Original Article

Enhancing cancer prognostics with group penalty models: a comparative study on radiomics feature selection in lung adenocarcinomas and meningiomas

Yonghan Kwon[1], Young Joo Suh[2], Yae Won Park[1], Sung Soo Ahn[1], Kyunghwa Han[1]*, Min Jin Ha[3]*

[1]Department of Radiology, Research Institute of Radiological Science, and Center for Clinical Imaging Data Science, Yonsei University College of Medicine, Seoul, Republic of Korea; [2]Department of Radiology, Research Institute of Radiological Science, Yonsei University College of Medicine, Seoul, Republic of Korea; [3]Department of Health Informatics and Biostatistics, Graduate School of Public Health, Yonsei University, Seoul, Republic of Korea

Contributions: (I) Conception and design: Y Kwon, K Han, MJ Ha; (II) Administrative support: K Han, MJ Ha; (III) Provision of study materials or patients: YJ Suh, YW Park, SS Ahn; (IV) Collection and assembly of data: YJ Suh, YW Park, SS Ahn; (V) Data analysis and interpretation: Y Kwon, K Han, MJ Ha; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

*These authors contributed equally to this work.

Correspondence to: Kyunghwa Han, PhD. Department of Radiology, Research Institute of Radiological Science, and Center for Clinical Imaging Data Science, Yonsei University College of Medicine, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea. Email: khhan@yuhs.ac; Min Jin Ha, PhD. Department of Health Informatics and Biostatistics, Graduate School of Public Health, Yonsei University, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea. Email: mjha@yuhs.ac.

Background: Radiomics, the extraction of features from medical images, has shown promise in cancer prognostics. However, its high-dimensional nature poses challenges for feature selection and model stability. Traditional least absolute shrinkage and selection operator (Lasso) tends to select only one feature from correlated groups, leading to unstable feature selection. This study aimed to compare group penalty models with Lasso in selecting radiomic features for cancer prognosis.

Methods: We analyzed 590 lung adenocarcinoma lesions for predicting early-stage spread through air spaces (STAS) and 194 meningioma cases for tumor grade prediction. Various group penalty models, including group Lasso, group minimax concave penalty (MCP), group smoothly clipped absolute deviation (SCAD), and adaptive penalization regression with external covariates using variational Bayes (graper), were employed alongside Lasso. Features were organized into natural groups based on segmentation regions and mathematical properties. Model performance was assessed using 10-fold cross-validation, evaluating area under the curve (AUC) and feature selection stability through Jaccard Index. The best-performing models were validated on independent test sets.

Results: For lung adenocarcinomas, group SCAD achieved the highest cross-validation AUC of 0.804 [standard deviation (SD) =0.056] with Jaccard Index of 0.613, compared to Lasso's AUC of 0.776 (SD =0.059) and Jaccard Index of 0.503. In the test set, both models showed comparable performance (group SCAD: AUC =0.874; Lasso: AUC =0.877; P=0.896). For meningiomas, group Lasso achieved the highest cross-validation AUC of 0.816 (SD =0.143) with Jaccard Index of 0.7, versus Lasso's AUC of 0.743 (SD =0.223) and Jaccard Index of 0.41. Test set validation showed no significant difference (group Lasso: AUC =0.877; Lasso: AUC =0.835; P=0.391).

Conclusions: Group penalty models demonstrated superior feature selection stability while maintaining comparable predictive performance to Lasso. By selecting biologically meaningful feature groups rather than individual features, these models enhance interpretability and align better with clinical reasoning, offering a robust framework for radiomics-based cancer prognostics.

Keywords: Cancer prognosis; group penalty models; lung adenocarcinoma; meningioma; radiomics

Submitted Sep 09, 2025. Accepted for publication Dec 23, 2025. Published online Jan 23, 2026.
doi: 10.21037/qims-2025-1951
View this article at: https://dx.doi.org/10.21037/qims-2025-1951

Introduction

Radiomics, the extraction of intricate features from diagnostic images such as computed tomography (CT) and magnetic resonance imaging (MRI), is revolutionizing cancer management by revealing patterns invisible to the naked eye (1). This technique enables quantitative analysis of tissue complexity, facilitating precise cancer prognosis prediction and personalized treatment planning.

Radiomics features inherently exhibit natural group structures based on their extraction methods and characteristics (2). These natural groupings reflect fundamental biological and technical relationships among features. For instance, features extracted from the same segmentation region share spatial context and biological relevance, while features computed using the same mathematical framework capture related aspects of tissue heterogeneity (3). This inherent group structure suggests that considering features collectively within their natural groups, rather than treating them as independent variables, may provide more robust and biologically interpretable results in cancer prognostics.

The least absolute shrinkage and selection operator (Lasso) (2) has been the standard method for feature selection in high-dimensional radiomics data, effectively reducing complexity into radiomics scores (rad-scores) (3-5). Despite the widespread use of Lasso, its effectiveness in analyzing complex, high-dimensional data remains uncertain, with critiques pointing to potential limitations in such scenarios (6). More critically, Lasso is challenged by the presence of highly correlated features. When predictors are highly correlated, Lasso tends to select only one feature from a group of correlated features and ignores the others (7), failing to leverage the collective information within feature groups. This limitation is particularly problematic in radiomics, where features naturally cluster into meaningful groups based on their mathematical properties or biological origins. Consequently, this arbitrary selection among correlated features leads to instability in feature selection—small changes in the data can result in different features being selected from the same correlated group, undermining the reliability and reproducibility of the model (8).

Recent work by Orton *et al.* (9) demonstrated that hierarchical correlation-based feature reduction could improve radiomics model interpretability in clear-cell renal cell carcinoma by considering the natural hierarchy of feature groups based on interpretability (shape features being more interpretable than first-order features, which in turn are more interpretable than texture features). However, their approach was limited to a single cancer type with predetermined, static group hierarchies and relied solely on group Lasso methodology.

Our study addresses these limitations by systematically comparing traditional Lasso with multiple group penalty approaches including group Lasso (10,11), group minimax concave penalty (group MCP) (11), group smoothly clipped absolute deviation (group SCAD) (11), and graper (12)—a Bayesian method that adaptively learns group-specific penalties from the data. Features within the same group often exhibit positive correlations, highlighting the potential of group-based analyses to enhance statistical power and improve reproducibility (13,14). To critically assess the effectiveness of these methods, we performed cross-validation based evaluation using existing radiomics datasets and validated the selected models through independent test sets. We hypothesized that group penalty models would provide more stable feature selection while maintaining predictive accuracy and improving interpretability of rad-scores. This comparative analysis was conducted through two case studies: the prediction of spread through air spaces (STAS) in early-stage lung adenocarcinomas (15) and the pre-surgical grading of meningiomas (16). We present this article in accordance with the TRIPOD+AI reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-2025-1951/rc).

Methods

Patients

Lung adenocarcinomas study

Lung adenocarcinoma is the most prevalent form of lung cancer and the leading cause of cancer-related deaths. STAS, introduced as a novel prognostic factor in the 2015

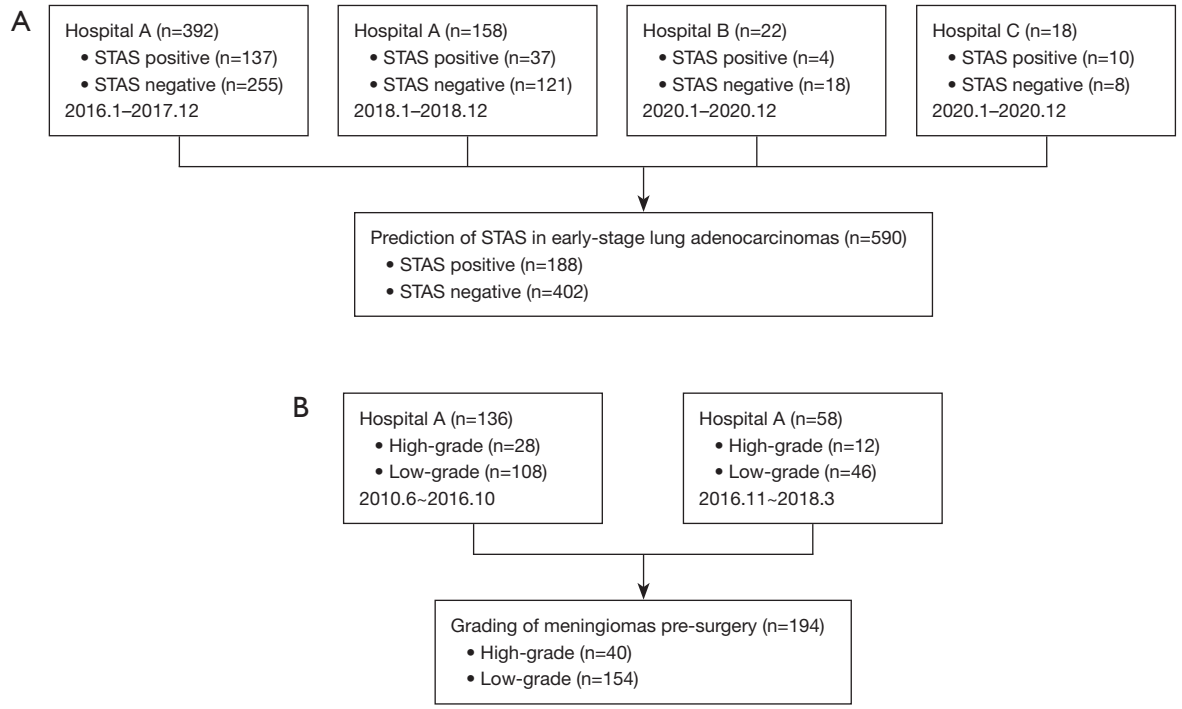

Figure 1 Selection flowchart for (A) STAS lung adenocarcinomas and (B) meningioma grading. Hospital A: Severance Hospital; Hospital B: Korea University Anam Hospital; Hospital C: Yongin Severance Hospital. STAS, spread through air spaces.

World Health Organization (WHO) classification, is correlated with increased recurrence risk and diminished survival rates. Detecting STAS is crucial for surgical planning; however, challenges exist in its identification using current preoperative and intraoperative methods. Radiomics offers a potential solution for these challenges through quantitative imaging analysis (15). This study was conducted in accordance with the Declaration of Helsinki and its subsequent amendments.

The cohort used in this research was initially collected in a prior study (15). Data from 590 lesions with stage I lung adenocarcinoma and presurgical chest CT images were used to develop and validate the radiomics model. The training set comprised 432 lesions (281 STAS-negative and 151 STAS-positive) from Severance Hospital (January 2016 to December 2017), Korea University Anam Hospital, and Yongin Severance Hospital, while the test set included 158 lesions (121 STAS-negative and 37 STAS-positive) from Hospital A (January to December 2018). The patient demographics are summarized in Table S1. A flowchart illustrating the study population is depicted in *Figure 1A*. Radiomics features were extracted from three distinct regions: the nodule, its 3-mm dilated periphery, and a combined region, yielding 279 features per region using Pyradiomics (RRID:SCR_026019) (17). These features included histogram, texture [gray level co-occurrence matrix (GLCM), gray level run length matrix (GLRLM), gray level size zone matrix (GLSZM), neighboring gray tone difference matrix (NGTDM)], and shape descriptors (Table S2).

Meningiomas study

Preoperative grading of meningiomas, the most prevalent primary intracranial tumors in adults, is critical for planning surgery and predicting prognosis. The WHO classification classifies these tumors into WHO grade 1, 2, and 3. While most meningiomas are WHO grade 1 ("low-grade"), more than one-fifth are WHO grade 2 or 3 ("high-grade"), demonstrating aggressive behavior and recurrence, leading to lower 5-year survival rates. Utilizing noninvasive radiomics, analyzing features from T1-weighted contrast (T1C)-enhanced images and diffusion tensor imaging (DTI), aids in this differentiation, especially for incidentally discovered tumors, which may not require histological verification and can be monitored without immediate intervention (16).

The cohort used in this research consists of data previously collected in prior study (16). Data from 194 patients with confirmed pathology and preoperative MRI scans were used to develop and validate the radiomics model. The training set comprised 136 patients (108 with low-grade and 28 with high-grade tumors) from June 2010 to October 2016, while the test set included 58 patients (46 with low-grade and 12 with high-grade tumors) from November 2016 to March 2018. The patient demographics are summarized in Table S1. A flowchart depicting the study population is shown in *Figure 1B*. After preprocessing with N4 bias correction and intensity normalization, regions of interest were delineated on T1C-enhanced images. A total of 98 radiomics features were extracted from T1C, apparent diffusion coefficient (ADC), and fractional anisotropy (FA) maps using 3D Slicer (RRID:SCR_005619), including histogram-based, co-occurrence matrix-based, and run-length matrix-based parameters, plus morphological features (Table S3).

Least absolute shrinkage and selection operator (Lasso)

Lasso is a widely used statistical method for feature selection in high-dimensional data analysis (2). In radiomics studies, where hundreds of features are extracted from medical images, Lasso helps identify the most relevant features while avoiding overfitting—a common problem when the number of features exceeds the number of patients.

The key advantage of Lasso is its ability to automatically select features by shrinking less important feature coefficients to exactly zero, effectively removing them from the model. This results in a simpler, more interpretable model that focuses only on the most predictive radiomics features. The degree of feature selection is controlled by a regularization parameter (λ), which is typically determined through cross-validation to optimize model performance. The detailed mathematical formulation is provided in the Appendix 1.

However, Lasso has important limitations, particularly when dealing with correlated features—a common scenario in radiomics data.

Limitations of Lasso with correlated features

When predictors are highly correlated, Lasso tends to select only one feature from a group of correlated features and ignore the others, which can be problematic when multiple correlated predictors are all relevant (7). This tendency introduces instability in feature selection, where small changes in the data can lead to large changes in the set of selected predictors. For instance, when data is split into different training and validation sets, Lasso may select different sets of features across the splits due to the inherent correlation structure among predictors (8). This instability undermines the reliability and reproducibility of the model, making it less robust for inference and prediction. Thus, the instability in feature selection by Lasso necessitates exploring alternative methods that can handle correlations more effectively and provide more stable and interpretable models, such as group penalty models which consider groups of correlated features.

Group penalty models

Group penalty models extend individual feature selection to groups of related features, which is particularly useful when features are naturally correlated or grouped. These models prioritize selecting groups of features over individual features based on predefined groupings from the study design or data characteristics (*Figure 2*). Unlike traditional Lasso, which can suffer from instability in feature selection, group penalty models handle correlations more effectively by considering the inherent group structures within the data (13,14).

The group penalty models evaluated in this study include group Lasso (10,11), group MCP (11), group SCAD (11), and graper (12). group Lasso applies a group-wise penalty that ensures either all features in a group are selected together or none at all. This all-or-nothing selection property makes it particularly suitable for radiomics applications where features within the same category are expected to be jointly relevant.

Group MCP and group SCAD employ non-convex penalties that address the bias issue inherent in group Lasso. While group Lasso applies constant shrinkage to all coefficients, potentially over-penalizing large coefficients, group MCP and group SCAD reduce this bias by applying minimal or no penalty to truly important feature groups. This property allows these methods to perform asymptotically as well as if the true model were known in advance.

Unlike the other group penalty methods that apply fixed penalties, graper takes a fundamentally different Bayesian approach. This method adaptively learns group-specific penalties from the data rather than applying the same penalty to all groups. Such flexibility allows automatic

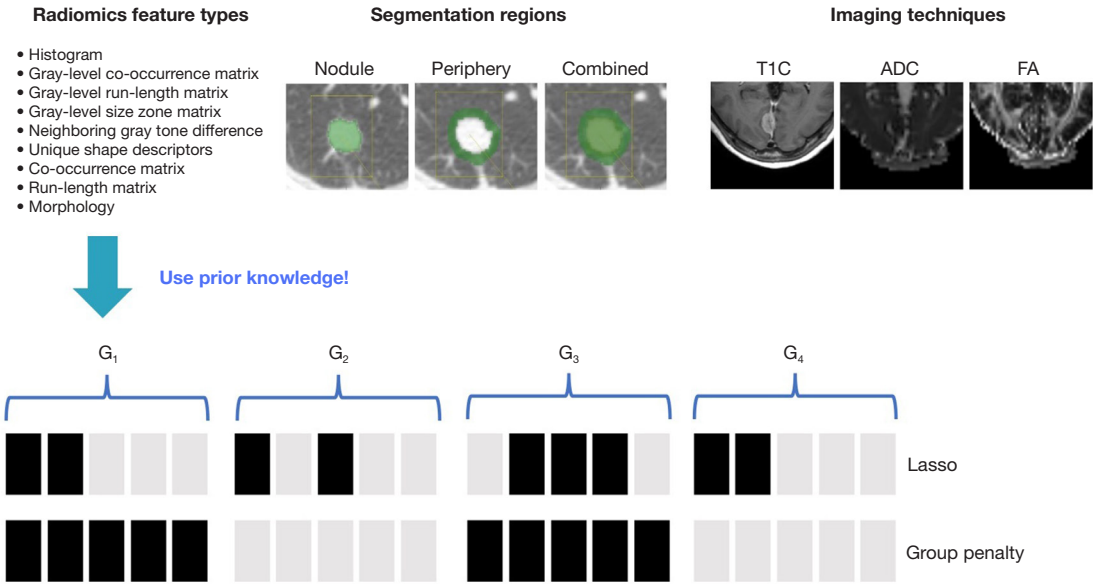

Figure 2 Contrast between Lasso's selection of individual features and group penalty models, which select clusters of correlated features in radiomics data analysis. graper makes feature selections with group information as a prior but does not make group-level selections. ADC, apparent diffusion coefficient; FA, fractional anisotropy; graper, adaptive penalization regression with external covariates using variational Bayes; Lasso, least absolute shrinkage and selection operator.

identification of important feature groups without extensive manual tuning. The method uses variational Bayes approximation to maintain computational efficiency while providing explicit importance scores for each feature group. Additionally, graper was implemented with Random Over-Sampling Examples (ROSE) (17) to address data imbalances in our cancer classification tasks.

The Appendix 1 provides detailed algorithmic descriptions and mathematical formulations for the group Lasso, group MCP, and group SCAD penalties, as well as the hierarchical Bayesian model and variational inference procedure employed in graper.

Feature grouping strategy for lung adenocarcinomas study

In the STAS prediction radiomics study, our methodological approach for prior group selection was anchored to three strategic feature classifications. (I) Segmentation regions: we grouped the features into three primary categories based on their distinct segmentation regions: nodule, peripheral, and a "nodule + peripheral" region. This classification establishes the foundation for our analysis by clearly differentiating the features according to their extraction areas. (II) Radiomics feature types: to further enhance the precision of our group penalty models, we performed a careful categorization of each radiomic feature into one of six distinct types based on their radiometric attributes: histogram, GLCM, GLRLM, GLSZM, NGTDM, and shape descriptors. This detailed categorization allowed for a more nuanced analysis of radiomic features. (III) Combination: this granular classification harmonized with the initial segmentation categories, leading to the creation of 18 unique groups. This sophisticated structure combines the segmentation regions (nodule, peripheral, and "nodule + peripheral") with radiomics feature types (histogram, GLCM, GLRLM, GLSZM, NGTDM, and shape). A summary of these groups is presented in *Figure 3* and Table S4, where the numbers in each cell indicate the count of features belonging to that specific group.

Feature grouping strategy for meningiomas study

In the study of pre-surgical grading of meningioma, our methodological approach to prior group selection was anchored in three strategic feature classifications. (I) Imaging techniques: initially, we organized the features according to medical imaging techniques, identifying three main categories: T1C, ADC, and FA. Additionally, we set the morphology as a separate group, laying the groundwork for our analysis. (II) Radiomics feature types: to further

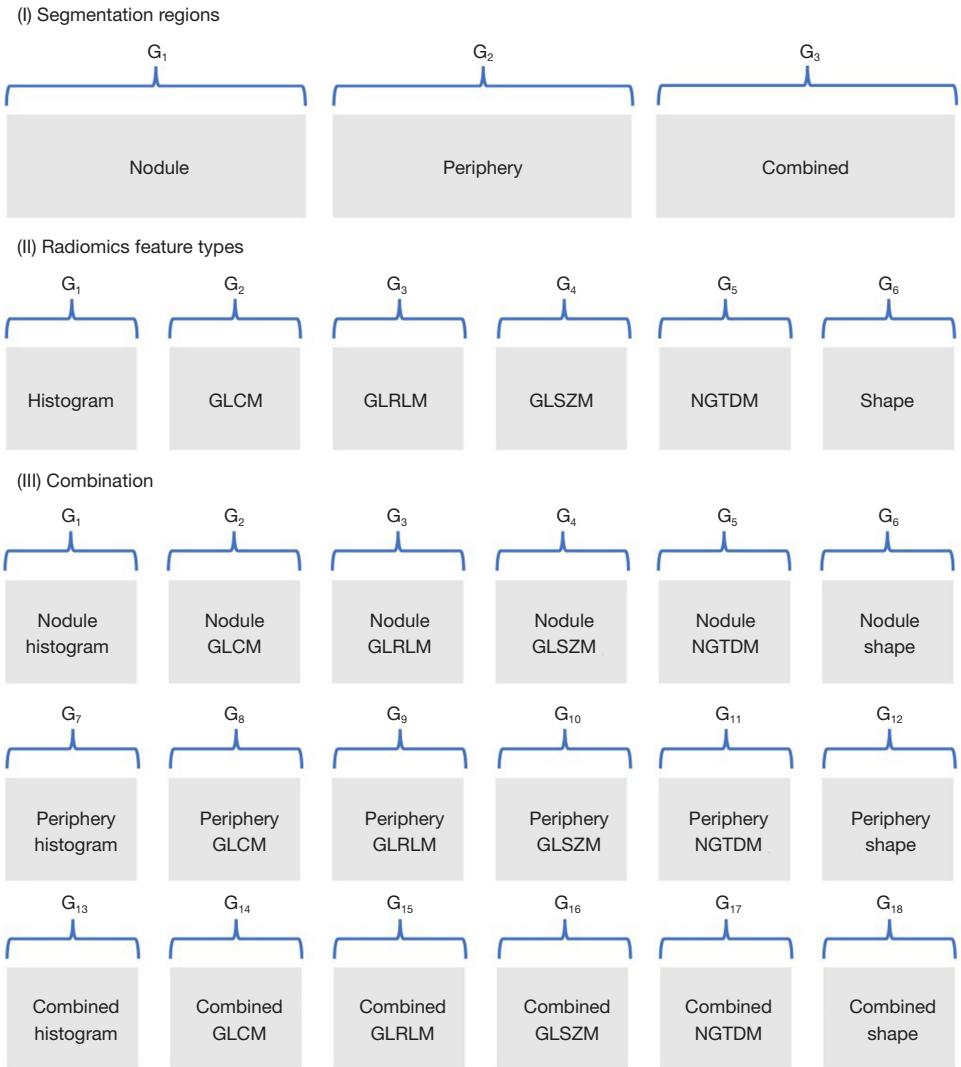

Figure 3 Group structure of radiomics feature groups for the lung adenocarcinomas study. Depicting the organization of features into three main levels: (I) segmentation regions; (II) radiomics feature types; (III) combination: combinations of the segmentation regions and radiomics feature types. GLCM, gray level co-occurrence matrix; GLRLM, gray level run length matrix; GLSZM, gray level size zone matrix; NGTDM, neighboring gray tone difference matrix.

refine our group penalty models, we classified each radiomic feature into one of four established types according to their mathematical and computational definitions: histogram, co-occurrence matrix (CM), run-length matrix (RLM), and morphology. This detailed categorization allowed for a more nuanced analysis of radiomic features. (III) Combination: this resulted in a comprehensive arrangement of ten distinct groups: three groups for each imaging techniques (T1C, ADC, and FA) combined with the three radiomics feature types (histogram, CM, and RLM) and one additional group for morphology. *Figure 4* and Table S5 present a summary of these groups, where the numbers in each cell indicate the count of features belonging to that specific group.

Statistical analysis

In our study, we evaluated model performance using nested 10-fold cross-validation (18). This method divides the entire dataset into 10 equal parts, each of which sequentially serves as the validation fold while the remaining nine parts collectively form the training fold. This process is repeated

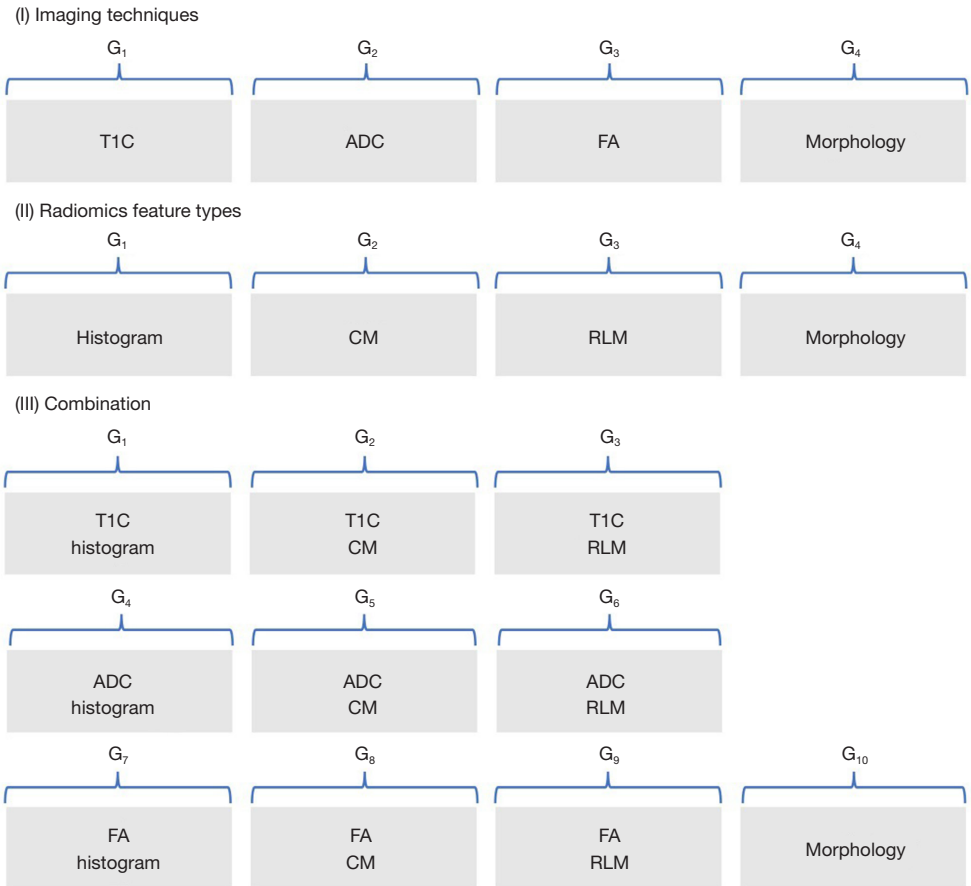

Figure 4 Group structure of radiomics feature groups for the meningiomas study. Depicting the organization of features into three main levels: (I) imaging techniques; (II) radiomics feature types; (III) combination: combinations of the imaging techniques and radiomics feature types, including morphology feature types. CM, co-occurrence matrix; RLM, run-length matrix; T1C, T1-weighted contrast; DTI, diffusion tensor imaging; FA, fractional anisotropy.

10 times, with each subset serving as the validation fold once. The process of the cross-validation is depicted in Figure S1. Within each training fold, we further performed a 10-fold cross-validation to determine the optimal value of the regularization parameter, λ, based on the "deviance" criterion. Specifically, we identified the λ_{min}, the value of λ that minimizes the deviance. For the group penalty models, we fixed the gamma values based on recommendations (11): γ=3 for group MCP and γ=4 for group SCAD, following the default suggestions by Fan and Li (19) and Zhang (20). For the graper method, model performance was evaluated using standard 10-fold cross-validation, reflecting its modeling strategy without the need for hyperparameter tuning. Features with posterior inclusion probability exceeding 0.5 were selected for the final model, following the median probability model criterion (21). ROSE-based oversampling was applied within each training fold to balance class distributions prior to model fitting, as shown in Figure S1.

The rad-score for each case was derived through a linear combination of selected features weighted by their coefficients, as determined by Lasso and group penalty models using the training folds. To assess the predictive performance of the rad-scores, we used the area under the receiver operating characteristic (ROC) curve (AUC) (22) as the primary metric. A high AUC value indicates the model's superior ability to accurately predict the presence or absence of a disease, including its progression stages. To provide a comprehensive assessment of the model's performance, we calculated and presented the mean and standard deviation (SD) of the AUC values obtained from the results of each

validation fold, thereby evaluating the model's consistency and reliability across different data subsets.

We employed the Jaccard Index (23) as a key metric for evaluating the stability of feature selection in our models. The Jaccard Index was calculated by dividing the number of common features selected in both folds (intersection) by the total number of unique features selected across both folds (union). For the 10-fold cross-validation, this computation was performed for every possible pair of folds, yielding 45 unique comparisons. The average Jaccard Index across these comparisons provided a quantitative measure of the consistency of the model in feature selection. This process is illustrated in Figure S2, which explains the Jaccard Index.

To assess the parsimony of our models, we examined the number and variability of selected features using the mean and SD across a 10-fold cross-validation. A more parsimonious model that selects fewer features is favored owing to its simplicity and interpretability.

To confirm the applicability in clinical practice, we selected the group penalty model that achieved the highest AUC for further evaluation. To compare these models, DeLong's test for two correlated ROC curves (24) was performed.

To evaluate model calibration, we calculated Brier scores (mean squared error between predicted probabilities and actual outcomes) and generated calibration curves comparing predicted probabilities against observed frequencies (25).

Statistical analysis was performed using the statistical software R version 4.3.1 (R Foundation for Statistical Computing, RRID:SCR_001905) (26). The Lasso analysis was conducted using the "glmnet" package (version 4.1.8) (27). The group Lasso, group MCP, and group SCAD analyses were conducted using the "grpreg" package (version 3.5.0) (11) while the graper analysis was implemented with the "graper" package (version 1.20.0) (12). The ROSE method was implemented using the "ROSE" package (version 0.0.4) (17).

Results

Cross-validation performance in lung adenocarcinomas study

In the STAS prediction study, the group SCAD model employing prior group selection (III) Combination achieved the highest AUC of 0.804 (SD =0.056). It exhibited a Jaccard Index of 0.613 (SD =0.293) and the Brier score of 0.167 (SD =0.035), while selecting an average of 29.6 features (SD =17.431). The Lasso model achieved an AUC of 0.776 (SD =0.059), a Jaccard Index of 0.503 (SD =0.084), and selected an average of 16.3 features (SD =2.669). The results are summarized in *Table 1*. To provide a detailed visualization of the groups of features selected by the group penalty methods in each fold, heatmaps have been organized in Figures S3-S5.

Cross-validation performance in meningiomas study

In the meningioma grading study, the group Lasso model employing the prior group selection (I) Imaging techniques achieved the highest AUC of 0.816 (SD =0.143). It also exhibited a Jaccard Index of 0.7 (SD =0.39) and the Brier score of 0.139 (SD =0.043), while selecting an average of 14 features (SD =12.649). The Lasso model achieved an AUC of 0.743 (SD =0.223), a Jaccard Index of 0.41 (SD =0.183), and selected an average of 4.9 features (SD =1.595). These findings are summarized in *Table 2*. To provide a detailed visualization of the groups of features selected by the group penalty methods in each fold, heatmaps have been organized in Figures S6-S8.

Application

Lung adenocarcinomas study

Based on the cross-validation results, we applied the best-performing group SCAD [(III) Combination] model and Lasso to an independent test set for validation. The Lasso model achieved an AUC of 0.877 [95% confidence interval (CI): 0.815–0.939], while the group SCAD [(III) Combination] model achieved an AUC of 0.874 (95% CI: 0.814–0.934). The test results showed no significant difference between the AUCs of the two models (P=0.896), with a 95% CI for the difference in AUCs ranging from −0.034 to 0.039. This suggests that both models have comparable performance in terms of AUC.

Unlike Lasso, which selects features individually across multiple groups—potentially leading to fragmented interpretations—group SCAD focuses on the "nodule shape" and "nodule + peripheral histogram" feature groups. These results are summarized in *Table 3*.

Meningiomas study

Based on the cross-validation results, we applied the best-performing group Lasso [(I) Imaging techniques] model and Lasso to an independent test set for validation. The

Table 1 Cross-validation performance metrics and number of selected features in lung adenocarcinomas study

Performance metrics	Lasso	Group Lasso	Group MCP	Group SCAD	graper
AUC					
(I) Segmentation regions	0.776 (0.059)	0.779 (0.078)*	0.773 (0.089)	0.775 (0.072)	0.768 (0.048)
(II) Radiomics feature types		0.790 (0.070)*	0.781 (0.066)*	0.779 (0.066)*	0.759 (0.057)
(III) Combination		0.790 (0.060)*	0.798 (0.063)*	0.804 (0.056)*	0.758 (0.058)
Jaccard Index					
(I) Segmentation regions	0.503 (0.084)	1 (0)*	0.533 (0.505)*	0.933 (0.135)*	0.157 (0.281)
(II) Radiomics feature types		0.787 (0.268)*	1 (0)*	0.651 (0.215)*	0.088 (0.146)
(III) Combination		0.644 (0.159)*	0.957 (0.088)*	0.613 (0.293)*	0.196 (0.139)
Brier score					
(I) Segmentation regions	0.172 (0.026)	0.179 (0.028)	0.186 (0.027)	0.180 (0.030)	0.195 (0.016)
(II) Radiomics feature types		0.171 (0.032)*	0.173 (0.030)	0.174 (0.032)	0.198 (0.019)
(III) Combination		0.172 (0.029)	0.169 (0.034)*	0.167 (0.035)*	0.198 (0.019)
The number of selected feature					
(I) Segmentation regions	16.3 (2.669)	279 (0)	93 (0)	269.7 (29.409)	53.5 (42.766)
(II) Radiomics feature types		229.2 (60.666)	69 (0)	167.4 (57.355)	28.4 (23.372)
(III) Combination		155.6 (42.594)	18.5 (1.581)	29.6 (17.430)	33.4 (13.558)

Values are presented as mean (standard deviation) for each performance metric. Improvements over Lasso are indicated with asterisks (*). The grouping strategy is threefold: (I) segmentation regions: grouping of features into three categories: nodule, peripheral, and 'nodule + peripheral' region; (II) radiomics feature types: categorization into six types (histogram, GLCM, GLRLM, GLSZM, NGTDM, and shape descriptors); and (III) combination: combinations of the segmentation regions and radiomics feature types. AUC, area under the curve; GLCM, gray level co-occurrence matrix; GLRLM, gray level run length matrix; GLSZM, gray level size zone matrix; graper, adaptive penalization regression with external covariates using variational bayes; Lasso, least absolute shrinkage and selection operator; MCP, minimax concave penalty; NGTDM, neighboring gray tone difference matrix; SCAD, smoothly clipped absolute deviation.

Lasso model achieved an AUC of 0.835 (95% CI: 0.706–0.965), while the group Lasso [(I) Imaging techniques] model achieved an AUC of 0.877 (95% CI: 0.779–0.974). The results showed no significant difference between the AUCs of the two models (P=0.391). The 95% CI for the difference in AUCs ranged from –0.137 to 0.054, indicating comparable performance between the two models.

Unlike Lasso, which selects features individually across multiple groups, group Lasso focuses on the morphology feature group. These results are summarized in *Table 4*.

Discussion

Our findings highlight the superiority of strict group penalty (group Lasso, group MCP and group SCAD) model over Lasso and graper in terms of both AUC and Jaccard Index, with only a few exceptions. This improvement highlights the benefits of integrating the group structures of features, which appear to enhance the models' discriminatory capability and consistency in feature selection. Moreover, group penalty models achieved similar or lower Brier scores compared to Lasso, suggesting that the integration of group structures does not compromise the predictive performance of the models. We observed a correlation between the size of the feature group and the Jaccard Index, particularly evident when comparing (I) the smaller groupings with (III) the extensive groupings. This suggests that models incorporating broader grouping of features tend to exhibit more stable feature selection across different cross-validation folds. Despite Lasso models selecting fewer features on average, their stability in feature selection is less consistent. Additionally, although graper is flexible in incorporating group information as a prior, it showed lower Jaccard Index values than Lasso, indicating

Table 2 Cross-validation performance metrics and number of selected features in meningiomas study

Performance metrics	Lasso	Group Lasso	Group MCP	Group SCAD	graper
AUC					
(I) Imaging techniques	0.743 (0.223)	0.816 (0.143)*	0.806 (0.147)*	0.808 (0.144)*	0.804 (0.164)*
(II) Radiomics feature types		0.808 (0.143)*	0.806 (0.147)*	0.808 (0.144)*	0.781 (0.175)*
(III) Combination		0.768 (0.151)*	0.776 (0.149)*	0.763 (0.154)*	0.773 (0.181)*
Jaccard Index					
(I) Imaging technique	0.41 (0.183)	0.7 (0.390)*	1 (0)*	1 (0)*	0.219 (0.189)
(II) Radiomics feature types		0.631 (0.349)*	1 (0)*	0.862 (0.280)*	0.239 (0.184)
(III) Combination		0.421 (0.239)*	1 (0)*	0.539 (0.312)*	0.190 (0.110)
Brier score					
(I) Imaging techniques	0.144 (0.051)	0.139 (0.043)*	0.134 (0.047)*	0.138 (0.043)*	0.180 (0.052)
(II) Radiomics feature types		0.140 (0.042)*	0.134 (0.047)*	0.139 (0.042)*	0.181 (0.051)
(III) Combination		0.148 (0.054)	0.143 (0.049)*	0.148 (0.050)	0.182 (0.057)
The number of selected feature					
(I) Imaging technique	4.9 (1.595)	14 (12.649)	8 (0)	8 (0)	26.4 (12.677)
(II) Radiomics feature types		18.8 (9.295)	8 (0)	9.8 (5.692)	31 (27.564)
(III) Combination		32.1 (21.533)	6 (0)	19.2 (11.003)	22.8 (7.757)

Values are presented as mean (standard deviation) for each performance metric. Improvements over Lasso are indicated with asterisks (*). The grouping strategy is threefold: (I) imaging techniques: grouping of features into three categories (T1C, ADC, and FA) and morphology; (II) radiomics feature types: categorization into four types (histogram, CM, RLM, and morphology); (III) combination: combinations of the imaging techniques and radiomics feature types, including morphology feature types. AUC, area under the curve; ADC, apparent diffusion coefficient; CM, cooccurrence matrix; FA, fractional anisotropy; graper, adaptive penalization regression with external covariates using variational bayes; Lasso, least absolute shrinkage and selection operator; MCP, minimax concave penalty; RLM, run-length matrix; SCAD, smoothly clipped absolute deviation; T1C, T1-weighted contrast.

less stable feature selection, and exhibited the highest Brier scores among all methods, suggesting poor calibration. Consistent feature selection in radiomics holds multifaceted importance, primarily because it enhances the reliability, interpretability, and generalizability of predictive models, which are crucial for clinical applicability. Therefore, this consistency is foundational for advancing personalized medicine by translating complex radiomics data into reliable, understandable, and widely applicable diagnostic and prognostic tools (23,28).

Further analysis through train-test split comparisons in a clinical setting yielded encouraging results. The group penalty models demonstrated comparable predictive performance to Lasso, with no statistically significant differences in their AUC values for both lung adenocarcinomas and meningiomas studies. This finding is particularly promising as it shows that group-based feature selection can maintain high predictive accuracy while offering additional benefits. The group-based feature selection significantly contributed to clinically relevant insights. For instance, in the lung adenocarcinoma study, the group SCAD model focused on the "nodule shape" and "nodule + peripheral histogram" feature groups, highlighting their importance for STAS prediction. This contrasts sharply with the Lasso model, which selected features individually across multiple groups. The Lasso approach often results in the selection of a scattered set of features from various categories such as histogram, GLCM, GLSZM, and shape descriptors, leading to fragmented interpretations. This fragmentation can hinder the ability to draw comprehensive clinical insights, as the model does not consider the contextual synergy among the selected features. The group SCAD model's emphasis on specific, clinically meaningful feature groups enables a more coherent

Table 3 Result of the clinical applications for lung adenocarcinomas study

Model	AUC (95% CI)	Selected features		
		Segmentation regions	Radiomics feature types	Feature name
Lasso	0.877 (0.815–0.939)	Nodule	Histogram	Mean
			GLCM	Imc2, InverseVariance, MaximumProbability
			GLSZM	SizeZoneNonUniformityNormalized
			Shape	Flatness
		Periphery	GLCM	ClusterShade, JointEntropy
			GLSZM	SmallAreaLowGrayLevelEmphasis
		Combined	Histogram	90Percentile, InterquartileRange, Range
			GLCM	DifferenceVariance
			GLSZM	LargeAreaLowGrayLevelEmphasis
			Shape	Flatness
Group SCAD [(III) Combination]	0.874 (0.814–0.934)	Nodule	Shape	Group selection
		Combined	Histogram	

AUC, area under the curve; CI, confidence interval; GLCM, gray level co-occurrence matrix; GLRLM, gray level run length matrix; GLSZM, gray level size zone matrix; Lasso, least absolute shrinkage and selection operator; NGTDM, neighboring gray tone difference matrix; SCAD, smoothly clipped absolute deviation.

Table 4 Result of the clinical applications for meningiomas study

Model	AUC (95% CI)	Selected features		
		Imaging techniques	Radiomics feature types	Feature name
Lasso	0.835 (0.706–0.965)	T1C	Histogram	Entropy
			RLM	Run-length nonuniformity
		ADC	CM	Variance
Group Lasso [(I) imaging techniques]	0.877 (0.779–0.974)	Morphology		Group selection

ADC, apparent diffusion coefficient; AUC, area under the curve; CI, confidence interval; CM, co-occurrence matrix; RLM, run-length matrix; T1C, T1-weighted contrast.

interpretation. By focusing on the "nodule shape" and "nodule + peripheral histogram" groups, the group SCAD model integrates relevant radiomic features that collectively contribute to the biological understanding of STAS in lung adenocarcinomas. This approach not only improves interpretability but also aligns better with clinical reasoning, which often considers related features in conjunction rather than in isolation. In the meningioma study, the group Lasso model's selection of the morphology group suggested that these features alone might be sufficient for accurate tumor grading, potentially simplifying preoperative evaluations. This is particularly significant as it implies that the detailed imaging techniques such as T1C and ADC, along with their respective complex feature types like histogram and co-occurrence matrix, might not be necessary for effective tumor grading. By focusing solely on morphology, which includes features such as volume, surface area, and sphericity, the group Lasso model simplifies the evaluation process, reducing the need for multiple imaging modalities and extensive feature extraction processes. This streamlined approach not only saves time and resources but also provides a more integrated understanding of tumor

properties, enhancing the clinical utility and interpretability of the model.

However, our study has some limitations. The artificial grouping of features within our methodology may not fully capture the inherent structural complexities of the data, suggesting the need for further exploration of methodologies representing natural feature groupings more accurately.

While Lasso tends to under-select correlated features by choosing only one representative from a correlated group, group penalty models exhibit the opposite tendency—they may over-select features by including all features within a selected group. For instance, in our lung adenocarcinoma study, selecting the segmentation-based grouping could result in all 279 features from a single region being included in the model. This "all-or-nothing" property inherent to strict group penalty methods (group Lasso, group MCP, and group SCAD) may lead to potential overfitting, reduced generalizability, and diminished interpretability when large feature groups are selected in their entirety. Future studies should consider hybrid approaches, such as sparse group Lasso, which applies both group-level and within-group sparsity penalties to select relevant groups while simultaneously identifying the most informative individual features within those groups.

Furthermore, it is important to clarify the interpretation of the stability metrics reported in this study. The Jaccard Index in our group penalty models primarily reflects the consistency of group-level selection across cross-validation folds, rather than the precise selection of specific individual features based on their information value. When a group penalty model achieves high stability (e.g., Jaccard Index approaching 1.0), this indicates that the same feature groups are consistently selected, but all features within those groups are included together regardless of their individual predictive contributions. Therefore, "perfect stability" in group penalty models should be interpreted as consistent group selection rather than optimal individual feature identification. This distinction is crucial for clinical applications where understanding the contribution of specific features may be important for biological interpretation.

In addition, model calibration analysis revealed important considerations for clinical implementation (Figures S9-S11). The graper model consistently showed poor calibration compared to other methods, with systematic overestimation of predicted probabilities particularly evident in the calibration curves where graper consistently fell below the diagonal line. This is likely attributable to the ROSE oversampling procedure applied exclusively to this Bayesian method. Although ROSE effectively addressed class imbalance to improve discrimination, it may have distorted the probability estimates by altering the underlying class distribution (29). Furthermore, calibration curves exhibited substantial fluctuations in high-probability regions (predicted probability >0.7) due to limited sample sizes in these bins, a common challenge in medical imaging studies with moderate sample sizes (30).

Although we validated our models using independent test sets, these sets were obtained from the same institutions but from different time periods (temporal validation). This approach, while commonly used in medical imaging research due to practical constraints of multi-institutional data sharing, may not fully capture the variability encountered across different clinical sites, imaging protocols, and patient populations. Future multi-institutional external validation studies are warranted to confirm the generalizability of our findings.

Conclusions

In conclusion, our study strongly advocates for the use of group penalty models in radiomics to enhance the accuracy and reliability of predictive models in medical imaging. These models significantly improve analytical stability and provide a deeper understanding of complex biological and clinical interactions that traditional methods often overlook. By leveraging the inherent group structures within features, group penalty models enhance the interpretability of radiomic analyses. This improved interpretability is crucial in clinical settings, where it can profoundly influence diagnostic, prognostic, and therapeutic decisions. Integrating group penalty models into radiomics not only facilitates more robust and consistent feature selection but also aligns the analysis more closely with clinical reasoning, thereby advancing the field of personalized medicine.

Acknowledgments

None.

Footnote

Reporting Checklist: The authors have completed the TRIPOD+AI reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-2025-1951/rc

Data Sharing Statement: Available at https://qims.

amegroups.com/article/view/10.21037/qims-2025-1951/dss

Funding: This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. NRF-2021R1I1A1A01059893); and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1A2C1091488).

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-2025-1951/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. This study was conducted in accordance with the Declaration of Helsinki and its subsequent amendments.

References

1. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, Sanduleanu S, Larue RTHM, Even AJG, Jochems A, van Wijk Y, Woodruff H, van Soest J, Lustberg T, Roelofs E, van Elmpt W, Dekker A, Mottaghy FM, Wildberger JE, Walsh S. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 2017;14:749-62.
2. Robert T. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol 1996;58:267-88.
3. Huang Y, Liu Z, He L, Chen X, Pan D, Ma Z, Liang C, Tian J, Liang C. Radiomics Signature: A Potential Biomarker for the Prediction of Disease-Free Survival in Early-Stage (I or II) Non-Small Cell Lung Cancer. Radiology 2016;281:947-57.
4. Zheng BH, Liu LZ, Zhang ZZ, Shi JY, Dong LQ, Tian LY, Ding ZB, Ji Y, Rao SX, Zhou J, Fan J, Wang XY, Gao Q. Radiomics score: a potential prognostic imaging feature for postoperative survival of solitary HCC patients. BMC Cancer 2018;18:1148.
5. Xie T, Wang X, Li M, Tong T, Yu X, Zhou Z. Pancreatic ductal adenocarcinoma: a radiomics nomogram outperforms clinical model and TNM staging for survival estimation after curative resection. Eur Radiol 2020;30:2513-24.
6. Rahnenführer J, De Bin R, Benner A, Ambrogi F, Lusa L, Boulesteix AL, Migliavacca E, Binder H, Michiels S, Sauerbrei W, McShane L; for topic group "High-dimensional data" (TG9) of the STRATOS initiative. Statistical analysis of high-dimensional biomedical data: a gentle introduction to analytical goals, common approaches and challenges. BMC Med 2023;21:182.
7. Hastie T, Tibshirani R, Wainwright M. Statistical learning with sparsity: The Lasso and Generalizations. 1st edition. New York: Chapman and Hall/CRC; 2015.
8. Roberts S, Nowak G. Stabilizing the lasso against cross-validation variability. Computational Statistics & Data Analysis 2014;70:198-211.
9. Orton MR, Hann E, Doran SJ, Shepherd STC, Ap Dafydd D, Spencer CE, López JI, Albarrán-Artahona V, Comito F, Warren H, Shur J, Messiou C, Larkin J, Turajlic S; Koh DM. Interpretability of radiomics models is improved when using feature group selection strategies for predicting molecular and clinical targets in clear-cell renal cell carcinoma: insights from the TRACERx Renal study. Cancer Imaging 2023;23:76.
10. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. J R Stat Soc Ser B Stat Methodol 2006;68:49-67.
11. Breheny P, Huang J. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. Stat Comput 2015;25:173-87.
12. Velten B, Huber W. Adaptive penalization in high-dimensional regression and classification with external covariates using variational Bayes. Biostatistics 2021;22:348-64.
13. Huang J, Zhang T. The benefit of group sparsity. Ann Stat 2010;38:1978-2004.
14. Huang J, Breheny P, Ma S. A Selective Review of Group Selection in High-Dimensional Models. 2012;27:10.1214/12-STS392.
15. Suh YJ, Han K, Kwon Y, Kim H, Lee S, Hwang SH, Kim MH, Shin HJ, Lee CY, Shim HS. Computed Tomography

Radiomics for Preoperative Prediction of Spread Through Air Spaces in the Early Stage of Surgically Resected Lung Adenocarcinomas. Yonsei Med J 2024;65:163-73.
16. Park YW, Oh J, You SC, Han K, Ahn SS, Choi YS, Chang JH, Kim SH, Lee SK. Radiomics and machine learning may accurately predict the grade and histological subtype in meningiomas using conventional and diffusion tensor imaging. Eur Radiol 2019;29:4068-76.
17. Lunardon N, Menardi G, Torelli N. ROSE: a package for binary imbalanced learning. R J 2014;6:79-89.
18. Borstelmann SM. Machine Learning Principles for Radiology Investigators. Acad Radiol 2020;27:13-25.
19. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 2001;96:1348-60.
20. Zhang CH. Nearly unbiased variable selection under minimax concave penalty. Ann Stat 2010;38:894-942.
21. Barbieri MM, Berger JO. Optimal predictive model selection. Ann Stat 2004;32:870-97.
22. Obuchowski NA. Receiver operating characteristic curves and their use in radiology. Radiology 2003;229:3-8.
23. Khaire UM, Dhanalakshmi R. Stability of feature selection algorithm: A review. J King Saud Univ Comput Inf Sci 2022;34:1060-73.
24. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837-45.
25. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21:128-38.
26. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2013.
27. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 2010;33:1-22.
28. Park JE, Park SY, Kim HJ, Kim HS. Reproducibility and Generalizability in Radiomics Modeling: Possible Strategies in Radiologic and Statistical Perspectives. Korean J Radiol 2019;20:1124-37.
29. van den Goorbergh R, van Smeden M, Timmerman D, Van Calster B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. J Am Med Inform Assoc 2022;29:1525-34.
30. Riley RD, Ensor J, Snell KIE, Archer L, Whittle R, Dhiman P, Alderman J, Liu X, Kirton L, Manson-Whitton J, van Smeden M, Moons KG, Nirantharakumar K, Cazier JB, Denniston AK, Van Calster B, Collins GS. Importance of sample size on the quality and utility of AI-based prediction models for healthcare. Lancet Digit Health 2025;7:100857.

Cite this article as: Kwon Y, Suh YJ, Park YW, Ahn SS, Han K, Ha MJ. Enhancing cancer prognostics with group penalty models: a comparative study on radiomics feature selection in lung adenocarcinomas and meningiomas. Quant Imaging Med Surg 2026;16(2):118. doi: 10.21037/qims-2025-1951

Supplementary

Appendix 1

Least absolute shrinkage and selection operator (Lasso)

Lasso is a popular technique for feature selection and regularization in regression models (2). It addresses the problem of overfitting by imposing a penalty on the absolute size of the regression coefficients. The Lasso optimization problem is defined as follows:

$$\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i x_i^T \beta\right)\right) + \lambda \parallel \beta \parallel_1 \right\} \quad [1]$$

where, $y = (y_1, y_2, \ldots, y_n)^T$ is the response vector, $X = (x_1, x_2, \ldots, x_n)^T$ is the $n \times p$ design matrix with x_i being the i-th observation, $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is the vector of regression coefficients, $\lambda \geq 0$ is the regularization parameter that controls the shrinkage of the coefficients. The Lasso penalty $\lambda \parallel \beta \parallel_1$ encourages sparsity in the coefficients, potentially setting some of them exactly to zero. This makes Lasso an effective tool for feature selection.

Generalized penalty form

Group penalty models fit models that fall into the penalized likelihood framework, in which we estimate β by minimizing the objective function:

$$Q(\beta \mid \mathrm{X}, \mathrm{y}) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp\left(-y_i x_i^T \beta\right)) + P\left(\beta_g\right) \quad [2]$$

where $\frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i x_i^T \beta\right)\right)$ is the loss (deviance) and $P(\beta_g)$ is the penalty. The following notation is used throughout (recall that the design matrix X is decomposed into groups X_1, X_2, ...): β denotes the entire vector of regression coefficients, β_g denotes the vector of regression coefficients corresponding to the g-th group, β_{gj} denotes the j-th regression coefficient in the g-th group, $\|\beta_g\|_2$ denotes the Euclidean (L_2) norm of β_g, $\|\beta_g\|_1$ denotes the L_1 norm of β_g.

The specific forms of $P(\beta_g)$ for group Lasso (10,11), group minimax concave penalty (MCP) (11), and group smoothly clipped absolute deviation (SCAD) (11) are as follows:

Group Lasso

For group Lasso, the penalty function $P(\beta_g)$ is the ℓ_2-norm of the coefficients in group g:

$$P_\lambda\left(\beta_g\right) = \lambda \left\|\beta_g\right\|_2 = \lambda \sqrt{\sum_{j \in G_g} \beta_j^2} \quad [3]$$

where λ is a tuning parameter. This ensures that either all features in a group are selected or none at all.

Group MCP

For group MCP, the penalty function $P_{\lambda,\gamma}(\theta)$ is defined as:

$$P_{\lambda,\gamma}(\theta) = \begin{cases} \lambda\theta - \frac{\theta^2}{2\gamma} & \text{if } \theta \leq \gamma\lambda, \\ \frac{1}{2}\gamma\lambda^2 & \text{if } \theta > \gamma\lambda, \end{cases} \quad [4]$$

where $\theta = \|\beta_g\|_2$ and λ, γ is a tuning parameter. This method encourages group-wise sparsity and reduces bias for large coefficients.

https://dx.doi.org/10.21037/qims-2025-1951

Group SCAD

For group SCAD, the penalty function $P_{\lambda,\gamma}(\theta)$ is defined as:

$$P_{\lambda,\gamma}(\theta) = \begin{cases} \lambda\theta & \text{if } \theta \le \lambda \\ \frac{\gamma\lambda\theta - 0.5(\theta^2+\lambda^2)}{\gamma-1} & \text{if } \lambda < \theta \le \gamma\lambda \\ \frac{\lambda^2(\gamma^2-1)}{2(\gamma-1)} & \text{if } \theta > \gamma\lambda \end{cases} \quad [5]$$

where $\theta = \|\beta_g\|_2$ and λ, γ is a tuning parameter. This reduces the bias in the estimation of large coefficients while encouraging group-wise sparsity.

Properties of group penalty methods

The choice among these group penalty methods involves important trade-offs between statistical properties and computational considerations. group Lasso maintains convexity, ensuring a unique global minimum and computational efficiency, but may introduce substantial bias for large coefficients due to its constant shrinkage rate. In contrast, group MCP and group SCAD employ non-convex penalties that nearly eliminate bias for large coefficients while maintaining the ability to produce sparse solutions.

The non-convex penalties achieve the oracle property under certain regularity conditions, meaning they perform asymptotically as well as if the true model were known in advance (11,19). Specifically, for coefficients with $\|\beta_g\|_2 > \gamma\lambda$, both group MCP and group SCAD apply minimal or no penalty, preserving the magnitude of truly important feature groups. This property is particularly valuable in radiomics applications where certain feature groups may contain strong predictive signals that should not be attenuated.

The parameter γ plays a crucial role in controlling the concavity of the penalty functions. For group MCP, γ determines the point at which the penalty becomes constant, while for group SCAD, it controls the smoothness of the transition between penalty regions.

From a computational perspective, while group Lasso can be solved efficiently using standard convex optimization algorithms, group MCP and group SCAD require specialized algorithms due to their non-convexity. We employ the group coordinate descent algorithm implemented in the grpreg package (11), which has been shown to converge reliably to good local minima when initialized appropriately. Despite the additional computational complexity, the potential for improved feature selection stability and reduced bias often justifies the use of these non-convex penalties in high-stakes medical applications.

graper

graper (12) is a Bayesian approach to group penalty regression that adaptively learns group-specific penalty parameters from the data. Unlike group Lasso, group MCP, and group SCAD which apply fixed penalty parameters across all groups, graper treats the penalty parameters as random variables to be inferred, allowing different groups to have different levels of regularization based on their importance.

graper employs a hierarchical Bayesian model with spike-and-slab priors. The spike-and-slab prior is a powerful Bayesian tool that performs feature selection and coefficient estimation simultaneously. The "spike" component (s_j) acts as a binary selector determining whether a feature is included (s_j =1) or excluded (s_j =0), effectively setting β_j =0 when excluded. The "slab" component (β_j) represents the continuous distribution of the actual coefficient magnitude for included features. This combination elegantly unifies the benefits of ridge regression (when π=1) and Lasso-type sparsity(when π<1), enabling simultaneous shrinkage and selection.

For logistic regression, the model uses:

$$P(y_i = 1|\mathrm{x}_i) = \sigma(\mathrm{x}_i^T\beta) = \frac{1}{1+\exp(-\mathrm{x}_i^T\beta)} \quad [6]$$

https://dx.doi.org/10.21037/qims-2025-1951

The coefficients are decomposed as $\beta_j = s_j b_j$ with group-specific priors:

$$b_j|\tau_{g(j)} \sim N(0, \tau_{g(j)}^{-1}),\ s_j|\pi_{g(j)} \sim \text{Bernoulli}(\pi_{g(j)}) \quad [7]$$

where $g(j)$ denotes the group membership of feature j, τ_k controls the shrinkage strength for group k with larger values indicating stronger shrinkage, and π_k controls the sparsity level for group k by determining the proportion of active features. The hyperparameters τ_k and π_k themselves follow conjugate priors: $\tau_k \sim Gamma(r_\tau, d_\tau)$ for group precisions and $\pi_k \sim Beta(r_\pi, d_\pi)$ for group sparsity levels, enabling closed-form updates in the variational inference framework. The implementation adopts weakly informative default settings based on empirical performance: $r_\tau = d_\tau = 0.001$ for the Gamma prior on group precisions and $r_\pi = d_\pi = 1$ for the Beta prior on group sparsity levels, corresponding to a uniform prior on the interval [0,1].

Instead of cross-validation to select penalty parameters, graper uses variational Bayes to simultaneously infer all parameters. The posterior distribution $p(\Theta|X, y)$ is approximated by minimizing the KL divergence $\text{KL}(q||p(\Theta|X, y))$ using a mean-field approximation:

$$q(\Theta) = \prod_{j=1}^{p} q(b_j, s_j) \cdot q(\tau) \cdot q(\pi) \quad [8]$$

Note that (b_j, s_j) are kept joint due to their strong dependencies, while other parameters are fully factorized for computational efficiency. This approach enables efficient coordinate ascent updates while maintaining the essential dependencies in the model.

graper exhibits several desirable properties that distinguish it from fixed-penalty methods. The method provides adaptive regularization by automatically learning which groups require stronger or weaker regularization. Groups with strong predictive signals receive less penalization through smaller values of τ_k, while irrelevant groups are heavily penalized with larger τ_k values. The learned hyperparameters offer automatic relevance determination with direct interpretability. The posterior mean $\hat{\tau}_k$ indicates the overall importance of group k, where smaller values signify greater importance. Similarly, $\hat{\pi}_k$ reveals the proportion of active features within the group, with larger values indicating denser groups, while $\hat{s}_j$ provides the posterior inclusion probability for individual feature j.

Despite the increased model complexity, graper maintains computational efficiency with $O(np)$ complexity per iteration through efficient coordinate ascent updates, making it scalable to high-dimensional problems typical in radiomics applications. The algorithm employs a maximum of 3,000 iterations with a convergence threshold of 0.01 based on the relative change in the evidence lower bound (ELBO), and features are standardized to unit variance by default to ensure comparability across different scales.

While group Lasso, MCP, and SCAD require careful tuning of penalty parameters through cross-validation, graper learns these automatically. This distinction becomes particularly advantageous when groups have heterogeneous characteristics, such as different scales or sparsity levels, which is common in radiomics where shape, texture, and intensity features naturally exhibit vastly different properties. The automatic learning also becomes crucial when the number of groups is large, making grid search computationally prohibitive.

In practical applications, several considerations enhance graper's performance. For classification tasks with imbalanced classes, integration with Random Over-Sampling Examples (ROSE) helps ensure that the adaptive penalty learning is not biased by class imbalance. The definition of feature groups significantly impacts results and should be guided by domain knowledge when available. The implementation has been optimized for radiomics applications while maintaining generality for other domains requiring group-structured regularization.

https://dx.doi.org/10.21037/qims-2025-1951

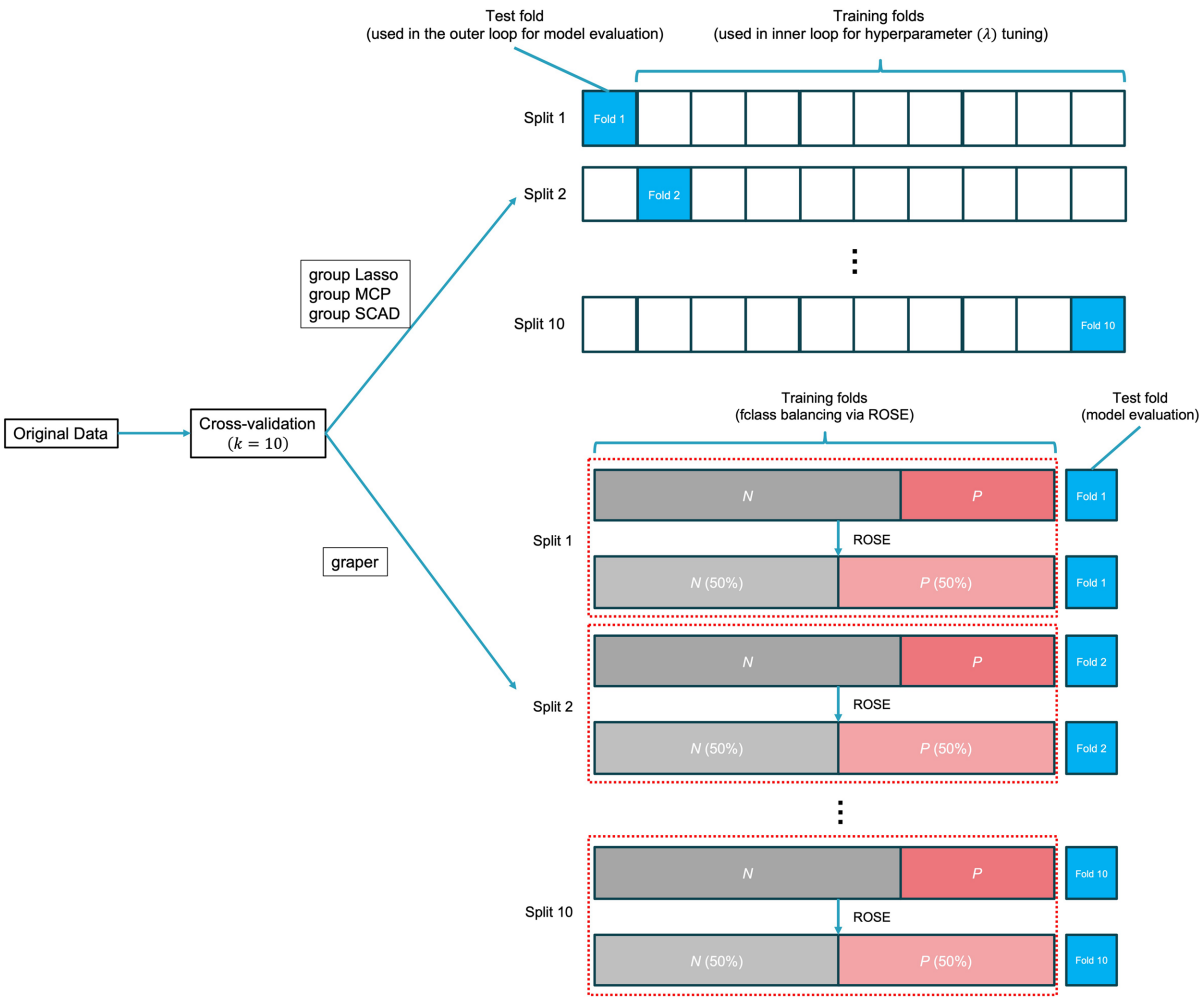

Figure S1 This diagram illustrates the nested 10-fold cross-validation framework used to evaluate the performance of group-regularized radiomics models and to tune the regularization parameter (λ). In the outer loop, the original dataset is split into 10 folds to assess model performance. Within each outer training fold, an inner 10-fold cross-validation is conducted to identify the optimal λ for the group Lasso, group MCP and group SCAD models. For the graper method, class balancing is performed within each outer training fold using the ROSE algorithm (shown in red), which generates a balanced dataset with equal numbers of negative (N) and positive (P) cases. Model performance is evaluated in the outer loop using the balanced dataset to assess the impact of oversampling on predictive accuracy. MCP, minimax concave penalty; ROSE, Random Over-Sampling Examples; SCAD, smoothly clipped absolute deviation.

https://dx.doi.org/10.21037/qims-2025-1951

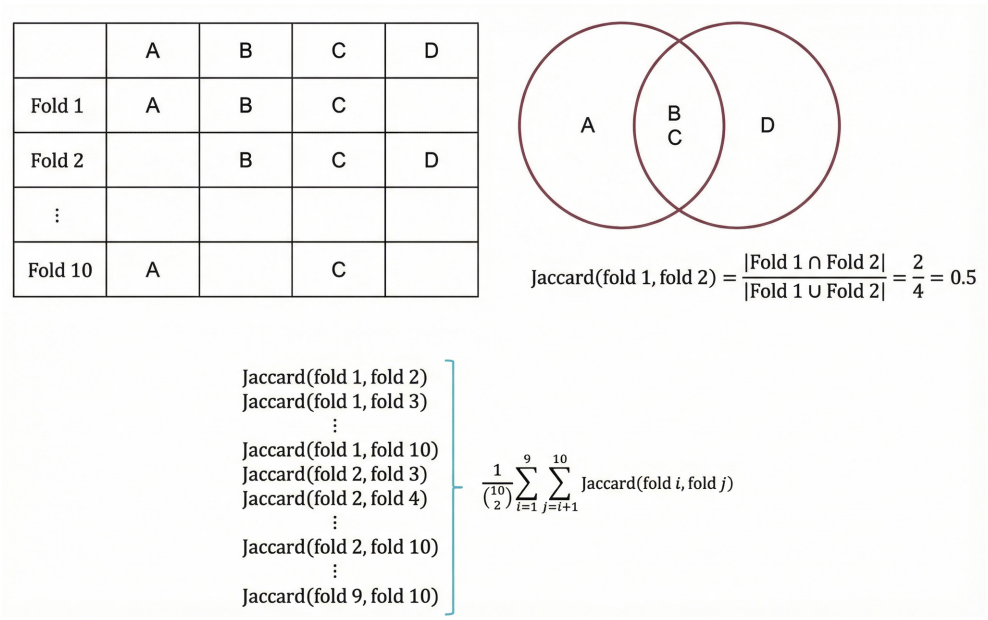

Figure S2 Illustration of the Jaccard Index calculation across 10 cross-validation folds, demonstrating the overlap of selected features between folds and the computation of feature selection stability.

https://dx.doi.org/10.21037/qims-2025-1951

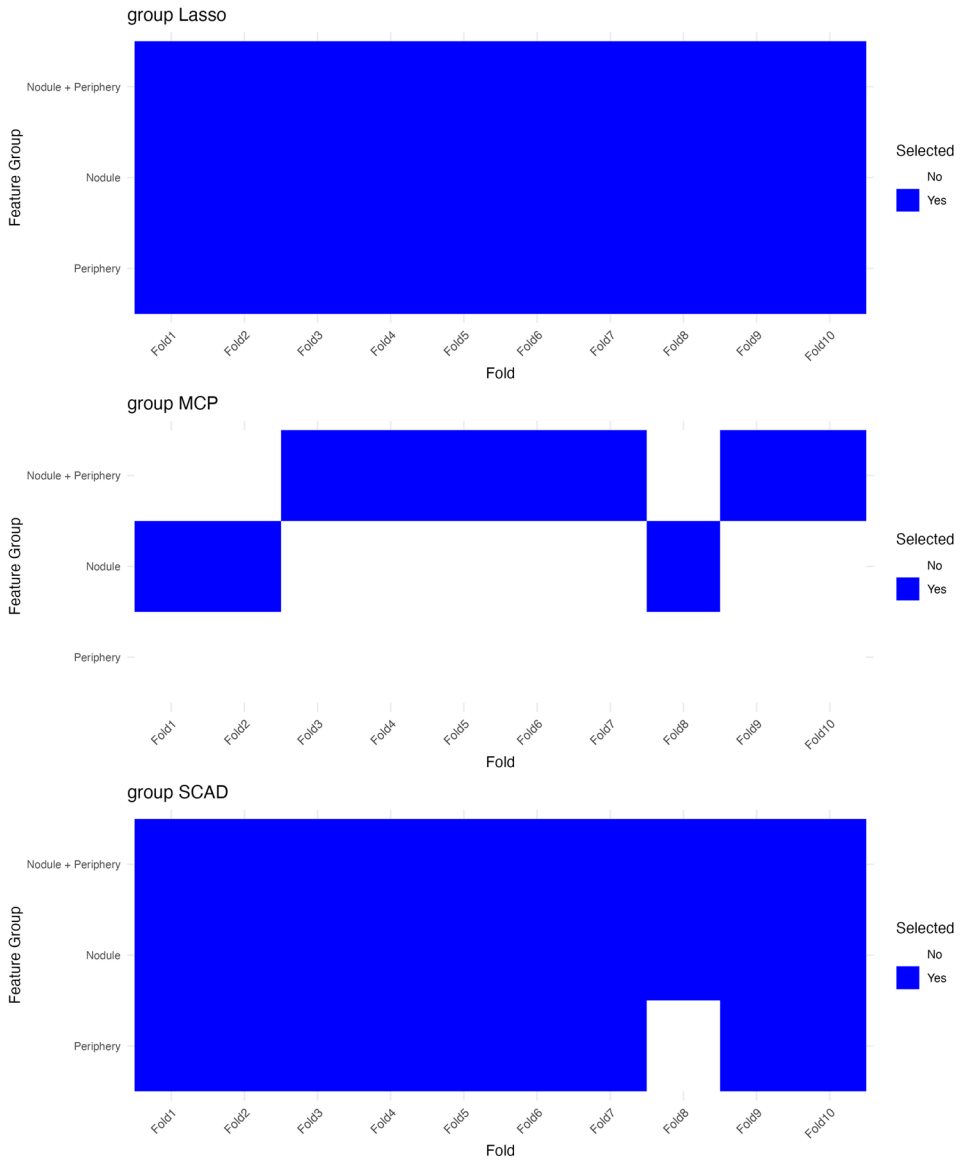

Figure S3 Heatmap of feature selection frequencies for group penalty models in the lung adenocarcinoma study using 10-fold cross-validation, grouped by (I) Segmentation regions. Each column represents one of the 10 folds, with blue cells indicating the feature group was selected in that fold and white cells indicating it was not. Lasso, least absolute shrinkage and selection operator; MCP, minimax concave penalty; SCAD, smoothly clipped absolute deviation.

https://dx.doi.org/10.21037/qims-2025-1951

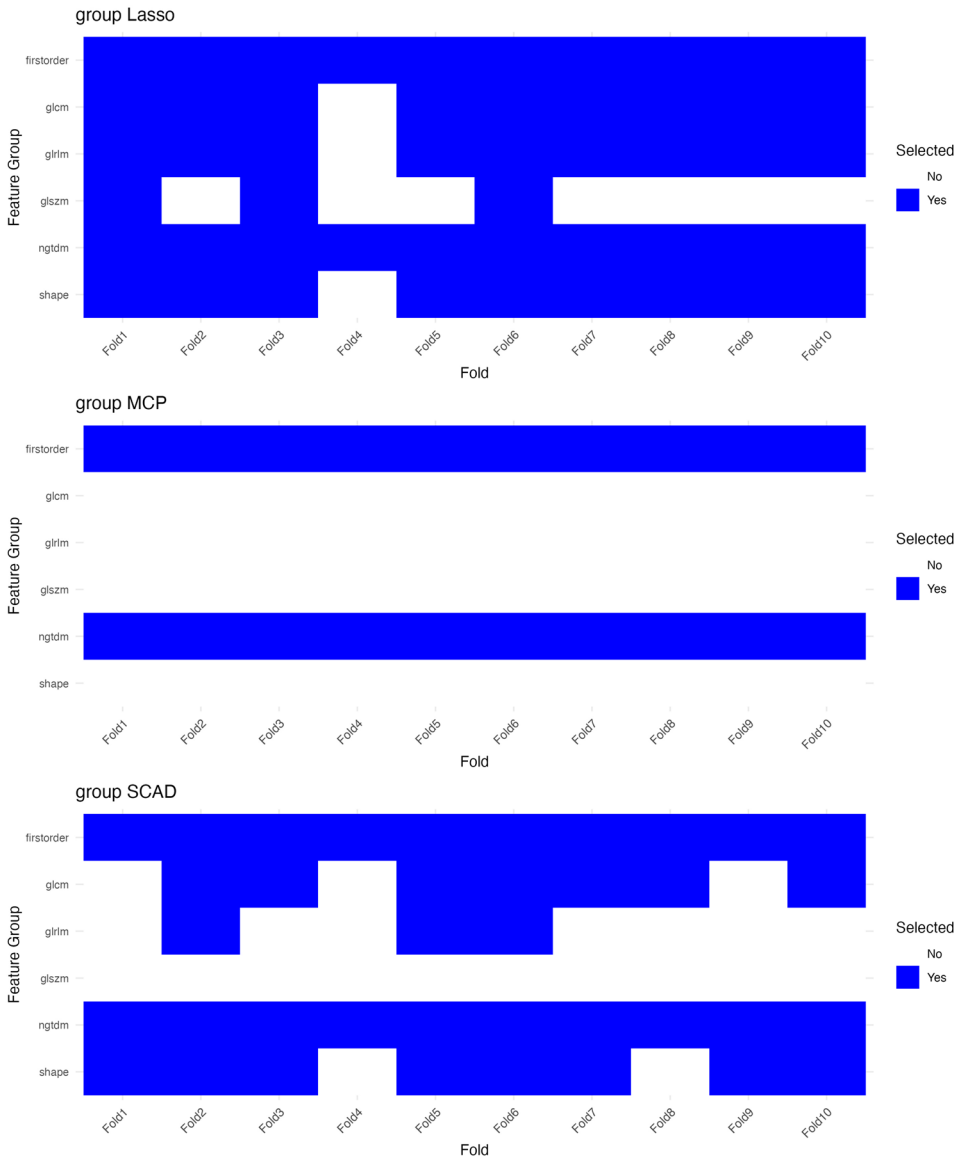

Figure S4 Heatmap of feature selection frequencies for group penalty models in the lung adenocarcinoma study using 10-fold cross-validation, grouped by (II) Radiomics feature types. Each column represents one of the 10 folds, with blue cells indicating the feature group was selected in that fold and white cells indicating it was not. GLCM, gray level co-occurrence matrix; GLRLM, gray level run length matrix; GLSZM, gray level size zone matrix; Lasso, least absolute shrinkage and selection operator; MCP, minimax concave penalty; NGTDM, neighboring gray tone difference matrix; SCAD, smoothly clipped absolute deviation.

https://dx.doi.org/10.21037/qims-2025-1951

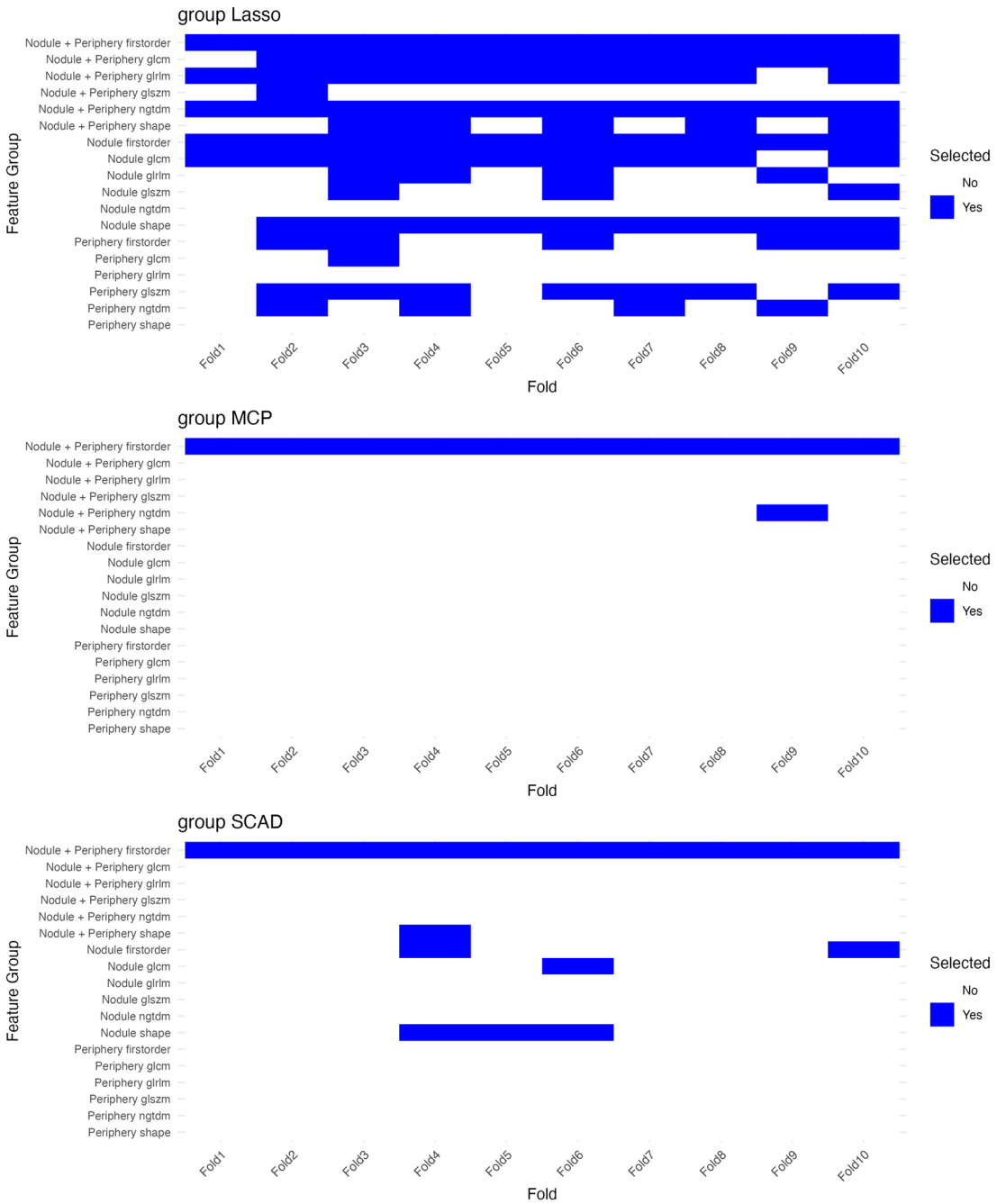

Figure S5 Heatmap of feature selection frequencies for group penalty models in the lung adenocarcinoma study using 10-fold cross-validation, grouped by (III) Combination. Each column represents one of the 10 folds, with blue cells indicating the feature group was selected in that fold and white cells indicating it was not. Lasso, least absolute shrinkage and selection operator; MCP, minimax concave penalty; SCAD, smoothly clipped absolute deviation.

https://dx.doi.org/10.21037/qims-2025-1951

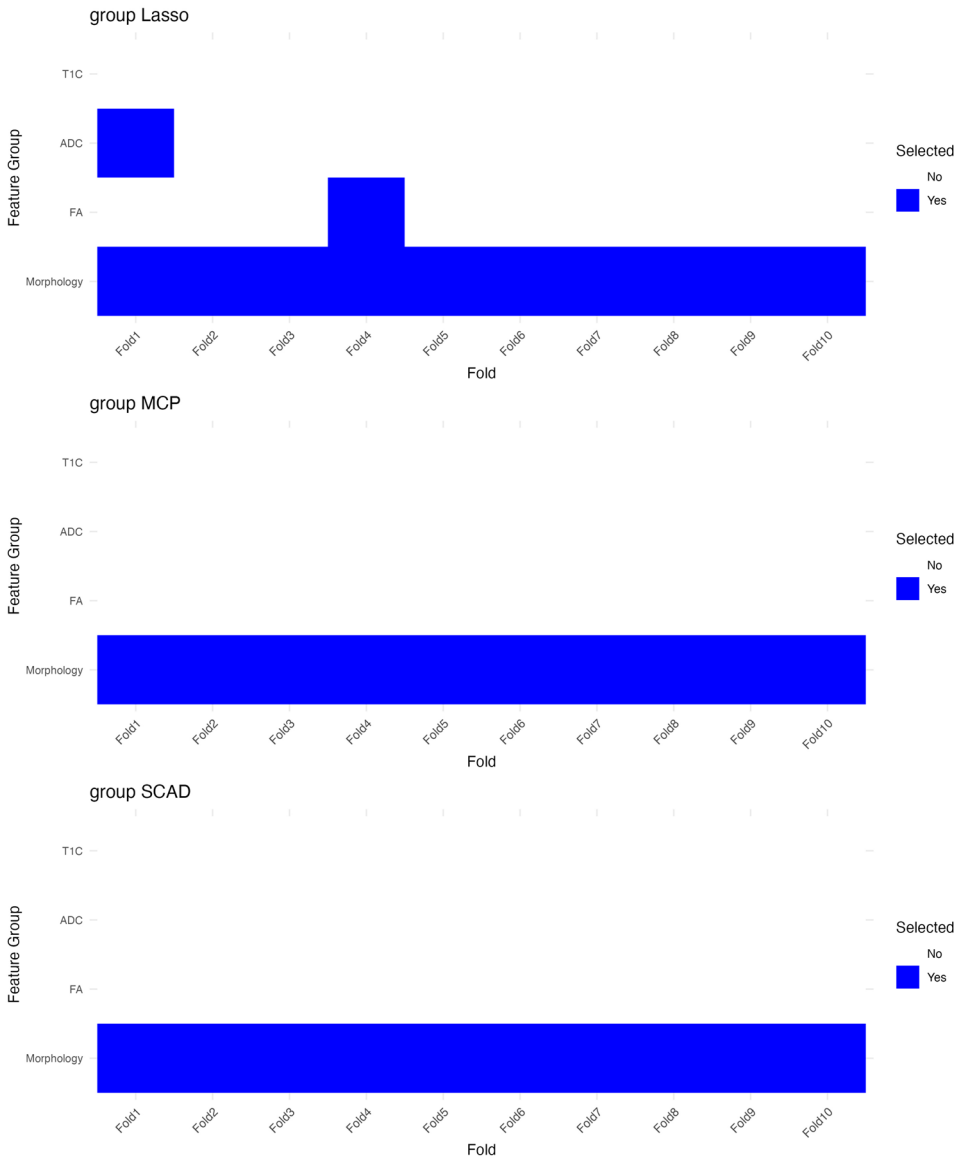

Figure S6 Heatmap of feature selection frequencies for group penalty models in the meningiomas study using 10-fold cross-validation, grouped by (i) Imaging techniques. Each column represents one of the 10 folds, with blue cells indicating the feature group was selected in that fold and white cells indicating it was not. ADC, apparent diffusion coefficient; FA, fractional anisotropy; Lasso, least absolute shrinkage and selection operator; MCP, minimax concave penalty; SCAD, smoothly clipped absolute deviation; T1C, T1-weighted contrast.

https://dx.doi.org/10.21037/qims-2025-1951

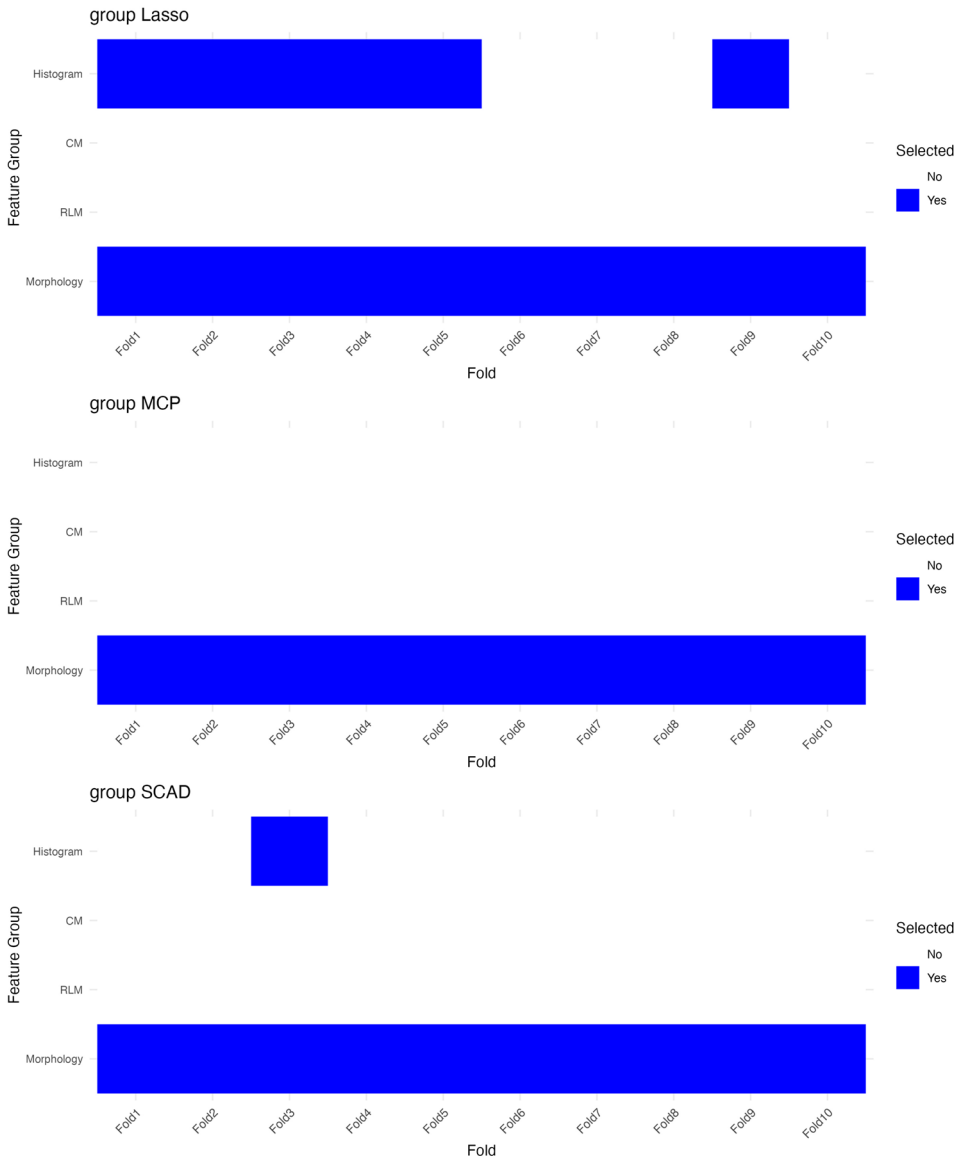

Figure S7 Heatmap of feature selection frequencies for group penalty models in the meningiomas study using 10-fold cross-validation, grouped by (II) Radiomics feature types. Each column represents one of the 10 folds, with blue cells indicating the feature group was selected in that fold and white cells indicating it was not. CM, cooccurrence matrix; Lasso, least absolute shrinkage and selection operator; MCP, minimax concave penalty; RLM, run-length matrix; SCAD, smoothly clipped absolute deviation.

https://dx.doi.org/10.21037/qims-2025-1951

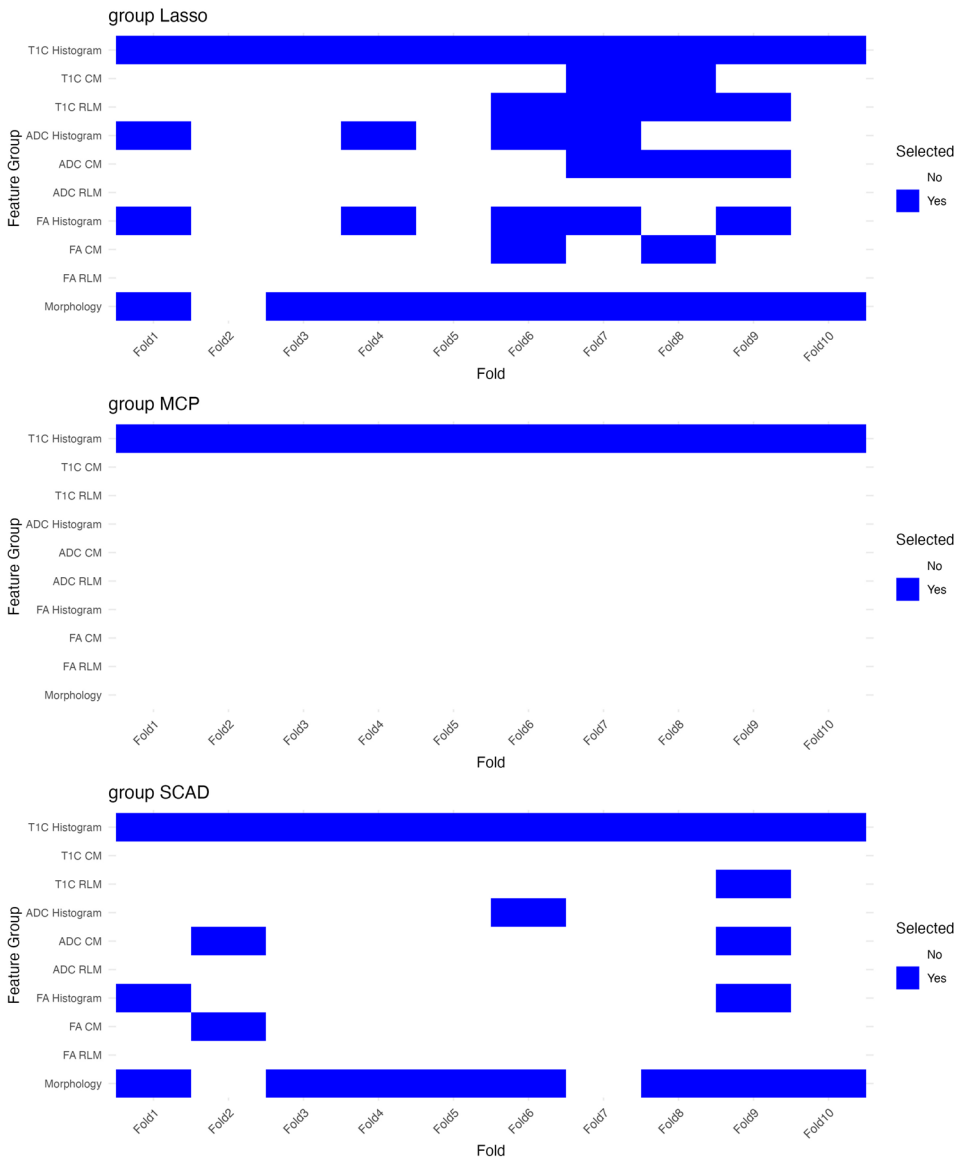

Figure S8 Heatmap of feature selection frequencies for group penalty models in the Meningiomas study using 10-fold cross-validation, grouped by (iii) Combination. Each column represents one of the 10 folds, with blue cells indicating the feature group was selected in that fold and white cells indicating it was not. ADC, apparent diffusion coefficient; FA, fractional anisotropy; Lasso, least absolute shrinkage and selection operator; MCP, minimax concave penalty; RLM, run-length matrix; SCAD, smoothly clipped absolute deviation; T1C, T1-weighted contrast.

https://dx.doi.org/10.21037/qims-2025-1951

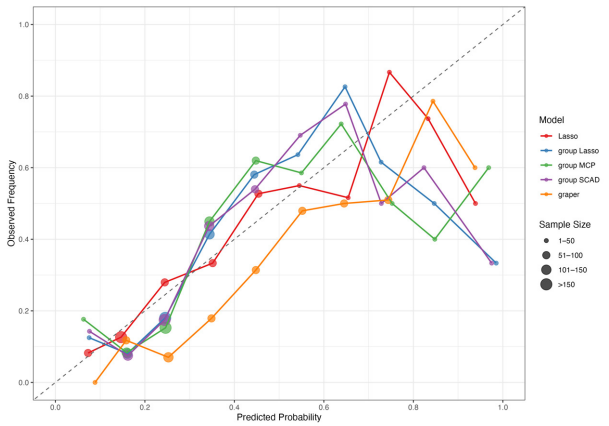

Figure S9 Calibration curves for lung adenocarcinoma models using the (I) Segmentation regions grouping. Each point represents the mean observed frequency plotted against the mean predicted probability within decile bins, with point size proportional to the number of samples in each bin. The dashed diagonal line represents perfect calibration. MCP, minimax concave penalty; SCAD, smoothly clipped absolute deviation.

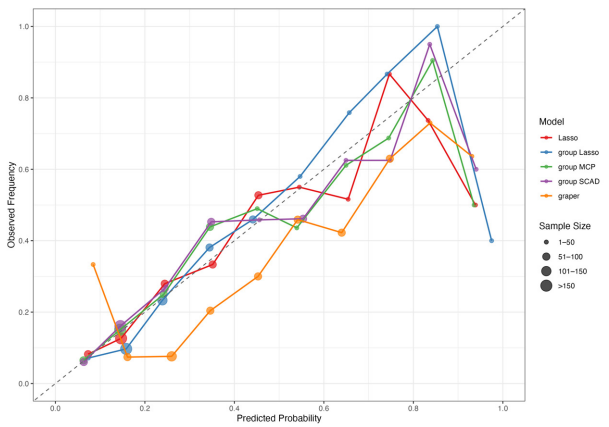

Figure S10 Calibration curves for lung adenocarcinoma models using the (II) Radiomics feature types type grouping. Each point represents the mean observed frequency plotted against the mean predicted probability within decile bins, with point size proportional to the number of samples in each bin. The dashed diagonal line represents perfect calibration. MCP, minimax concave penalty; SCAD, smoothly clipped absolute deviation.

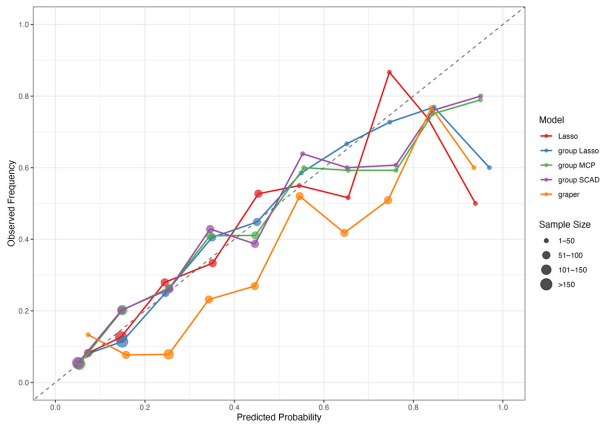

Figure S11 Calibration curves for lung adenocarcinoma models using the (III) Combination grouping. Each point represents the mean observed frequency plotted against the mean predicted probability within decile bins, with point size proportional to the number of samples in each bin. The dashed diagonal line represents perfect calibration. MCP, minimax concave penalty; SCAD, smoothly clipped absolute deviation.

https://dx.doi.org/10.21037/qims-2025-1951

Table S1 Patient characteristics in the studies of lung adenocarcinomas study and meningiomas study

Characteristics	Lung adenocarcinomas study			Meningiomas study		
	STAS negative (n=402)	STAS positive (n=188)	Overall (n=590)	Low-grade (n=154)	High-grade (n=40)	Overall (n=194)
Sex						
Female	252 (62.7)	85 (45.2)	337 (57.1)	125 (81.2)	27 (67.5)	152 (78.4)
Male	150 (37.3)	103 (54.8)	253 (42.9)	29 (18.8)	13 (32.5)	42 (21.6)
Age (years)	64.4 [9.7]	63.9 [11.8]	64.2 [10.4]	55.9 [12.5]	61.0 [14.6]	57.0 [13.1]

Data are presented as n (%) and mean [standard deviation]. STAS, spread through air spaces.

Table S2 Radiomics features extracted from lung adenocarcinomas study

Feature category	Description	Number
Histogram	10th percentile, 90th percentile, energy, entropy, interquartile range, kurtosis, maximum, mean, mean absolute deviation, median, minimum, range, robust mean absolute deviation, root mean squared, skewness, total energy, uniformity, variance	18
GLCM	Autocorrelation, cluster prominence, cluster shade, cluster tendency, contrast, correlation, difference average, difference entropy, difference variance, ID, IDM, IDMN, IDN, IMC1, IMC2, inverse variance, joint average, joint energy, joint entropy, MCC, maximum probability, sum average, sum entropy, sum squares	24
GLRLM	Gray level non-uniformity, gray level non-uniformity normalized, gray level variance, high gray level run emphasis, long run emphasis, long run high gray level emphasis, long run low gray level emphasis, low gray level run emphasis, run entropy, run length non-uniformity, run length non-uniformity normalized, run percentage, run variance, short run emphasis, short run high gray level emphasis, short run low gray level emphasis	16
GLSZM	Gray level non-uniformity, gray level non-uniformity normalized, gray level variance, high gray level zone emphasis, large area emphasis, large area high gray level emphasis, large area low gray level emphasis, low gray level zone emphasis, size zone non-uniformity, size zone non-uniformity normalized, small area emphasis, small area high gray level emphasis, small area low gray level emphasis, zone entropy, zone percentage, zone variance	16
NGTDM	Busyness, coarseness, complexity, contrast, strength	5
Shape	Elongation, flatness, least axis length, major axis length, maximum 2D diameter column, maximum 2D diameter row, maximum 2D diameter slice, maximum 3D diameter, mesh volume, minor axis length, sphericity, surface area, surface volume ratio, voxel volume	14

GLCM, gray level co-occurrence matrix; GLRLM, gray level run length matrix; GLSZM, gray level size zone matrix; ID, inverse difference; IDM, inverse difference moment; IDMN, inverse difference moment normalized; IDN, inverse difference normalized; IMC, informational measure of correlation; MCC, maximal correlation coefficient; NGTDM, neighboring gray tone difference matrix.

https://dx.doi.org/10.21037/qims-2025-1951

Table S3 Radiomics features extracted from meningiomas study

Feature category	Description	Number
Histogram	Entropy, mean, standard deviation, skewness, kurtosis, uniformity	6
CM	Cluster prominence, cluster shade, contrast, correlation, dissimilarity, energy, entropy, homogeneity 1, homogeneity 2, maximum probability, mean, sum variance, standard deviation	13
RLM	Short-run emphasis, long-run emphasis, gray-level nonuniformity, run-length nonuniformity, run percentage, low gray-level run emphasis, high gray-level run emphasis, short-run low-gray-level emphasis, short-run high-gray-level emphasis, long-run low-gray-level emphasis, long-run high-gray-level emphasis	11
Morphology	Volume (cc), surface area (mm^2), surface:volume ratio, compactness 1, compactness 2, maximum 3D diameter, spherical disproportion, sphericity	8

CM, cooccurrence matrix; RLM, run-length matrix.

Table S4 Categorization of radiomics features for lung adenocarcinomas study

Segmentation regions	Radiomics feature types						Total
	Histogram	GLCM	GLRLM	GLSZM	NGTDM	Shape	
Nodule	18	24	16	16	5	14	93
Periphery	18	24	16	16	5	14	93
Combined	18	24	16	16	5	14	93
Total	54	72	48	48	15	42	279

GLCM, gray level co-occurrence matrix; GLRLM, gray level run length matrix; GLSZM, gray level size zone matrix; NGTDM, neighboring gray tone difference matrix.

Table S5 Categorization of radiomics features for meningiomas study

Imaging techniques	Radiomics feature types			Total
	Histogram	CM	RLM	
T1C	6	13	11	30
ADC	6	13	11	30
FA	6	13	11	30
Total	18	39	33	90

With 8 morphology radiomics features. ADC, apparent diffusion coefficient; T1C, T1-weighted contrast; FA, fractional anisotropy; CM, co-occurrence matrix; RLM, run-length matrix.

https://dx.doi.org/10.21037/qims-2025-1951