Technische Hochschule Ingolstadt

Technische Hochschule Ingolstadt

Seminar Künstliche Intelligenz

Studiengang Künstliche Intelligenz, B.Sc.
Fakultät Informatik

MRI Radiomics for IDH Genotype Prediction in Glioblastoma Diagnosis

Vor- und Zuname: **Stanislav Kozák**

Matrikelnummer: **00135825**

Technische Hochschule Ingolstadt

Contents

List of Abbreviations

Medical terminology

ADC - apparent diffusion coefficient (maps)
ASL - arterial spin labeling
ATRX - alpha-thalassemia/mental retardation syndrome X-linked
CNS - central nervous system
CT - computed tomography
DWI - diffusion weighted imaging
EGFR - epidermal growth factor receptor
GBMA - glioblastoma
HARDI - high angular resolution diffusion imaging
IDH - isocitrate dehydrogenase
MGMT - 06-methylguanine-DNA methyltransferase
MRI - magnetic resonance imaging
PET - positron emission tomography
ROI - region of interest
SPECT - single-photon emission computed tomography
SWI - susceptibility weighted imaging
T1C - post-contrast T1WI sequence
T1WI - T1-weighted imaging, pre-contrast
T2WI - T2-weighted imaging
FLAIR - fluid attenuated inversion recovery
VOI - volume of interest

Radiomic Features and Tools

ANT - Advanced normalization tool
BET - Brain extraction tool
FSL - FMRIB software library
GLCM - grey level co-occurrence matrix
GLDM - grey level dependence matrix
GLRLM - grey level run length matrix
GLSZM - grey level size zone matrix
IBSI - Image Biomarker Standardization Initiative
NGTDM - neighbouring grey tone difference matrix

Machine Learning Terminology

Adam - adaptive moment estimation (optimiser)
AUC - area under the curve (ROC)
BR - binary relevance (classifier)
(E)CC - (ensemble) classifier chain
LASSO - least absolute shrinkage and selection operator
MCC - Matthews' correlation coefficient
ML-SMOTE - multi-label synthetic minority oversampling technique
RFECV - recursive feature elimination with cross-validation
ROC - receiver operator characteristics (curve)

1 Introduction to Grade IV Gliomas

Radiomics is a relatively new field which utilises automatically identified features from radiological imaging results based on MRI, CT, PET, PET/CT or SPECT scans or combinations thereof. It has found a widespread application, particularly in oncology because many of the important oncological biomarkers are not visible to the naked eye. (van Timmeren et al. 2020) The relatively recent advent of big data, including in medical imaging, and the development of new machine learning techniques brought the possibility of faster and more accurate oncological diagnosis. Furthermore, standardised mathematical feature extraction helps to eliminate possible radiologist bias.

This paper reviews the recent development in the oncological use of MRI radiomic features. More specifically, it focuses on the identification of the isocitrate dehydrogenase (IDH) mutation status, which is an important biomarker for the diagnosis of grade IV glioma (glioblastoma and grade IV astrocytoma).

1.1 Glioblastoma

Glioblastoma is the most common and aggressive type of primary brain tumour (grade IV glioma), presumably arising from neural progenitor cells. Its appearance and internal genotype are highly heterogeneous, so that biopsies taken from different parts of the tumour can be very different. (Hooper GW 2023) Non-invasive diagnosis of glioblastoma through conventional radiological methods can therefore be very difficult.

Glioblastoma stem cells are prone to epithelial-mesenchymal transition, making them more flexible and invasive. It also contains non-malignant cells that are immunosuppressive and create a supporting microenvironment for the tumour growth (Hooper GW 2023). These factors make the treatment very challenging, resulting in a median patient survival of less than 2 years. Recent improvements in treatment, which can include surgery, radiotherapy, chemotherapy and targeted therapy, have improved short-term survival. However, 5-year survival has remained relatively constant at 5.8%, due to the recurrent nature of glioblastoma in most cases. As the causes are not yet well understood, there is no known way to prevent it (Tan et al. 2020).

1.2 Genotype Alterations in Grade IV Gliomas

Grade IV gliomas can undergo different mutations and molecular alterations. In about 60% of cases, mutations of the epidermal growth factor receptor (EGFR) have been identified, leading to a more aggressive tumour behaviour. Another important predictor of patient survival is the methylation status of 06-methylguanine DNA methyltransferase (MGMT). Methylation inhibits the production of this enzyme and slows down its DNA repair function.

This may lead to a better prognosis and response to alkylating chemotherapy (usually done with temozolomide). (Tan et al. 2020, Hooper GW 2023)

1.3 IDH Mutation

Isocitrate dehydrogenase (IDH) mutation on the chromosome 2 is one of the most important biomarkers of high-grade glioma as it significantly changes the tumour behaviour and therefore affects the survival prediction. (Tan et al. 2020)

The IDH mutation has also been included by the World Health Organization (WHO) in the Classification of Tumors of the Central Nervous System. In 2016, glioblastoma was divided into IDH-mutant and IDH-wildtype subgroups. In the fifth edition of the WHO-classification (2021), glioblastoma with IDH-1 or IDH-2 mutation was reclassified as astrocytoma, IDH-mutant. Therefore, a reliable method for predicting the mutation status of IDH is crucial for differentiating glioblastoma from grade IV astrocytoma and for subsequent treatment planning. (Sohn et al. 2021)

IDH-wildtype (glioblastoma) is the more common and more aggressive variant, occuring predominantly in older population (median age 62 years). It develops mostly de novo, without any identifiable precursor lesion (Tan et al. 2020). It is more prone to EGFR amplification (making the tumour more aggressive), but also to MGMT promoter methylation (leading to a more favourable prognosis). Subsequently, astrocytoma with IDH1 or IDH2 mutation (affecting between 5 to 13% patients (Calabrese et al. 2020)) has overall a better prognosis and is more common in younger patients. (Tan et al. 2020)

Currently, the mutation status is usually determined by immunohistochemical staining with the R132H mutant IDH antibody based on tumour resection or biopsy. Non-invasive diagnosis methods based on MRI sequences continue to be explored as demonstrated in the following sections. (Cui et al. 2023)

2 MRI Radiomics Workflow

In this chapter, the commonly used radiomic pipeline for feature extraction from MRI images is introduced, as shown in figure 1. This includes specific examples of radiomic features which are the essential components of the pipeline.

2.1 Image Acquisition

Although image acquisition itself is not a direct part of the radiomics pipeline, it has a major impact on the quality of the input data. The starting point is the output of magnetic resonance imaging (MRI). The images are produced by an MRI scanner emitting and measuring

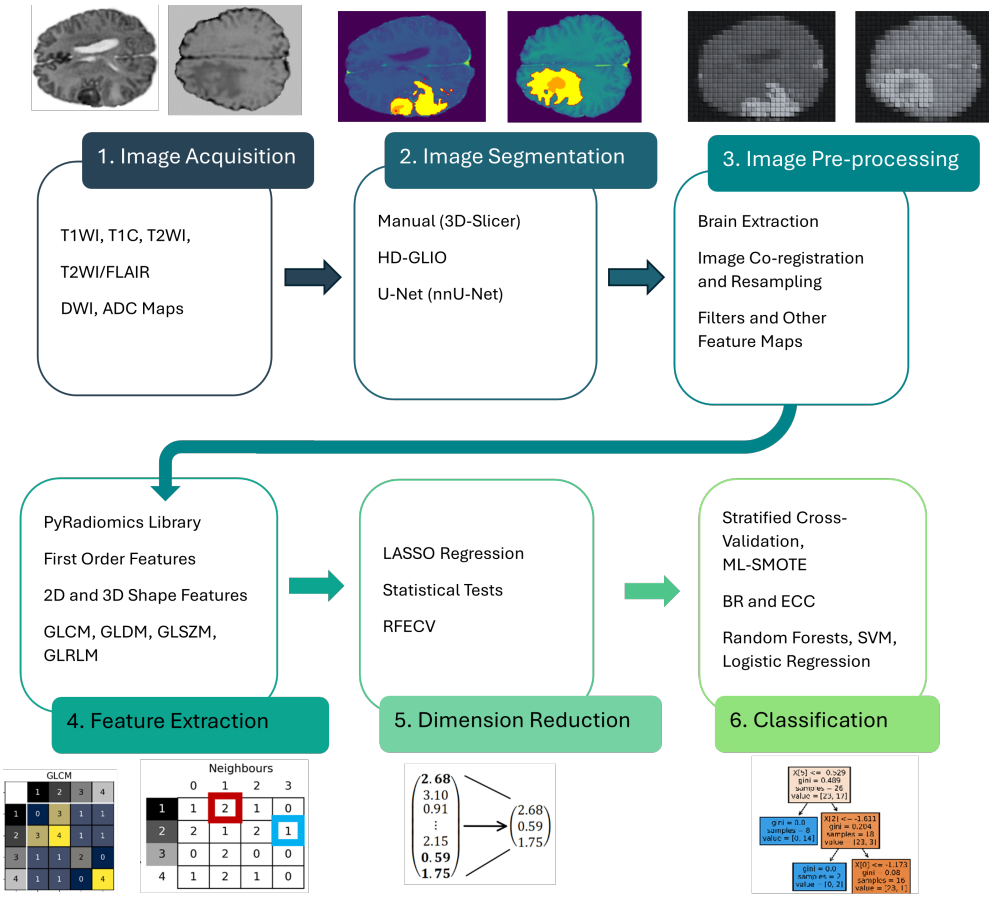

Figure 1: Radiomics pipeline based on traditional ML models.

magnetic fields. MRI takes longer than computed tomography (CT) or positron emission tomography (PET) and is usually more expensive, but it does not require patients to be exposed to ionising radiation so it can be done more often and on patients with more critical health conditions. (Preston 2016)

MRI detects unusual structural forms, such as torn ligaments or tumours. It produces a number of 2D scans that can be assembled to create a 3D representation.This visualisation can be examined in all three planes (axial, sagittal and coronal). (Preston 2016)

An MRI scanner extracts a number of different sequences. T1-weighted (T1WI) highlights anatomical structures. A gadolinium-based contrast agent can be applied through an intravenous line to obtain a post-contrast sequence (T1C). (Preston 2016)

With modified scan settings, T2WI sequences can be acquired, which are used to detect pathological regions. Other sequences acquired by MRI or derived from the mentioned ones include T2WI/FLAIR, diffusion-weighted imaging (DWI), arterial spin labeling (ASL) or apparent diffusion coefficient (ADC) maps (Calabrese et al. 2020, Cui et al. 2023).

2.2 Image Segmentation

Segmentation is the process of delineating the region of interest (ROI) or volume of interest (VOI) from 2D or 3D images. This can be done manually, semi-automatically or automatically. Manual segmentation is the most commonly used method, but it can be time consuming, resource intensive and it can subject to observer bias. Semi-automatic methods inlude computer algorithms (e.g. region growing, tresholding) whose results are manually corrected. (van Timmeren et al. 2020)

Automatic image segmentation uses deep learning models, mostly based on the U-Net or nnU-Net architecture (for example HD-GLIO). This approach is faster and does not need expert supervision but it needs large datasets for training. The path of fully automatic segmentation seems promising but its generalisability to different datasets is still under intensive research. (van Timmeren et al. 2020)

2.3 Image Pre-processing

Through image pre-processing, segmented images are homogenised to provide input for feature extraction that is consistent in its characteristics. General image pre-processing includes the following procedures:

- *interpolation to isotropic voxel spacing (resampling)* – ensures the same voxel size for each sequence in the three dimensions. Because MRI yields multiple outputs, one sequence is often used as a reference scale for the rest.
- *intensity outlier filtering* – filters out grey values outside of a predefined range.
- *discretisation* – the scale of image intensities is divided into bins to reduce the number of possible values. (van Timmeren et al. 2020)

For MRI, image pre-processing also includes skull stripping (brain extraction) to remove non-brain tissue which is not relevant for the analysis (Cui et al. 2023).

Various filters or transformations can also be applied to the sequences. An open-source Python library PyRadiomics, for example, supports the following: wavelet-filters (spatial low-frequency and high-frequency filtering), Gaussian-filters (producing images with enhanced edges), square, square root, exponential and logarithmic images (Cui et al. 2023). The role of PyRadiomics in image pre-processing and feature extraction is illustrated in the figure 2.

2.4 Feature Extraction

Feature extraction uses various mathematical calculations and algorithms to quantitatively analyse the greyscale images. PyRadiomics divides extracted features into following categories: (van Griethuysen et al. 2017)

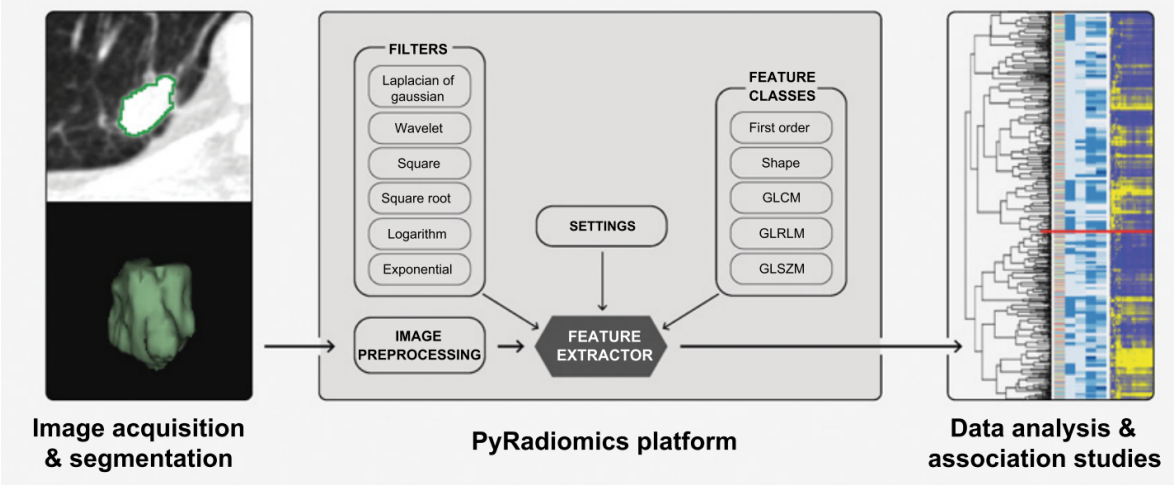

Figure 2: Usage of PyRadiomics in radiomic studies. Taken from van Griethuysen et al. (2017), edited.

- *first order (histogram based) statistics* describe the intensity distribution, for example through mean, percentiles, entropy or skewness. Entropy is calculated by the formula shown in equation 1, with N_g being the number of non-zero intensity bins, $p(i)$ the normalized histogram of intensities and ϵ an arbitrary small number:

$$entropy = -\sum_{i=1}^{N_g} p(i) * log_2(p(i) + \epsilon) \tag{1}$$

- *3D shape features* are morphology descriptors, independent on the grey level intensities, computed only from a derived 3D mesh. (compactness, sphericity or surface area to volume ratio).

- *2D shape features* are derived from a circumference mesh (perimeter, surface, elongation).

Furthermore, PyRadiomics uses following helping matrices to extract *texture-based features*: (van Griethuysen et al. 2017)

- *grey level co-occurrence matrix (GLCM)* quantifies, how often each combination of grey values appears together (within a specified distance and angle). An example with unit distance and zero angle (left-right direction) is shown in figure 3. The value of the combination 1-2 in the GLCM (red highlighting) is three, because the bin value 2 appears three times left or right to bin value 1 (and vice versa). The same applies for single-value combinations (blue highlighting for bin value 4).

  GLCM is by default symmetrical and can be used to derive contrast, autocorrelation, joint average, difference entropy, etc. Contrast is computed by the formula 2, iterating

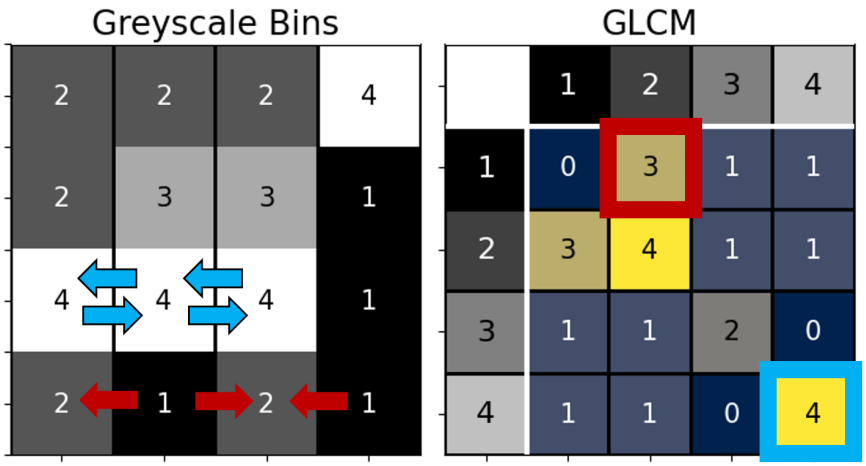

Figure 3: Computation of GLCM from binned greyscale values. Left: simplified original image with binned greyscale values. Right: computed GLCM matrix.

over the normalised GLCM entries $p(i, j)$ for each grey level combination (i and j). The contrast is low, when the highest values of the GLCM are close to the main diagonal (i and j are similar):

$$contrast = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i - j)^2 * p(i, j) \tag{2}$$

- *grey level dependence matrix (GLDM)* counts for each grey value, how often it appears with a given number of similar or equal neighbours. It is therefore used to identify clusters of similar intensities. A simplified calculation of GLDM is shown in the figure 4. For example, there are two occurrences of bin value 1 with exactly one equal neighbour in the original picture (red highlighting) and one occurrence of bin value 2 with three equal neighbours (blue highlighting). Non-relevant neighbours are marked with white arrows.

  GLDM can be used to calculate grey level variance, high and low grey level emphasis, etc. Small and large dependence emphasis (SDE and LDE) are calculated by equations 3 and 4, respectively. Here, N_d is the number of neighbours and $p(i, j)$ is the normalised GLDM matrix. SDE is small if the image has small clusters (high values on the left side of the GLCM, with low number of neighbours j) and therefore low homogeneity. Conversely, LDE identifies large clusters and high homogeneity.

$$SDE = \sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \frac{p(i, j)}{j^2} \tag{3}$$

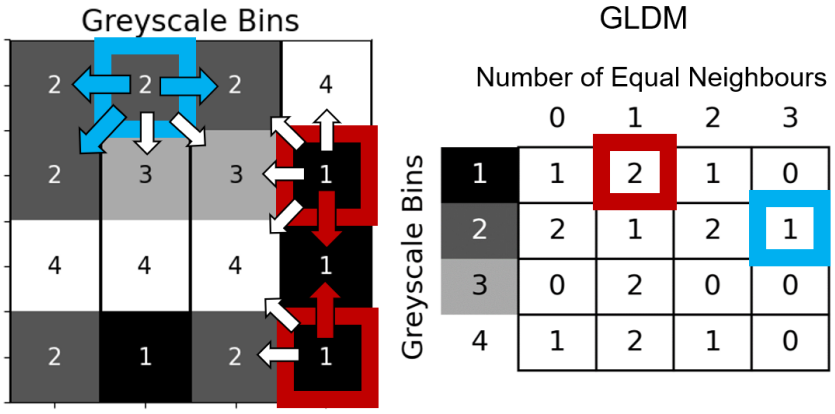

Figure 4: Computation of GLDM from binned greyscale values

$$LDE = \sum_{i=1}^{N_g} \sum_{j=1}^{N_d} p(i,j) * j^2 \tag{4}$$

- *grey level size zone matrix (GLSZM)* quantifying zone sizes of voxels with the same grey values (directionally independent).
- *grey level run length matrix (GLRLM)* describing length of equal consecutive pixels along a given angle.
- *neighbouring gray tone difference matrix (NGTDM)* represents the differences between grey value of a pixel and its neighbours withing a given distance.

Further categories, such as *model-based* (e.g. texture regularity analysis using autoregressive models) or *transform-based* (e.g. obtained by discrete Haar wavelet transform) are also sometimes referenced. (Mayerhoefer et al. 2020)

Some studies also differentiate between semantic features that can be recognised by a radiologist (e.g. size or morphology of the tumour) and agnostic features. These are derived purely mathematically and cannot be perceived by a human from the images. (Hooper GW 2023)

To address the lack of reproducibility and validation of feature extraction methods in radiomic studies, the Image Biomarker Standardization Initiative (IBSI) has published guidelines and definitions for the acquisition of radiomic biomarkers (Zwanenburg et al. 2020). PyRadiomics also follows those guidelines.

2.5 Feature Selection and Dimension Reduction

Feature selection follows directly after feature extraction, as this can produce from hundred to several thousands of different features for each patient. The excessive amount of training

features contains noise and can lead to overfitting of the classification model if not enough data points are available. If known, the non-reproducible features (with high intra- or inter-observer variability) should be excluded. The importance of the remaining features can be evaluated by multiple methods.

The correlation between features can be calculated to identify and remove highly correlated features. Using statistical tests, the distribution of a given feature is analysed for each class to eliminate features that are similar in all classes.

Least Absolute Shrinkage and Selection Operator (LASSO) can also be trained on normalised input features to predict the target class. It uses L1 regularisation to penalise large coefficient values, which also reduces overfitting. After fitting the model, the features with the largest respective coefficients (in absolute terms) are selected as they contribute the most to the selection.

Recursive feature elimination with cross-validation (RFECV) also identifies the important features by performing the classification itself and iteratively discarding features which do not significantly improve the accuracy.

In traditional ML research, dimension reduction (e.g. through linear discriminant analysis, principal component analysis or t-SNE) is also utilised. But because it combines multiple features together, its outputs are less interpretable, so it is not widely used in radiomic research.

Data visualisation can also help to identify important correlations. Using the tools mentioned above, the features can be grouped into correlation clusters. For each cluster, the most representative features are selected for the model fitting. This can reduce the number of features from thousands to less than 10-20. (van Timmeren et al. 2020)

2.6 Classification Models

Classification or regression can be performed after the previous steps have been completed. Common classification tasks using radiomic features are prediction of the survival rate and treatment response or risk assessment, presence or stage of a particular tumour type or prediction of the tumour recurrence time. Several machine learning model types are widely used, including support vector machines, random forests, logistic regression or neural networks. Deep neural networks can only be used when larger datasets are available.

3 Methodology

This chapter provides a detailed review of the radiomic processes presented by three recent studies that focused on predicting IDH mutation status through radiomics from MRI sequences. Due to hypothesised associations between glioblastoma genotypes, the studies

trained models to predict several different genetic biomarkers. These studies were selected in particular because they used different approaches to the prediction task while achieving comparable results. They all utilised Scikit-learn for the data preparation and prediction task, which is a standard Python library for traditional machine learning.

In 2020, Calabrese et al. used a radiomic approach to predict nine different genetic biomarkers in patients with glioblastoma. From a total cohort size of 199, 195 patients were tested for the presence of IDH mutation and 190 on the MGMT promoter methylation. (Calabrese et al. 2020)

A similar study was performed by Sohn et al. in 2021. This study focused solely on predicting EGFR, IDH mutation, MGMT methylation and ATRX loss status based on a cohort of 418 patients. (Sohn et al. 2021)

Finally, Cui et al. investigated in 2023 predictive models for IDH mutation, histological phenotype (differentiation between low-grade and high-grade glioma) and Ki-67 expression level with a contrast analysis. The dataset used included 150 patients with glioblastoma and other types of glioma. (Cui et al. 2023)

3.1 Image Acquisition

Most of the preoperative MRI scans were performed on patients diagnosed between 2015 and 2020. All three research groups used only patient cohorts from their respective medical institutions, based on the 2016 WHO classification. Therefore, the datasets are rather small and may possibly lead to model overfitting.

Table 1 provides a detailed overview of the datasets and MRI sequence acquisition protocols. Some of the datasets were originally larger, but cases without preoperative MRI or biomarker information had to be excluded. The dataset of the third research group contained only 55 cases of diagnosed glioblastoma with a known IDH mutation status.

Due to the naturally occurring prevalence of IDH1 mutation in grade IV gliomas (between 5 and 13% (Calabrese et al. 2020)), the classes are moderately unbalanced which needs to be compensated for in the following steps. The second group identified only 3.6% of IDH1 mutation cases. Next-generation genetic sequencing based on biopsy or tumour resection was used to determine the IDH mutation status. Cui et al. used immunohistochemical staining with R132H mutant antibody.

Each group used different scanners and configurations. Using more scanners may make the classification task more difficult but may also help to make the dataset more generalisable. The scan produced several different sequences (4-8 in each group), including T1WI (pre-contrast and with a gadolinium-based contrast agent), T2WI, T2WI/FLAIR and DWI. Cui et al. used the DWI results to calculate ADC maps.

Table 1: Patient cohorts and image acquisition details of the reviewed studies

Research group	**Calabrese et al. (2020)**	**Sohn et al. (2021)**	**Cui et al. (2023)**
Classified biomarkers	IDH, ATRX, CDKN2, EGFR, MGMT, PTEN, TERT, TP53, aneuploidy of chromosomes 7 and 10	IDH, ATRX, EGFR, MGMT	IDH, Ki-67-expressions, histological phenotype
Cohort size	199 patients	418 patients	150 patients
IDH tested	195 patients	418 patients	55 patients (GBMA), 125 (all)
IDH-mutant	18 patients (9.2%)	15 patients (3.6%)	6 patients GBMA (10.9%), 51 patients all (40.8%)
IDH diagnosis technique	Next-generation genetic sequencing	Next-generation genetic sequencing	Immunohistochemical staining
Used scanners	One 3.0 T scanner	Two 3.0 T scanners	7 scanners, mainly 1.5 T Signa HDxt
Extracted MRI sequences	T1WI, T1C, T2WI, T2WI/FLAIR, SWI, DWI, ASL, HARDI	T1WI, T1C, T2WI, T2WI/FLAIR	T1WI, T1C, T2WI, DWI, ADC maps

3.2 Image Segmentation

The group of Cui et al. chose the traditional method of manual tumour segmentation. Two radiologists, blinded to the histological and immunohistochemical results, performed the segmentation using 3D-slicer. Each 2D slice from the T2WI sequence was segmented and then assembled to reconstruct a 3D model. (Cui et al. 2023)

The other two research groups performed automatic tumour segmentation using pre-trained deep convolutional neural networks in order to automate the entire pipeline. Segmentation was carried out using 2D T1WI (pre and post-contrast), T2WI and T2WI/FLAIR sequences. The segmentation results were then manually examined but not corrected. (Calabrese et al. 2020, Sohn et al. 2021)

Sohn et al. used the HD-GLIO algorithm, which separates contrast-enhancing tumour from non-enhancing T2/FLAIR signal abnormalities. The model used by Calabrese et al. consisted of three binary sub-models and segmented the images into enhancing tumour, non-enhancing tumour, surrounding tumour-related edema and background. They used Adam

optimiser with learning rate decay and binary softmax cross-entropy loss for the training. (Calabrese et al. 2020, Sohn et al. 2021)

3.3 Image Pre-processing Techniques

After segmentation, brain extraction was performed, for example with BET (Brain Extraction Tool) from the FMRIB Software Library (FSL) (Cui et al. 2023). To ensure the same dimension of each sequence, the resulting images were co-registered (based on T1C or T2WI sequences) and resampled to isotropic voxel spacing (with 1x1x1 mm voxel size).

N4 bias correction with advanced normalisation tool was then applied to remove low frequency intensities caused by magnetic field inhomogeneity. Finally, the image intensities across all sequences were normalised (with $\mu = 0$ and $\sigma^2 = 1$). (Sohn et al. 2021)

Calabrese et al. computed four additional diffusivity maps from the HARDI data: mean, axial and radial diffusivity, and fractional anisotropy, giving a total of 11 inputs for the feature extraction.

Cui et al. applied 14 filters and transformations supported by PyRadiomics to derive wavelet-filtered, Gaussian-filtered, etc. images.

3.4 Feature Extraction

All studies used PyRadiomics to extract radiomic features from the segmented and normalised sequences. The respective number of features in each study is listed in table 2. The 2D and 3D shape features were extracted independent of the sequences. First order and higher order features were extracted either for each sequence (Sohn et al. 2021), each combination of sequence and filter (Cui et al. 2023) or for each combination of sequence and segmented region (whole tumour, tumour core, 3 tumour compartments; Calabrese et al. (2020)). This resulted in very different output sizes (660 for Sohn et al., 5300 for Calabrese et al., 6580 for Cui et al.).

Calabrese et al. extracted all the features provided by the PyRadiomics library. The filtering criteria for shape and first-order features by Cui et al. and Sohn et al. is unknown. This lack of transparency is unfortunate as it makes the results more difficult to reproduce.

3.5 Feature Selection

Sohn et al. and Cui et al. used LASSO regression to discard features with a low predictive value.

Cui et al. also performed Mann-Whitney U test for each class, selecting only input features whose distributions varied. They then calculated the Pearson correlation coefficient between the remaining features and removed highly correlated features.

Table 2: Extracted features

Research group	**Calabrese et al. (2020)**	**Sohn et al. (2021)**	**Cui et al. (2023)**
Extracted from	11 sequences, 5 segmented regions (55 combinations)	4 sequences	5 sequences, 14 image types
Shape features	26 per segmentation	Not specified	14 per sequence
First order features	19 per combination	Not specified	18 per image type
Higher order (texture) features	75 per combination	Not specified	75 per image type
Total features per patient	$26 * 5 + (19 + 75) * 55 =$ **5300**	**660**	$(14 + (18 + 75) * 14) * 5 =$ **6580**

Calabrese et al. and Cui et al. used RFECV to select the final features. In each iteration step, RFECV trained a simple classification model to quantify their importance. The least important features were then removed and the process was repeated several times for different dataset splits to avoid overfitting.

3.6 Classification Models

The low natural prevalence of the IDH mutation makes the data set unbalanced. To prevent the model from underperforming on the minority class, Calabrese et al. used 10-fold stratified cross validation (with 60/40 train-test split), which ensures the same class distribution in each set. Cui et al. trained the classifier for the data set with all gliomas which was not imbalanced (40.8% patients with IDH mutation). They used repeated k-fold cross validation (80/20 split repeated 30 times).

Another method is synthetic data generation which was used by Sohn et al. They run the Multi-Label SMOTE algorithm to generate new data points of minority classes while maintaining the associations between the labels. They used a 70/30 ratio for the train test split. (Sohn et al. 2021)

There are two main approaches to multi-label classification: binary relevance (BR) and classifier chain (CC). BR uses separate independent binary classifiers trained on each label. In CC, each label is also predicted by a binary model, but it can use the result of previous labels and therefore takes into account the correlation between different labels. When the optimal order of classification is not known, an ensemble classifier chain (ECC) can be used, which iterates over chains with different classifier orders. Calabrese et al. and Cui et al.

chose the binary relevance approach, while Sohn et al. compared both approaches. (Sohn et al. 2021)

The choice of model and loss function can have a significant impact on the training time and on the accuracy. Linear models are usually faster to train but they learn less representative decision boundary than non-linear models.

Calabrese et al. treated each label prediction as a binary regression task, predicting the probabilities for both the negative and positive class. They used a random forest regressor and randomised search for hyperparameter tuning. (Calabrese et al. 2020)

Sohn et al. used a linear kernel support vector machine trained by SGD with manual hyperparameter tuning. They also tried out 10 different classifier orders for the ECC approach, evaluated by the mean absolute Shapley values. The optimal classifier order found was IDH-ATRX-MGMT-EGFR.

Cui et al. built independent classifiers for each MRI sequence. They applied support vector machines and logistic regression with 9 different loss functions, including hinge, logarithmic, Huber or epsilon insensitive. They also tried L1 and L2-regularisation and their combination (Elastic Net). Based on the results of the single sequence classification, they built final classifiers combining T1C and ADC sequences because those two sequences reached the most accurate predictions. (Cui et al. 2023)

4 Evaluation and Discussion

In this chapter, a comparison and discussion of the results and conclusions of the three studies under review will be presented. This includes a comparison of the model performance, the optimal radiomic features found, and the associations between the biomarkers themselves.

The following sections conclude the paper by describing the potential and limitations of predicting IDH mutation status based on radiomic features in clinical practice.

4.1 Result Comparison

Table 3 compares the results obtained by the final classification and regression models for the IDH mutation prediction. It is important to note that Cui et al. predicted the IDH mutation status for all gliomas, not just grade IV. Based on the ROC curve, they probably also considered IDH-wild-type as a positive class, unlike the other studies. This discrepancy may be due to the fact that the positive class usually indicates more rare and dangerous cases at the same time. In other words, Cui et al. chose the more dangerous class as positive, whereas the other studies chose the more rare class as positive. To make it easier to compare the results between studies, the metrics using the IDH mutation as a positive class were also included (shown in brackets).

The high recall values indicate the ability of the model to detect most cases of IDH mutation (minimising false negatives and the type II error). This is offset by more false positives and therefore lower precision (and higher type I error). Optimising for recall (type II error) is helpful for the clinical practice because it avoids higher costs and exposing patients to riskier treatment, as the diagnosis is rarely overestimated. However, it results in more cases being underestimated.

Labelling the minority class as positive also validates the usage of precision, recall and their harmonic mean (F1-Score) as metrics. However, MCC and AUC are often more preferable metrics for imbalanced data sets as they are symmetric under class labelling.

Very good predictive performance for distinguishing glioblastoma from IDH-mutated astrocytoma was achieved by Calabrese et al. (MCC 0.62). Cui et al. achieved an even better performance (MCC 0.68) for classifying IDH mutation and IDH wild-type, but this could be caused by including lower grade gliomas in the data set. (Calabrese et al. 2020, Cui et al. 2023)

Table 3: Final model evaluation

Research	**Calabrese et al. (2020)**	**Sohn et al. (2021)**	**Cui et al. (2023), all gliomas**
Precision	0.50	0.26	0.93 (0.73)
Recall	0.93	1.0	0.81 (0.89)
F1-Score	0.62	0.42	0.87 (0.80)
MCC	0.62	0.48	0.68 (0.68)
AUC	0.95	0.96	0.88 (0.88)

4.2 Found Features and Biomarker Associations

Surprisingly, the features with the highest predictive value differed significantly between the three studies. There was little to no overlap in the imaging sequences (except for T1C), segmented regions (tumour core, contrast enhancing tumour, etc.) or the feature groups (first order, shape features, etc.) in the top selected features.

Features extracted from T1C sequences were among the most efficient. Calabrese et al. identified variance on the whole tumour segmented region, Sohn et al. identified coarseness (NGTDM feature), surface area to volume ratio (shape feature) and maximum correlation coefficient (GLCM feature) on contrast-enhancing segmented regions as important. Cui et al. did not rank the features by the predictive performance, but only listed the 21 features

selected for the final classifier. They identified several first-order features and features from GLRLM, GLSZM or GLCM on logarithmic or wavelet-filtered images from both T1C and ADC sequences.

For Calabrese et al., diffusivity metrics had also a high predictive value. One of those was high grey level emphasis, a feature based on GLDM that quantifies the diffusivity of the non-enhancing tumour produced by the DWI sequence. Another was the kurtosis of tumour-related edema from the mean diffusivity mask.

The optimal classifier order found by Sohn et al. highlights a significant effect of IDH prediction on ATRX and MGMT classification (as mentioned in chapter 1.1): *"IDH mutation increases the overall genomic CpG methylation and is strongly associated with MGMT promoter methylation."* (Sohn et al. 2021)

The found correlation of IDH and MGMT prediction is particularly helpful as MGMT alone was difficult to predict, for example in Calabrese et al. also found that ATRX mutations are more common in IDH mutant gliomas but rare in IDH wild-type.

In conclusion, the studies reviewed all achieved satisfactory classification performance, but did not find any fully overlapping radiomic features that would allow unambiguous identification of the IDH mutation. This is due to the different segmentation, image processing and feature extraction techniques used.

It is also possible that the models were highly dependent on the small and unbalanced datasets, as the performance was worse on independent validation data.

4.3 Potential and Limitations of Radiomics for IDH Genotype Prediction

The presented radiomic approach may contribute to a non-invasive and faster diagnosis when differentiation between glioblastoma and IDH-mutated astrocytoma is required for further treatment planning. The development of a highly accurate classification model could validate or replace biopsy or radiologist diagnosis.

Further research is needed to address several issues. Most importantly, the procedures and techniques used in the radiomic pipeline should be standardised and consistently documented. The efforts of the IBSI to standardise feature extraction have been effective as there were only small differences between the reviewed studies in the extracted features (all have used standard features by PyRadiomics).

The second issue is the generalisability of the classification model. Larger datasets with different patient demographics, health conditions and scanner types are needed to develop a reliable classifier. Furthermore, IDH2 mutation could be included as it is also a indicator of grade IV astrocytoma (albeit with a lower prevalence).

This also raises the question of the overall suitability of the radiomic approach. When

larger datasets are available, the traditional machine learning classifiers could be replaced by a deep learning network. The whole pipeline could also be replaced by an end-to-end framework, for example using a convolutional neural network for classification. In this case, the interpretability of the model's diagnosis would suffer. However, as has been shown, the optimal radiomics features found also convey characteristics that are not visible to humans and they vary significantly between studies, so there is no ultimate feature combination that could unambiguously solve the classification task while being fully interpretable.

Appendix

Reproduction of Classification Results

We tried to reproduce the results discussed in chapter 4 with multiple machine learning models trained on a dataset presented in Calabrese et al. (2022). The source code and description are publicly available at `https://github.com/standakozak/Glioma-Classification`.

References

Calabrese, E., Villanueva-Meyer, J. E. & Cha, S. (2020), 'A fully automated artificial intelligence method for non-invasive, imaging-based identification of genetic alterations in glioblastomas', *Scientific reports* **10**(1), 11852.
URL: *https://pubmed.ncbi.nlm.nih.gov/32678261/*

Calabrese, E., Villanueva-Meyer, J., Rudie, J., Rauschecker, A., Baid, U., Bakas, S., Cha, S., Mongan, J. & Hess, C. (2022), 'The University of California San Francisco Preoperative Diffuse Glioma MRI (UCSF-PDGM)(Version 4) [Dataset]'.

Cui, Y., Dang, Y., Zhang, H., Peng, H., Zhang, J., Li, J., Shen, P., Mao, C., Ma, L. & Zhang, L. (2023), 'Predicting isocitrate dehydrogenase genotype, histological phenotype, and Ki-67 expression level in diffuse gliomas with an advanced contrast analysis of magnetic resonance imaging sequences', *Quantitative imaging in medicine and surgery* **13**(6), 3400–3415.

Hooper GW, G. D. (2023), 'MRI radiomics and potential applications to glioblastoma', *Frontiers in Oncology* **13**, 1134109.

Mayerhoefer, M. E., Materka, A., Langs, G., Häggström, I., Szczypiński, P., Gibbs, P. & Cook, G. (2020), 'Introduction to Radiomics', *Journal of Nuclear Medicine* **61**(4), 488–495.
URL: *http://jnm.snmjournals.org/lookup/doi/10.2967/jnumed.118.222893*

Preston, D. C. (2016), 'Magnetic resonance imaging (MRI) of the brain and spine: Basics'. Accessed: 2024-07-02.
URL: *https://case.edu/med/neurology/NR/MRI%20Basics.htm*

Sohn, B., An, C., Kim, D., Ahn, S. S., Han, K., Kim, S. H., Kang, S.-G., Chang, J. H. & Lee, S.-K. (2021), 'Radiomics-based prediction of multiple gene alteration incorporating mutual genetic information in glioblastoma and grade 4 astrocytoma, IDH-mutant', *Journal of neuro-oncology* **155**(3), 267–276.
URL: *https://pubmed.ncbi.nlm.nih.gov/34648115/*

Tan, A. C., Ashley, D. M., López, G. Y., Malinzak, M., Friedman, H. S. & Khasraw, M. (2020), 'Management of glioblastoma: State of the art and future directions', *CA: a cancer journal for clinicians* **70**(4), 299–312.
URL: *https://pubmed.ncbi.nlm.nih.gov/32478924/*

van Griethuysen, J. J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R. G., Fillion-Robin, J.-C., Pieper, S. & Aerts, H. J. (2017), 'Computational Radiomics System to Decode the Radiographic Phenotype', *Cancer Research*

77(21), e104–e107.
URL: *https://aacrjournals.org/cancerres/article/77/21/e104/662617/Computational-Radiomics-System-to-Decode-the*

van Timmeren, J. E., Cester, D., Tanadini-Lang, S., Hatem Alkadhi, Alkadhi, H. & Baeßler, B. (2020), 'Radiomics in medical imaging-"how-to" guide and critical reflection', *Insights Into Imaging* **11**(1), 1–16.

Zwanenburg, A., Vallières, M., Abdalah, M. A., Aerts, H. J. W. L., Andrearczyk, V., Apte, A., Ashrafinia, S., Bakas, S., Beukinga, R. J., Boellaard, R., Bogowicz, M., Boldrini, L., Buvat, I., Cook, G. J. R., Davatzikos, C., Depeursinge, A., Desseroit, M.-C., Dinapoli, N., Dinh, C. V., Echegaray, S., El Naqa, I., Fedorov, A. Y., Gatta, R., Gillies, R. J., Goh, V., Götz, M., Guckenberger, M., Ha, S. M., Hatt, M., Isensee, F., Lambin, P., Leger, S., Leijenaar, R. T. H., Lenkowicz, J., Lippert, F., Losnegård, A., Maier-Hein, K. H., Morin, O., Müller, H., Napel, S., Nioche, C., Orlhac, F., Pati, S., Pfaehler, E. A. G., Rahmim, A., Rao, A. U. K., Scherer, J., Siddique, M. M., Sijtsema, N. M., Socarras Fernandez, J., Spezi, E., Steenbakkers, R. J. H. M., Tanadini-Lang, S., Thorwarth, D., Troost, E. G. C., Upadhaya, T., Valentini, V., van Dijk, L. V., van Griethuysen, J., van Velden, F. H. P., Whybra, P., Richter, C. & Löck, S. (2020), 'The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping', *Radiology* **295**(2), 328–338.
URL: *http://arxiv.org/pdf/1612.07003*