary objectives: 1. To develop an automated workflow applicable to various intracranial tumor types for predicting Ki-67 expression levels. 2. To investigate the enhancement of predictive performance through a multimodal strategy integrating clinical information, hand-crafted features, and deep learning features. 3. To elucidate the association between key imaging features and Ki-67 status using interpretable AI techniques, thereby providing a basis for clinical translation of the model.

2. Materials and methods

The overall workflow of our study is illustrated in **Figure 1**. The process includes four main stages: (I) Data Collection and Preprocessing, (II) Feature Extraction, (III) Feature Selection, and (IV) XGBoost Model Development, Validation, and Evaluation[20-22].

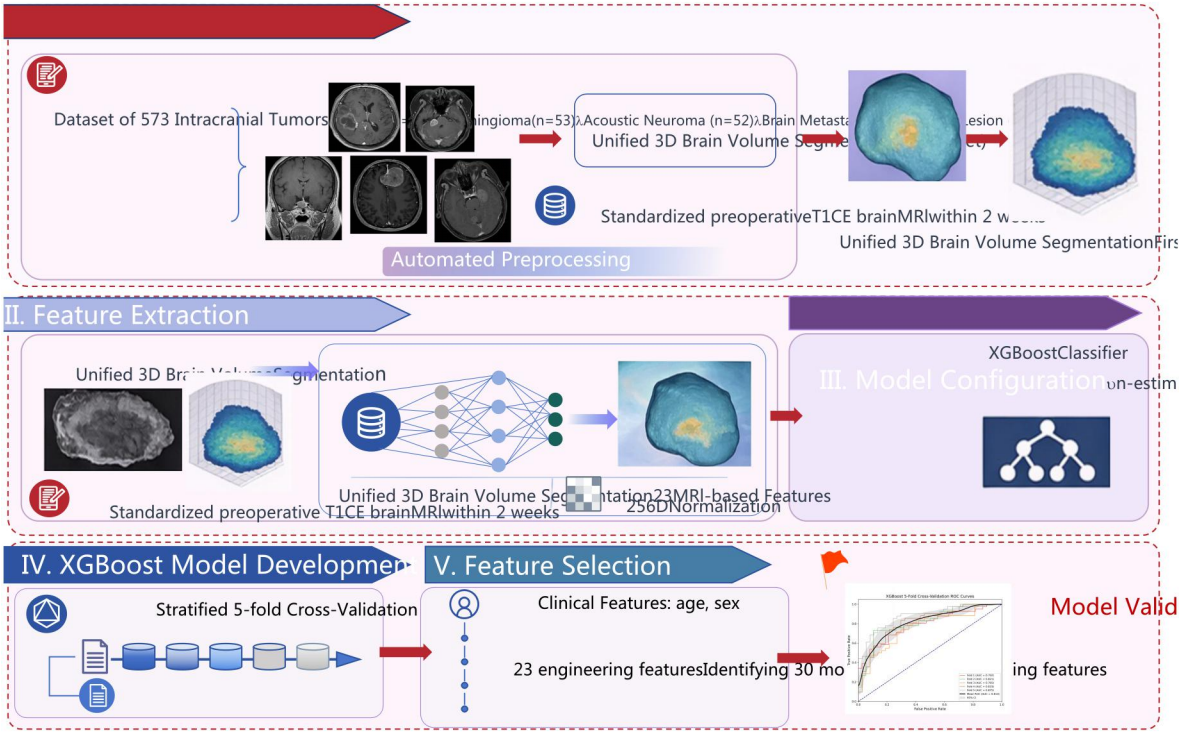

Figure 1: The overall study workflow.

This flowchart illustrates the key steps of our methodology. (I) We start with a dataset of 573 intracranial tumor cases, perform automated preprocessing and unified 3D brain volume segmentation using nnU-Net. (II) We then extract 23 MRI-based hand-crafted features and 256 deep learning features. (III) A feature selection process identifies the most relevant 14 hand-crafted and 30 deep learning features, along with 2 clinical features. (IV) Finally, these 46 selected features are used to train and evaluate an XGBoost classifier using 5-fold cross-validation, with performance assessed by metrics such as AUC.

2.1. Patient Cohort and Demographics

This retrospective study was approved by the Institutional Review Board, and the requirement for written informed consent was waived. We identified a final cohort of 573 patients with pathologically confirmed intracranial tumors who underwent surgical resection or biopsy. Among these 573 cases, 394 were from the First Affiliated Hospital of USTC and 179 were from the First Affiliated Hospital of Anhui Medical University.

Inclusion criteria: (1) Postoperative pathology confirmed as intracranial tumors, with tumor types including glioma, meningioma, acoustic neuroma, metastatic tumor, or sellar lesions (primarily pituitary adenomas); (2) Availability of preoperative MRI with at least T1CE sequence within 2 weeks prior to surgery/biopsy, with image quality meeting diagnostic requirements (no significant motion artifacts or magnetic susceptibility

artifacts); (3) Pathological report explicitly documenting Ki-67 proliferation index test results; (4)Availability of core clinical and pathological records required for analysis.

Table 1: Baseline characteristics of the patient cohort.

Tumor Type	N	Age (Mean±SD)	Male n(%)	Female n(%)	Ki-67 High n(%)	Ki-67 Low n(%)
Overall	573	51.9±13.8	318 (55.49%)	255 (44.5%)	317 (55.3%)	256 (44.7%)
Glioma	369	48.6±13.7	221(59.9%)	148 (40.1%)	254 (68.8%)	115 (31.2%)
Acoustic Neuroma	52	57.4±12.4	29 (55.8%)	23 (44.2%)	5 (9.6%)	47 (90.4%)
Meningioma	53	57.5±11.1	16 (30.2%)	37 (69.8%)	11 (20.8%)	42 (79.2%)
Metastatic Tumor	50	63.5±8.8	32 (64.0%)	18 (36.0%)	44 (88.0%)	6 (12.0%)
Sellar Lesion	49	52.3±13.1	20 (40.8%)	29 (59.2%)	3 (6.1%)	46 (93.9%)

This table summarizes the demographic data, including age and gender, and the distribution of Ki-67 expression levels across the five different types of intracranial tumors included in the study. Exclusion criteria: (1) Previous antitumor treatments such as radiotherapy, chemotherapy, targeted therapy, or immunotherapy; (2) Concurrent central nervous system disorders (e.g., cerebral infarction, cerebral hemorrhage, multiple sclerosis); (3) Severe artifacts in MRI images or incomplete scanning parameters; (4) Substandard pathological specimen quality or non-standard Ki-67 detection methods; (5) Unavailable key pathological records, non-evaluable imaging, or unavailable follow-up information.

The final cohort comprised 573 patients, including 369 cases of glioma, 53 of meningioma, 52 of acoustic neuroma, 50 of metastatic tumor, and 49 of sellar lesions. The overall mean age was 51.9 +/- 13.8 years, and the cohort included 318 males , 255 females. For the analysis, patients were divided into two groups based on the Ki-67 labeling index: a low-expression group (Ki-67 < 10%, n=256) and a high-expression group (Ki-67 >= 10%, n=317). Detailed demographic and histopathological information for the overall cohort and stratified by tumor type is summarized in Table 1.

2.2. MRI Acquisition and Preprocessing

Patients underwent preoperative brain MRI according to clinical protocols at participating data sources. To ensure consistency and comparability across the merged dataset, a standardized preprocessing pipeline was applied to all MRI scans using tools from ITK-SNAP software. The pipeline consisted of the following steps:

1. **Image Registration and Standardization**: All MRI sequences for each patient were co-registered to the T1CE volume. Subsequently, all images were resampled to a uniform isotropic resolution of 1.0 × 1.0 × 1.0 mm^3.
2. **Bias Field Correction**: The N4 bias field correction algorithm was applied to mitigate low-frequency intensity non-uniformities caused by magnetic field inhomogeneities.
3. **Intensity Normalization**: Image intensities were normalized by converting them to Z-scores (zero mean, unit variance) relative to the mean and standard deviation of intensities within the brain mask.

2.3. Tumor Segmentation

Regions of interest (ROIs) were independently manually delineated by two board-certified radiologists using ITK-SNAP software. In cases of inter-observer disagreement, adjudication was carried out by a senior radiologist with more than two decades of clinical experience. To standardize image resolution across all volumetric datasets, voxel spacing was resampled to 1 mm^3 isotropic dimensions using a fixed-resolution interpolation technique[23]. Accurate and consistent delineation of the tumor volume is critical for feature extraction, especially in a multi-observer setting. As the ground truth masks in our cohort were annotated by multiple radiologists, we employed the nnU-Net framework to develop a unified, automated segmentation model[24]. The purpose of this was to create a standardized segmentation pipeline that mitigates potential inter-observer variability, ensuring that the subsequent feature extraction is based on consistent and reproducible tumor masks. The nnU-Net model was configured to use the `3d_fullres` mode and was trained using a 5-fold cross-validation scheme. This single, robust model was capable of segmenting all five tumor types, providing a harmonized basis for our analysis.

2.4. Feature Extraction

To ensure the robustness and reproducibility of our results across different scanners and acquisition protocols, we adopted a specific strategy for feature extraction based on a quality assessment of our cohort's imaging data. Our analysis revealed significant variability in slice thickness and anisotropy, particularly in T2-weighted sequences. In contrast, T1-weighted contrast-enhanced (T1CE) sequences more frequently offered thin-slice or near-isotropic data, providing clearer tumor boundaries and higher tumor-to-background contrast.

Therefore, to minimize biases introduced by resampling thick-slice, anisotropic data, we restricted the extraction of quantitative features—both hand-crafted and deep learning—to the **T1CE sequence only**. The roles of T1 and T2 sequences were primarily auxiliary:

1. **To Assist in Feature Engineering**: The peritumoral edema, most sensitively visualized on T2 images, was segmented and used in conjunction with the T1CE tumor mask to calculate the Edema Index (Edema Volume / Tumor Volume), which was included as a hand-crafted feature.
2. **To Inform Phenotypic Labels**: Radiologists used T2 images to determine binary phenotypic features (e.g., presence of a CSF cleft), which are less sensitive to slice thickness variations.

This strategy maximizes the stable, high-resolution information from T1CE for quantitative analysis while retaining the diagnostic value of T2 for assessing peritumoral changes and qualitative phenotypes. From this foundation, we extracted three categories of features:

2.4.1. Clinical Features

Two fundamental clinical variables were collected from patient medical records:

- **Age**: Patient's age at the time of MRI scan (continuous variable).
- **Gender**: Patient's gender (binary variable, encoded as Male=1, Female=0).

2.4.2. Hand-Crafted Features

A set of 23 quantitative, hypothesis-driven features were engineered to capture macroscopic properties of the tumor and its surrounding environment[25, 26]. These features were designed to quantify aspects of tumor volume, enhancement patterns, necrosis, spatial location, and relationship with adjacent structures. The features were grouped into categories including:

- **Volume and Ratios**: e.g., tumor_volume_ml, edema_volume_ml, edema_index.
- **Enhancement/Necrosis**: e.g., necrosis_ratio, enhancing_ratio.
- **Rim/Core Contrast**: e.g., rim_core_mean_ratio, ring_continuity_pct.
- **Location and Geometry**: e.g., laterality, center_distance_mm.
- **Cerebrospinal Fluid(CSF) Adjacency**: e.g., csf_cleft_present, min_csf_distance_mm.

2.4.3. Deep Learning (DL) Features

To capture intricate and hierarchical image patterns not easily described by hand-crafted features, we developed a task-driven feature learning pipeline using a custom 3D U-Net architecture[27]. This approach is distinct from using the segmentation model's (nnU-Net) features.

The process was as follows:

1. **Model Architecture**: A simplified 3D U-Net model (SimpleUNet3D_256) was constructed with an encoder-decoder structure. The model takes a T1CE image patch as input and is trained to produce a corresponding tumor segmentation mask.
2. **Supervised Training**: Within each fold of the 5-fold cross-validation, the network was trained on the training-split T1CE images with the tumor masks (generated by the nnU-Net pipeline) serving as the ground truth. The training objective was to minimize a Dice Loss function, effectively teaching the network to perform the segmentation task.

3. **Feature Extraction**: After 30 epochs of training, the model's weights were frozen. For each patient's T1CE image, we performed a forward pass and extracted the activation map from the penultimate layer of the decoder (d1). A global average pooling operation was applied to this activation map, followed by a linear projection layer, to produce a final, fixed-length feature vector of 256 dimensions.

This supervised pre-training strategy allows the network to learn a rich representation of tumor morphology and texture relevant to its structure. These learned features are then transferred for the downstream task of Ki-67 prediction.

2.4.4. Radiomics Features (for comparison)

For comparative purposes, a comprehensive set of 1,131 standard radiomics features were also extracted from the tumor region on T1CE images using the PyRadiomics library. These included first-order statistics, shape features, and texture features (e.g., Gray Level Co-occurrence Matrix [GLCM], Gray Level Run Length Matrix [GLRLM], Gray Level Size Zone Matrix [GLSZM]). These features were not used for the final XGBoost model training but were leveraged for a correlation analysis against the DL features[28-31].

2.5. Feature Selection

With a high-dimensional initial feature set (2 clinical + 23 hand-crafted + 256 DL = 281 features), a multi-step feature selection process was implemented to reduce dimensionality, mitigate the risk of overfitting, and select the most informative predictors for the XGBoost model.

1. **Clinical Features**: Both age and gender were retained *a priori* due to their established clinical relevance.
2. **Hand-crafted Features**: A univariate analysis was performed by calculating the point-biserial correlation coefficient between each of the 23 features and the binary Ki-67 label. Features with a p-value less than 0.1 were selected, resulting in a reduced set of **14** hand-crafted features.

3. **Deep Learning Features**: A two-stage selection was applied. First, to remove redundant or non-informative features, a variance threshold was used to select the top 100 DL features with the highest variance. Second, from this subset, we calculated the absolute correlation of each feature with the Ki-67 label and selected the top **30** features.

This process yielded a final, consolidated feature set of **46 dimensions** (2 clinical + 14 hand-crafted + 30 DL) to be used as input for the final classification model. A separate selection was performed on the radiomics features, where the top 10 features with the highest absolute correlation to the Ki-67 label were identified solely for the comparative heatmap analysis (Figure 8).

2.6. Model Development and Evaluation

2.6.1. XGBoost Classifier

An XGBoost (Extreme Gradient Boosting) classifier was chosen for the classification task due to its high performance, efficiency, and inherent regularization capabilities, which help prevent overfitting. The model was configured with a set of optimized hyperparameters, including `n_estimators=100`, `max_depth=4`, and `learning_rate=0.1`. The objective was set to `binary:logistic` for binary classification.

2.6.2. Cross-Validation and Performance Metrics

To ensure a robust and generalizable evaluation of the model, we employed a 5-fold stratified cross-validation strategy. The dataset was partitioned into five folds, ensuring that the proportion of high- and low-expression Ki-67 cases was consistent across all folds. In each iteration, the model was trained on four folds and validated on the held-out fold. The predictions from the held-out fold in each iteration were collected to form a complete set of out-of-fold (OOF) predictions for the entire dataset.

The model's performance was assessed using several standard metrics:

- **Area Under the Receiver Operating Characteristic Curve (AUC)**: To evaluate the model's overall discriminative ability.

- **Accuracy**: The proportion of correctly classified instances.
- **Sensitivity (Recall)**: The ability to correctly identify high-expression cases.
- **Specificity**: The ability to correctly identify low-expression cases.
- **F1 Score**: The harmonic mean of precision and sensitivity.
- **Matthews Correlation Coefficient (MCC)**: A balanced measure of classification quality.

Confidence intervals (95% CI) for the performance metrics were calculated using the bootstrap method with 1000 resamples on the OOF predictions.

- The performance was evaluated using the following equations:

$$Accuracy=\frac{(TP+TN)}{(TP+TN+FP+FN)}\#(1)$$

$$\text{Sensitivity(Recall)}=\frac{TP}{(TP+FN)}\#(2)$$

$$\text{Specificity}=\frac{TN}{(TN+FP)}\#(3)$$

$$F1Score=2\times\frac{(\text{P r e c i s i o n}\times\text{R e c a l l})}{(\text{P r e c i s i o n}+\text{R e c a l l})}\#(4)$$

$$MCC=\frac{(TP\times TN-FP\times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}\#(5)$$

Where:

- True Positives (TP) are the high-expression cases correctly predicted as high.
- True Negatives (TN) are the low-expression cases correctly predicted as low.
- False Positives (FP) are the low-expression cases incorrectly predicted as high.
- False Negatives (FN) are the high-expression cases incorrectly predicted as low.

2.7. Model Interpretability with SHAP

To understand the inner workings of the "black-box" XGBoost model and to identify which features were most influential in predicting Ki-67 status, we utilized the SHapley

Additive exPlanations (SHAP) technique. SHAP is a game theory-based approach that assigns an importance value (SHAP value) to each feature for each individual prediction, representing its contribution to pushing the model's output from the base value to the final prediction.

We generated several SHAP visualizations to interpret the model both globally and locally:

- **SHAP Beeswarm Plots**: To summarize the importance and impact of the top features across the entire dataset. These plots show the distribution of SHAP values for each feature and how high or low feature values correlate with the prediction.
- **SHAP Waterfall Plots**: To explain individual predictions for representative cases, illustrating how different features collaboratively contribute to the final classification for a specific patient.

This analysis was performed separately for the hand-crafted features (including clinical) and the deep learning features to assess the relative importance of each feature category.

3. Results

3.1. Patient Characteristics

The baseline characteristics of the 573 patients included in the study are detailed in Table 1. The cohort was stratified by the five tumor types. Significant differences in Ki-67 expression, age, and gender distribution were observed across the tumor types. Metastatic tumors and gliomas exhibited the highest proportions of Ki-67 high-expression cases (88.0% and 68.8%, respectively), consistent with their aggressive nature. Conversely, sellar lesions and acoustic neuromas, which are typically benign, had the lowest rates of high Ki-67 expression (6.1% and 9.6%, respectively).

Overall, 317/573 patients (55.3%) were labeled as high Ki-67 and 256/573 (44.7%) as low. Tumor-type counts were: Glioma (n=369), Meningioma (n=53), Acoustic Neuroma (n=52), Metastatic tumor (n=50), and Sellar lesion (n=49).

3.2. Model Performance

The XGBoost model, trained on the 46 selected features, demonstrated good predictive performance in distinguishing between high and low Ki-67 expression levels in the out-of-fold (OOF) validation set. The pooled out-of-fold AUC was 0.804 (95% CI: 0.769-0.838), indicating good discriminative ability. The ROC curves for each fold and the mean ROC curve are shown in Figure 2.

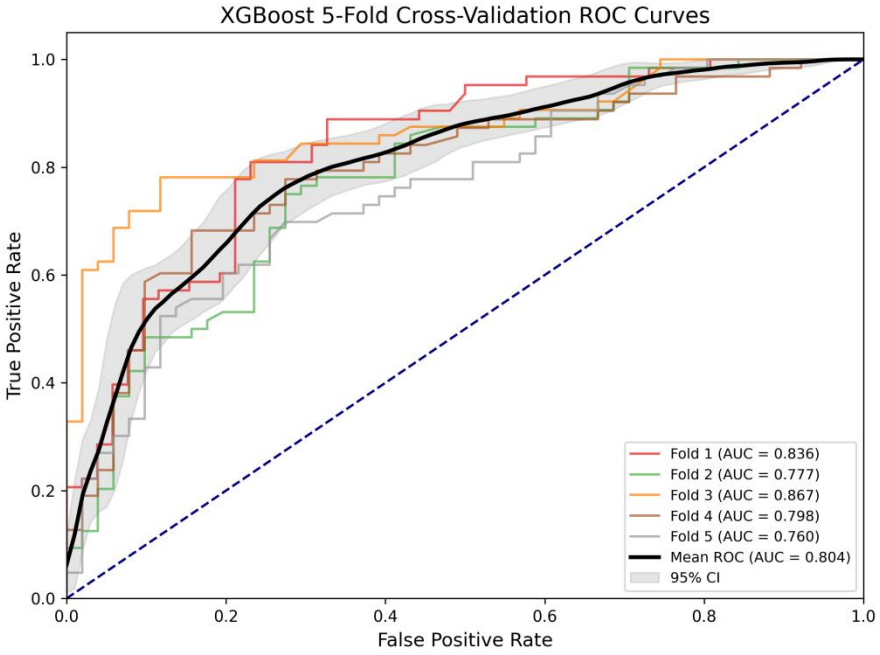

Figure 2: Receiver Operating Characteristic (ROC) Curves. The plot shows the ROC curves for each of the 5 cross-validation folds, along with the mean ROC curve (solid blue line) and its 95% confidence interval (shaded area). The pooled out-of-fold AUC of 0.804 demonstrates the model's overall performance.

The overall classification performance metrics are summarized in Table 2. The model achieved an accuracy of 0.743, a sensitivity of 0.785, and a specificity of 0.691. The F1 score was 0.772 and the MCC was 0.479, reflecting a well-balanced classification performance.

Table 2: Overall Diagnostic Performance of the XGBoost Model (OOF Validation).

Dataset	AUC (95% CI)	Accuracy	Sensitivity	Specificity	F1 Score	MCC
Validation (OOF)	0.804 [0.769-0.838]	0.743	0.785	0.691	0.772	0.479

The stability of the model was supported by consistent fold-wise ROC behavior and a relatively narrow 95% confidence interval around the mean ROC curve (Figure 2).

The confusion matrix for the OOF predictions (Figure 3) further details the classification results, showing that the model correctly identified 249 of the 317 high-expression cases and 177 of the 256 low-expression cases. The model's calibration, assessed via a calibration curve (Figure 4), indicated acceptable agreement between the predicted probabilities and the observed frequencies.

Numerically, the OOF predictions consisted of 249 true positives, 177 true negatives, 79 false positives, and 68 false negatives (n=573). The reliability curve closely follows the diagonal over most probability bins, with only mild underestimation at the highest-probability bin (Figure 4).

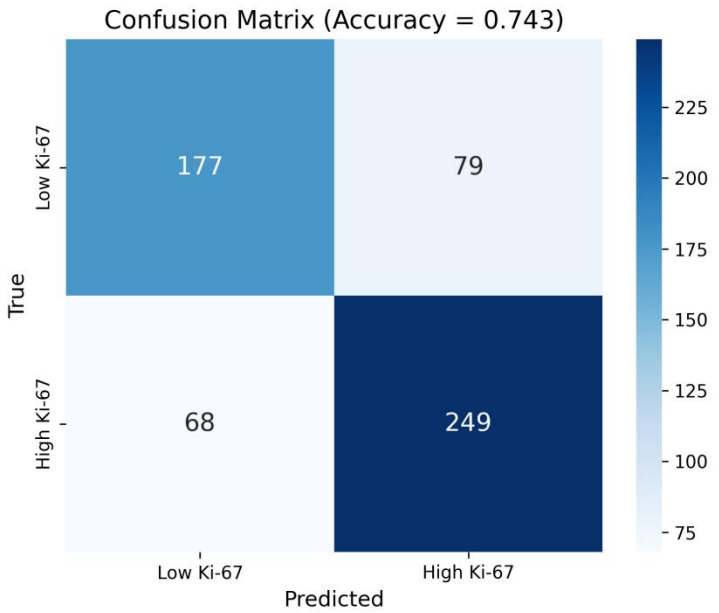

Figure 3: Overall Confusion Matrix. The matrix visualizes the performance on the out-of-fold predictions for all 573 patients, detailing true positives, true negatives, false positives, and false negatives.

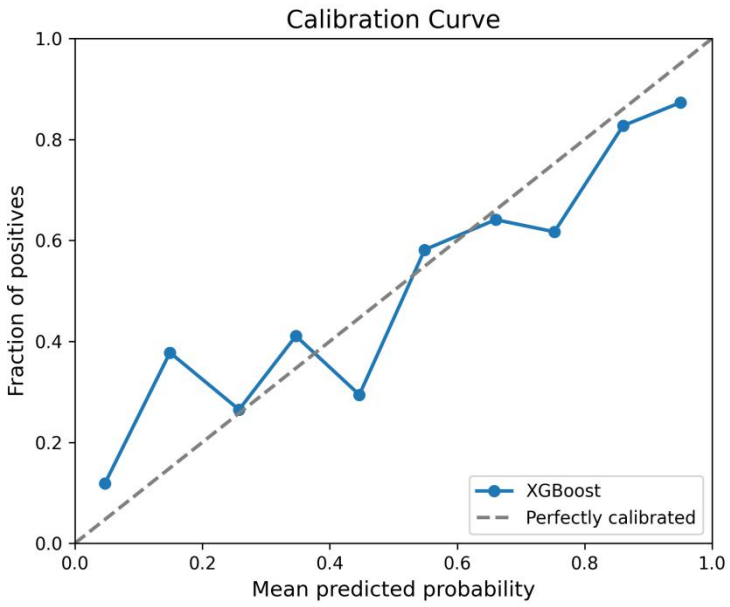

Figure 4: Calibration Curve. The plot shows the relationship between the predicted probabilities from the model and the actual observed proportion of high Ki-67 cases, demonstrating acceptable model calibration.

3.2.1. Performance by Tumor Type

To assess the model's utility across different pathologies, we analyzed its performance stratified by each of the five tumor types. The confusion matrices for each subgroup are presented in **Figure 5 and Table 3**. The model showed the highest sensitivity for gliomas and metastatic tumors, consistent with their high prevalence of elevated Ki-67, while also identifying the low-proliferative status of sellar lesions and acoustic neuromas.

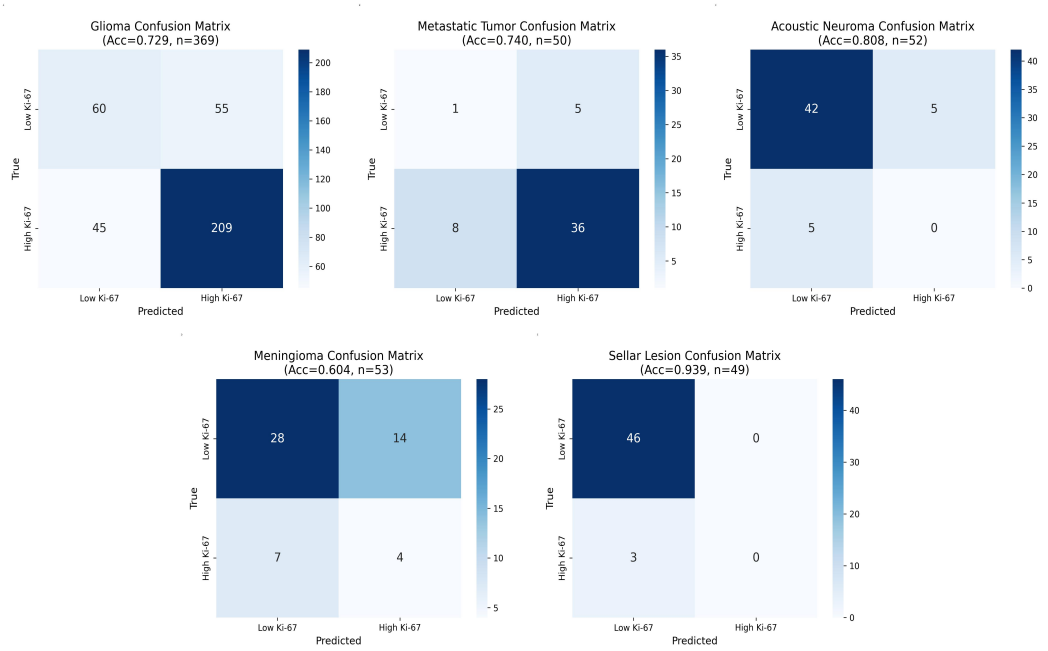

Figure 5: Confusion Matrices by Tumor Type. These matrices show the classification performance for each of the five tumor subtypes: Glioma (A), Metastatic Tumor (B), Acoustic Neuroma (C), Meningioma (D), and Sellar Lesion (E).

Table 3: Performance by Tumor Type (OOF).

Tumor type	n	Accuracy	Sensitivity	Specificity
Glioma	369	0.729	0.823 (209/254)	0.522 (60/115)
Meningioma	53	0.604	0.364 (4/11)	0.667 (28/42)
Acoustic neuroma	52	0.808	0.000 (0/5)	0.894 (42/47)
Metastatic tumor	50	0.740	0.818 (36/44)	0.167 (1/6)
Sellar lesion	49	0.939	0.000 (0/3)	1.000 (46/46)

These results indicate a high sensitivity for detecting elevated Ki-67 in glioma and metastatic tumors, while predictions for benign-predominant entities (sellar lesions, acoustic neuroma) favor specificity, reflecting their very low prevalence of high Ki-67 in our cohort.

3.3. Model Interpretability using SHAP

SHAP analysis was conducted to provide insight into the model's decision-making process and to identify the most influential features.

3.3.1. Global Feature Importance

The SHAP beeswarm plots (**Figures 6 and 7**) illustrate the global importance of the top hand-crafted and deep learning features, respectively. Each point on the plot represents a single patient's prediction, with the color indicating the feature's value (red for high, blue for low) and its position on the x-axis showing its impact on the prediction (positive SHAP values push the prediction towards high Ki-67).

For the hand-crafted features (Figure 6), skullbase_distance_mm, age, rim_core_entropy_diff, and enhancement-related descriptors emerged as highly influential predictors. Clinical features such as age showed strong contribution, and location-/ring-related characteristics consistently affected the probability of Ki-67 >=10%.

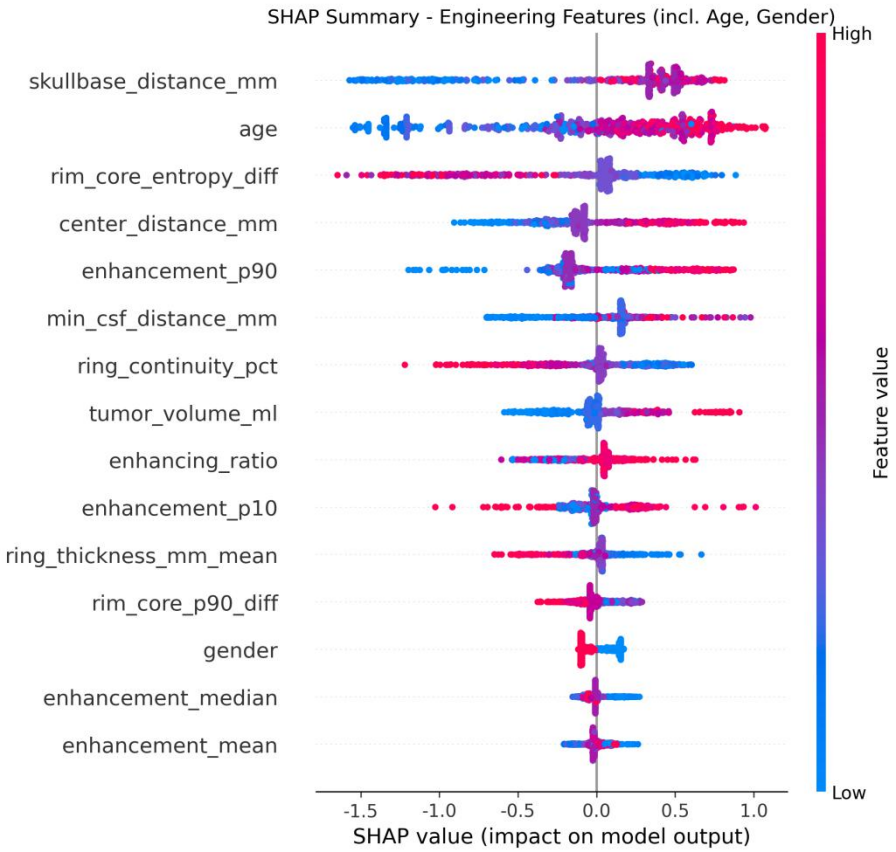

Figure 6: SHAP Beeswarm Plot for Hand-crafted and Clinical Features. The plot ranks the most important features in this category. Features are ordered by their mean absolute SHAP value.

Among the deep learning features (Figure 7), several abstract features learned by the model demonstrated significant predictive power. For instance, dl_feat_78, dl_feat_41, and dl_feat_92 showed notable contributions, capturing complex textural and morphological patterns beyond the scope of the hand-crafted features.

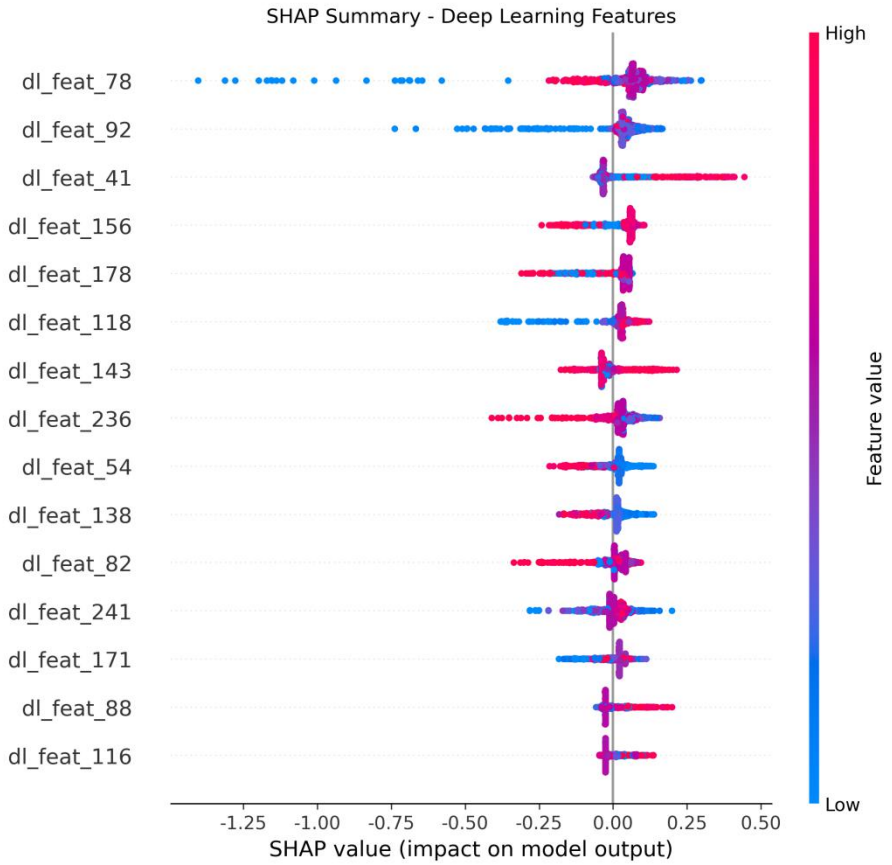

Figure 7: SHAP Beeswarm Plot for Deep Learning Features. The plot ranks the most important DL features, revealing the complex image patterns captured by the model that are critical for prediction.

Top global features by mean absolute SHAP (selected): skullbase_distance_mm (0.529), age (0.521), rim_core_entropy_diff (0.418), enhancement_p90 (0.258), center_distance_mm (0.248), ring_continuity_pct (0.224), min_csf_distance_mm (0.202), tumor_volume_ml (0.159), enhancing_ratio (0.128), enhancement_p10 (0.115). Values in parentheses denote mean |SHAP| from the validation set.

3.3.2. Correlation between DL and Radiomics Features

To better understand the nature of the abstract DL features, we performed a correlation analysis between the top 10 most predictive DL features and the top 10 most predictive standard radiomics features. The resulting heatmap (**Figure 8**) reveals moderate to strong correlations between certain DL features and established radiomic concepts like tumor intensity, texture, and shape, suggesting that the DL model independently learned to identify patterns analogous to well-known radiomic markers.

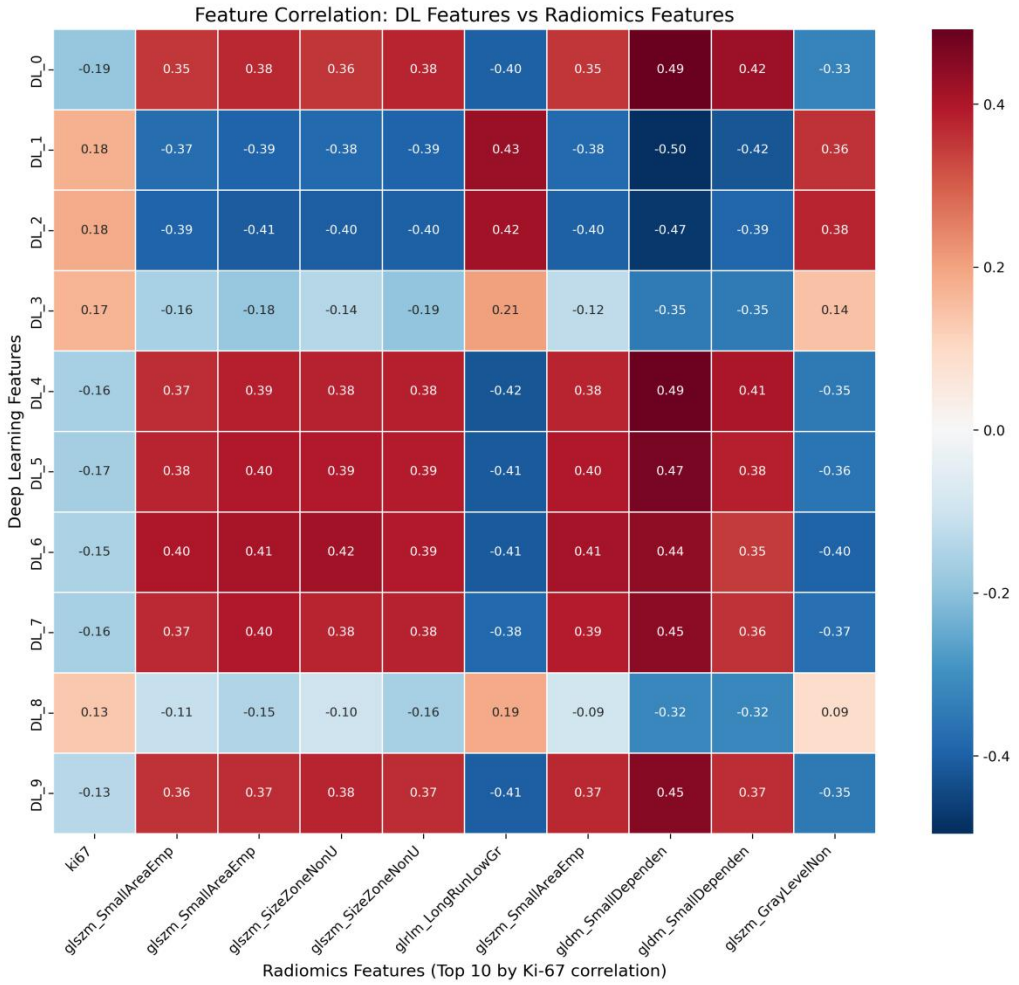

Figure 8: Feature Correlation Heatmap. This heatmap shows the Pearson correlation between the top 10 DL features and top 10 radiomics features (ranked by their correlation with the Ki-67 label).

3.3.3. Local (Individual) Prediction Explanation

SHAP waterfall plots provide a transparent view of how the model arrives at a decision for an individual patient. Figure 9A shows a case correctly predicted as having high Ki-67 expression. The plot details how features such as rim_core_entropy_diff,

skullbase_distance_mm, and ring_continuity_pct cumulatively increase the SHAP value, pushing the prediction above the threshold. Conversely, Figure 9B illustrates a case correctly predicted as low-expression, where strong negative contributions from ring_continuity_pct, enhancing_ratio, center_distance_mm, and age drive the prediction toward low Ki-67.

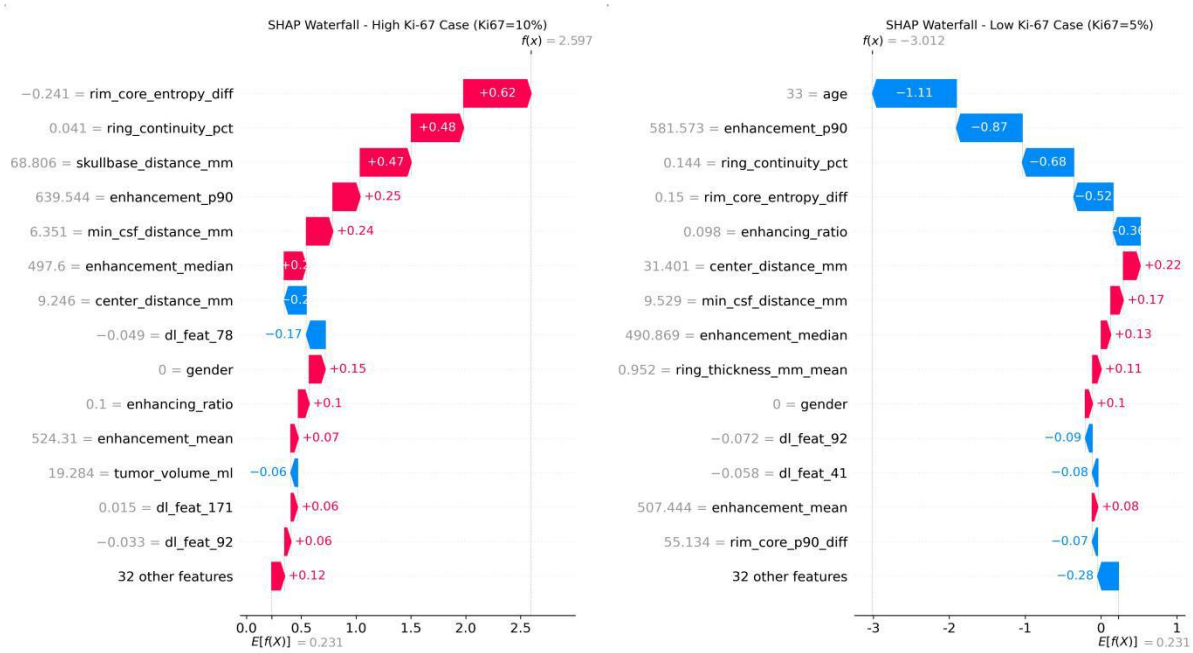

Figure 9A: SHAP Waterfall Plot for a High Ki-67 Prediction. This plot explains the prediction for a single patient, showing how each feature contributes to the final output.

Figure 9B: SHAP Waterfall Plot for a Low Ki-67 Prediction. This plot provides a local explanation for a patient correctly classified as having low Ki-67 expression.

In a representative high-expression case (Figure 9A), positive contributions were dominated by rim_core_entropy_diff (+0.90), skullbase_distance_mm (+0.73), ring_continuity_pct (+0.57), and min_csf_distance_mm (+0.35), partially offset by negatives from center_distance_mm (−0.18) and enhancement_p90 (−0.14). In a representative low-expression case (Figure 9B), large negative pushes came from ring_continuity_pct (−0.99), enhancing_ratio (−0.39), center_distance_mm (−0.37), and age (−0.25), with only small positive offsets from several DL features.

3.4. Error Analysis

Across all OOF predictions, there were 147 misclassifications (79 false positives, 68 false negatives). By tumor type, misclassifications were most frequent in glioma and meningioma (reflecting larger sample sizes and class imbalance), with far fewer in metastasis and sellar lesions. In acoustic neuroma and sellar lesions the model prioritized specificity over sensitivity (no true positives in our cohort), which mirrors the rarity of high Ki-67 in these entities. Threshold tuning by subtype or cost-sensitive training may improve sensitivity for these rare high-expression cases without materially affecting AUC.

4. Discussion

In this study, we developed and validated a unified, explainable machine learning framework for the non-invasive prediction of Ki-67 proliferation status across five common types of intracranial tumors. By integrating multimodal features-clinical, hand-crafted radiomic, and deep learning descriptors-extracted solely from routine preoperative T1CE MRI, our XGBoost model achieved an AUC of 0.804. This approach demonstrates the feasibility of obtaining a key molecular prognostic marker non-invasively, which could inform preoperative planning, risk stratification, and patient counseling.

A primary contribution of this work is the development of a single model applicable to a heterogeneous spectrum of intracranial tumors. Most prior radiomics or deep learning studies have focused on predicting Ki-67 within a specific tumor type, particularly glioma[3, 32]. While such models achieve notable accuracy, their clinical utility is limited when the pathological diagnosis remains uncertain preoperatively. Our model, trained on a mixed cohort of gliomas, meningiomas, acoustic neuromas, metastatic tumors, and sellar lesions, more closely mirrors the real-world diagnostic workflow. Its maintained performance across entities with vastly different biological behaviors—from aggressive metastases to typically benign sellar lesions—suggests its potential for clinical application as an imaging-based triage tool.

The multimodal feature fusion strategy contributed to the model's overall performance. SHAP analysis revealed that predictions were driven by a synergistic combination of human-interpretable, hypothesis-driven features and abstract, data-driven deep learning features. Macroscopic and interpretable characteristics such as

skullbase_distance_mm, rim_core_entropy_diff, enhancement_p90, ring_continuity_pct, and tumor_volume_ml were strongly associated with Ki-67 risk stratification, aligning with established radiological signs of aggressiveness[26]. Concurrently, the high-ranking deep learning features captured subtler, hierarchical patterns of texture and morphology beyond manual quantification. The moderate-to-strong correlation between top DL features and conventional radiomics features (Figure 8) suggests the DL network independently learned representations analogous to known radiomic concepts, yet potentially more comprehensive. This suggests that hand-crafted and DL features encode complementary biological and phenotypic information, and their integration provides a more complete characterization of tumor proliferation.

The implementation of explainable AI (XAI) through SHAP fundamentally enhances the translational potential of our model. Moving beyond a "black-box" prediction fosters clinical trust and facilitates integration into decision-making. Globally, SHAP plots validate that the model learns clinically intuitive relationships, such as the association between deep location (e.g., small skullbase_distance_mm), disrupted enhancement patterns (e.g., high rim_core_entropy_diff, low ring_continuity_pct), and high proliferation risk. Locally, SHAP waterfall plots (Figure 9) offer a powerful mechanism for clinicians to interrogate individual predictions. Understanding why a specific tumor is predicted as high-risk-whether due to location-related, enhancement-related, or morphology-related imaging patterns-can provide actionable imaging-derived evidence to support or question a treatment plan, paving the way for personalized imaging biomarker reports.

Our stratified analysis revealed nuanced performance across tumor types. The model showed high sensitivity in gliomas and metastases, where accurate identification of high proliferation is clinically critical. For typically benign entities like acoustic neuromas and sellar lesions, it prioritized high specificity, correctly classifying the vast majority of low-Ki-67 cases. This behavior appropriately reflects the prior probability in our cohort. The lower sensitivity for high-Ki-67 cases in these rare subgroups suggests that future iterations could benefit from tumor-type-specific probability threshold tuning or cost-

sensitive learning to better identify the occasional aggressive variant without compromising overall accuracy.

Despite these promising results, our study has limitations. First, its retrospective design and merged multi-source cohort may introduce selection bias and heterogeneity. Although data expansion improved diversity, further prospective validation on broader multi-institutional datasets with varying MRI protocols is still essential to confirm robustness. Second, while we leveraged multiple sequences for segmentation and qualitative assessment, quantitative feature extraction was restricted to T1CE to ensure consistency. Incorporating advanced sequences like diffusion-weighted imaging (DWI; for cellularity) or dynamic susceptibility contrast (DSC; for perfusion) could capture additional biological dimensions and potentially improve accuracy. Third, the binary threshold of Ki-67 >=10%, while clinically common, is a simplification. Proliferation exists on a continuum, and its prognostic significance can be tumor-subtype-dependent. Future work could explore regression models to predict continuous Ki-67 values or multi-class stratification. Finally, the clinical utility of this model must be prospectively evaluated to determine its impact on actual surgical planning, adjuvant therapy decisions, and patient outcomes[33, 34].

In conclusion, this study presents a transparent, T1CE-based XGBoost model capable of non-invasively predicting Ki-67 status across multiple intracranial tumor types. By integrating clinical, radiomic, and deep-learning features, the model achieves moderate overall performance (AUC 0.804), particularly for gliomas, and provides interpretable imaging correlates of tumor proliferation, offering a potential preoperative tool to aid clinical decision-making. Further validation and refinement of subtype-specific thresholds are warranted for clinical translation.

Funding

This work was supported by the General Program of the National Natural Science Foundation of China (Grant No. 12375328), Joint Fund for Synchrotron Radiation of University of Science and Technology of China (USTC-NSRL2025-KY2310000900), Research Project of Leading Medical Institute and Frontier Technology Institute (2025IHM02018).

Author Contributions All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by Yang Yang, Yang Lv, Zhongying Li, Dasheng Wang, Xianchao Hu, Longfei Hu, Yong Guan, Fei Wang, Zheng Jiang and Yong-fei Dong. The first draft of the manuscript was written by Yang Yang, Yong-fei Dong, Zheng Jiang and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

[1] Lu Y, Huang Y, Zhu C, et al. Cancer brain metastasis: molecular mechanisms and therapeutic strategies. Mol. Biomed. 6, 12 (2025). https://doi.org/10.1186/s43556-025-00251-0

[2] Chu T P C, Shah A, Walker D, et al. How Do Biological Characteristics of Primary Intracranial Tumors Affect Their Clinical Presentation in Children and Young Adults? J. Child. Neurol. 33, 503-511(2018). https://doi.org/10.1177/0883073818767562

[3] Bhuiyan E H, Khan M M, Hossain S A, et al. Classification of glioma grade and Ki-67 level prediction in MRI data: A SHAP-driven interpretation. Comput. Med. Imaging. Graph. 124, 102578 (2025). https://doi.org/10.1016/j.compmedimag.2025.102578

[4] Meng Y, Bernstein K, Mashiach E, et al. Outcomes of Radiosurgery for WHO Grade 2 Meningiomas: The Role of Ki-67 Index in Guiding the Tumor Margin Dose. Neurosurgery, 10-1227 (2024). https://doi.org/10.1227/neu.0000000000003255

[5] Mizrachi M, Hartley B, Saleem S, et al. Ki-67 index as a predictive marker of meningioma recurrence following surgical resection. J. Clin. Neurosci. 124, 15-19(2024). https://doi.org/10.1016/j.jocn.2024.04.015

[6] Salle H, Durand S, Duchesne M, et al. Influence of clinical and histological criteria on meningioma recurrence: The decisive role of Ki-67. Clin. Neuropathol. 44, 180-192 (2025). https://doi.org/10.5414/NP301681

[7] Gürsoy G. CRP/albumin ratio and WBC values correlate with Ki-67 and survival in glioblastoma multiforme. Front. Oncol. 15, 1612212(2025). https://doi.org/10.3389/fonc.2025.1612212

[8] Narendra R N, Vijayakumar C, Haritha G, et al. Ki-67 Levels and Their Association With Response to Neoadjuvant Chemotherapy in Triple-Negative Breast Cancer: A Prospective Observational Study. Cureus. 17, e83207 (2025). https://doi.org/10.7759/cureus.83207

[9] Lin X, Zhu S, Wang D, et al. Correlation of dynamic contrast-enhanced ultrasonography and the Ki-67 labelling index in pancreatic ductal adenocarcinoma. World. J. Gastroenterol. 30, 4697-4708 (2024). https://doi.org/10.3748/wjg.v30.i44.4697

[10] Fares J, Wan Y, Li Y, et al. Magnetic Resonance Imaging Physics in Brain Tumor Imaging: A Primer for Neurosurgeons. World. Neurosurg. 204, 124591 (2025). https://doi.org/10.1016/j.wneu.2025.124591

[11] Spaanderman D J, Hakkesteegt S N, Hanff D F, et al. Multi-center external validation of an automated method segmenting and differentiating atypical lipomatous tumors from lipomas using radiomics and deep-learning on MRI. EClinicalMedicine, 76, 102802 (2024). https://doi.org/10.1016/j.eclinm.2024.102802

[12] Zhu Y, Wang J, Xue C, et al. Deep Learning and Habitat Radiomics for the Prediction of Glioma Pathology Using Multiparametric MRI: A Multicenter Study. Acad. Radiol. 32, 963-975 (2025). https://doi.org/10.1016/j.acra.2024.09.021

[13] Iannella G, de Vincentiis M, Di Gioia C, et al. Subtotal resection of vestibular schwannoma: Evaluation with Ki-67 measurement, magnetic resonance imaging, and long-term observation. J. Int. Med. Res. 45, 1061-1073 (2017). https://doi.org/10.1177/0300060516686873

[14] La Rosa S. Diagnostic, Prognostic, and Predictive Role of Ki67 Proliferative Index in Neuroendocrine and Endocrine Neoplasms: Past, Present, and Future. Endocr. Pathol. 34, 79-97 (2023). https://doi.org/10.1007/s12022-023-09755-3

[15] Li H, Liu Z, Li F, et al. Preoperatively Predicting Ki67 Expression in Pituitary Adenomas Using Deep Segmentation Network and Radiomics Analysis Based on Multiparameter MRI. Acad. Radiol. 31, 617-627 (2024). https://doi.org/10.1016/j.acra.2023.05.023

[16] Prueter J, Norvell D, Backous D. Ki-67 index as a predictor of vestibular schwannoma regrowth or recurrence. J. Laryngol. Otol. 133, 205-207 (2019). https://doi.org/10.1017/S0022215119000549

[17] Akgündoğdu A, Çelikbaş Ş. Explainable deep learning framework for brain tumor detection: Integrating LIME, Grad-CAM, and SHAP for enhanced accuracy. Med. Eng. Phys. 144, 104405 (2025). https://doi.org/10.1016/j.medengphy.2025.104405

[18] Rahman A, Hayat M, Iqbal N, et al. Enhanced MRI brain tumor detection using deep learning in conjunction with explainable AI SHAP based diverse and multi feature analysis. Sci. Rep. 15, 29411 (2025). https://doi.org/10.1038/s41598-025-14901-4

[19] Yan X, Duan F, Chen L, et al. A Multimodal MRI-Based Model for Colorectal Liver Metastasis Prediction: Integrating Radiomics, Deep Learning, and Clinical Features with SHAP Interpretation. Curr. Oncol. 32, 431 (2025). https://doi.org/10.3390/curroncol32080431

[20] Farzipour A, Elmi R, Nasiri H. Detection of Monkeypox Cases Based on Symptoms Using XGBoost and Shapley Additive Explanations Methods. Diagnostics (Basel). 13, 2391 (2023). https://doi.org/10.3390/diagnostics13142391

[21] Wang R, Liu Q, You W, et al. A transformer-based deep learning survival prediction model and an explainable XGBoost anti-PD-1/PD-L1 outcome prediction model based on the cGAS-STING-centered pathways in hepatocellular carcinoma. Brief. Bioinform. 26, bbae686 (2024). https://doi.org/10.1093/bib/bbae686

[22] Xu Q, Lu X. Development and validation of an XGBoost model to predict 5-year survival in elderly patients with intrahepatic cholangiocarcinoma after surgery: a SEER-based study. J. Gastrointest. Oncol. 13, 3290-3299 (2022). https://doi.org/10.21037/jgo-22-1238

[23] Yang Y, Wu J, Li J, et al. Transformer-based multimodal fusion framework for predicting postoperative cognitive improvement in glioma: integrating radiomics and pathomics. Int. J. Surg. 10-1097 (2025). https://doi.org/10.1097/JS9.0000000000004453

[24] Isensee F, Jaeger P F, Kohl S A A, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods. 18, 203-211 (2021). https://doi.org/10.1038/s41592-020-01008-z

[25] Chen C, Zhao Y, Cai L, et al. A multi-modal deep learning model for prediction of Ki-67 for meningiomas using pretreatment MR images. NPJ. Precis. Oncol. 9, 21 (2025). https://doi.org/10.1038/s41698-025-00811-1

[26] Kibriya H, Amin R, Kim J, et al. A Novel Approach for Brain Tumor Classification Using an Ensemble of Deep and Hand-Crafted Features. Sensors (Basel). 23, 4693 (2023). https://doi.org/10.3390/s23104693

[27] Zunair H, Ben Hamza A. Sharp U-Net: Depthwise convolutional network for biomedical image segmentation. Comput. Biol. Med. 136, 104699 (2021). https://doi.org/10.1016/j.compbiomed.2021.104699

[28] Salmanpour M R, Piri S M, Mehrnia S S, et al. Pathobiological Dictionary Defining Pathomics and Texture Features: Addressing Understandable AI Issues in Personalized Liver Cancer; Dictionary Version LCP1.0. J. Imaging. Inform. Med. 10-1007(2026). https://doi.org/10.1007/s10278-025-01817-8

[29] Shin H, Sheen H, Oh J, et al. Evaluating feature extraction reproducibility across image biomarker standardization initiative-compliant radiomics platforms using a digital phantom. J. Appl. Clin. Med. Phys. 26, e70110 (2025). https://doi.org/10.1002/acm2.70110

[30] Yun Y C, Jende J M E, Garhöfer F, et al. Combined peritumoral radiomics and clinical features predict 12-month progression free survival in glioblastoma. J. Neurooncol. 174, 111-120 (2025). https://doi.org/10.1007/s11060-025-05037-6

[31] Zhong S, Ren J, Yu Z, et al. Predicting glioblastoma molecular subtypes and prognosis with a multimodal model integrating convolutional neural network,

radiomics, and semantics. J. Neurosurg. 139, 305-314 (2022). https://doi.org/10.3171/2022.10.JNS22801

[32] Ni J, Zhang H, Yang Q, et al. Machine-Learning and Radiomics-Based Preoperative Prediction of Ki-67 Expression in Glioma Using MRI Data. Acad. Radiol. 31, 3397-3405 (2024). https://doi.org/10.1016/j.acra.2024.02.009

[33] Freitas N R, Vieira P M, Cordeiro A, et al. Detection of bladder cancer with feature fusion, transfer learning and CapsNets. Artif. Intell. Med. 126, 102275 (2022). https://doi.org/10.1016/j.artmed.2022.102275

[34] Huang Y, Yao Z, Li L, et al. Deep learning radiopathomics based on preoperative US images and biopsy whole slide images can distinguish between luminal and non-luminal tumors in early-stage breast cancers. EBioMedicine. 94, 104706 (2023). https://doi.org/10.1016/j.ebiom.2023.104706