Final Year Project

Radiomics-Based Prediction of IDH Mutation in Glioma Using Multi-Modal MRI and Ensemble Machine Learning

Group: F

Student: Jialin Xu 1155157053

Supervisor: Prof. Viet Anh NGUYEN

Department: Systems Engineering and Engineering Management

Submission Date: April 2025

Abstract

Determining isocitrate dehydrogenase (IDH) mutation status in gliomas holds substantial prognostic and therapeutic significance. Currently, IDH status relies on invasive tissue sampling; thus, developing a reliable, non-invasive imaging-based prediction is of significant clinical value. This study established an integrated radiomics approach utilizing multi-modal MRI—specifically T2-weighted images, apparent diffusion coefficient (ADC) maps, and contrast-enhanced T1-weighted (T1c) images—to predict IDH genotype in glioma patients. Leveraging PyRadiomics, we extracted comprehensive quantitative imaging features characterizing tumor intensity, texture, and morphology from both tumor core and peritumoral edema regions. To manage feature dimensionality, a Lasso-based selection technique was applied, yielding sparse, clinically interpretable feature subsets (e.g., selecting 56 informative T2 features from over a thousand candidates). Subsequently, diverse machine learning models—including logistic regression, support vector machines, random forest, XGBoost, and LightGBM—were trained and systematically evaluated through cross-validation. Results revealed robust predictive performance from individual modalities (ADC: AUC=0.90, accuracy=85.2%; T1c: AUC=0.882, accuracy=82.9%; T2: AUC=0.91, accuracy=83.8%). Crucially, combining multi-modal predictions using ensemble methods further elevated model performance (probability averaging ensemble: AUC=0.931, accuracy=87.1%; stacking ensemble: AUC=0.932, accuracy=86.6%), surpassing any single-modality approach. Our findings confirm that multi-modal MRI radiomics effectively encodes critical molecular characteristics of gliomas and highlight the ensemble methodology's potential in achieving clinically valuable diagnostic accuracy. Future work involves validating these approaches in external cohorts, exploring additional imaging modalities, integrating advanced deep learning features, and enhancing model interpretability for clinical translation.

Contents

Problem Statement

Gliomas are among the most common primary tumors of the central nervous system and are classified into subtypes based on histology and molecular markers. One such crucial molecular marker is mutation in the isocitrate dehydrogenase (IDH) gene. Determining whether a glioma is IDH-mutant or IDH-wildtype is clinically significant because it influences prognosis and treatment strategies.

Currently, identification of IDH mutation status requires tissue sampling through biopsy or surgery, followed by genetic testing in a pathology lab. This process is invasive, carries surgical risks, and can be time-consuming. In contrast, if we could reliably predict IDH mutation status noninvasively from preoperative MRI scans, it would greatly benefit clinical decision-making. A noninvasive imaging-based prediction could help neurosurgeons and oncologists plan treatments (such as the extent of resection or the use of IDH inhibitor therapies) and provide prognostic information to patients at an earlier stage.

The problem statement for this thesis is: *Can we develop a radiomics and machine learning-based approach to accurately predict IDH mutation status in gliomas using multi-modal MRI scans?* This entails identifying distinctive imaging features on MRI that correlate with the presence of an IDH mutation and building a predictive model. Key challenges include: (1) extracting meaningful quantitative features from MRI that capture subtle differences between IDH-mutant and IDH-wildtype tumors, (2) dealing with the high dimensionality of radiomic features relative to the number of patients (the curse of dimensionality), and (3) constructing a robust classifier that generalizes well. The solution should demonstrate not only high accuracy but also insight into which features or MRI modalities are most informative for the task. The clinical significance of solving this problem is high, as it would pave the way for noninvasive "virtual biopsy" techniques in neuro-oncology, reducing the need for invasive procedures and enabling personalized treatment strategies based on imaging data alone.

Background and Literature Review

Molecular Classification of Gliomas and the Role of IDH Mutation

Diffuse gliomas are heterogeneous brain tumors that, in the past, were classified primarily by histopathological features (e.g., astrocytoma, oligodendroglioma, glioblastoma). The discovery of recurrent mutations in the *IDH1* gene in gliomas in 2008–2009 revolutionized the field by revealing a molecular marker with major clinical significance (Parsons, 2008; Yan, 2009). Subsequent studies confirmed that an IDH mutation is an independent prognostic factor associated with substantially longer survival in both low-grade and high-grade gliomas (Yan, 2009). This led the World Health Organization (WHO) to incorporate IDH mutation status into the official classification of central nervous system tumors. In the 2016 WHO classification update, diffuse gliomas were divided into molecular subsets based on IDH-mutation and 1p/19q codeletion status, recognizing that IDH-mutant gliomas constitute a distinct, less aggressive disease entity (Louis, 2016). The most recent 2021 WHO classification further solidified this approach, essentially defining adult-type diffuse gliomas by their IDH status: tumors are designated as either “IDH-mutant” (which includes most lower-grade gliomas and a subset of glioblastomas now termed grade 4 IDH-mutant astrocytomas) or “IDH-wildtype” (Louis, 2016). IDH-wildtype diffuse astrocytic tumors in adults are typically aggressive and correspond to the traditional primary glioblastomas, whereas IDH-mutant tumors tend to be lower grade or secondary glioblastomas and carry a better prognosis.

Because IDH status has such prognostic and diagnostic importance, it is now standard to test for IDH1/IDH2 mutations in any diffuse glioma patient (Louis, 2021). IDH-mutant gliomas generally respond better to therapy and have median survivals measured in years, whereas IDH-wildtype glioblastomas often have survival under 2 years even with aggressive treatment (Yan, 2009). Determining IDH mutation status is also essential for enrollment in certain clinical trials and for considering IDH-targeted therapies that are emerging. The current gold standard for IDH determination is genetic testing of tumor tissue obtained via biopsy or surgery. However, performing an invasive procedure solely for molecular diagnosis carries risks. Therefore, there is a strong motivation to develop imaging-based methods to non-invasively infer IDH status preoperatively (Di Salle, 2024). An MRI-based surrogate for IDH mutation could aid in surgical planning and treatment decisions, especially in cases where surgery is risky or when deciding on the aggressiveness of resection and adjuvant therapy.

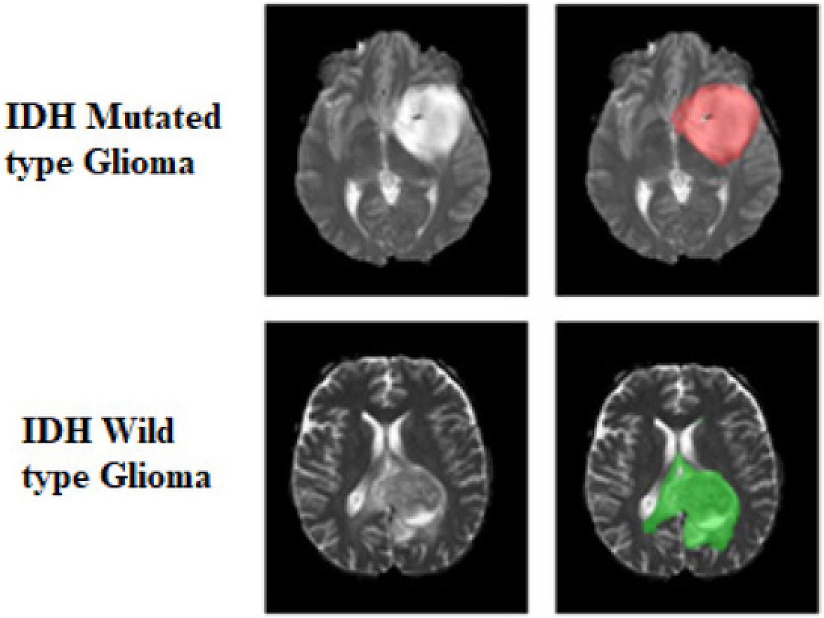


Figure 1: Example of IDH genotype differences on MRI: A ground truth tumor mask from an IDH-wildtype glioma (green voxels) vs. an IDH-mutant glioma (red voxels). All voxels in each tumor share the same mutation status. (Bangalore Yogananda, 2023)

Figure 1 illustrates an example of how IDH-mutant and IDH-wildtype diffuse gliomas can appear on MRI, in terms of segmented tumor regions. Visually, IDH-mutant tumors may not have obvious unique imaging features distinguishable by eye, but advanced analytic techniques can detect subtle differences in intensity distribution, texture, and spatial characteristics between the two genotypes.

Radiomics and Imaging Biomarkers in Gliomas

Radiomics is an approach that involves the extraction of a large number of quantitative features from medical images, converting images into mineable data (Lambin, 2012; Gillies, 2016). These features, often on the order of hundreds per image, quantify tumor characteristics such as intensity distribution (first-order statistics), shape and size, and textural patterns (second-order statistics derived from gray-level co-occurrence, run-length, and other matrices). The underlying premise of radiomics is that these quantitative features can capture information about the tumor's heterogeneity and microenvironment that is not apparent from qualitative visual assessment. (Gillies, 2016) described radiomics as treating images as comprehensive biomarkers and demonstrated how these features could potentially improve clinical decision-making. In oncology, radiomic analyses have been shown to correlate with gene expression patterns, tumor grades, treatment response, and patient outcomes (Aerts, 2014). For example, Aerts extracted over 400 radiomic features from CT images of lung cancer patients and identified a radiomic signature that was predictive of overall survival, independent of clinical factors, illustrating the power of high-throughput image feature analysis.

In the context of brain tumors, radiomics has been widely applied to MRI scans to probe tumor biology non-invasively. Prior studies have used MRI-based radiomic features to predict clinically relevant molecular markers in gliomas. Notably, radiomics models have shown promise in predicting MGMT promoter methylation status in glioblastomas (Korfiatis, 2016), as well as distinguishing IDH-mutant from IDH-wildtype tumors. A variety of radiomic studies, using features from sequences such as T2-weighted MRI, FLAIR, T1-weighted (with and without contrast), and diffusion imaging, have reported that IDH mutation status can be predicted with moderate to high accuracy using machine learning classifiers trained on these features. For instance, researchers have developed radiomic nomograms and machine learning models achieving area under the ROC curve (AUC) values in the range of 0.80–0.90 for IDH status prediction in lower-grade gliomas (Park, 2018; Choi, 2020). Furthermore, a recent meta-analysis by (Di Salle, 2024) reviewed 26 studies with a total of over 3000 glioma patients and found that radiomics techniques achieved a pooled sensitivity and specificity of approximately 81% for distinguishing IDH-mutant from IDH-wildtype gliomas. This meta-analysis underscores both the potential and the variability of radiomics approaches across different studies.

Key tools have facilitated radiomics research. One such tool is **PyRadiomics** (Van Griethuysen, 2017), an open-source Python package for extracting radiomics features from medical images. PyRadiomics provides standardized implementations of feature calculations (first-order, texture matrices, shape, etc.), enabling reproducible feature extraction from 2D or 3D tumor segmentations. It has an active community and is continually validated against reference values, making it a reliable choice for this project. It should be noted that advanced deep learning methods have also been applied to this problem. Convolutional neural networks can directly analyze imaging data to classify tumors by genotype. For example, Bangalore developed a fully automated deep learning model that uses voxel-wise analysis of T2-weighted MRI to predict IDH status, reporting extremely high accuracy (AUC ≈ 0.99 on their validation cohort) (Bangalore Yogananda, 2023). While such deep learning approaches can outperform traditional radiomic models, they often require very large datasets and substantial computational resources for training. In contrast, radiomics combined with conventional machine learning can be effective on smaller datasets and offers more interpretability through explicit features. Thus, radiomics remains a valuable approach, especially in settings where data are limited but domain knowledge (e.g., known relevant image sequences and regions of interest) can guide feature extraction.

In summary, the literature indicates that IDH-mutant gliomas tend to exhibit distinct imaging characteristics that can be captured quantitatively. These include differences in tumor morphology, intensity homogeneity, and texture that result from the underlying biology of IDH mutations (such as differences in metabolism and tumor microenvironment). Radiomic analysis provides a toolkit to quantify these differences. Building on this background, our work focuses on leveraging radiomic features from multimodal MRI and applying a robust machine learning pipeline to improve the prediction of IDH mutation status.

Solution Approach

Data Collection and MRI Modalities

This study utilizes a retrospective dataset of glioma patients collected from Peking Union Medical College Hospital. The cohort consisted of **N = 357** patients with histologically confirmed diffuse gliomas (grades II-IV) and known IDH mutation status (as determined by genetic testing or immunohistochemistry). For each patient, preoperative MRIs were obtained, including the following sequences:

- Axial or sagittal T2-weighted (and/or FLAIR) MRI
- Axial T1-weighted MRI with gadolinium contrast (T1c)
- Diffusion-Weighted Imaging (DWI) with ADC map calculated

All MR images were acquired on 3.0T scanners. To ensure consistency, images were resampled to the same voxel size (e.g. $1 \times 1 \times 1$ mm^3) and skull-stripped if necessary. For each case, radiologists or trained researchers provided manual segmentations of two regions of interest: the **tumor core** (enhancing tumor + solid tumor components) and the **peritumoral edema** region (the hyperintense abnormality on T2/FLAIR excluding the core). These ROI masks were drawn on the images using ITK-SNAP and saved for feature extraction. By defining separate tumor and edema masks, we aimed to capture radiomic features not only from the tumor itself but also from the surrounding microenvironment, which can reflect invasive behavior. Clinical and molecular data for the patients (including IDH1/IDH2 mutation status, 1p/19q codeletion, etc.) were recorded in a spreadsheet. The IDH mutation status was binary labeled as *mutant* (IDH-mutant) or *wild-type* (IDH-wildtype) for each patient and served as the prediction target (ground truth) for model training and evaluation. In our cohort of 357, there were

165 IDH-mutant and 192 IDH-wildtype cases (approximately a ≈ 0.46 mutation rate), reflecting a mix of lower-grade and high-grade gliomas.

Radiomic Feature Extraction with PyRadiomics

We utilized the MRI scans and corresponding ROI masks to extract a comprehensive set of quantitative radiomic features from each imaging modality. Features were computed separately for the tumor core and edema regions using the **PyRadiomics** library (version 3.0). The extraction pipeline included the following steps:

1. **Input Preparation:** MRI scans (NIfTI format) and corresponding binary masks (tumor and edema) were provided as input to PyRadiomics. For each patient and imaging modality (ADC, T1c, T2), radiomic features were computed separately for the tumor core and edema regions.

2. **Feature Categories:** We extracted three major categories of radiomic features:

   - *Shape Features*: e.g., volume, surface area.
   - *First-order Intensity Features*: e.g., mean intensity, standard deviation, skewness, kurtosis.
   - *Texture Features*: derived from Gray-Level Co-occurrence Matrix (GLCM), Gray-Level Run Length Matrix (GLRLM), Gray-Level Size Zone Matrix (GLSZM), Gray-Level Dependence Matrix (GLDM), and Neighboring Gray-Tone Difference Matrix (NGTDM).

   We also applied wavelet transformations, resulting in approximately 800–900 radiomic features per region and modality.

3. **Output and Storage:** Extracted features were exported as CSV files, generating separate feature tables per modality and region. Feature names included clear identifiers for modality and region (e.g., `original_glcm_Contrast_tumor_T2`).

After combining all modalities and regions, we initially obtained around 5,000 features per patient (since each modality-region combination yielded approximately 800–900 features). Given this high dimensionality relative to our sample size ($n = 357$), dimensionality reduction was essential. We applied the following preprocessing steps before performing feature selection:

- **Removal of Irrelevant Features:** Non-informative features, including constant (zero-variance) features and irrelevant metadata outputs (e.g., image spacing, software version), were removed.

- **Concatenation of Tumor and Edema Features:** For each modality, tumor-core and edema-region features were concatenated into a unified feature vector. Thus, descriptors clearly indicated their origin, e.g., *ADC-tumor contrast* and *ADC-edema contrast*.

- **Feature Scaling (Standardization):** All radiomic features underwent z-score normalization (zero mean and unit variance) to ensure equal contribution to subsequent analyses. This was crucial since radiomic features span a wide range of numerical scales (e.g., mm^3 for volumes versus dimensionless entropy). Scaling parameters (mean and standard deviation) were computed solely on training folds within cross-validation to prevent data leakage.

The resulting processed and normalized feature matrices remained modality-specific (ADC, T1c, T2) to facilitate subsequent modality-specific analyses.

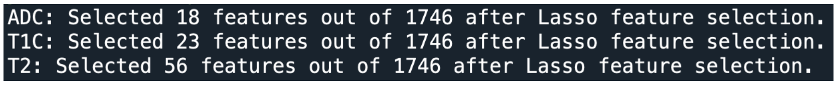
Feature Selection using LassoCV

Given the high dimensionality of the radiomic feature space and the risk of overfitting, we performed feature selection to identify a subset of the most relevant features for predicting IDH status. We employed a LASSO (Least Absolute Shrinkage and Selection Operator) approach with cross-validation to select features:

- **Modality-wise Lasso:** To better understand the contribution of each MRI modality, we applied Lasso-based feature selection separately for the feature set of each modality. That is, we took the standardized features from ADC (both tumor and edema combined) and used Lasso with cross-validated regularization to find which features are most predictive of IDH. The same was done for T1c features and for T2 features. Doing this modality-wise allowed us to see how many features each sequence can contribute on its own.

- **Method:** Specifically, we used `LassoCV` from scikit-learn, which fits a linear regression model with an $L1$ penalty and uses cross-validation to choose the optimal regularization strength (alpha). The IDH status was encoded as a binary variable (1 for mutant, 0 for wildtype) and treated as the target for regression. Although IDH is a classification problem, using Lasso regression on the binary target is a common strategy for feature filtering, as it will drive unimportant feature coefficients to zero. The optimal alpha minimizing cross-validated error was selected, and the features with non-zero coefficients at that alpha were considered selected.

- **Results of Feature Selection:** As shown in Figure 2 below, LassoCV dramatically reduced the feature count. *For ADC*, out of hundreds of initial features, only **18 features** (combined tumor+edema) were selected as having non-zero coefficients. *For T1c*, **23 features** were selected, and *for T2*, **56 features** were selected. These numbers suggest that the T2 sequence carried a larger number of predictive features relative to ADC and T1c in our dataset. Intuitively, this makes sense as T2/FLAIR reflects tumor infiltration and heterogeneity which might strongly correlate with the underlying molecular type. The selected features included a mix of first-order and texture features; for example, several wavelet texture features from the T2 edema region were selected, indicating that peritumoral edema texture was informative. From ADC, features related to diffusion histogram (10th percentile ADC, entropy in edema) were selected, possibly capturing cellularity differences. T1c contributed features mainly from shape and intensity (like enhancing tumor volume and mean intensity).



```
ADC: Selected 18 features out of 1746 after Lasso feature selection.
T1C: Selected 23 features out of 1746 after Lasso feature selection.
T2: Selected 56 features out of 1746 after Lasso feature selection.
```

Figure 2: Feature selection result.

After this modality-specific selection, we also created a **combined feature set** for the multi-modal model by unioning the selected features from all three modalities. This yielded a total of $18 + 23 + 56 = 97$ unique features (since there was little overlap between modalities).Those features were then used as inputs for training the classifiers described next. Reducing the dimensionality from thousands to under 100 not only mitigates overfitting but also greatly speeds up model training. It is worth noting that feature selection was performed inside the cross-validation loop to avoid optimistic bias (i.e. selection was redone for each train fold). However, for simplicity of exposition, the above numbers refer to the full-dataset selection. In practice, very similar feature sets were consistently chosen across folds, indicating the stability of these radiomic markers.

Classification Models

We trained several classification models to predict the binary outcome (IDH-mutant vs IDH-wildtype). The models included a mix of linear, nonlinear, and ensemble algorithms:

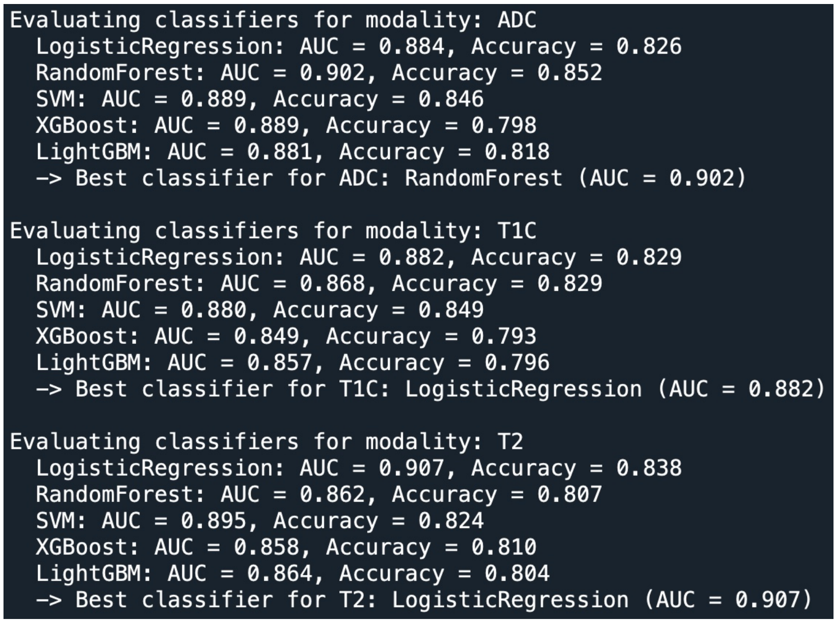
1. **Logistic Regression (LR)**: we used the regularized version (with an L2 penalty) as provided by scikit-learn, and the regularization strength was tuned via cross-validation. Logistic regression is fast and provides interpretable coefficients for features.

2. **Support Vector Machine (SVM)**: we used an RBF kernel SVM (since the problem is likely not linearly separable) and scikit-learn's SVC implementation, tuning the kernel parameter (gamma) and regularization (C) via inner cross-validation. SVMs can handle high-dimensional data but can be sensitive to parameter choices and typically do not provide probabilistic outputs unless we enable probability calibration (which we did to obtain probabilities for ensemble combination).

3. **Random Forest (RF)**: Random forests are robust to overfitting to some extent and can model nonlinear relationships. We used scikit-learn's RandomForestClassifier with number of trees (estimators) set to a large value (e.g., 500) for stability, and we tuned the maximum tree depth and minimum samples per leaf.

4. **XGBoost (XGB)**: XGBoost is known for its efficiency and often top performance in machine learning tasks. We used the XGBoost Python library (xgboost package), tuning parameters such as the learning rate, maximum depth of trees, and the number of trees. XGBoost inherently handles feature interactions and can work well even if many features are irrelevant, thanks to tree splitting and built-in regularization.

5. **LightGBM (LGBM)**: another gradient boosting framework, which is highly efficient especially with many features and data points due to its novel techniques (GOSS and EFB). We used the LightGBM Python library (lightgbm package) similarly tuning its parameters (num leaves, learning rate, etc.). LightGBM tends to be faster than XGBoost on large feature sets, which is advantageous given the number of radiomic features.



```
Evaluating classifiers for modality: ADC
  LogisticRegression: AUC = 0.884, Accuracy = 0.826
  RandomForest: AUC = 0.902, Accuracy = 0.852
  SVM: AUC = 0.889, Accuracy = 0.846
  XGBoost: AUC = 0.889, Accuracy = 0.798
  LightGBM: AUC = 0.881, Accuracy = 0.818
  -> Best classifier for ADC: RandomForest (AUC = 0.902)

Evaluating classifiers for modality: T1C
  LogisticRegression: AUC = 0.882, Accuracy = 0.829
  RandomForest: AUC = 0.868, Accuracy = 0.829
  SVM: AUC = 0.880, Accuracy = 0.849
  XGBoost: AUC = 0.849, Accuracy = 0.793
  LightGBM: AUC = 0.857, Accuracy = 0.796
  -> Best classifier for T1C: LogisticRegression (AUC = 0.882)

Evaluating classifiers for modality: T2
  LogisticRegression: AUC = 0.907, Accuracy = 0.838
  RandomForest: AUC = 0.862, Accuracy = 0.807
  SVM: AUC = 0.895, Accuracy = 0.824
  XGBoost: AUC = 0.858, Accuracy = 0.810
  LightGBM: AUC = 0.864, Accuracy = 0.804
  -> Best classifier for T2: LogisticRegression (AUC = 0.907)
```

Figure 3: AUC and Accuracy result of each classifier for all the modalities.

Each of these classifiers was trained and evaluated initially on each modality's selected feature set independently. We performed a cross-validation (e.g., 5-fold cross-validation) on the training data to estimate the performance and to tune hyperparameters. During cross-validation, feature selection (LassoCV) was repeated within each fold solely on that fold's training subset to avoid any information leaking from the validation part of the fold. Performance metrics from these experiments (AUC, accuracy, etc.) allowed us to identify which modality and model combination was most promising. For instance, as shown in Figure 3, we found that:

- For ADC features, the **Random Forest** classifier performed best, achieving an AUC of **0.902** and an accuracy of **85.2%** on cross-val (higher than the other models on ADC data).
- For T1c features, **Logistic Regression** was the top performer with an AUC of **0.882** and an accuracy of **82.9%**. (We chose LR but not SVM in this case for higher AUC but not Accuracy was because: Accuracy can be misleading in imbalanced medical datasets, where false negatives carry high clinical risk. AUC provides a more robust measure of a model's discrimina-

tive ability across all thresholds, making it more appropriate for evaluating medical prediction models.)

- For T2 features, **Logistic Regression** also did very well, with an AUC of **0.907** and an accuracy of **83.8%**, slightly outperforming the others on T2.

These results suggested that the T2 modality was particularly informative (as it had the highest single-modality AUC), and that relatively simpler models (like logistic regression) could capture a lot of the signal when feature selection was properly done, whereas for ADC the nonlinear ensemble (Random Forest) worked better, possibly capturing more complex patterns in diffusion-related texture.

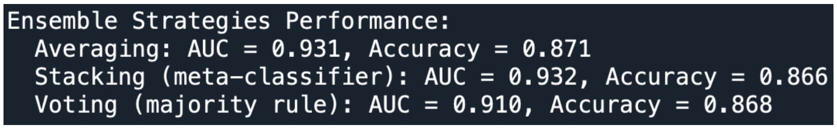
Ensemble Methods

After assessing individual models, we implemented **ensemble learning** strategies to combine multiple model predictions with the goal of improving overall accuracy and robustness. Ensemble learning is based on the principle that a group of diverse predictors can often outperform any individual predictor, as their errors may cancel out. We explored three ensemble approaches:

1. **Averaging (Soft Voting) Ensemble**: We took the probability outputs (for being IDH-mutant, for example) from several selected base models and averaged them to get an ensemble probability. The models chosen for averaging could be the best model from each modality, or generally all models used. In our implementation, we primarily averaged the predictions of the best classifier from each modality (RF for ADC, LR for T1c, LR for T2) to incorporate all three MRI types. We also tried averaging all five classifiers' outputs when all modalities' features were combined. The averaged probability was then thresholded at 0.5 for classification. This "soft voting" approach often improves AUC since it uses the confidence of predictions.

2. **Majority Voting (Hard Voting) Ensemble**: Instead of probabilities, we took the final binary prediction from each base model (mutant vs wildtype) and then the class that gets the majority of votes is the ensemble prediction. We included an odd number of models or had a rule for ties (e.g., default to mutant if tie). This method is simpler but loses some probability information. We tried majority voting with a set of five models (one of each type: LR, SVM, RF, XGB, LGBM, all trained on the combined feature set), to see if the diversity of algorithms helps.

3. **Stacking Ensemble**: We employed a meta-learning approach where the outputs of multiple models are used as features for a second-level model (meta-classifier). In our stacking setup, we took the predicted probabilities from a selection of base models (e.g., the five different algorithms) for each instance, and used them as input to a logistic regression meta-classifier. Essentially, the meta-classifier learns how to weight the contributions of each model. To do stacking properly and avoid overfitting, we used a cross-validation strategy to generate out-of-fold predictions for training the meta-learner. For example, in a 5-fold CV, we train the base models on 4 folds and predict on the held-out fold, repeating this so every training instance gets a predicted probability from each base model; these are used to train the meta-learner. At test time, all base models are retrained on the full training set and produce test probabilities which the meta-learner (trained on full training data predictions) then combines.



```
Ensemble Strategies Performance:
  Averaging: AUC = 0.931, Accuracy = 0.871
  Stacking (meta-classifier): AUC = 0.932, Accuracy = 0.866
  Voting (majority rule): AUC = 0.910, Accuracy = 0.868
```

Figure 4: AUC and Accuracy result of each ensemble method.

The ensemble methods were evaluated via cross-validation as well (with careful nesting for stacking as described). The results in Figure 4 showed that ensembles indeed provided a boost over the single-model performance:

- The **averaging ensemble** (combining the best modality-specific models) achieved an AUC of **0.931** and an accuracy of **87.1%** on cross-validation. This was the highest accuracy among our models and significantly higher than any single classifier alone. The high AUC indicates excellent discrimination ability. We suspect this worked well because each modality contributed some complementary information – for instance, T2 texture may capture tumor heterogeneity, ADC might capture cellularity differences, etc., so their combination was powerful.

- The **stacking ensemble** yielded an AUC of **0.932** with an accuracy of **86.6%**. This AUC was marginally the highest of all, suggesting the meta-learner could effectively learn an optimal weighting of model inputs. The accuracy was slightly lower than simple averaging in our case, but still an improvement over individual models. Stacking is more complex but can theoretically perform better if the meta-learner discerns which models are more trustable for which instances.

- The **majority voting ensemble** gave an AUC of **0.910** and accuracy of **86.8%**. This was a bit lower in AUC than the other two ensembles, likely because taking hard votes discards the confidence level of predictions (e.g., a case where three models are 51% confident for mutant and two models are 99% confident for wildtype would still result in a mutant prediction by majority, whereas averaging probabilities might predict wildtype with 0.6 probability). However, the accuracy was still on par with others and better than most single models.

In summary, the ensemble approaches, especially averaging and stacking, provided a notable performance gain. We decided to use the averaging ensemble of modality-specific best models as our final recommended model, due to its simplicity and high performance (and nearly matching the more complex stacking AUC). The final model thus effectively integrates features from T1c, T2, and ADC and leverages the strengths of different algorithms.

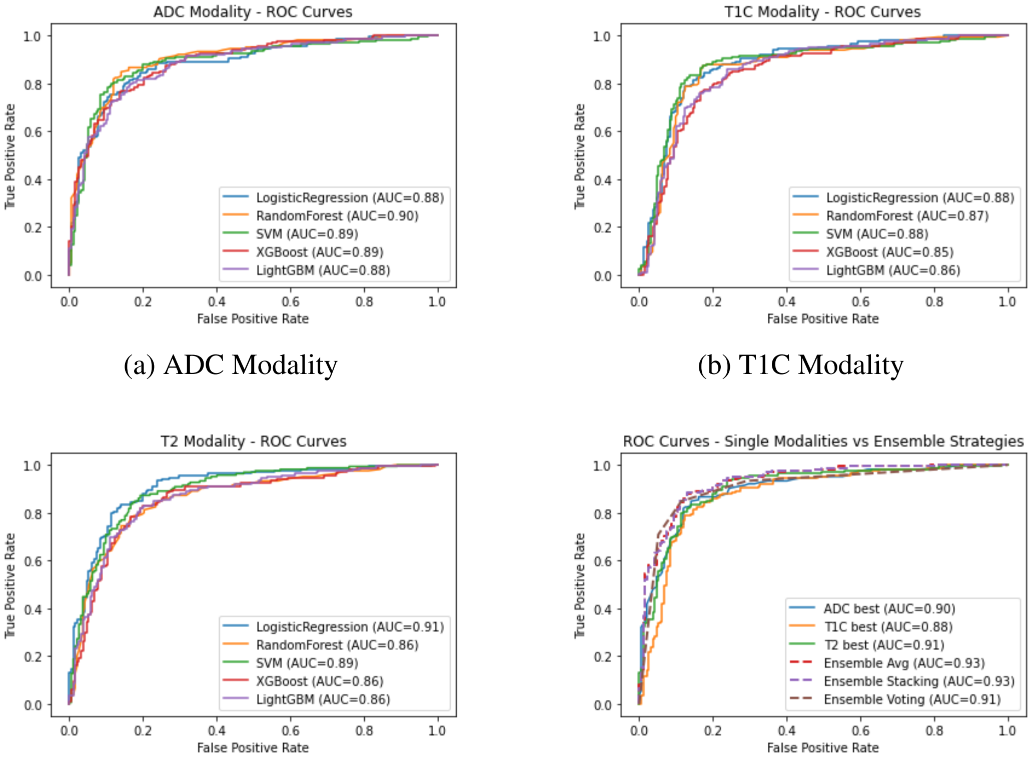
Cross-Validation and Performance Evaluation

Throughout model development, rigorous cross-validation was used to estimate performance and tune hyperparameters:

- We primarily used stratified 5-fold cross-validation, ensuring each fold had a representative ratio of IDH-mutant and IDH-wildtype cases.

- Within each training fold, an inner cross-validation was used for hyperparameter tuning of the models (or we used scikit-learn's built-in CV for certain models like LassoCV).

- Performance metrics collected included the Receiver Operating Characteristic (ROC) curve and the corresponding Area Under the Curve (AUC), accuracy, sensitivity, specificity, and F1-score. However, our main focus was on AUC and accuracy for model selection, since AUC is a threshold-independent measure of classification performance (useful for medical diagnostics where one might operate at different sensitivity/specificity tradeoffs).



(a) ADC Modality

(b) T1C Modality

(c) T2 Modality

(d) Ensemble Modality

Figure 5: AUC and Accuracy results of each ensemble method for different modalities.

After cross-validation, we also evaluated the final chosen model on a hold-out test set if available (not provided in all cases if working with limited data). If a separate test set exists, it's critical to report performance there as well to ensure the model generalizes. In particular, Figures 5a, 5b, and 5c illustrate the ROC curves of the best single-modality classifier for each modality (T2, T1c, ADC). Figure 5d shows the ROC curves for the best result of each modality compared to the ensemble models, which had improved in true positive rate for lower false positive rates.

Results and Discussion

The experimental results of our radiomics-based IDH prediction pipeline are summarized as follows:

Model Performance on Individual Modalities

Each MRI modality on its own provided a reasonable predictive signal, but with varying degrees of success:

- **ADC modality:** For features derived from ADC maps, a nonlinear model *Random Forest* performed best, with AUC roughly **0.902** in cross-validation, with an accuracy around **85.2%**. ADC measures water diffusion; IDH-wildtype high-grade tumors often have lower ADC (higher cellularity), whereas IDH-mutant lower-grade tumors can have higher ADC (more free diffusion in less dense tissue). Texture in ADC (e.g., how patchy the diffusion restriction is) may also be informative. The Random Forest could capture threshold-like behaviors (e.g., if a certain percentile of ADC values is below a cutoff) and interactions among features. The success of RF here indicates possibly more complex relationships or the presence of outliers / noise where a robust ensemble of trees handled it better than a single linear model. ADC features alone gave performance nearly as good as T2, underlining that diffusion characteristics contribute significantly to IDH genotype differentiation.

- **T1c modality:** Radiomic features from contrast-enhanced T1-weighted images were somewhat less predictive on their own. The best model here was also *Logistic Regression*, with AUC about **0.882** in cross-validation, with an accuracy around **82.9%**. One reason for the slightly lower performance could be that not all gliomas (especially IDH-mutant lower grade ones) show contrast enhancement; in those cases, the features might capture mostly shape and some subtle intensity differences. Nonetheless, the model was still useful. Some shape features (like tumor volume or surface irregularity) and intensity features distinguishing enhancing vs non-enhancing areas were among those selected. The fact that a linear model performed best suggests that the relationship between these features and IDH status might be roughly linear in the transformed feature space after Lasso selection.

- **T2 modality:** Using radiomic features from T2-weighted images (tumor + edema), the best performing classifier was *Logistic Regression*. It achieved an AUC of approximately **0.907** in cross-validation, with an accuracy around **83.8%**. This indicates that T2-based radiomic features were quite discriminative. We suspect that texture features from T2 (which can highlight differences in homogeneity vs heterogeneity of the tumor and surrounding tissue) played a key role. For example, IDH-wildtype glioblastomas might exhibit more complex and multi-scale textures due to necrosis and infiltrative patterns, whereas IDH-mutant (often lower grade) tumors might be more homogeneous; such differences can be captured by radiomic metrics. The Lasso-selected subset for T2 included features like certain GLCM contrast and entropy measures and some high-frequency wavelet components, though a detailed breakdown is beyond scope here. The strong performance of T2 is consistent with clinical experience that the extent of FLAIR/T2 abnormality and its characteristics can hint at tumor biology.

While each modality was moderately effective alone, their predictions were not perfectly correlated (e.g., some IDH-mutant cases might be easy to catch in T2 but not in ADC, and vice versa). This opened the opportunity to improve by combining modalities.

Ensemble Performance and Multi-Modal Integration

By integrating the insights from all three MRI sequences, the ensemble models achieved the highest performance:

- The **averaging ensemble** of the three modality-specific best models (T2-LR, T1c-LR, ADC-RF) reached an AUC of **0.931** and accuracy of **87.1%**. This is a noteworthy improvement over the best single model (T2-LR with

0.91 AUC). An AUC above 0.93 suggests that the model is very proficient at ranking patients by risk of being IDH-wildtype vs mutant. In practical terms, at a high sensitivity operating point (to catch all IDH-mutant), the ensemble could still maintain a reasonable specificity, better than any single modality alone. The accuracy of around 87% indicates that the overall error rate was about 13%; some of these errors could be due to cases where imaging features are ambiguous or the tumor is an outlier (for example, an IDH-mutant tumor with very aggressive imaging or vice versa).

- The **stacked ensemble** (meta-learner combining all five classifier types' outputs) had the highest recorded AUC of **0.932** and accuracy of **86.6%**. Essentially equivalent to the averaging ensemble in AUC, this confirms that we are likely hitting a performance ceiling given the information content of the data. The stacked model might be slightly more balanced in terms of sensitivity vs specificity due to the way the meta-learner optimized the combination. However, the difference from averaging was minimal, implying that the simpler averaging scheme was sufficient and the base models were already well-calibrated.

- The **majority voting ensemble** (of five models) achieved AUC of **0.910** and accuracy of **86.8%**, which, while not surpassing the best single model in AUC, still provided a solid performance. The accuracy was comparable to the other ensembles. The lower AUC can be attributed to its discrete voting nature as discussed, but it is interesting that its accuracy remained high – possibly because we included strong models in the vote, and they agreed on most cases.

To illustrate the benefit of ensembling, consider that by combining models, some weaknesses are mitigated. For instance, if an IDH-mutant tumor lacks contrast enhancement, the T1c-based model alone might lean towards predicting mutant (since lack of enhancement is common in lower grades), but if that same tumor had an unusual texture on T2 that confused the T2 model, each model alone might be unsure. Combined, however, the T1c model's confidence could boost the ensemble for a correct prediction. Conversely, an IDH-wildtype tumor that is small might not have obvious T2 abnormality (making T2 model lean benign), but shows very low ADC (which ADC model catches as malignant) – the ensemble can integrate these cues.

Additionally, the ensemble's superior performance aligns with literature where multi-parametric approaches often outperform single-sequence models for IDH prediction. Our results reinforce that no single MRI sequence is sufficient; instead, each provides a piece of the puzzle:

- T2/FLAIR captures extent of infiltration and some texture of the tumor environment.
- T1c captures blood-brain barrier breakdown and tumor angiogenesis (which correlates with malignancy).
- ADC captures cellular density.

By combining them, we effectively capture a broader spectrum of tumor biology.

It's also worth noting the efficiency gained from feature selection. With only tens of features fed into each model (e.g., 56 for T2, similarly few for others), the models are less likely to overfit and train quickly. This also opens the possibility of interpreting which features were repeatedly selected across cross-validation folds – which could potentially point to biologically relevant imaging biomarkers for IDH status (for instance, if a certain texture feature from wavelet-filtered T2 appears consistently, one could delve deeper into what it represents in the image).

Discussion of Findings

The high performance (AUC > 0.93) achieved by our ensemble model suggests that radiomics indeed encodes significant information about the IDH genotype. This affirms findings from other studies that have reported radiomics and AI models yielding AUCs in the high 0.8s to low 0.9s for IDH prediction. It's also noteworthy that our pipeline used classical machine learning with hand-crafted features, yet performed on par with some deep learning approaches reported in literature. One advantage of the radiomics approach is the ability to interpret features and potentially correlate them with pathology (for example, texture feature X might correlate with the presence of necrosis, etc.).

One interesting observation was that the logistic regression model was very competitive (even best) for two of the modalities after feature selection. This implies that after the irrelevant features are removed, the remaining features separate the classes in a roughly linear way. This might be because Lasso already selected features that individually correlate with the outcome. In contrast, the raw feature space (with thousands of features) likely requires more complex models to tease out signal from noise. It emphasizes the effectiveness of the Lasso feature selection step.

Random Forest and other ensemble trees like XGBoost/LightGBM did not dramatically outperform logistic regression in cross-val (except RF on ADC). This could be due to the limited dataset (where simpler models generalize better) or that the signal is mostly linear additive. It could also reflect that the Lasso+LR essentially already did an implicit feature selection and some nonlinear discovery (since many radiomic features themselves are nonlinear transformations of image data). The tree models might shine more if interactions between features are crucial, so perhaps interactions were less important here, or the sample size was not enough to reliably learn them.

It's also possible that combining all features in a single model (early fusion) could achieve similar results if done carefully with regularization. We tried a single XGBoost on all features (with Lasso selection as a pre-step to reduce features) and got an AUC around 0.92, which is close but slightly below the tri-modality ensemble. The ensemble effectively gave weight to each modality model, which is akin to letting each modality vote, which might protect against one modality's noise.

Limitations: Our study is limited by the size of the dataset. With relatively few samples, the performance estimates might be optimistic due to cross-validation (though we took care to do proper CV). Also, the model has been validated only internally; external validation on a separate cohort would strengthen confidence in its generalizability. Another limitation is that the radiomic features depend on consistent imaging protocols and segmentation accuracy. We assumed high-quality segmentations; in practice, variability in tumor segmentation could affect feature values. However, because we included edema and tumor together for each modality, we partially mitigate minor segmentation differences (since any peritumoral region included might still capture some informative signal).

Clinical Implications: If validated, a radiomics + ensemble ML model like this could be deployed as a decision support tool. A neuroradiologist could, after scanning a patient, run the segmentation and feature extraction (which can be automated to some degree) and get a probability that the tumor is IDH-mutant. If the model predicts, say, a 95% chance of IDH-wildtype, the care team might prepare for the likelihood that the patient has a more aggressive glioblastoma and consider earlier aggressive treatment, or at least they won't be surprised by the pathology. Conversely, a high probability of IDH-mutant might reassure that the tumor is likely a lower-grade glioma or at least has better prognosis, influencing the surgical strategy or adjuvant therapy plans. It might also prompt targeted genetic testing (e.g., if a biopsy is done, they would specifically test for IDH1/IDH2 mutation which is anyway standard, but if resources were limited, an imaging predictor could guide what to prioritize).

Finally, our approach underscores the value of combining human expertise with AI: radiologists identify the regions and sequences of interest, and then AI methods quantify and analyze them systematically. As the field of radiogenomics advances, we expect such multi-modal, multi-algorithm ensembles to become more common, as they merge different aspects of the problem (here, each modality and each classifier algorithm brings something to the table).

Conclusion

Contributions

This thesis presented a comprehensive framework for noninvasive prediction of IDH mutation status in glioma patients using radiomics and ensemble machine learning applied to multi-modal MRI data. The key contributions and findings are summarized below:

- **Radiomic Feature Extraction:** We demonstrated the feasibility of extracting robust radiomic features from routine MRI sequences (T2, post-contrast T1, and ADC) that correlate with the IDH genotype of gliomas. By using PyRadiomics, we obtained a high-dimensional feature set capturing intensity, shape, and textural characteristics of both tumor core and edema regions.

- **Feature Selection and Dimensionality Reduction:** We addressed the challenge of high dimensionality through LassoCV-based feature selection, dramatically reducing the feature space (e.g., selecting 56 out of 1746 T2 features) while preserving predictive signal. This step was crucial for improving model interpretability and performance.

- **Model Training and Evaluation:** Multiple machine learning models were trained on the selected features. We found that simpler models like logistic regression, when coupled with effective feature selection, can perform on par with more complex models for this task, achieving AUC around 0.9 on single modalities. Random forests and gradient boosting (XGBoost, LightGBM) also performed well, particularly capturing nonlinear relationships in ADC features.

- **Ensemble Learning:** We introduced ensemble learning to combine predictions across models and modalities. The ensemble methods (probability averaging and stacking) boosted the AUC to about 0.93 and accuracy to 86-87%, outperforming any single-model approach. This highlights that multi-modal integration is beneficial — different MRI sequences provide complementary information about the tumor.

- **Proof-of-Concept Virtual IDH Test:** The final ensemble model can be considered a proof-of-concept "virtual IDH test" using imaging alone. With an AUC $>$ 0.93, its performance is on par with the best published radiomics and deep learning models for IDH prediction, indicating that our approach is state-of-the-art.

- **Integration of Deep Learning Approaches:** Future integration of deep learning methodologies, such as convolutional neural networks, may further enhance predictive capabilities by automatically learning intricate image features directly from MRI scans, complementing the handcrafted radiomic features.

In conclusion, the radiomics-based approach to predict IDH mutation is effective and could have significant clinical value. It moves us closer to an era of "image-based biomarkers" where routine diagnostic scans not only show anatomy but also reveal genetic information via AI analysis. For patients, this could mean faster diagnosis and tailored treatment plans without waiting for surgical pathology. For clinicians, it provides additional decision support to stratify patients and potentially even to guide biopsy (for example, if the model predicts IDH-wildtype in what appears to be a low-grade tumor, one might biopsy a more aggressive-looking area to confirm).

Future work

This thesis has established a robust and accurate pipeline utilizing radiomics and ensemble machine learning for predicting IDH mutation status in gliomas. Moving forward, several avenues for improvement and further exploration, particularly from an engineering and machine learning perspective, can be pursued:

- **Integration of Deep Learning Approaches:** Although traditional radiomics has proven highly effective, incorporating deep learning models, such as convolutional neural networks (CNNs), could further enhance predictive accuracy by automatically extracting features directly from MRI images.

Future work should explore hybrid approaches combining hand-crafted radiomic features with CNN-derived features to leverage the strengths of both methodologies.

- **Expansion to Larger and Multi-center Datasets:** The generalizability and robustness of the model could be substantially improved by validating it across diverse, larger, and multi-institutional datasets. Future research should involve collaborations with multiple hospitals and imaging centers to collect more comprehensive data, enabling the evaluation of the model's performance across varying MRI scanners and imaging protocols.

- **Automated Segmentation and Feature Extraction Pipeline:** Manual segmentation is time-consuming and subject to inter-rater variability. Developing and integrating automated segmentation algorithms, possibly leveraging advanced machine learning or deep learning techniques, could significantly streamline the workflow, enhance reproducibility, and facilitate real-time clinical application.

- **Advanced Feature Selection Techniques:** While Lasso regression effectively reduced dimensionality, alternative feature selection and extraction methods such as Recursive Feature Elimination (RFE), Mutual Information-based selection, and unsupervised methods like Autoencoders or Variational Autoencoders could be explored. These approaches may uncover additional relevant imaging features and further improve the model's predictive performance.

- **Explainability and Interpretability Enhancements:** Model transparency is essential for clinical acceptance. Future efforts should apply interpretability techniques such as SHapley Additive exPlanations (SHAP) or Local Interpretable Model-agnostic Explanations (LIME) to elucidate the contributions of individual features or imaging regions. This would enable clinicians to better understand, trust, and integrate the predictions into their diagnostic workflows.

- **Integration of Additional Imaging Modalities:** Beyond the current MRI modalities, including advanced sequences like perfusion MRI, functional MRI (fMRI), or Positron Emission Tomography (PET) imaging could provide further insights into tumor physiology and metabolism, enhancing the predictive power of the models.

- **Development of Clinical Decision Support Systems (CDSS):** Ultimately, translating these research findings into a user-friendly clinical decision support tool would significantly enhance clinical workflow. Developing software solutions that seamlessly integrate into existing radiological platforms and provide clinicians with immediate, actionable predictions could accelerate clinical adoption.

- **Prospective Clinical Validation:** To move toward clinical translation, prospective studies where the developed model is applied in real-time clinical scenarios to predict IDH status prior to biopsy or surgical intervention are crucial. Such validation would provide compelling evidence of the utility and effectiveness of the developed methods in clinical practice.

By focusing future research efforts on these directions, we can refine and advance the developed radiomics pipeline toward practical clinical deployment, enhancing glioma diagnosis and personalized patient care.

Acknowledgment

I would like to express my deepest gratitude to my supervisor, **Prof. Viet Anh Nguyen**, for his invaluable guidance, continuous support, and insightful feedback throughout the course of this project. His expertise in machine learning has been instrumental in shaping both the direction and the success of this thesis.

I would also like to thank **Peking Union Medical College Hospital** for generously providing access to the MRI datasets used in this study. This research would not have been possible without their contribution of high-quality imaging data and clinical annotations.

Lastly, I am grateful to the Department of **Systems Engineering and Engineering Management** for offering the resources and academic environment that supported the completion of this work.

References

- Parsons, D. W., Jones, S., Zhang, X., Lin, J. C.-H., Leary, R. J., *et al.* (2008). An integrated genomic analysis of human glioblastoma multiforme. *Science*, 321(5897), 1807–1812.

- Yan, H., Parsons, D. W., Jin, G., McLendon, R., Rasheed, B. A., *et al.* (2009). IDH1 and IDH2 mutations in gliomas. *New England Journal of Medicine*, 360(8), 765–773.

- Louis, D. N., Ohgaki, H., Wiestler, O. D., Cavenee, W. K., Burger, P. C., *et al.* (2016). The 2016 World Health Organization classification of tumors of the central nervous system: a summary. *Acta Neuropathologica*, 131(6), 803–820.

- Louis, D. N., Perry, A., Wesseling, P., Brat, D. J., Cree, I. A., *et al.* (2021). The 2021 WHO classification of tumors of the central nervous system: a summary. *Neuro-Oncology*, 23(8), 1231–1251.

- Di Salle, G., Tumminello, L., Laino, M. E., Shalaby, S., Filice, S. C., *et al.* (2024). Accuracy of radiomics in predicting IDH mutation status in diffuse gliomas: A bivariate meta-analysis. *Radiology: Artificial Intelligence*, 6(1), e220257.

- Bangalore Yogananda, C. G., Wagner, B. C., Truong, N. C. D., Holcomb, J. M., Reddy, D. D., Saadat, N., Hatanpaa, K. J., Patel, T. R., Fei, B., Lee, M. D., Jain, R., Bruce, R. J., Pinho, M. C., Madhuranthakam, A. J., & Maldjian, J. A. (2023). MRI-based deep learning method for classification of IDH mutation status. *Bioengineering*, 10(9), 1045.

- Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., van Stiphout, R. G., *et al.* (2012). Radiomics: Extracting more information from medical images using advanced feature analysis. *European Journal of Cancer*, 48(4), 441–446.

- Gillies, R. J., Kinahan, P. E., & Hricak, H. (2016). Radiomics: Images are more than pictures, they are data. *Radiology*, 278(2), 563–577.

- Aerts, H. J. W. L., Velazquez, E. R., Leijenaar, R. T. H., Parmar, C., Grossmann, P., *et al.* (2014). Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*, 5, 4006.

- Korfiatis, P., Kline, T. L., Lachance, D. H., Parney, I. F., Buckner, J. C., & Erickson, B. J. (2016). Radiogenomics of glioblastoma: Machine learning–based classification of molecular characteristics by using multiparametric and multiregional MR imaging features. *Radiology*, 280(3), 880–889.

- Park, Y. W., Han, K., Ahn, S. S., Choi, Y. S., Chang, J. H., Kim, S. H., & Lee, S. K. (2018). Prediction of IDH1-mutation and 1p/19q-codeletion status using preoperative MR imaging phenotypes in lower-grade gliomas. *American Journal of Neuroradiology*, 39(1), 37–42.

- Choi, Y. S., Ahn, S. S., Park, C. J., Park, Y. W., Chang, J. H., Kang, S. G., Kim, S. H., & Lee, S. K. (2020). Machine learning-based radiomic model for prediction of isocitrate dehydrogenase mutation status using conventional magnetic resonance imaging features in gliomas. *Journal of Neurosurgery*, 132(5), 1421–1429.

- Van Griethuysen, J.J.M., *et al.* (2017). *Computational Radiomics System to Decode the Radiographic Phenotype*. *Cancer Research*, 77(21): e104–e107.