Cureus
Part of SPRINGER NATURE

Open Access Original Article

# Preoperative Prediction of Fibrous Meningioma Using Contrast-Enhanced T1-Weighted Radiomics and a Tabular Foundation Model: A Single-Center Retrospective Study

**Review began** 10/21/2026
**Review ended** 11/06/2026
**Published** 11/10/2026

**© Copyright** 2026
Zhu et al. This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**DOI:** 10.7759/cureus.119408

Fengzhang Zhu [1], Hao Ma [1], Shizhou Tang [1], Shiyin Zou [1], Heji Ma [1], Hui Guo [1]

1. Radiology, The First Affiliated Hospital of Jinzhou Medical University, Jinzhou, CHN

**Corresponding author:** Hui Guo, huiguo198509@163.com

## Abstract

### Background and objective

Preoperative identification of fibrous meningioma may help inform surgical planning. We compared contrast-enhanced T1-weighted imaging (T1CE) and fluid-attenuated inversion recovery (FLAIR) radiomics, TabPFN and voxel-input three-dimensional convolutional neural networks (CNNs), and imaging-clinical candidate models. We also evaluated training-only controls and regional-clinical fusion.

### Methods

This single-center retrospective study screened 529 records from May 2018 to September 2026 and included 64 cases of World Health Organization (WHO) grade 1 meningioma (18 fibrous and 46 non-fibrous), of which 44 were used for training and 20 for internal evaluation. Automated segmentation and manual quality control defined the tumor and interface regions. In group-aware validation, clinical-only and matched-input classifiers, probability ensembles, and BIG-Huang regional fusion all used a fixed 44/20 split, and all preprocessing and selection were confined to the training data. Performance was assessed at a threshold of 0.5 with 2,000 outcome-stratified bootstrap resamples. Agreement between repeated volume measurements was assessed in 30 cases.

### Results

In the sequence/CNN benchmark, the six-feature and two-feature imaging-clinical candidate models had areas under the curve (AUCs) of 0.810 (95% conditional confidence interval (CI): 0.524-1.000) and 0.905 (0.726-1.000), respectively. Both models had true negatives (TN) = 12, false positives (FP) = 2, false negatives (FN) = 2, and true positives (TP) = 4. In group-aware validation, the five-feature combined TabPFN model had an AUC of 0.726 (0.464-0.929). In the structurally matched BIG-Huang TabPFN regional/fusion models, adding nine clinical variables increased the AUC from 0.702 (0.429-0.905) to 0.833 (0.619-0.976) and reduced the Brier score from 0.234 to 0.151. The 11-input candidate model had an outer out-of-fold (OOF) AUC of 0.422 (0.231-0.625) in the training set. The intra-observer and inter-observer volume intraclass correlation coefficients (ICCs) were 0.9836 and 0.9655, respectively.

### Conclusions

This study describes the complementarity of MRI and clinical information through direct-input and regional-score models. The BIG-Huang analysis supports further investigation of regional-clinical integration. Model probabilities require independent calibration assessment before they are used in individual clinical risk communication.

**Categories:** Neurology, Radiology, Medical Physics
**Keywords:** fibrous subtype, magnetic resonance imaging, meningioma, radiomics, tabpfn, tabular foundation model

## Introduction

Preoperative assessment of meningioma can provide information on lesion type and histological characteristics to help guide surgical planning. MRI depicts tumor location, enhancement, peritumoral changes, and relationships with adjacent structures, whereas definitive histological diagnosis relies on pathological examination [1]. The fifth edition of the World Health Organization (WHO) Classification of Central Nervous System Tumours, published in 2021, incorporates morphology and grade into a unified diagnostic framework [2]. Fibrous meningioma is characterized by spindle cells and collagenous tissue. Preoperative identification of this subtype may complement conventional MRI assessment of internal tumor structure. This task distinguishes histological subtypes within the same grade and should be evaluated separately from high- versus low-grade classification.

Quantitative MRI features have shown associations with meningioma histology. Using conventional contrast-enhanced MRI and diffusion tensor imaging, Park et al. found differences in some texture parameters between fibrous and non-fibrous meningiomas [3]. Han et al. developed a clinical-radiomics nomogram in 423 patients to distinguish these two subtypes, achieving a validation-set area under the curve (AUC) of 0.756 [4]. These studies provide a basis for subtype prediction and emphasize the need to consider population composition, imaging sequences, and validation design when comparing model performance.

Radiomics converts gray-level distributions, spatial textures, and tumor geometry into computable features [5]. PyRadiomics provides an open-source implementation, while the Image Biomarker Standardisation Initiative (IBSI) promotes standardization of feature definitions and processing parameters [6,7]. In studies with limited patient numbers, engineered features compress high-dimensional voxel information into fewer candidate variables. Redundancy, unstable feature selection, and class imbalance may still affect performance. Therefore, whether additional sequences or features improve prediction requires comparisons using the same evaluation samples and metrics.

TabPFN is a transformer pretrained on a large number of synthetic tabular tasks. It uses training samples as context to predict outcomes for new records, offering an approach to small-sample tabular modeling [8]. Starting from radiomics of contrast-enhanced T1-weighted imaging (T1CE) and fluid-attenuated inversion recovery (FLAIR) imaging, this study compared sequence combinations and TabPFN versus voxel-input three-dimensional convolutional neural network (CNN) strategies and examined the effect of adding clinical variables. Study reporting was guided by TRIPOD+AI [9]. This study focuses on the integration of compact radiomics features with a pretrained tabular model to investigate its value for preoperative prediction of fibrous meningioma and to systematically evaluate model performance in terms of discrimination and the quality of probability predictions.

# Materials And Methods

## Study design and population

This retrospective single-center study consecutively screened 529 records collected between May 2018 and September 2026 at the First Affiliated Hospital of Jinzhou Medical University. Cases were included if they met all of the following conditions: (1) surgically resected intracranial meningioma with a postoperative histopathological diagnosis at this hospital; (2) a final diagnosis of either fibrous meningioma or another non-fibrous histological subtype, classified as WHO grade 1 according to the 2021 WHO Classification of Central Nervous System Tumours [2]; (3) preoperative MRI performed within two months before surgery; (4) available T1CE, T2-weighted, and FLAIR images; (5) a solitary primary lesion, rather than a recurrent, multiple, or postoperative one; and (6) image quality adequate for segmentation and feature extraction. The exclusion criteria were absence of postoperative pathology; recurrent, multiple, or postoperative lesions; missing or indeterminate WHO grade or histological subtype; missing required MRI sequences or otherwise incomplete imaging; and severe artifacts or other image-quality problems that precluded analysis. In the sequential exclusion process, 420 records were excluded based on pathology and lesion status, leaving 109. A further 32 were excluded for not meeting the requirements for grade, subtype, and imaging completeness, leaving 77. Another 12 were excluded for insufficient image quality, leaving 65, and one postoperative examination was then excluded, yielding a final total of 64 analysis records. The screening process is shown in Figure *1*.

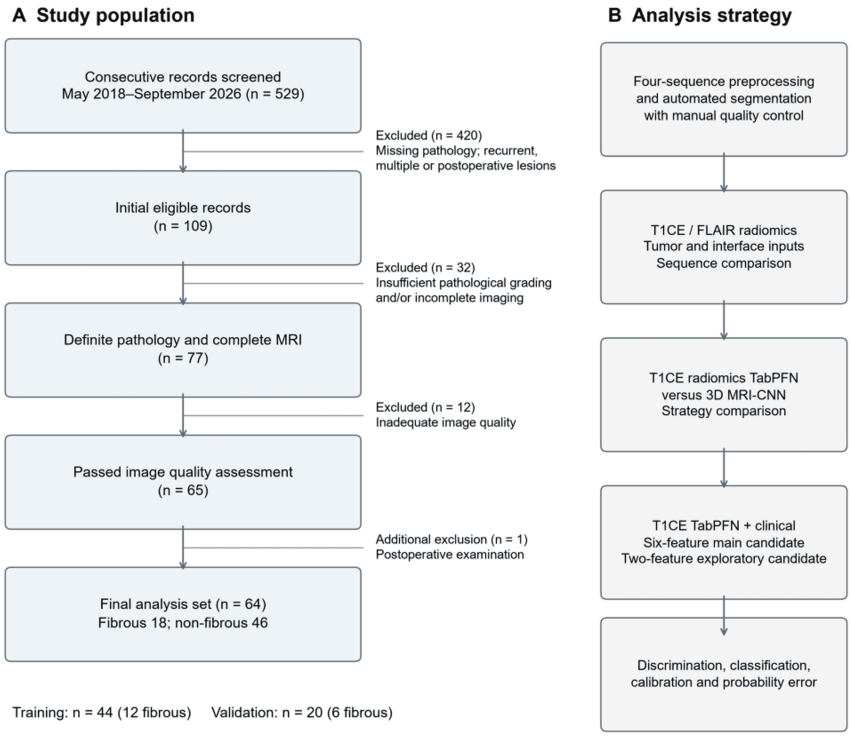



**FIGURE 1: Study population and analysis strategy**

(A) Consecutive screening and sequential exclusions, resulting in 64 analysis records (18 fibrous and 46 non-fibrous), with 44 training and 20 internal-validation records. (B) Sequence comparison, imaging-strategy comparison, and imaging-clinical candidate development. Grouped exclusion counts do not imply separate counts for each listed reason

T1CE: contrast-enhanced T1-weighted imaging; FLAIR: fluid-attenuated inversion recovery; MRI: magnetic resonance imaging; CNN: convolutional neural network

Created by Hao Ma and Fengzhang Zhu

All included cases had postoperative pathology, and MRI was performed within two months before surgery. Two senior pathologists reviewed the diagnoses according to the 2021 WHO classification [2]. Both groups were restricted to WHO grade 1. Fibrous meningioma was the positive outcome, and all other grade 1 histological subtypes were pooled as non-fibrous. The final dataset included 18 fibrous and 46 non-fibrous cases. The training set included 44 cases (12 fibrous and 32 non-fibrous), and the internal validation set included 20 cases (six fibrous and 14 non-fibrous). Models in the sequence/CNN benchmark were evaluated using the same validation records and outcomes. The group-aware validation models used a separately fixed 44/20 split. The Medical Research Ethics Committee of the First Affiliated Hospital of Jinzhou Medical University approved this study on April 1, 2026 (approval number: 2026LL-KY-099), and waived informed consent.

## MRI acquisition and conventional imaging assessment

MRI was acquired using 1.5 T and 3.0 T systems, including MAGNETOM Vida and MAGNETOM Verio (Siemens Healthineers, Erlangen, Germany), Signa HDxt (GE Healthcare, Milwaukee, WI), and Vantage Elan (Toshiba Medical Systems Corporation, Otawara, Japan). Sequences included T1-weighted, T1CE, T2-weighted, and FLAIR imaging. Magnevist (gadopentetate dimeglumine; Bayer Pharma AG, Berlin, Germany) was administered at a dose of 0.1 mmol/kg and a rate of 2 mL/s. Original slice thickness and sequence parameters varied by scanner and year, and a common preprocessing pipeline was applied before analysis. Figure *2* provides author-created T1CE and FLAIR examples of cases from the two histological categories. These screenshots were used for display and were not used for feature extraction.

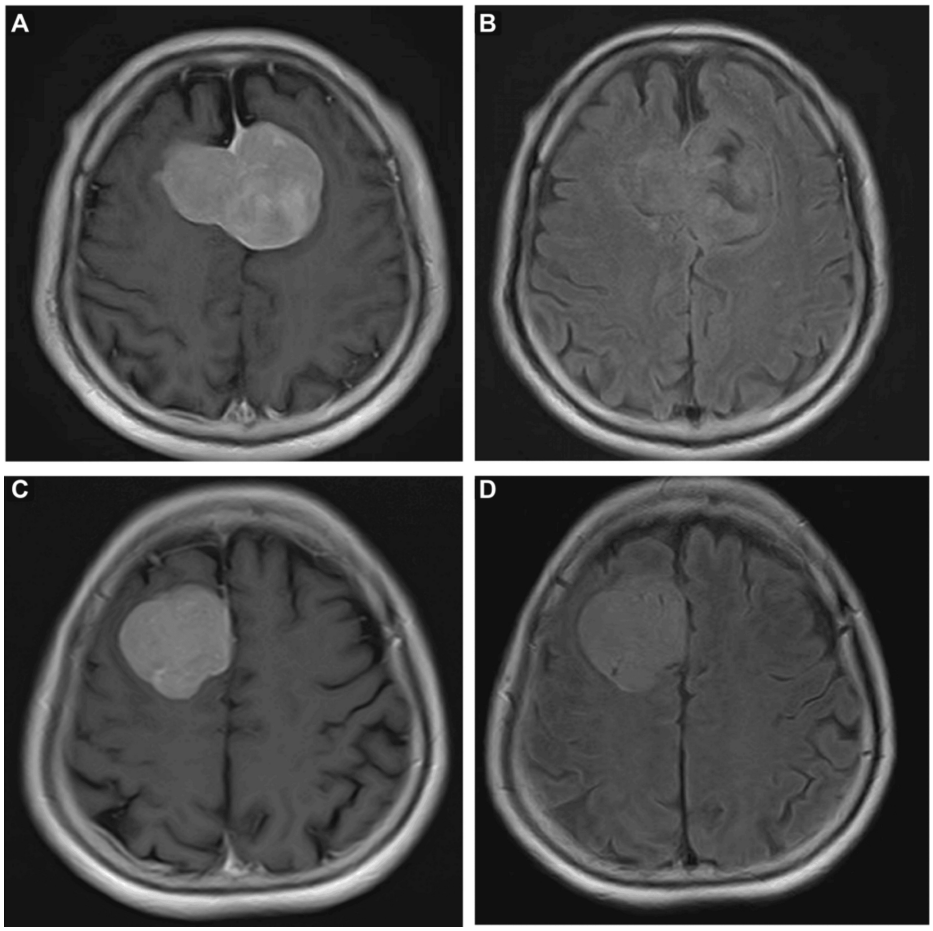



**FIGURE 2: Representative MRI appearances of fibrous and non-fibrous meningiomas**

(A) Fibrous T1CE. (B) Fibrous FLAIR. (C) Non-fibrous T1CE. (D) Non-fibrous FLAIR. Panels may represent different slice levels; quantitative features were extracted from the original MRI volumes

MRI: magnetic resonance imaging; T1CE: contrast-enhanced T1-weighted imaging; FLAIR: fluid-attenuated inversion recovery

Created by Hao Ma and Fengzhang Zhu

Age, sex, and seven conventional MRI findings were collected: supratentorial location, shape, T2 hypointensity, marked enhancement, peritumoral edema, cystic change, and dural tail sign. Two senior radiologists independently assessed the images while blinded to the pathological results, and disagreements were adjudicated by a chief physician. In the sequence/CNN benchmark, findings were classified qualitatively according to the source-table coding: marked enhancement was coded as 2 and a lower degree of enhancement as 1; presence of other binary findings was coded as 1 and absence as 2. Age was treated as a continuous variable, and sex, shape, and other imaging categories as categorical variables.

## Image preprocessing and region construction

DICOM conversion was performed using 3D Slicer version 5.12.4 (Slicer Community, Boston, MA). MRI preprocessing used the default BrainLesion pipeline: each MRI sequence was registered to T1CE, and T1CE was then registered to the SRI24 atlas using an ANTs rigid transformation. In the default ANTs pipeline, nearest-neighbor interpolation was used for spatial resampling of both registered images and label maps. Processing included N4 bias-field correction and HD-BET brain extraction (German Cancer Research Center, Heidelberg, Germany) [10-12]. Intracranial intensities used for radiomics were processed using the 0.1st-99.9th percentiles and linearly normalized to 0-1. Automated segmentation used the BraTS23_2 meningioma algorithm in BraTS Orchestrator (BraTS Consortium, Philadelphia, PA), corresponding to the Blackbean/STU-Net method, with T1-weighted, T1CE, T2-weighted, and FLAIR images as inputs. The software implementation is available at https://github.com/BrainLesion/BraTS. Raters performed segmentation quality control and manual correction while blinded to the histopathological subtype; examples are shown in Figures *3*-*4*. Agreement of ROI volumes from repeated delineation was evaluated in

30 cases [13].

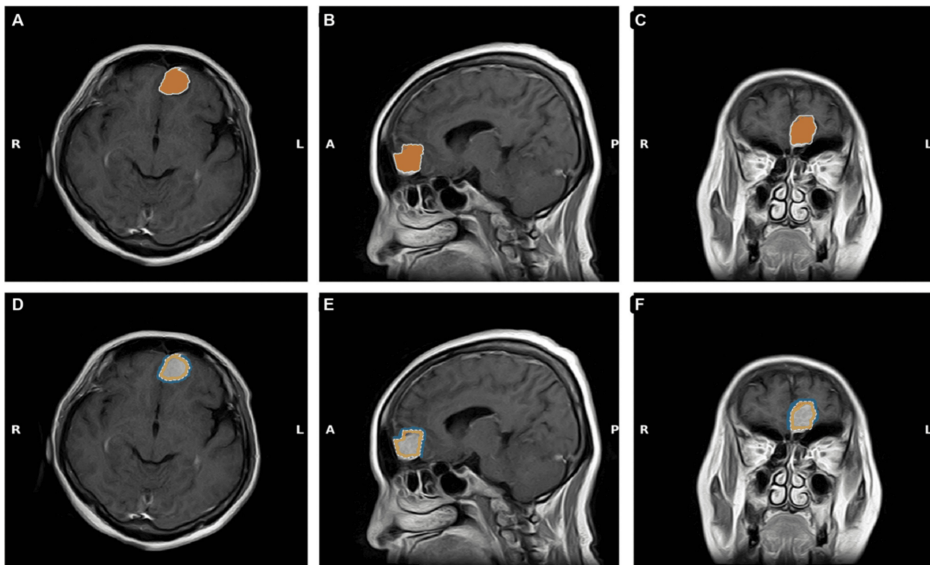



**FIGURE 3: Automated segmentation quality-control example in a fibrous meningioma**

(A-C) Tumor overlays on axial, sagittal, and coronal T1CE images, respectively. (D-F) Corresponding inner 2.5-mm and brain-restricted outer 2.5-mm interface overlays. Orange: tumor; yellow: inner interface; blue: outer interface

T1CE: contrast-enhanced T1-weighted imaging

Created by Hao Ma and Fengzhang Zhu

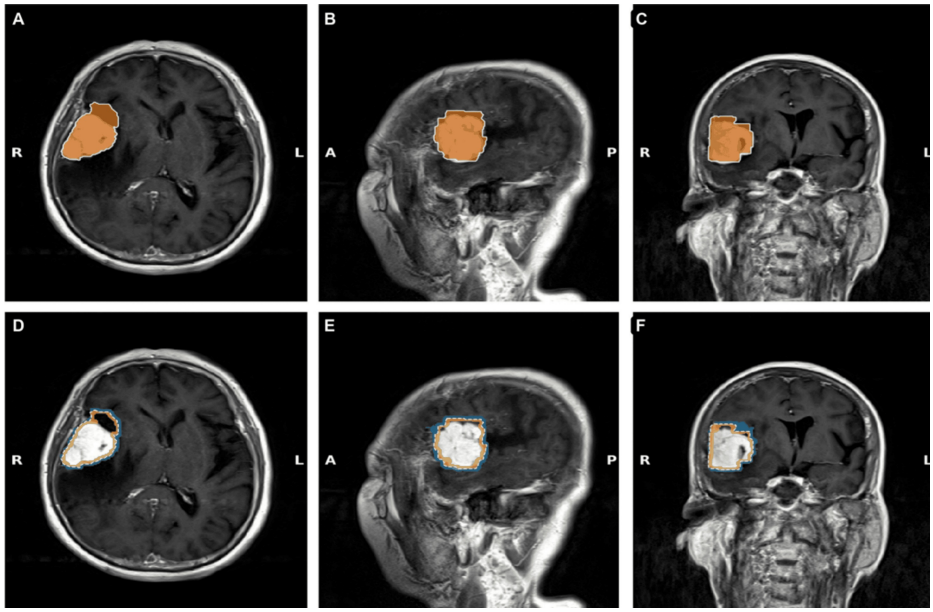



**FIGURE 4: Automated segmentation quality-control example in a non-fibrous meningioma**

(A-C) Tumor overlays on axial, sagittal, and coronal T1CE images, respectively. (D-F) Corresponding inner 2.5-mm and brain-restricted outer 2.5-mm interface overlays. Orange: tumor; yellow: inner interface; blue: outer interface

T1CE: contrast-enhanced T1-weighted imaging

Created by Hao Ma and Fengzhang Zhu

Segmentation labels 1 and 3 were merged to form the whole-tumor mask. Three-dimensional Euclidean distances were calculated using physical voxel spacing. The inner interface comprised tumor voxels within 2.5 mm of the boundary; the outer interface comprised voxels within 2.5 mm outside the boundary and within the brain mask. Their union was defined as the interface region. Thus, the outer interface was restricted by the brain mask, and the inner interface overlapped the whole-tumor region. All regions were constructed on a common 1 mm isotropic grid. After preprocessing, automated segmentation, and manual quality control, features were separately extracted in Python using PyRadiomics from the registered images and their corresponding masks on this grid. Mapping back to the original space was used for quality-control display. The technical roadmap and radiomics workflow are shown in Figures *5*-*6*.

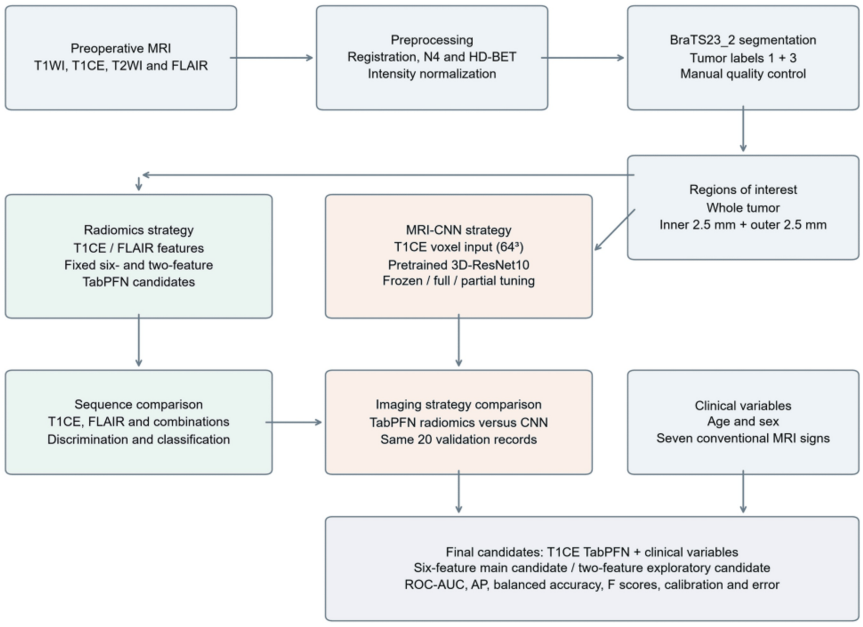



**FIGURE 5: Technical roadmap from four-sequence MRI preprocessing and segmentation to engineered-radiomics TabPFN, voxel-input MRI-CNN, and imaging-clinical candidate evaluation**

T1CE: contrast-enhanced T1-weighted imaging; FLAIR: fluid-attenuated inversion recovery; MRI: magnetic resonance imaging; CNN: convolutional neural network

Created by Hao Ma and Fengzhang Zhu

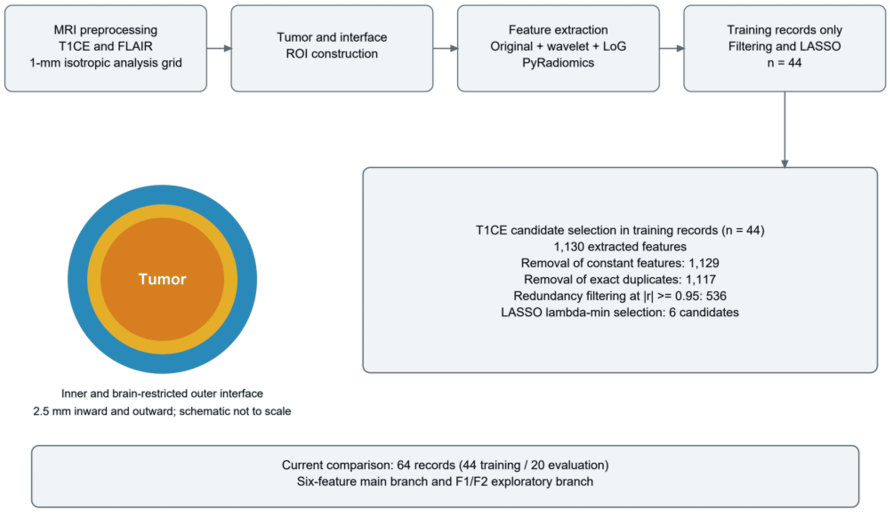



**FIGURE 6: Radiomics workflow and ROI geometry**

The schematic illustrates the intended training workflow. The six-feature benchmark inventory matched a 65-record selection output; group-aware feature selection was confined to 44 training records. Orange: tumor; yellow: inner interface; blue: outer interface. Regions are schematic and not to scale, and the outer interface is brain-restricted

ROI: region of interest; MRI: magnetic resonance imaging; T1CE: contrast-enhanced T1-weighted imaging; FLAIR: fluid-attenuated inversion recovery

Created by Hao Ma and Fengzhang Zhu

## Feature extraction and candidate feature lists

PyRadiomics version 3.0.1 (open-source PyRadiomics developers, Boston, MA) was used to extract features from original images, eight wavelet-decomposition images, and Laplacian-of-Gaussian-filtered images at scales of 1.0, 2.0, and 3.0 mm. The fixed gray-level bin width was 0.05. Each single-sequence, single-region configuration produced 1,130 features, covering shape, first-order statistics, gray-level co-occurrence matrix (GLCM), gray-level run-length matrix (GLRLM), gray-level size-zone matrix, gray-level dependence matrix, and neighboring gray-tone difference matrix features. Combined-region models concatenated tumor and interface features.

Feature extraction was performed on registered three-dimensional image and mask volumes. Gray-level discretization used normalized image intensities and the fixed bin width described above. The Original image type refers to the unfiltered registered image; Wavelet and Laplacian-of-Gaussian features were calculated from its filtered derivatives. Shape features used the mask of the corresponding region. Compliance with the recommendations of the IBSI has not been verified. The reported PyRadiomics version, intensity normalization, bin width, and preprocessing grid describe the pipeline actually performed but do not establish IBSI compliance. This analysis did not verify the correctMask and geometryTolerance settings used during extraction or benchmarking against IBSI reference values [7,14].

The sequence/CNN benchmark used an exploratory six-feature T1CE inventory; the minimum-deviance penalty selected six features, while the one-standard-error rule selected none (Figure 7). Group-aware validation separately fitted a training-only selection pipeline, yielding five and four features under the two rules, respectively [15,16].

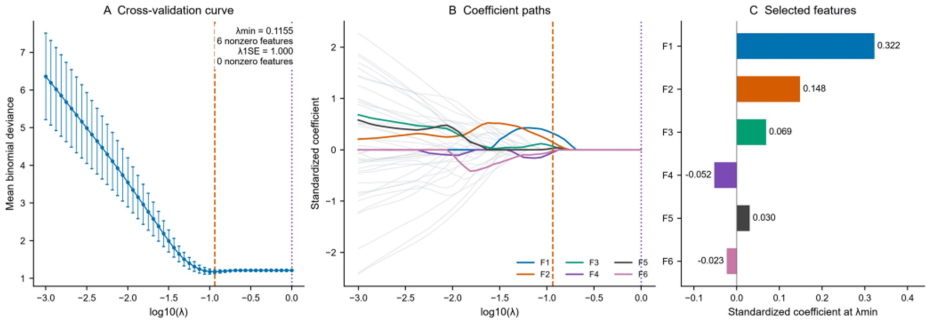



**FIGURE 7: Exploratory LASSO selection for the sequence/CNN benchmark**

(A) Mean binomial deviance; (B) coefficient paths; (C) the six coefficients at the minimum-deviance penalty. The group-aware training-only fit selected five minimum-deviance and four one-SE features

CNN: convolutional neural network

Created by Hao Ma and Fengzhang Zhu

The six candidate features were wavelet-HHL_glcm_MCC, wavelet-HLL_glcm_MCC, original_glrlm_LongRunHighGrayLevelEmphasis, original_shape_Maximum3DDiameter, original_glcm_Id, and wavelet-LLH_firstorder_Energy, designated F1-F6, respectively (Table *1*). The first two were GLCM maximal correlation coefficient (MCC) features. The sequence/CNN benchmark used the fixed six-feature list as the main branch and F1/F2 as an exploratory branch. F1 and F2 are feature identifiers, distinct from the classification F-score. The FLAIR candidate features included five tumor and nine interface features (Table *1*). After the duplicate tumor maximum three-dimensional diameter was identified from the training data, combining the six T1CE and 14 FLAIR features yielded 19 distinct features; combining the two T1CE features with FLAIR yielded 16 features. Table *1* separately lists the feature inventories for the sequence/CNN benchmark and the within-training selection results for group-aware validation.

| Analysis | Region/ID | Feature | λ-min coefficient | 1-SE coefficient | Selection frequency |
|---|---|---|---|---|---|
| Benchmark | F1 | wavelet-HHL_glcm_MCC | 0.3223 | 0.0000 | 94% |
| Benchmark | F2 | wavelet-HLL_glcm_MCC | 0.1482 | 0.0000 | 77% |
| Benchmark | F3 | original_glrlm_LongRunHighGrayLevelEmphasis | 0.0686 | 0.0000 | 45% |
| Benchmark | F4 | original_shape_Maximum3DDiameter | −0.0525 | 0.0000 | 40% |
| Benchmark | F5 | original_glcm_Id | 0.0305 | 0.0000 | 21% |
| Benchmark | F6 | wavelet-LLH_firstorder_Energy | −0.0234 | 0.0000 | 29% |
| Benchmark | Tumor | original_shape_Maximum3DDiameter | — | — | — |
| Benchmark | Tumor | wavelet-LHL_firstorder_Skewness | — | — | — |
| Benchmark | Tumor | wavelet-HHL_ngtdm_Complexity | — | — | — |
| Benchmark | Tumor | wavelet-HHL_glszm_GrayLevelVariance | — | — | — |
| Benchmark | Tumor | original_firstorder_Skewness | — | — | — |
| Benchmark | Interface | original_shape_Maximum3DDiameter | — | — | — |
| Benchmark | Interface | wavelet-HLH_firstorder_Median | — | — | — |
| Benchmark | Interface | wavelet-HHL_glcm_ClusterProminence | — | — | — |
| Benchmark | Interface | wavelet-HLH_glszm_LowGrayLevelZoneEmphasis | — | — | — |
| Benchmark | Interface | wavelet-LLL_glcm_MaximumProbability | — | — | — |
| Benchmark | Interface | original_glcm_Imc1 | — | — | — |
| Benchmark | Interface | wavelet-LLH_glcm_Correlation | — | — | — |
| Benchmark | Interface | wavelet-LLL_firstorder_Minimum | — | — | — |
| Benchmark | Interface | original_shape_Flatness | — | — | — |
| Group-aware | T1CE tumor | original_shape_Sphericity | 0.184 | 0.090 | — |
| Group-aware | T1CE tumor | original_gldm_LargeDependenceHighGrayLevelEmphasis | 0.179 | 0.128 | — |
| Group-aware | T1CE tumor | wavelet-HLL_glcm_ClusterShade | -0.031 | 0.000 | — |
| Group-aware | T1CE tumor | wavelet-HHL_firstorder_Maximum | 0.037 | 0.000 | — |
| Group-aware | T1CE tumor | wavelet-HHL_glcm_MCC | 0.512 | 0.440 | — |
| Group-aware | T1CE tumor | original_glrlm_LongRunHighGrayLevelEmphasis | 0.000 | 0.013 | — |

**TABLE 1: Benchmark radiomics inventories and group-aware T1CE coefficients**

The benchmark comprises six T1CE and 14 FLAIR features; its T1CE one-SE selection selected no features. Group-aware coefficients were fitted using only the fixed 44 training records; minimum-deviance and one-SE rules selected five and four features. Dashes indicate statistics not reported in the inventory or not estimated for that selection procedure

T1CE: contrast-enhanced T1-weighted imaging; FLAIR: fluid-attenuated inversion recovery

Created by Hao Ma and Fengzhang Zhu

## TabPFN modeling

The sequence/CNN benchmark used TabPFN version 9.2.0 (Prior Labs, Freiburg, Germany) with the official default v2 classification weights, four estimators, random seed 20261015, and balance_probabilities = True. Continuous inputs included age and radiomics features; eight columns of categorical clinical variables were explicitly specified. Text-category indices were derived only from the training set, while numeric categories retained their source-data meanings. Five clinical cells were missing in three records and were handled natively by TabPFN. Constant and numerically duplicate features were identified using only the 44 training

records. Class-balanced outputs were used to compare balanced accuracy, with 0.5 as the common threshold for classification scoring. Calibration metrics describe the original output probabilities.

This study used a pretrained TabPFN model to predict probabilities for patients in the evaluation set based on the features and pathological labels of patients in the training set, keeping the pretrained network weights unchanged during inference. Separate models were constructed for six T1CE features, two T1CE features, FLAIR tumor features, FLAIR interface features, all FLAIR features, and the corresponding cross-sequence combinations. The nine clinical variables were then concatenated with the six or two T1CE features to form a 15-input main imaging-clinical candidate model and an 11-input exploratory candidate model. Building on the six-feature model, a compact model was further constructed using the two GLCM maximal correlation coefficient (MCC) features to compare predictive performance with different feature counts and assess the feasibility of compact imaging inputs.

## MRI-CNN and score-level fusion models

The voxel-input models used MedicalNet-pretrained 3D-ResNet10 (Tencent AI Lab, Shenzhen, China), implemented through MONAI (MONAI Consortium; initiated by NVIDIA, Santa Clara, CA, and King's College London, London, UK). The software sources were MedicalNet (https://github.com/Tencent/MedicalNet) and MONAI (https://github.com/Project-MONAI/MONAI). Separate branches were established for the whole tumor and interface, with T1CE and the corresponding regional mask as inputs. Intensities within the nonzero brain region were z-score standardized and clipped to −5 to 5. Images were reoriented to RAS and, based on a common regional bounding cube, resized to $64^3$ voxels using trilinear interpolation for intensities and nearest-neighbor interpolation for masks. Voxels outside the region were set to zero. Interface width was defined on the original 1 mm grid; millimeter distances were not recalculated after resizing. This comparison evaluated the implemented image-processing and modeling pipelines. The CNN received cropped images and masks resized to $64^3$ voxels, whereas the radiomics representation included an explicit maximum three-dimensional diameter feature. These input representations encode tumor geometry in different ways.

Three training strategies were compared. With the encoder frozen, the 512-dimensional representation was first standardized within the training set, reduced to a maximum of 16 components using principal component analysis, and then entered into a regularized classifier. Grouped four-fold cross-validation repeated three times within the training set was used to select the regularization parameter. Full-network fine-tuning was performed for 10 epochs, with encoder and classifier learning rates of $10^{-5}$ and $10^{-4}$, respectively. Partial fine-tuning froze the network through layer3 and updated only layer4 and the classifier. AdamW was used with learning rates of $10^{-6}$ and $10^{-3}$, respectively, weight decay of 0.001, batch size of 8, gradient clipping of 1.0, and left-right flip augmentation. Grouped four-fold cross-validation within the training set allowed a maximum of 30 epochs, with early stopping when the weighted validation cross-entropy failed to improve by at least 0.001 for five consecutive epochs. The median of the best epoch counts across folds determined the final training duration: 8 epochs for the tumor branch and 11 for the interface branch. Batch-normalization running statistics were fixed.

The CNN-clinical models used class-balanced ridge logistic regression to fuse the prediction logit of the partially fine-tuned CNN with the nine clinical variables, with C fixed at 1. The fusion layer was trained using grouped out-of-fold CNN scores within the training set; stopping epochs were determined from the corresponding internal validation losses. This was a staged score-fusion pipeline. Clinical imputation and encoding parameters were estimated from the training records, and the 20 evaluation records were not involved in model fitting. Model construction and processing settings are summarized in Table *2*. Clinical-only and matched-input logistic regression, TabPFN, and CatBoost models, together with CNN-clinical score fusion, provided classifier and input controls. Their fitting procedures are described in the group-aware modeling subsection.

| Category | Parameter | Value |
|---|---|---|
| Study sample | Analysis records | 64 |
| Study sample | Training records | 44 |
| Study sample | Validation records | 20 |
| Outcome | Positive class | Fibrous meningioma |
| Outcome | Negative class | Non-fibrous meningioma |
| Outcome | Eligible WHO grade | Grade 1 |
| Radiomics | Software version | PyRadiomics 3.0.1 |
| Radiomics | Fixed bin width | 0.05 |
| Radiomics | Image type | Original |

| | | |
|---|---|---|
| Radiomics | Image type | Wavelet |
| Radiomics | LoG sigma | 1.0 mm |
| Radiomics | LoG sigma | 2.0 mm |
| Radiomics | LoG sigma | 3.0 mm |
| Interface | Distance method | Spacing-aware Euclidean |
| Interface | Inner width | 2.5 mm |
| Interface | Outer width | 2.5 mm |
| Interface | Outer constraint | Brain mask |
| TabPFN | Version | 9.2.0 |
| TabPFN | Classifier weights | Official v2 default |
| TabPFN | Estimators | 4 |
| TabPFN | Random seed | 20261015 |
| TabPFN | Hardware | CPU |
| TabPFN | Probability option | balance_probabilities = True |
| TabPFN | Adaptation method | In-context inference |
| TabPFN | Task-specific gradient updates | None |
| Imaging–clinical TabPFN | Main radiomics inputs | 6 |
| Imaging–clinical TabPFN | Exploratory radiomics inputs | 2 |
| Imaging–clinical TabPFN | Clinical inputs | 9 |
| Imaging–clinical TabPFN | Main total inputs | 15 |
| Imaging–clinical TabPFN | Exploratory total inputs | 11 |
| CNN | Backbone | 3D-ResNet10 |
| CNN | Pretraining source | MedicalNet 23-dataset |
| CNN | Implementation | MONAI |
| CNN | MRI sequence | T1CE |
| CNN | Region input | Tumor or interface mask |
| CNN | Orientation | RAS |
| CNN | Input size | 64 × 64 × 64 voxels |
| CNN | Intensity normalization | Foreground z-score |
| CNN | Lower clipping limit | −5 |
| CNN | Upper clipping limit | 5 |
| Frozen CNN | Representation size | 512 |
| Frozen CNN | PCA maximum components | 16 |
| Frozen CNN | Selection folds | 4 |
| Frozen CNN | Selection repeats | 3 |
| Frozen CNN | Tumor alpha | 1 |
| Frozen CNN | Interface alpha | 10 |
| Full fine-tuning | Epochs | 10 |
| Full fine-tuning | Encoder learning rate | $1 \times 10^{-5}$ |
| Full fine-tuning | Head learning rate | $1 \times 10^{-4}$ |

| | | |
|---|---|---|
| Full fine-tuning | BatchNorm running statistics | Fixed |
| Partial fine-tuning | Frozen layers | Through layer3 |
| Partial fine-tuning | Updated layers | Layer4 and classifier |
| Partial fine-tuning | Optimizer | AdamW |
| Partial fine-tuning | Encoder learning rate | $1 \times 10^{-6}$ |
| Partial fine-tuning | Head learning rate | $1 \times 10^{-3}$ |
| Partial fine-tuning | Weight decay | 0.001 |
| Partial fine-tuning | Batch size | 8 |
| Partial fine-tuning | Head dropout | 0.25 |
| Partial fine-tuning | Gradient clipping | 1.0 |
| Partial fine-tuning | Augmentation | Left–right flip |
| Early stopping | Internal folds | 4 |
| Early stopping | Maximum epochs | 30 |
| Early stopping | Patience | 5 epochs |
| Early stopping | Minimum loss improvement | 0.001 |
| Final partial model | Tumor epochs | 8 |
| Final partial model | Interface epochs | 11 |
| CNN–clinical fusion | CNN input | Partial-CNN logit |
| CNN–clinical fusion | Clinical inputs | 9 |
| CNN–clinical fusion | Classifier | Balanced ridge logistic regression |
| CNN–clinical fusion | C | 1 |
| Clinical variables | Continuous variable | Age |
| Clinical variables | Categorical variables | Sex and seven MRI findings |
| Clinical coding | Marked enhancement: marked | 2 |
| Clinical coding | Marked enhancement: not marked | 1 |
| Clinical coding | Other finding: present | 1 |
| Clinical coding | Other finding: absent | 2 |
| Clinical missingness | Missing cells | 5 |
| Clinical missingness | Affected records | 3 |
| Evaluation | Decision threshold | 0.5 |
| Evaluation | Bootstrap resamples | 2,000 |
| Evaluation | Bootstrap unit | Validation record |
| Evaluation | Interval type | Percentile conditional interval |
| Evaluation | Calibration model | Unpenalized logistic regression |
| Evaluation | CITL slope | Fixed at 1 |
| Evaluation | Prediction recalibration | None |

**TABLE 2: Sequence/CNN benchmark construction and processing settings**

Group-aware classifier configurations and their clinical coding are specified in the Methods section

CNN: convolutional neural network

Created by Hao Ma and Fengzhang Zhu

## Model evaluation and statistical analysis

Balanced accuracy was the primary evaluation metric. Other metrics included ROC-AUC, macro-averaged F1, accuracy, sensitivity, specificity, positive predictive value, and negative predictive value. Probability quality was evaluated using the Brier score and log loss; precision-recall curves were summarized by average precision (AP). Calibration intercepts and slopes were estimated by unpenalized logistic regression of evaluation outcomes on prediction logits. Calibration-in-the-large (CITL) was estimated separately, with the slope fixed at 1. Calibration plots and expected calibration error (ECE) used three equal-frequency groups. Calibration fitting was used only for evaluation and did not alter the original predictions [17,18].

The classification threshold was fixed at 0.5. The 20 validation records underwent 2,000 outcome-stratified bootstrap resamples, each drawing six positive and 14 negative records with replacement. The 2.5th and 97.5th percentiles defined the 95% conditional confidence intervals (CIs). Paired differences between models used the same resampling indices. These intervals describe sampling variation conditional on the fitted models and validation-set composition. All ROC curves were calculated from the saved probabilities and retained their empirical steps. Radar plots displayed seven metrics, transforming the Brier score to 1−Brier so that higher values consistently represented better performance. The complete heatmap covered all 17 models. Decision curve analysis (DCA) described net benefit over a hypothetical threshold range of 0.05-0.95 [19].

This study considered model performance in the 20-case evaluation set, feature counts, and processing requirements together to determine the final model configuration and sequence priority, and evaluated and reported model performance in the same evaluation set.

An analytical probability-scale sensitivity calculation examined the effect of removing TabPFN probability balancing. In this implementation, after averaging across estimators, the balancing step divides each class probability by its training frequency and then renormalizes the output. With 12 fibrous and 32 non-fibrous records in the training set, the balanced probability q and its unbalanced counterpart p satisfy $\text{logit}(p) = \text{logit}(q)+\ln(12/32)$. This strictly increasing transformation preserves AUC. For the fitted calibration relationship $\text{logit}\Pr(Y = 1) = a+b\ \text{logit}(q)$, the transformed intercept is $a+b\ln(32/12)$, the slope remains b, and the transformed CITL equals the original CITL+ln(32/12). Calculations used the reported calibration estimates rounded to three decimal places and are therefore approximate. The implementation of probability balancing is available at https://github.com/PriorLabs/TabPFN/blob/v9.2.0/src/tabpfn/utils.py.

## SHAP model explanation

Post hoc Shapley additive explanations (SHAP) analysis was performed on the two saved final imaging-clinical TabPFN candidate models [20]. This analysis explained the probability of class 1 (fibrous) in the original class-balanced outputs for all 20 validation records. The background comprised 12 records randomly sampled from the 44 training records with outcome stratification, including three fibrous and nine non-fibrous cases. SHAP version 0.52.0 (open-source SHAP developers, Redmond, WA), PermutationExplainer, and the Independent background masker were used. The two random seeds, 20261016 and 20261017, each generated four pairs of forward/reverse permutations, for a total of eight pairs. Global importance was ranked by mean absolute SHAP value, and rank agreement between the two runs was assessed. Model predictions and the sum of the background value and all feature attributions were each checked against the original probabilities. Models, feature lists, and the classification threshold were held fixed. The sequence/CNN benchmark analysis used Python 3.11.9 (Python Software Foundation, Beaverton, OR); the group-aware classifier and fusion analyses used Python 3.12.14.

## Group-aware modeling

Group-aware modeling used 64 examination records (18 fibrous and 46 non-fibrous cases), with a fixed split into 44 training records (12 fibrous and 32 non-fibrous cases) and 20 evaluation records (six fibrous and 14 non-fibrous cases). Known repeat-examination groups were kept together, and records with unresolved identities were assigned to the training set. Group-aware validation and benchmark results correspond to different evaluation partitions.

Radiomics preprocessing and LASSO selection for the group-aware models were fitted exclusively in the training data. Outcome-stratified, patient-group-aware five-fold cross-validation was repeated 20 times; each fit repeated imputation, removal of constant and duplicate columns, correlation filtering at $|r|\geq 0.95$, and standardization. Forty-nine penalty parameters from 0.001 to 1 were evaluated, and the minimum-deviance and 1-SE rules determined the number of selected features. For the specified complete modeling pipelines, group-aware five-fold outer cross-validation was performed within the 44 training records; outer held-out groups were excluded from preprocessing, feature selection, and fitting.

Matched inputs were used to compare clinical-only, radiomics-only, and imaging-clinical models. Ridge logistic regression used C = 1; TabPFN used the official v2 classification checkpoint, 4 estimators, and balance_probabilities = False; CatBoost used 500 iterations, depth 4, a learning rate of 0.03, L2 regularization of 3, and ordered boosting. All random seeds were set to 20261105. The nine clinical variables were age, sex, supratentorial location, shape, T2 hypointensity, marked enhancement, peritumoral edema, cystic change, and dural tail sign. Marked enhancement was coded as 1 and non-marked enhancement as 2. These input and probability settings applied to the group-aware models [21,22].

Probability fusion included fixed 40% CatBoost/60% TabPFN weighting and a normalized-weight search. Raw weights for each learner ranged from 0.1 to 0.9, yielding 81 pairs and 55 distinct normalized ratios, ranked by higher training outer OOF AUC and then lower Brier score. Another 18 combinations added clinical-only TabPFN or ridge regression scores with weights of 0.1-0.9. Eleven-input stacking combined the weighted probabilities of CatBoost and TabPFN, the imaging-clinical TabPFN probability, and the nine clinical variables, using ridge logistic regression and TabPFN as the respective fusion layers. The stacking pipeline used five-fold grouped outer cross-validation and five-fold inner cross-fitting; each inner training subset independently completed preprocessing, LASSO selection, and base-model fitting, and out-of-fold probabilities were used to train the fusion layer. Two schemes were compared: fixed 3/7 CatBoost+4/7 TabPFN weights and weight selection within the training partition. The latter selected a ratio from the 55 ratios using only internal OOF predictions from each outer training partition, ordered by higher AUC, lower Brier score, and lower CatBoost weight. Outer held-out groups were used only for evaluation.

The BIG-Huang regional-clinical fusion framework extended the regional-score integration approach of Huang et al. [23] to MRI-based fibrous meningioma classification. Radiomics from the tumor and interface regions were separately converted into regional probabilities, which were then entered into a fusion learner alone or together with the nine clinical variables. Ten configurations compared CatBoost and TabPFN as regional and fusion learners and included two direct-feature input schemes; four additional extensions incorporated clinical-only ridge regression or TabPFN scores. The complete pipeline used the same five grouped outer folds. Within each outer training partition, classifiers were further cross-fitted to generate regional scores for fusion training; regional features were selected in that outer training partition and held fixed during internal classifier cross-fitting. Final selection on the full training set retained two T1CE regional-score inputs. Thus, all four clinical-score extensions had twelve final inputs: two regional scores, nine clinical variables, and one clinical score.

Group-aware validation used a fixed probability threshold of 0.5 and 2,000 outcome-stratified bootstrap resamples to obtain 95% conditional confidence intervals; paired comparisons used the same resampling indices. The intervals were conditional on saved predictions and did not repeat feature selection, fitting, or configuration selection; training OOF intervals resampled complete identity-constrained groups as the unit. For the five-feature imaging-clinical TabPFN and the 11-input BIG-Huang fusion function, interventional permutation Shapley estimation was performed on the positive-class probability scale: 16 training background records were sampled uniformly, and 16 feature permutations (8 orders and their reverses) were used [24]. SHAP analysis quantified the contributions of individual input variables to the predicted probabilities of the five-feature imaging-clinical TabPFN model and the 11-input BIG-Huang fusion model.

Decision curves calculated net benefit as $TP/n - FP/n \times t/(1-t)$ over a fixed threshold range of 0.05-0.80 in steps of 0.01, with act-on-all and act-on-none references [19]. Clinical impact curves displayed $(TP+FP)/n \times 1{,}000$ and $TP/n \times 1{,}000$; this scale standardized the observed proportions. Calibration used three equal-frequency probability groups formed without outcomes (7, 7, and 6 records), with Wilson intervals reported. The three-group Hosmer-Lemeshow diagnostic used 1 degree of freedom, and small expected counts affected its chi-square approximation [25]. Risk cutoffs were the medians of each model's outer OOF probabilities for the 44 training records, fixed and then applied to the 20 evaluation records; proportions between groups were compared using two-sided Fisher's exact tests.

Repeated delineation and volume measurement were performed in 30 patients. The intra-observer and inter-observer ICCs were 0.9836 and 0.9655, respectively, supporting high agreement of ROI volume measurements.

# Results

## Baseline characteristics

The cohort comprised 64 examination records, including 18 fibrous and 46 non-fibrous meningiomas, with 23 male and 41 female patients. Table *3* summarizes the clinical characteristics of the group-aware 44/20 partition.

| Characteristic | Overall | Training | Evaluation |
|---|---|---|---|
| Examination records, n | 64 | 44 | 20 |
| Fibrous, n | 18 | 12 | 6 |
| Non-fibrous, n | 46 | 32 | 14 |
| Age, median, years | 62.0 | 61.0 | 65.5 |
| Age, Q1, years | 52.8 | 53.0 | 51.0 |
| Age, Q3, years | 67.0 | 67.0 | 67.5 |
| Male, n | 23 | 14 | 9 |
| Female, n | 41 | 30 | 11 |
| Supratentorial, n | 59 | 42 | 17 |
| Irregular shape, n | 18 | 14 | 4 |
| T2 hypointensity, n | 30 | 22 | 8 |
| Marked enhancement, n | 36 | 26 | 10 |
| Peritumoral edema, n | 32 | 21 | 11 |
| Cystic change, n | 13 | 8 | 5 |
| Dural tail, n | 57 | 42 | 15 |

**TABLE 3: Clinical characteristics of the group-aware 44/20 partition**

The cohort contains 64 examination records; case-level correspondence with the benchmark partition was not established

Created by Hao Ma and Fengzhang Zhu

## Feature selection results

In the sequence/CNN benchmark, the minimum-deviance rule selected six T1CE features, whereas the one-standard-error rule selected no features (Figure *7*). Table *1* presents the benchmark inventory and the training-only selection results for the group-aware validation, which selected five and four features, respectively.

## Sequence and feature combinations

The main six-feature T1CE branch had an AUC of 0.798 (95% conditional bootstrap confidence interval: 0.500-1.000), a balanced accuracy of 0.702, and a macro F1 of 0.642. The exploratory two-feature branch had an AUC of 0.869 (95% conditional bootstrap CI: 0.690-1.000), a balanced accuracy of 0.726, and a macro F1 of 0.715. The AUCs of the FLAIR tumor, interface, and combined-region models were 0.702 (0.393-0.952), 0.821 (0.548-1.000), and 0.762 (0.452-1.000), respectively; all intervals were 95% conditional bootstrap CIs (Table *4*).

| Model | AUC (95% conditional CI) | BA | Macro F1 | Sensitivity | Specificity | Brier |
|---|---|---|---|---|---|---|
| TabPFN T1CE six features | 0.798 (0.500–1.000) | 0.702 | 0.642 | 0.833 | 0.571 | 0.207 |
| TabPFN T1CE two features | 0.869 (0.690–1.000) | 0.726 | 0.715 | 0.667 | 0.786 | 0.184 |
| TabPFN FLAIR tumor | 0.702 (0.393–0.952) | 0.738 | 0.688 | 0.833 | 0.643 | 0.217 |
| TabPFN FLAIR interface | 0.821 (0.548–1.000) | 0.738 | 0.688 | 0.833 | 0.643 | 0.203 |
| TabPFN FLAIR combined | 0.762 (0.452–1.000) | 0.702 | 0.642 | 0.833 | 0.571 | 0.225 |
| TabPFN T1CE six + FLAIR | 0.786 (0.464–1.000) | 0.774 | 0.733 | 0.833 | 0.714 | 0.208 |
| TabPFN T1CE two + FLAIR | 0.845 (0.500–1.000) | 0.774 | 0.733 | 0.833 | 0.714 | 0.196 |
| Frozen CNN tumor | 0.750 (0.500–0.929) | 0.619 | 0.583 | 0.667 | 0.571 | 0.213 |
| Frozen CNN interface | 0.762 (0.488–0.964) | 0.655 | 0.627 | 0.667 | 0.643 | 0.193 |
| Fully tuned CNN tumor | 0.714 (0.464–0.905) | 0.619 | 0.583 | 0.667 | 0.571 | 0.218 |
| Fully tuned CNN interface | 0.738 (0.452–0.952) | 0.607 | 0.601 | 0.500 | 0.714 | 0.197 |
| Partially tuned CNN tumor | 0.738 (0.500–0.929) | 0.571 | 0.560 | 0.500 | 0.643 | 0.196 |
| Partially tuned CNN interface | 0.738 (0.452–0.952) | 0.571 | 0.560 | 0.500 | 0.643 | 0.201 |
| TabPFN six + clinical | 0.810 (0.524–1.000) | 0.762 | 0.762 | 0.667 | 0.857 | 0.152 |
| TabPFN two + clinical | 0.905 (0.726–1.000) | 0.762 | 0.762 | 0.667 | 0.857 | 0.121 |
| CNN tumor score + clinical | 0.798 (0.571–0.976) | 0.726 | 0.715 | 0.667 | 0.786 | 0.171 |
| CNN interface score + clinical | 0.821 (0.619–0.976) | 0.726 | 0.715 | 0.667 | 0.786 | 0.172 |

**TABLE 4: Sequence/CNN benchmark performance across 17 model configurations**

Intervals are 95% conditional intervals for fixed predictions. These configurations and the group-aware models use distinct evaluation allocations

CI: confidence interval; BA: balanced accuracy; T1CE: contrast-enhanced T1-weighted imaging; FLAIR: fluid-attenuated inversion recovery; CNN: convolutional neural network

Created by Hao Ma and Fengzhang Zhu

After FLAIR was added to the main six-feature branch, the AUC changed from 0.798 (95% conditional bootstrap CI: 0.500-1.000) to 0.786 (0.464-1.000), balanced accuracy from 0.702 to 0.774, and macro F1 from 0.642 to 0.733. The number of correctly classified fibrous cases remained 5/6, whereas that of non-fibrous cases changed from 8/14 to 10/14. In the two-feature branch, the AUC changed from 0.869 (0.690-1.000) to 0.845 (0.500-1.000), and balanced accuracy changed from 0.726 to 0.774; correctly classified fibrous cases changed from 4/6 to 5/6, and non-fibrous cases from 11/14 to 10/14. All AUC intervals were 95% conditional bootstrap CIs. After these preliminary comparisons, T1CE radiomics was prioritized for subsequent modeling, considering the inconsistent changes in AUC, the number of features, and processing requirements.

## Comparison of TabPFN and MRI-CNN

The tumor and interface CNNs with frozen encoders had evaluation AUCs of 0.750 (95% conditional bootstrap CI: 0.500-0.929) and 0.762 (0.488-0.964), respectively. Full-network fine-tuning increased apparent training discrimination, with evaluation AUCs of 0.714 (0.464-0.905) and 0.738 (0.452-0.952), respectively. With partial fine-tuning and early stopping, the tumor and interface AUCs were 0.738 (0.500-0.929) and 0.738 (0.452-0.952), respectively, and both had a balanced accuracy of 0.571. All intervals were 95% conditional bootstrap CIs. These values describe the performance of the specified procedures in the same evaluation cases (Table 4). The models exhibited different performance characteristics in discrimination and probability quality. The frozen interface CNN had a Brier score of 0.193, lower than the 0.207 of the six-feature T1CE TabPFN; the fully fine-tuned interface CNN had an ECE of 0.102, lower than the 0.210 of the six-feature TabPFN. The AUC difference between the six-feature T1CE TabPFN and the frozen tumor CNN was 0.048 (paired 95% conditional CI: −0.250 to 0.334).

The main six-feature imaging-clinical candidate had an AUC of 0.810 (95% conditional bootstrap CI: 0.524-

1.000), and the exploratory two-feature candidate had an AUC of 0.905 (0.726-1.000). Relative to their respective imaging-only branches, the changes in AUC were 0.012 (95% conditional bootstrap CI: −0.250 to 0.238) and 0.036 (−0.167 to 0.238), respectively. Brier scores changed from 0.207 to 0.152 and from 0.184 to 0.121, respectively; log losses changed from 0.620 to 0.449 and from 0.551 to 0.383, respectively. The tumor CNN-clinical and interface CNN-clinical models had AUCs of 0.798 (95% conditional bootstrap CI: 0.571-0.976) and 0.821 (0.619-0.976), with Brier scores of 0.171 and 0.172, respectively (Table *4*). The six-feature and two-feature imaging-clinical candidates had the same balanced accuracy at the 0.5 threshold. The difference in AUC point estimates does not establish the superiority of the two-feature candidate.

At the 0.5 threshold, the two combined TabPFN candidates produced identical confusion matrices: 12 true negatives (TN), two false positives (FP), two false negatives (FN), and four true positives (TP). Both had an accuracy of 0.800, sensitivity of 0.667, specificity of 0.857, and balanced accuracy and macro F1 of 0.762. Compared with the six-feature imaging-only branch, the main combined candidate increased correctly classified non-fibrous cases from 8/14 to 12/14, while correctly classified fibrous cases changed from 5/6 to 4/6, reflecting a classification trade-off at a fixed threshold.

The main six-feature combined candidate had an AP of 0.802, calibration intercept of −0.796, slope of 0.893, CITL of −0.810, and ECE of 0.126. The corresponding values for the two-feature model were 0.843, −0.686, 0.966, −0.687, and 0.078. Mean probabilities were 0.404 and 0.378, exceeding the observed proportion of 0.300. The sequence/CNN benchmark is shown in Figures *8*-*13*.

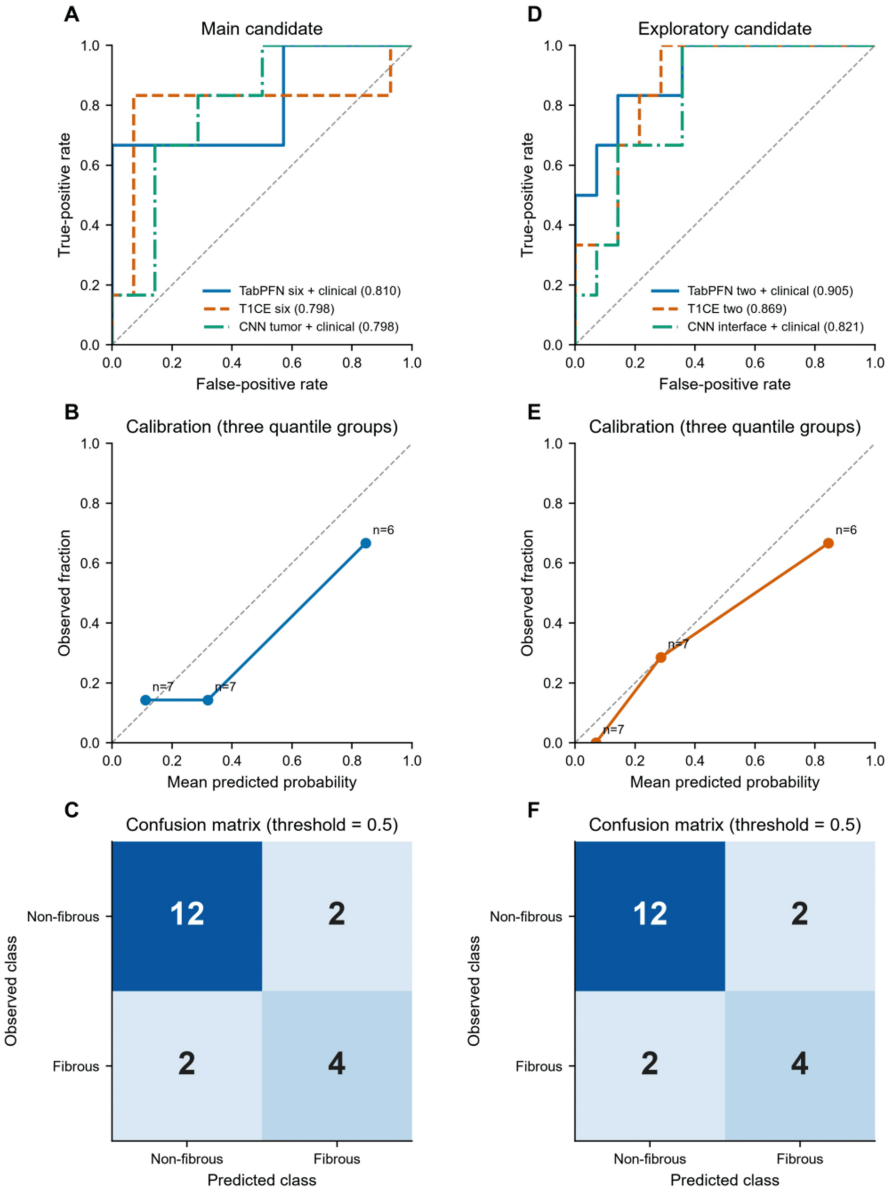



**FIGURE 8: Discrimination, calibration, and classification of the benchmark candidates**

(A-C) Six-feature main candidate; (D-F) two-feature exploratory candidate. Confusion matrices use a fixed threshold of 0.5

T1CE: contrast-enhanced T1-weighted imaging; CNN: convolutional neural network

Created by Hao Ma and Fengzhang Zhu

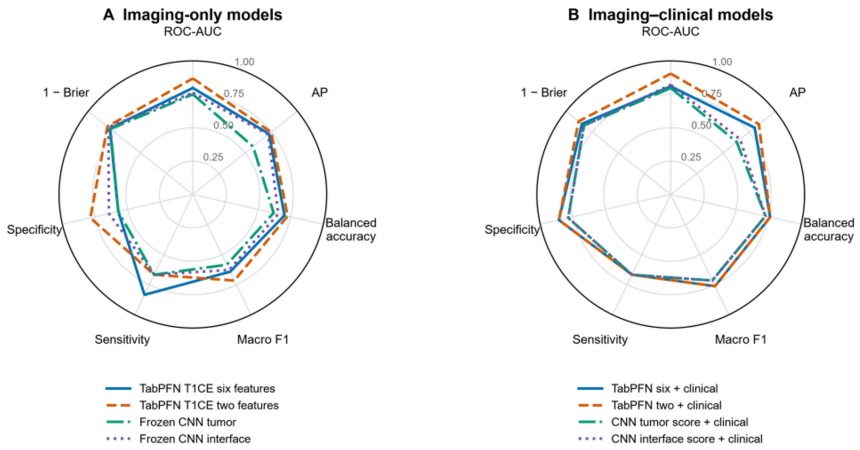



**FIGURE 9: Multidimensional validation comparison using radar plots**

(A) Imaging-only models; (B) imaging-clinical models. Seven axes show ROC-AUC, average precision, balanced accuracy, macro F1, sensitivity, specificity, and 1 minus the Brier score

T1CE: contrast-enhanced T1-weighted imaging; CNN: convolutional neural network

Created by Hao Ma and Fengzhang Zhu

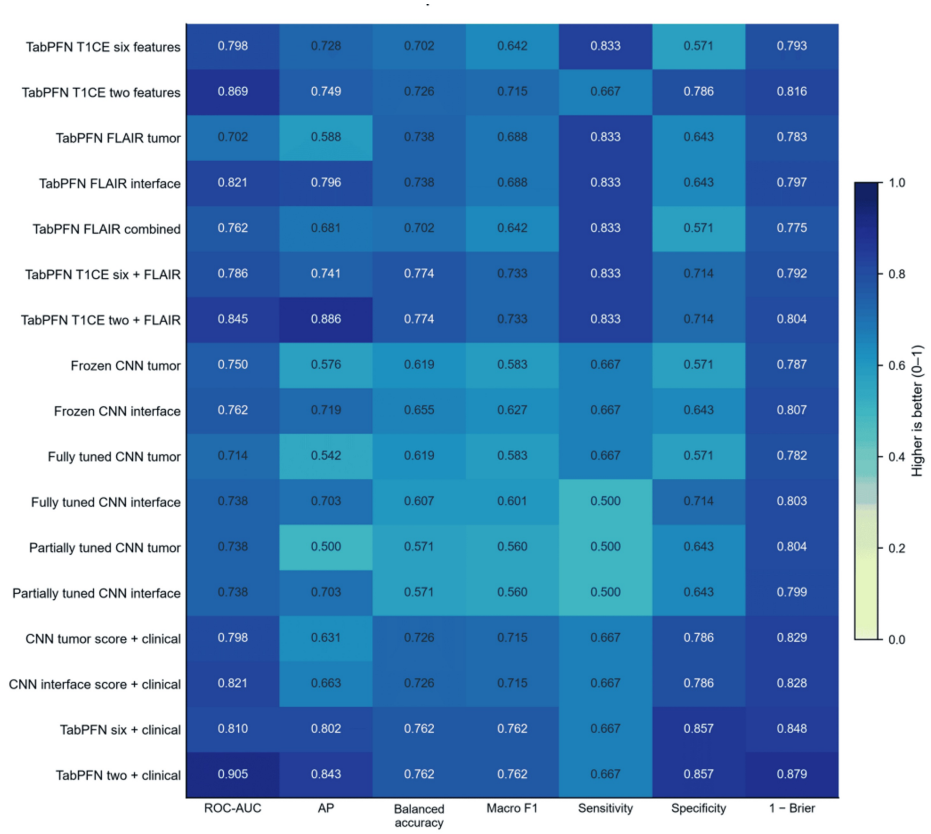



**FIGURE 10: Validation heatmap across the 17 benchmark models**

Cells show metric values to three decimals, except that Brier is displayed as 1 minus Brier

T1CE: contrast-enhanced T1-weighted imaging; FLAIR: fluid-attenuated inversion recovery; CNN: convolutional neural network

Created by Hao Ma and Fengzhang Zhu

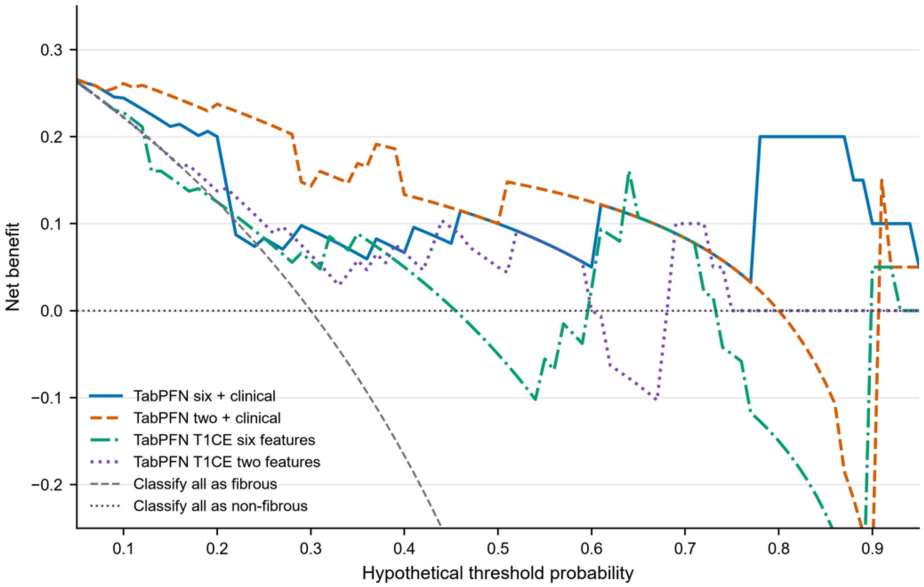



**FIGURE 11: Descriptive decision curves of the two benchmark TabPFN candidates and their imaging-only references**

Threshold probabilities range from 0.05 to 0.95

T1CE: contrast-enhanced T1-weighted imaging

Created by Hao Ma and Fengzhang Zhu

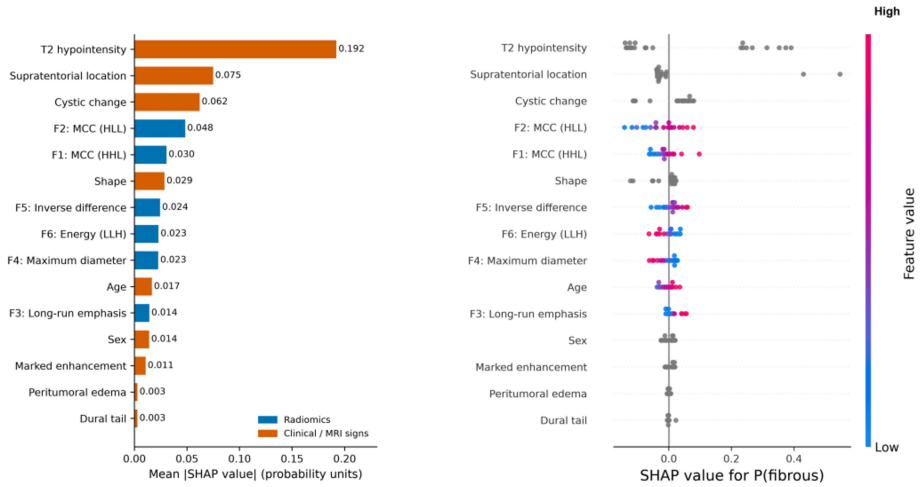



**FIGURE 12: SHAP analysis of the six-feature imaging-clinical TabPFN main candidate**

(A) Mean absolute SHAP values; (B) beeswarm distributions in probability units for fibrous meningioma

Created by Hao Ma and Fengzhang Zhu

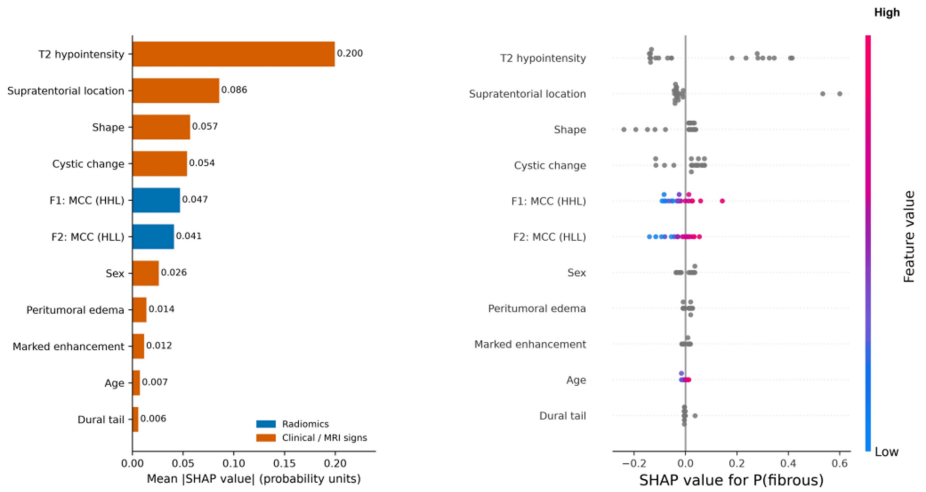



**FIGURE 13: SHAP analysis of the two-feature imaging-clinical exploratory TabPFN candidate**

(A) Global mean absolute attribution. (B) Validation beeswarm

Created by Hao Ma and Fengzhang Zhu

After probability balancing was removed analytically, the AUCs of the final candidates remained 0.810 (95% conditional bootstrap confidence interval, 0.524-1.000) and 0.905 (0.726-1.000). For the six-feature candidate, the calibration intercept changed from −0.796 to approximately 0.080, CITL changed from −0.810 to 0.171, and the slope remained 0.893. For the two-feature candidate, the corresponding intercept changed from −0.686 to approximately 0.261, CITL changed from −0.687 to 0.294, and the slope was 0.966. These values resulted from a probability-scale transformation of the reported calibration fits, rather than refitting the prediction models.

## SHAP attribution analysis

In the main six-feature combined candidate, the mean absolute SHAP values for T2 hypointensity, supratentorial location, and cystic change were 0.192, 0.075, and 0.062, respectively. The two MCC features, F2 and F1, had values of 0.048 and 0.030, ranking fourth and fifth among all 15 inputs. T2 hypointensity also ranked first in the exploratory two-feature combined candidate (0.200); F1 and F2 had values of 0.047 and 0.041, ranking fifth and sixth, respectively. In the evaluation records, higher MCC values generally corresponded to more positive SHAP contributions. The Spearman correlations between the rankings from the two permutation runs were 0.989 and 1.000, respectively; the top three features were consistent for both models. The maximum additive reconstruction errors were $1.11\times10^{-16}$ and $5.55\times10^{-17}$, respectively. Global SHAP feature importance and beeswarm distributions for the six-feature and two-feature imaging-clinical TabPFN models are shown in Figures *12* and 13, respectively.

## Performance of group-aware models

In the training-only selection for the group-aware validation, T1CE tumor LASSO retained five features at the minimum-deviance penalty $\lambda = 0.115478$ and four features under the 1-SE rule at $\lambda = 0.133352$; neither FLAIR region retained any features. The numbers of T1CE features in the five outer training folds were 3, 0, 3, 4, and 0, respectively. These results correspond to training-only selection within the group-aware procedure and are presented separately from the six-feature and fourteen-feature inventories of the sequence/CNN benchmark.

Among the 20 evaluation records in the group-aware validation, imaging-only TabPFN had an AUC of 0.655 (95% conditional confidence interval, 0.357-0.905), clinical-only TabPFN had an AUC of 0.643 (0.321-0.905), and five-feature imaging-clinical TabPFN had an AUC of 0.726 (0.464-0.929); ridge regression with the same inputs yielded corresponding values of 0.679 (0.393-0.905), 0.583 (0.274-0.881), and 0.655 (0.333-0.929). Combined CatBoost had an AUC of 0.631 (0.345-0.881). The paired AUC difference after adding clinical variables to the five-feature TabPFN was 0.071 (−0.084 to 0.262). At the 0.5 threshold, the five-feature combined TabPFN had TN = 13, FP = 1, FN = 4, and TP = 2, whereas the two-feature combined TabPFN in the group-aware validation had TN = 12, FP = 2, FN = 4, and TP = 2. These counts correspond to the group-aware partition; the two combined benchmark candidates in Figure *8* both had TN = 12, FP = 2, FN = 2, and TP = 4. In the reverse comparison, adding radiomics to clinical-only TabPFN yielded a paired ΔAUC of 0.083 (95% conditional confidence interval, −0.298 to 0.429). An overview of classifier and fusion model performance is

shown in Figure 14.

| | AUC 95% conditional CI | AP | BA | Macro F1 | Sensitivity | Specificity | 1 − Brier |
|---|---|---|---|---|---|---|---|
| Clinical only / CatBoost | 0.631 (0.381–0.869) | 0.389 | 0.524 | 0.524 | 0.333 | 0.714 | 0.710 |
| Clinical only / TabPFN | 0.643 (0.321–0.905) | 0.554 | 0.548 | 0.531 | 0.167 | 0.929 | 0.804 |
| T1CE (5) / CatBoost | 0.595 (0.310–0.846) | 0.514 | 0.560 | 0.561 | 0.333 | 0.786 | 0.705 |
| T1CE (5) / TabPFN | 0.655 (0.357–0.905) | 0.545 | 0.560 | 0.561 | 0.333 | 0.786 | 0.786 |
| T1CE (5) + clinical / CatBoost | 0.631 (0.345–0.881) | 0.576 | 0.631 | 0.642 | 0.333 | 0.929 | 0.759 |
| **Five-feature combined TabPFN** | 0.726 (0.464–0.929) | 0.625 | 0.631 | 0.642 | 0.333 | 0.929 | 0.805 |
| Fixed 40% Cat + 60% PFN | 0.714 (0.440–0.917) | 0.616 | 0.595 | 0.600 | 0.333 | 0.857 | 0.794 |
| OOF-selected 3/7 Cat + 4/7 PFN | 0.714 (0.440–0.917) | 0.616 | 0.595 | 0.600 | 0.333 | 0.857 | 0.792 |
| Score stacking / ridge | 0.548 (0.250–0.845) | 0.483 | 0.595 | 0.600 | 0.333 | 0.857 | 0.746 |
| Score stacking / TabPFN | 0.619 (0.298–0.905) | 0.534 | 0.631 | 0.642 | 0.333 | 0.929 | 0.807 |
| BIG-Huang / MRI scores only | 0.702 (0.429–0.905) | 0.543 | 0.607 | 0.601 | 0.500 | 0.714 | 0.766 |
| **BIG-Huang / MRI + clinical** | 0.833 (0.619–0.976) | 0.666 | 0.643 | 0.643 | 0.500 | 0.786 | 0.849 |
| BIG-Huang + clinical ridge score CatBoost fusion | 0.821 (0.583–0.976) | 0.750 | 0.690 | 0.670 | 0.667 | 0.714 | 0.818 |
| BIG-Huang + clinical ridge score TabPFN fusion | 0.798 (0.571–0.952) | 0.682 | 0.595 | 0.600 | 0.333 | 0.857 | 0.835 |
| BIG-Huang + clinical PFN score CatBoost fusion | 0.798 (0.583–0.952) | 0.631 | 0.643 | 0.643 | 0.500 | 0.786 | 0.799 |
| BIG-Huang + clinical PFN score TabPFN fusion | 0.774 (0.524–0.940) | 0.609 | 0.595 | 0.600 | 0.333 | 0.857 | 0.833 |

Same 20 evaluation records; threshold 0.5. Cat = CatBoost; PFN = TabPFN. Higher is favorable in each column.

**FIGURE 14: Annotated performance heatmap for 16 classifier and fusion configurations on the same 20 evaluation examinations, at a fixed classification threshold of 0.5**

AUC is accompanied by its 95% conditional bootstrap interval. The final column is 1 minus the Brier score. Colors encode each displayed metric on a common 0–1 scale. All configurations shown use the group-aware partition

AP: average precision; BA: balanced accuracy; Macro F1: the unweighted mean of class-specific F1 scores; T1CE: contrast-enhanced T1-weighted imaging

Created by Hao Ma and Fengzhang Zhu

The fixed 40% CatBoost/60% TabPFN fusion had an evaluation AUC of 0.714 (95% conditional confidence interval, 0.440-0.917) and a Brier score of 0.206. Ranking 55 candidate proportions by training OOF performance selected 3/7 CatBoost and 4/7 TabPFN, with an OOF AUC of 0.633 (0.396-0.824), an evaluation AUC of 0.714 (0.440-0.917), and a Brier score of 0.208. The clinical-score search selected effective weights of 0.08 CatBoost, 0.12 TabPFN, and 0.80 clinical-only ridge regression, yielding an evaluation AUC of 0.619 (0.309-0.905). TabPFN and ridge regression stacking had evaluation AUCs of 0.619 (0.298-0.905) and 0.548 (0.250-0.845), respectively. These combinations describe probability fusion within the group-aware partition.

In five-fold group-aware outer cross-validation within the training set, with fixed probability weights (3/7 for CatBoost and 4/7 for TabPFN), ridge logistic regression stacking had an AUC of 0.677 (0.457-0.853) and a Brier score of 0.199; TabPFN stacking had an AUC of 0.516 (0.268-0.737) and a Brier score of 0.211. When selection among the 55 candidate weight combinations was restricted to each outer training partition, the two stacking models had AUCs of 0.677 (0.457-0.849) and 0.490 (0.256-0.693), and Brier scores of 0.198 and 0.215, respectively. Fixed-weight results are presented in Table 6; all intervals are 95% conditional

confidence intervals based on outer out-of-fold predictions.

Among the ten BIG-Huang configurations, the configuration with regional TabPFN models, a TabPFN fusion layer, and nine clinical variables achieved the highest observed evaluation AUC of 0.833 (95% conditional confidence interval, 0.619-0.976), with a Brier score of 0.151, TN = 11, FP = 3, FN = 3, and TP = 3. Its eleven fusion inputs comprised two regional probabilities and nine raw clinical variables. The structurally identical configuration without clinical variables had an AUC of 0.702 (0.429-0.905) and a Brier score of 0.234, with a paired AUC difference of 0.131 (−0.071 to 0.393) and a Brier difference of −0.083 (−0.190 to 0.009). The eleven-input configuration had a training outer OOF AUC of 0.422 (0.231-0.625); the top-ranked configuration in training was the score-only fusion of regional TabPFN models with CatBoost, with an OOF AUC of 0.477 (0.282-0.683) and an evaluation AUC of 0.762 (0.524-0.940). The highest evaluation result and the candidate ranked first in training are therefore reported separately.

None of the four twelve-input clinical-score extensions surpassed the observed AUC or Brier performance of the eleven-input candidate. Among these extensions, the clinical ridge regression score with CatBoost fusion had the highest evaluation AUC of 0.821 (95% conditional confidence interval, 0.583-0.976) and a Brier score of 0.182. The XGBoost/Boruta configuration based on the reference structure retained two imaging scores, with no clinical variables confirmed, and had an AUC of 0.518 (0.226-0.792). The compact overview displays 119 evaluation configurations and training OOF results for 92 of them; complete outer cross-validation results for stacking are presented in Table *6*. Proportion aliases and branches with identical inputs do not represent independent models. The complete configuration grid is shown in Figure *15*.

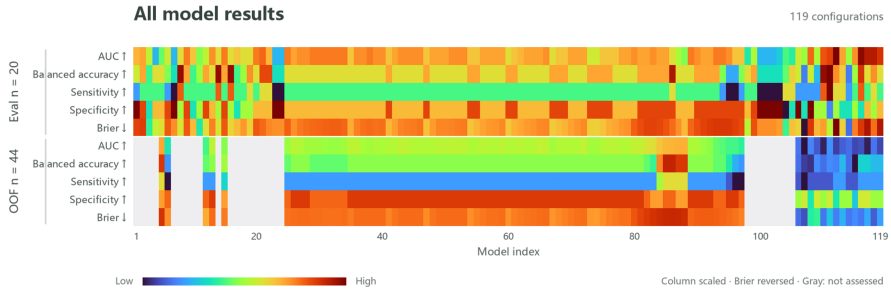
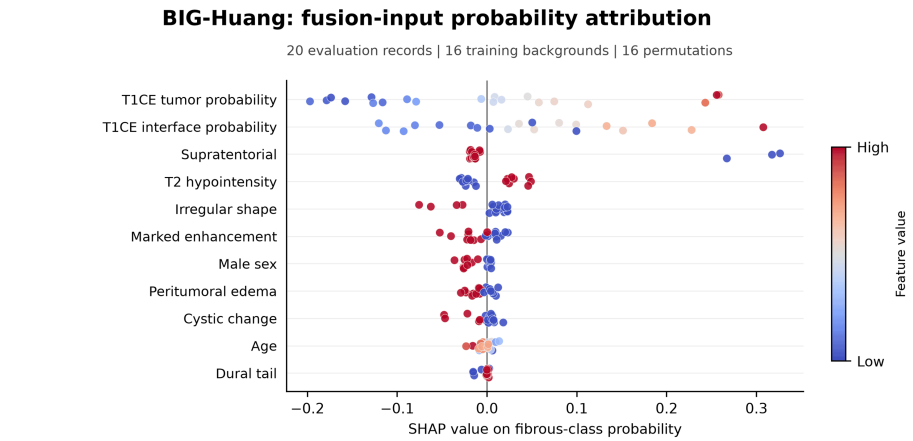



**FIGURE 15: Overview of all 119 registered configurations, including aliases**

The upper block shows 119 fixed 20-record evaluations; the lower block shows 92 training-OOF evaluations. Columns retain the same model order. Each metric is independently normalized for color; Brier is reversed so higher color scores indicate lower probability error. Gray denotes 27 configurations not displayed in the OOF block. Complete stacking outer-CV results are reported in Table 6. No cell values are imputed, and the two evaluation contexts are not pooled

Created by Hao Ma and Fengzhang Zhu

At the 0.30 threshold, net benefit in the structurally matched clinical comparison for BIG-Huang increased from 0.1214 to 0.1714, with a paired difference of 0.0500 (pointwise 95% conditional confidence interval, −0.0429 to 0.1714). Among the 20 evaluation cases, BIG-Huang classified 12 as high risk, including all 6 fibrous cases; standardized per 1,000, these counts corresponded to 600 and 300, respectively. At the 0.50 threshold, 6 cases were classified as high risk, including 3 fibrous cases. In the BIG-Huang fusion function, the T1CE tumor-region score (mean absolute estimate, 0.1166), interface-region score (0.0968), and supratentorial location (0.0577) made the largest contributions; the direct five-feature combined TabPFN placed greater emphasis on MCC (0.1537), sphericity (0.0504), and HHL maximum intensity (0.0464). Attribution, decision curves, and clinical-impact curves are shown in Figures *16*-*18*.




Model-agnostic interventional permutation estimates; background sampled only from the 44 training records.

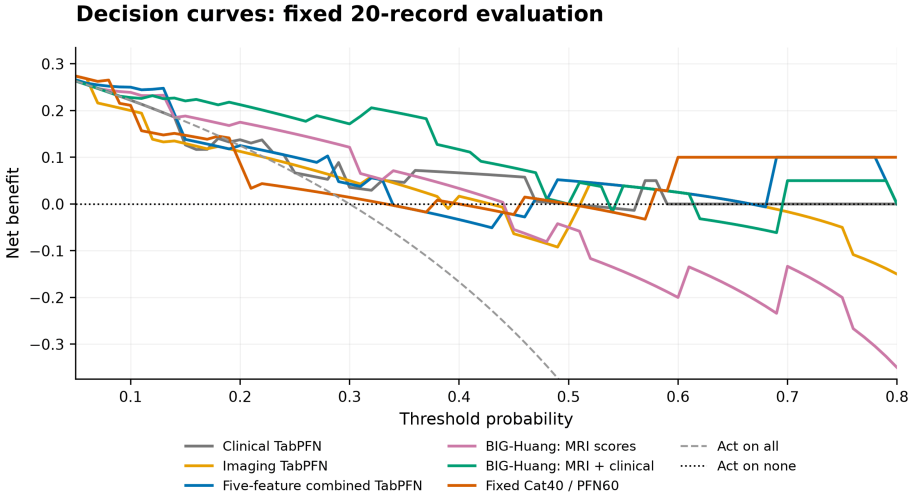
Features retain their fitted input scale. Regional probabilities are fusion inputs, not individual texture effects.

## FIGURE 16: Probability-scale attribution for the BIG-Huang TabPFN fusion function with two regional probabilities and nine clinical inputs

Each dot represents one of the 20 evaluation records; inputs are ordered by mean absolute interventional Shapley estimate, and color indicates the observed feature value within that feature. The reference distribution comprises 16 uniformly sampled training records, and 16 antithetic feature permutations were used. Regional probability attribution is at the fusion-input level and is not an attribution to individual underlying textures

T1CE: contrast-enhanced T1-weighted imaging

Created by Hao Ma and Fengzhang Zhu




Display thresholds fixed at 0.05-0.80, step 0.01; no threshold optimization.

Act-on-all continues below the plotted range. Net benefit uses each dataset's event proportion.

## FIGURE 17: Decision curves on the fixed 20-record evaluation set (six fibrous and 14 non-fibrous examinations)

Net benefit is TP/n − FP/n × t/(1−t). The threshold grid was fixed at 0.05–0.80 in steps of 0.01. Curves represent six designated saved models and the act-on-all/act-on-none references; the act-on-all curve continues below the displayed y range. The saved group-aware model probabilities were held fixed for this curve analysis

MRI: magnetic resonance imaging

Created by Hao Ma and Fengzhang Zhu

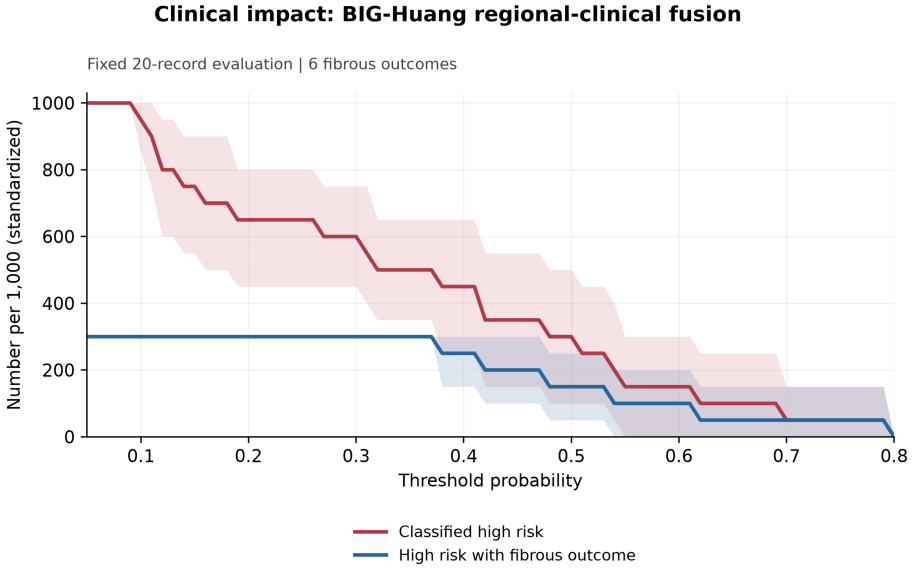
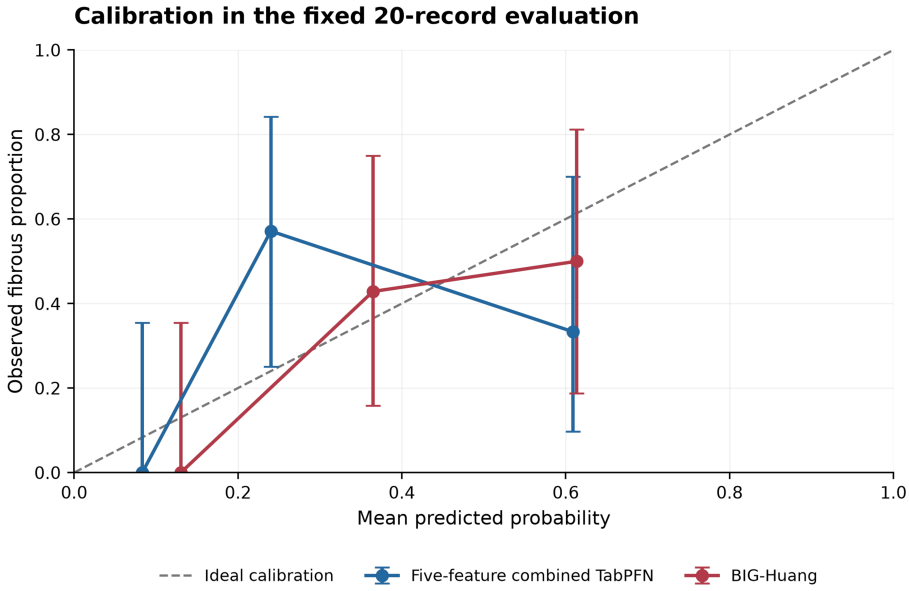



**FIGURE 18: Clinical-impact curve for BIG-Huang regional-clinical fusion in the fixed 20-record evaluation partition**

The red curve shows the number classified as high risk, and the blue curve shows the number classified as high risk with a fibrous outcome, each standardized to 1,000 using the observed record proportions. Shading indicates 95% pointwise conditional percentile intervals from 2,000 outcome-stratified record bootstrap resamples. The threshold range was fixed at 0.05–0.80 with a step of 0.01; no threshold was optimized. The cohort comprised 20 observed records, not 1,000 individuals

Created by Hao Ma and Fengzhang Zhu

In the group-aware validation, combined TabPFN had a mean prediction of 0.296, compared with an observed fibrous proportion of 0.300, a CITL of 0.028, and a calibration slope of 0.631; combined ridge regression had a mean prediction of 0.362 and a CITL of −0.643. The three-group Hosmer-Lemeshow statistics were 6.7573 (df = 1, p = 0.0093) for five-feature combined TabPFN and 1.4976 (df = 1, p = 0.2210) for BIG-Huang, with minimum expected cell counts of 0.5821 and 0.9117, respectively. Using the training OOF median cutoffs, the five-feature combined model had 2/10 fibrous cases in the low-risk group and 4/10 in the high-risk group (cutoff, 0.2229; Fisher p = 0.6285); the corresponding figures for BIG-Huang were 0/7 and 6/13 (cutoff, 0.2018; Fisher p = 0.0515). Grouped analysis and risk stratification were used to evaluate agreement between predicted probabilities and observed fibrous meningioma proportions. Grouped calibration curves are shown in Figure *19*.




Three equal-frequency probability groups (n = 7, 7, 6); vertical bars are Wilson 95% intervals.

**FIGURE 19: Grouped calibration in the fixed 20-record evaluation partition**

Each model is represented by three equal-frequency probability groups of 7, 7, and 6 records; plotted values are the group mean predicted probability and observed fibrous proportion. Vertical bars are Wilson 95% binomial intervals for observed proportions. The dashed line denotes ideal calibration. Grouping is descriptive and is based on predicted probabilities without using outcomes

Created by Hao Ma and Fengzhang Zhu

In the volume analysis of 30 cases, 90 A1/A2/B1 readings yielded an intra-observer ICC of 0.9836 and an inter-observer ICC of 0.9655, supporting the consistency of ROI volumes. Results for the group-aware validation and fusion models are presented in Tables *5*-*6* and Figures *14*-*19*.

| Model/inputs | AUC | CI lower | CI upper | Brier | TN | FP | FN | TP |
|---|---|---|---|---|---|---|---|---|
| Clinical: ridge | 0.583 | 0.274 | 0.881 | 0.241 | 12 | 2 | 4 | 2 |
| Clinical: TabPFN | 0.643 | 0.321 | 0.905 | 0.196 | 13 | 1 | 5 | 1 |
| Clinical: CatBoost | 0.631 | 0.381 | 0.869 | 0.290 | 10 | 4 | 4 | 2 |
| T1CE five: ridge | 0.679 | 0.393 | 0.905 | 0.225 | 10 | 4 | 4 | 2 |
| T1CE five: TabPFN | 0.655 | 0.357 | 0.905 | 0.214 | 11 | 3 | 4 | 2 |
| T1CE five: CatBoost | 0.595 | 0.310 | 0.846 | 0.295 | 11 | 3 | 4 | 2 |
| T1CE five + clinical: ridge | 0.655 | 0.333 | 0.929 | 0.244 | 11 | 3 | 2 | 4 |
| T1CE five + clinical: TabPFN | 0.726 | 0.464 | 0.929 | 0.195 | 13 | 1 | 4 | 2 |
| T1CE five + clinical: CatBoost | 0.631 | 0.345 | 0.881 | 0.241 | 13 | 1 | 4 | 2 |
| T1CE five + clinical: CatBoost 40%/TabPFN 60% | 0.714 | 0.440 | 0.917 | 0.206 | 12 | 2 | 4 | 2 |
| T1CE two: ridge | 0.738 | 0.452 | 0.964 | 0.200 | 12 | 2 | 3 | 3 |
| T1CE two: TabPFN | 0.690 | 0.393 | 0.940 | 0.209 | 12 | 2 | 3 | 3 |
| T1CE two + clinical: ridge | 0.655 | 0.345 | 0.929 | 0.234 | 11 | 3 | 3 | 3 |
| T1CE two + clinical: TabPFN | 0.726 | 0.452 | 0.940 | 0.194 | 12 | 2 | 4 | 2 |

**TABLE 5: Classifier and input comparisons on the group-aware 20-record evaluation partition**

AUC intervals are 95% conditional bootstrap intervals, and classification uses 0.5. These models use a different allocation from the sequence/CNN benchmark

AUC: area under the curve; CI: confidence interval; TN: true negative; FP: false positive; FN: false negative; TP: true positive

Created by Hao Ma and Fengzhang Zhu

| Configuration | Training OOF AUC (95% CI) | Evaluation AUC (95% CI) | Evaluation Brier |
|---|---|---|---|
| TabPFN regions/CatBoost fusion | 0.477 (0.282–0.683) | 0.762 (0.524–0.940) | 0.255 |
| CatBoost regions/TabPFN fusion + clinical | 0.466 (0.294–0.629) | 0.738 (0.464–0.940) | 0.182 |
| CatBoost regions/CatBoost fusion + clinical | 0.456 (0.267–0.637) | 0.643 (0.369–0.893) | 0.235 |
| Direct regional features+clinical / CatBoost | 0.451 (0.259–0.634) | 0.607 (0.333–0.845) | 0.344 |
| TabPFN regions/TabPFN fusion | 0.443 (0.259–0.639) | 0.702 (0.429–0.905) | 0.234 |
| CatBoost regions/TabPFN fusion | 0.438 (0.273–0.606) | 0.619 (0.333–0.857) | 0.224 |
| Direct regional features + clinical/TabPFN | 0.427 (0.261–0.615) | 0.714 (0.452–0.917) | 0.207 |
| TabPFN regions/TabPFN fusion + clinical | 0.422 (0.231–0.625) | 0.833 (0.619–0.976) | 0.151 |
| TabPFN regions/CatBoost fusion + clinical | 0.404 (0.207–0.622) | 0.750 (0.500–0.940) | 0.215 |
| CatBoost regions/CatBoost fusion | 0.393 (0.240–0.562) | 0.607 (0.333–0.845) | 0.372 |
| 3/7 CatBoost + 4/7 TabPFN | 0.633 (0.396–0.824) | 0.714 (0.440–0.917) | 0.208 |
| 8% CatBoost + 12% TabPFN + 80% clinical ridge | 0.693 (0.463–0.890) | 0.619 (0.309–0.905) | 0.211 |
| Clinical TabPFN score/CatBoost fusion | 0.456 (0.271–0.654) | 0.798 (0.583–0.952) | 0.201 |
| Clinical ridge score/CatBoost fusion | 0.432 (0.247–0.636) | 0.821 (0.583–0.976) | 0.182 |
| Clinical TabPFN score/TabPFN fusion | 0.427 (0.238–0.634) | 0.774 (0.524–0.940) | 0.167 |
| Clinical ridge score/TabPFN fusion | 0.401 (0.218–0.603) | 0.798 (0.571–0.952) | 0.165 |
| Ridge logistic stacking (fixed weights) | 0.677 (0.457–0.853) | 0.548 (0.250–0.845) | 0.254 |
| TabPFN v2 stacking (fixed weights) | 0.516 (0.268–0.737) | 0.619 (0.298–0.905) | 0.193 |

**TABLE 6: BIG-Huang, probability-combination, and clinical-score models**

The 10 regional configurations are ordered by training-OOF rank; further configurations follow below. AUC intervals are 95% conditional intervals from fixed predictions. For the two stacking rows, training OOF AUC is from complete five-fold group-aware outer cross-validation with fixed 3/7 CatBoost + 4/7 TabPFN weights; evaluation columns report the 20-record results for the same configurations. Weight-search training OOF results are tuning scores. The same group-aware 44/20 partition is used throughout

AUC: area under the curve; CI: confidence interval; OOF: out-of-fold

# Discussion

In 64 WHO grade 1 meningiomas, this study compared MRI sequences, imaging modeling strategies, and the effects of adding clinical variables. The branch using two T1CE MCC features achieved a relatively high AUC, whereas the main six-feature branch retained a broader range of texture, intensity, and shape information. After clinical variables were added, their AUCs were 0.905 (95% conditional confidence interval 0.726-1.000) and 0.810 (0.524-1.000), respectively, and both had lower Brier scores than their corresponding imaging-only branches. These results indicate the potential of combining a compact set of radiomics features with pretrained TabPFN for the preoperative prediction of fibrous meningioma. Adding MRI sequences or expanding the extent of CNN fine-tuning had different effects on discrimination, classification performance, and probability quality.

## Objective radiomics features and their information content

The six T1CE candidate features represent different aspects of tumor imaging. F1 and F2 are GLCM maximal correlation coefficients from different wavelet subbands and describe the dependence structure of the transformed gray-level co-occurrence distribution. GLRLM long-run high-gray-level emphasis characterizes contiguous runs of high intensity; maximum 3D diameter reflects geometric scale; GLCM inverse difference reflects local gray-level homogeneity; and first-order energy summarizes the sum of squared intensities in a filtered image [6,7]. Together, these features provide quantitative information about tumor size, intensity magnitude, and spatial texture, complementing visual assessment of internal structure.

The performance of F1/F2 in the exploratory branch suggests that these two directional wavelet subbands contain co-occurrence information that warrants further investigation. At a fixed $\lambda$, their selection

frequencies were 94% and 77%, respectively, compared with 21%-45% for the other four features, yielding a compact candidate feature list. Wavelet subbands are mathematical decompositions; HHL and HLL have not been directly linked to collagen orientation, cellular arrangement, or specific pathological components. SHAP analysis shows their contributions after inclusion in the final combined models (Figures *12*-*13*).

The two-feature combined model used fewer radiomics inputs and achieved the same classification counts as the six-feature combined model at a threshold of 0.5, correctly identifying 4 fibrous and 12 non-fibrous meningiomas in each case. The two models nevertheless differed in AUC and probability error, reflecting differences in predicted probabilities and case ranking.

For the six-feature branch, the minimum-deviance LASSO rule retained six features, whereas the more conservative one-standard-error rule retained none. This indicates that the candidate feature set is sensitive to the regularization criterion. When positive events are few, correlated features may substitute for one another, and the selected feature set may change with sample composition; the observed differences should therefore be evaluated in new patients rather than further optimized on the current validation sample [17]. The six-feature list defines the inputs for the main comparison, whereas the two-feature list examines whether a smaller representation can maintain discrimination. Subsequent validation should assess both the reproducibility of feature computation and model performance to determine whether simplification provides a stable benefit.

SHAP further describes the role of the radiomics features. In the main six-feature combined candidate model, F2 and F1 ranked fourth and fifth among all inputs and first and second among the radiomics features. Both also retained measurable contributions in the exploratory two-feature combined candidate model. MCC dependence plots showed that higher values generally corresponded to more positive contributions to fibrous predictions. Wavelet transformation enables GLCM features to characterize gray-level dependence at different frequencies and orientations, providing information that complements visual assessment of tumor internal structure. SHAP describes the contributions of these features in the final nonlinear combined model, extending the information provided by the LASSO coefficients.

T2 hypointensity had the largest mean absolute SHAP contribution in both combined candidate models. This is consistent with the intratumoral T2 hypointensity sometimes seen in fibrous meningiomas and with imaging-histology studies. Soyama et al. reported lower T2 signal in fibrous tumors and an association with collagen content in 40 meningiomas [26]. In a prospective study of 20 patients, Yoneoka et al. associated shorter T2 with collagen-rich fibrous tissue [27]. A lower proportion of mobile water and restricted water motion in collagen-rich tissue may explain the reduced T2 signal; relationships among tissue water, T2/proton-density signal, and tumor consistency have also been reported [28]. These observations provide histological context for the T2 finding, while the quantitative focus of this study remains T1CE texture and shape. Qualitative T2 hypointensity complements the spatial gray-level information quantified by MCC and other features.

## The contribution of FLAIR and the selection of T1CE

Adding FLAIR to both T1CE branches improved balanced accuracy and macro F1, whereas the corresponding T1CE-only models had higher AUC point estimates. This illustrates the distinction between score ranking and classification at a fixed threshold. In the six-feature branch, adding FLAIR maintained fibrous identification at 5/6 and correctly identified two additional non-fibrous cases. In the two-feature branch, it correctly identified one additional fibrous case and one fewer non-fibrous case. These results describe the complementary information that FLAIR provided for threshold-based classification in this cohort.

The FLAIR interface model alone had an AUC of 0.821 (95% conditional bootstrap confidence interval 0.548-1.000), compared with 0.798 (0.500-1.000) for six-feature T1CE radiomics, supporting further investigation of interface signals. T1CE was selected because cross-sequence AUC gains were inconsistent, and FLAIR added 14 candidate features as well as requirements for sequence quality and processing. This choice reflects a balance between performance and input size in this cohort. Scan duration and economic costs were not measured; simplification therefore primarily concerns the sources of radiomics data and the feature-processing stage. The combined models still use conventional multisequence findings, including T2 hypointensity and edema, and automated segmentation also requires four modalities. The complete workflow therefore remains dependent on multiple sequences.

AP provides another perspective. The two-feature T1CE-FLAIR combination had an AP of 0.886, higher than that of its T1CE-only counterpart (0.749) and the two-feature imaging-clinical candidate model (0.843), despite having a lower AUC than both. AP emphasizes precision at different levels of recall and is particularly meaningful when evaluating the minority class in imbalanced datasets, whereas ROC-AUC summarizes the ranking of positive and negative cases [29]. If future studies prioritize identifying a small number of cases strongly suspected to be fibrous, the FLAIR combination remains a useful comparator, but the evaluation metric should be prespecified before assessment in a new validation set.

## The tabular foundation model and MRI-CNN

TabPFN directly accepts a limited number of continuous and categorical variables and uses inductive structure acquired through pretraining for in-context prediction [8]. This study included only 44 training records, of which 12 were fibrous. Engineered features compress spatial information before classification, providing more constrained inputs than estimating discriminative structure from $64^3$ voxels. TabPFN is itself a neural network; this study compares tabular radiomics and voxel-input CNN strategies.

Under the implemented training and preprocessing settings, different CNN strategies produced different point estimates. Full-network fine-tuning increased apparent training discrimination, whereas evaluation discrimination was lower than that of the frozen encoder; point estimates for partial fine-tuning with internal early stopping are presented in Table *4*. The wide intervals did not establish a ranking of these strategies. In small samples, model complexity, training variation, and input processing can all affect observed performance [17]. The thin interface was first defined on a 1 mm grid and then resized to $64^3$ voxels; interpolation and scale changes may affect its spatial information. The pretrained backbone, spatial scale, and regional inputs were not separately evaluated in ablation analyses.

TabPFN has previously been used for multicenter meningioma grading [30], whereas the present study addresses identification of the fibrous subtype within WHO grade 1. Compared with the clinical-radiomics model reported by Han et al., this study included fewer cases and candidate features and incorporated voxel-input CNN comparators [4]. Differences in the task, population, and evaluation design limit direct ranking across studies by AUC. By using the same validation records, this study illustrates trade-offs among sequences, modeling strategies, and combinations of variables.

## Probability quality and multidimensional evaluation

The combined models showed a trade-off between sensitivity and specificity at the fixed threshold of 0.5. Lower Brier scores and log loss indicate smaller overall probability errors for these records. In practice, different misclassification costs may require different thresholds; the fixed threshold provided a consistent basis for this comparison.

Calibration analysis showed mean predicted probabilities of 0.404 and 0.378 for the six-feature and two-feature combined models, respectively, exceeding the observed proportion of fibrous meningiomas in the evaluation set (0.300); their CITL values were −0.810 and −0.687, respectively, indicating overall overprediction. The two-feature model had a calibration slope close to 1 but still showed an overall calibration shift. Class balancing may alter probability calibration while preserving discrimination [18,31]. Analytical removal of probability balancing was used to examine the effects of adjusting the probability scale on calibration metrics. Its effects can be evaluated in an independent cohort and combined with recalibration, when needed, to further improve individual probability estimates. DCA compared net benefit at different hypothetical thresholds; subsequent work can assess the suitability of these thresholds in relation to specific clinical actions [19].

Multidimensional displays make these trade-offs visible. In the radar plots, the six-feature imaging-only model extended farther along the sensitivity axis, whereas the two-feature model performed better in specificity, AUC, and probability error. After clinical variables were added, the two candidate models overlapped in balanced accuracy, macro F1, sensitivity, and specificity, while differences remained in AUC, AP, and Brier score. The heatmap of all models showed that the multisequence combinations had the highest balanced accuracy point estimate (0.774), slightly above that of the final combined candidate models (0.762). Model selection therefore depends on the intended task and input requirements. Radar polygon area is not an overall performance score.

## Regional and clinical fusion

For the preoperative prediction of fibrous meningioma, this study developed the BIG-Huang fusion framework, which integrates MRI regional modeling and clinical information into a hierarchical prediction process. Its design was inspired by the study by Huang et al. on the fusion of CT habitat scores and clinicopathological information [23]. The present study extends this approach to MRI and binary classification of fibrous meningioma: regional prediction scores are extracted separately from the tumor and interface regions and then entered into a fusion layer together with clinical indicators, with systematic comparisons of CatBoost, TabPFN, and different fusion approaches. The innovation focuses on organizing multiregional imaging and clinical information for this clinical task, rather than merely adding another classifier.

The BIG-Huang configuration that most warrants further development combines regional TabPFN models with a TabPFN fusion layer and retains nine clinical indicators in that layer. In the evaluation of 20 cases under group-aware validation, the structurally matched regional fusion model without clinical indicators had an AUC of 0.702 (95% conditional confidence interval 0.429-0.905); after clinical indicators were added, the AUC was 0.833 (0.619-0.976), and the Brier score decreased from 0.234 to 0.151. The paired AUC difference was 0.131 (95% conditional confidence interval −0.071 to 0.393), and the paired Brier difference was −0.083 (−0.190 to 0.009). This comparison within the same structure directly shows the observed increment after adding clinical information and is more informative about the role of clinical information

than comparisons of the highest values across algorithms or inputs.

Regional scores and clinical indicators have different roles in fusion. The tumor and interface models convert high-dimensional imaging features into two regional probabilities, allowing the fusion layer to use information from both the interior and the boundary region of the lesion. Entering age and conventional MRI findings directly into the fusion layer preserves potential nonlinear relationships between individual clinical variables and regional scores. This hierarchical structure provides a clear route for jointly interpreting imaging phenotypes and conventional assessment. At a threshold of 0.5, the above 11-input configuration produced TN = 11, FP = 3, FN = 3, and TP = 3, further describing its classification behavior at the current threshold.

To determine how clinical information should enter the fusion layer, this study additionally incorporated prediction scores from clinical-only models. The nine raw clinical indicators were retained, and either a clinical logistic regression or a clinical TabPFN score was added, with CatBoost or TabPFN as the fusion layer, yielding four 12-input configurations. Among these, the combination of the clinical logistic regression score with CatBoost fusion had the highest AUC in the 20 cases, at 0.821 (95% conditional confidence interval 0.583-0.976), with a Brier score of 0.182; none of the four extended configurations exceeded the AUC or Brier performance of the 11-input configuration. This comparison suggests that retaining raw clinical variables directly in the regional fusion layer is currently the form of information representation that most warrants priority development; generating an additional prediction score from the same variables is not equivalent to adding new clinical information.

The five-feature comparison under group-aware validation also showed complementarity between clinical and imaging information. Imaging-only TabPFN had an AUC of 0.655 (95% conditional confidence interval 0.357-0.905), increasing to 0.726 (0.464-0.929) with clinical information, with a paired difference of 0.071 (−0.084 to 0.262). BIG-Huang represents the same information-integration task through regional probabilities and clinical inputs. Direct imaging-clinical inputs, regional score fusion, and clinical score extensions provide complementary ways of organizing existing information. The five-feature branch under group-aware validation differs from the six-feature and dual-MCC branches of the benchmark in both the selected representations and the evaluation partition.

## Probability weighting and other fusion strategies

Probability weighting provides another route for combining CatBoost and TabPFN. The fixed 40% CatBoost and 60% TabPFN model had an AUC of 0.714 (95% conditional confidence interval 0.440-0.917) and a Brier score of 0.206 in the 20 cases. Normalization of 81 raw weight combinations yielded 55 distinct ratios. The training OOF criterion selected the adjacent ratio of 3/7 CatBoost and 4/7 TabPFN, with a training OOF AUC of 0.633 (95% conditional confidence interval 0.396-0.824) and an evaluation AUC of 0.714 (0.440-0.917). Systematic weighting therefore identified a range around the fixed 40%/60% ratio with similar probability-ranking performance. Proportional raw weights correspond to the same normalized ensemble and should be understood as equivalent configurations.

Introducing a clinical-only score and a second fusion layer further examined how existing information could be organized. A clinical score search across 18 candidates selected a combination with effective weights of 0.08 for combined CatBoost, 0.12 for combined TabPFN, and 0.80 for clinical-only ridge logistic regression. Its training OOF AUC was 0.693 (95% conditional confidence interval 0.463-0.890), and its evaluation AUC was 0.619 (0.309-0.905). Another stacking configuration combined the 3/7-4/7 ensemble score, the imaging-clinical TabPFN score, and nine clinical variables. With TabPFN fusion, its evaluation AUC was 0.619 (95% conditional confidence interval 0.298-0.905) and its Brier score was 0.193; with ridge regression fusion, the AUC was 0.548 (0.250-0.845), and the Brier score was 0.254.

In full group-aware outer validation, fixed-weight ridge logistic regression stacking had a paired AUC difference of 0.078 (95% conditional confidence interval −0.044 to 0.205) and a Brier difference of −0.006 (−0.055 to 0.046) relative to the base imaging-clinical TabPFN model. Group-aware outer cross-validation further described the performance of the different stacking configurations. Compared with the TabPFN fusion layer, the ridge logistic regression fusion layer achieved a higher AUC point estimate and a lower Brier score point estimate. Results with fixed weights were similar to those with weights selected within the training partition, and additional weight searching showed no further benefit, supporting priority use of the simpler fixed-weight stacking configuration. In the probability ensembles tested, the base combined models already included clinical variables; BIG-Huang instead retains separate regional imaging summaries and supplies raw clinical variables directly to the fusion layer.

CatBoost and TabPFN were treated as interchangeable regional learners and fusion learners in the regional fusion comparisons. In addition to direct combined modeling, this study examined fixed probability weighting, a weight grid, late fusion of clinical-only models, and hierarchical stacking. The results of different combinations provide specific evidence for framework design: in the evaluation of 20 cases, the two-layer TabPFN configuration that retained regional information and directly incorporated clinical variables showed the most favorable observed AUC and Brier performance among the regional configurations tested; the additional clinical prediction scores and stacking approaches tested did not

further improve these two metrics. Further development of BIG-Huang can therefore focus on effective integration of regional representations and clinical information. This 11-input candidate had an AUC of 0.422 (95% conditional confidence interval 0.231-0.625) in outer cross-validation within the training set, differing from its performance in the evaluation of 20 cases; the next phase should prioritize testing the reproducibility of this configuration in a new cohort.

## Data boundaries for feature selection and measurement reproducibility

This study first fixed the training set of 44 cases and the evaluation set of 20 cases, then estimated preprocessing parameters, selected features, and fitted models within the training data. Missing-value imputation, filtering of constant and duplicate features, correlation filtering, and standardization in LASSO cross-validation were all refitted within the corresponding training folds, preventing evaluation labels from contributing to feature formation. Because supervised feature selection uses pathological labels to determine predictors, this study treats it as part of model fitting and consistently observes the training-data boundary [15,16].

Within the 44 training cases, five T1CE tumor features were retained under the minimum-deviance rule and four under the 1-SE rule; no features were retained from either the FLAIR tumor or interface region. These results support compact modeling based on T1CE features and show how different regularization rules affect feature-set size. Building on this, the study further incorporated clinical variables and compared predictive performance across input combinations and classification algorithms.

This study assessed the agreement of ROI volume measurements after repeated delineation in 30 cases. Each case had three sets of measurements, A1, A2, and B1, yielding 90 volume readings. The intra-observer and inter-observer ICCs were 0.9836 and 0.9655, respectively, supporting high agreement of ROI volume measurements [13].

## Attribution, probability quality, and clinical interpretation

Attribution on the probability scale further distinguishes regional fusion from direct feature inputs [24]. Tumor and interface probabilities had the largest mean contributions to the BIG-Huang fusion layer, followed by supratentorial location. In the five-feature imaging-clinical model under group-aware validation, MCC, sphericity, and filtered maximum intensity had the three largest mean absolute contributions. These input-level attributions explain the fitted prediction functions and complement the SHAP analysis of the benchmark; they do not represent causal effects or directly measure histological components.

Decision curves and clinical impact curves extend the comparison from discrimination to threshold-based decisions [19]. At a threshold of 0.30, adding nine clinical variables to the structurally matched BIG-Huang model increased observed net benefit from 0.1214 to 0.1714, with a paired difference of 0.0500 (95% conditional confidence interval −0.0429 to 0.1714). At this threshold, the clinical impact curve classified 12 of the 20 cases as high risk and captured all six fibrous outcomes; on a common population scale, this corresponds to 600 high-risk classifications and 300 captured fibrous outcomes per 1,000 individuals. This is a complementary expression of the same threshold-based predictions and does not represent a separate cohort of 1,000 individuals. Defining the corresponding clinical actions and their benefit-harm thresholds will be the next step toward clinical application.

Probability balancing and model calibration need to be interpreted together [18,31]. The five-feature combined TabPFN model under group-aware validation had a mean predicted probability of 0.296 and a calibration slope of 0.631; the corresponding combined logistic regression model had a mean predicted probability of 0.362 and a CITL of −0.643. Grouped calibration and exploratory Hosmer-Lemeshow diagnostics provide direct assessments of probability quality [25]. Using risk strata defined from the training data, 6 of the 13 cases in the BIG-Huang high-risk group were fibrous, whereas none of the 7 cases in the low-risk group were fibrous. Independent calibration assessment, followed by recalibration when necessary, remains required before using these probabilities to communicate individual clinical risk.

## Limitations and further validation

This single-center retrospective study included MRI acquired across different years and scanners. Differences in acquisition and reconstruction may introduce variation in radiomics measurements even after common preprocessing [7]. The sample and number of positive events were small: correctly classifying one more or one fewer fibrous case changes validation sensitivity by 16.7 percentage points. AUC and calibration estimates therefore have substantial uncertainty. SHAP for the six-feature and dual-MCC combined candidate models used a fixed set of 12 training background records and marginal masking; when predictors are dependent or correlated, attribution may be redistributed among features and should be interpreted as a description of the fitted model rather than as a causal effect [20,32].

This study did not directly measure collagen content, tissue water, or T2 relaxation time. Feature selection under group-aware validation used only training data; it has not yet been established that the feature

inventories of the sequence/CNN benchmark were selected using only training data. Model prioritization drew on performance in the evaluation set of 20 cases, and selection of favorable models may inflate reported performance estimates. Evaluation-set bootstrap intervals resampled cases with fixed predictions, whereas training OOF intervals used outcome-stratified resampling of patient-constrained groups, with records within each group sampled together; these intervals do not include all uncertainty arising from repeating feature selection, model fitting, and strategy selection [17].

The requirements for complete MRI and adequate image quality yielded a surgical cohort suitable for imaging evaluation. Applicability to patients with incomplete sequences or lower image quality requires separate assessment. This study compares the specified TabPFN radiomics and voxel-input CNN workflows. Clinical-only and same-input classifiers provide comparators within the group-aware validation partition. An independent cohort can evaluate the generalizability of these complementary representations. ROI volume agreement and the reproducibility of texture or intensity features assess different measurement properties [13].

Strategy comparisons informed by evaluation performance remain exploratory. The conditional confidence intervals do not encompass uncertainty in the complete fitting and selection process. Volume ICCs do not establish feature-level repeatability; IBSI compliance has not been verified.

## Conclusions

This single-center study compared T1CE and FLAIR radiomics, voxel-input CNNs, and imaging-clinical TabPFN candidate models for identifying fibrous meningioma. The six-feature and two-feature combined candidate models in the benchmark had identical confusion counts at a threshold of 0.5 (TN = 12, FP = 2, FN = 2, TP = 4). Training-only group-aware models and BIG-Huang regional-clinical fusion describe complementary inputs; in structurally matched BIG-Huang models, nine clinical variables improved observed discrimination and the Brier score. The complete workflow warrants evaluation in new patients, and independent calibration assessment, with recalibration where appropriate, is required before clinical risk communication.

## Additional Information

### Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

**Concept and design:** Fengzhang Zhu, Hao Ma, Hui Guo

**Acquisition, analysis, or interpretation of data:** Fengzhang Zhu, Hao Ma, Shizhou Tang, Shiyin Zou, Heji Ma

**Drafting of the manuscript:** Fengzhang Zhu, Hao Ma

**Critical review of the manuscript for important intellectual content:** Fengzhang Zhu, Hao Ma, Shizhou Tang, Shiyin Zou, Heji Ma, Hui Guo

**Supervision:** Hao Ma, Hui Guo

### Disclosures

**Human subjects:** Informed consent for treatment and open access publication was obtained or waived by all participants in this study. Medical Ethics Committee of the First Affiliated Hospital of Jinzhou Medical University issued approval 2026LL-KY-099. The Medical Ethics Committee waived the requirement for informed consent. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

### Acknowledgements

Thank you to all the staff of the Department of Radiology at the First Affiliated Hospital of Jinzhou Medical University. Best wishes to all medical postgraduate students of the Class of 2024 for a bright future!

## References


1. Goldbrunner R, Stavrinou P, Jenkinson MD, et al.: EANO guideline on the diagnosis and management of meningiomas. Neuro Oncol. 2021, 23:1821-34. 10.1093/neuonc/noab150
2. Louis DN, Perry A, Wesseling P, et al.: The 2021 WHO classification of tumors of the central nervous system: a summary. Neuro Oncol. 2021, 23:1231-51. 10.1093/neuonc/noab106
3. Park YW, Oh J, You SC, et al.: 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. 10.1007/s00330-018-5830-3
4. Han T, Liu X, Xu Z, et al.: Preoperative prediction of meningioma subtype by constructing a clinical-radiomics model nomogram based on magnetic resonance imaging. World Neurosurg. 2024, 181:e203-13. 10.1016/j.wneu.2023.09.119
5. Gillies RJ, Kinahan PE, Hricak H: Radiomics: images are more than pictures, they are data . Radiology. 2016, 278:563-77. 10.1148/radiol.2015151169
6. van Griethuysen JJ, Fedorov A, Parmar C, et al.: Computational radiomics system to decode the radiographic phenotype. Cancer Res. 2017, 77:e104-7. 10.1158/0008-5472.CAN-17-0339
7. Zwanenburg A, Vallières M, Abdalah MA, et al.: The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology. 2020, 295:328-38. 10.1148/radiol.2020191145
8. Hollmann N, Müller S, Purucker L, et al.: Accurate predictions on small data with a tabular foundation model. Nature. 2025, 637:319-26. 10.1038/s41586-024-08328-6
9. Collins GS, Moons KG, Dhiman P, et al.: TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024, 385:e078378. 10.1136/bmj-2023-078378
10. Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, Gee JC: N4ITK: improved N3 bias correction. IEEE Trans Med Imaging. 2010, 29:1310-20. 10.1109/TMI.2010.2046908
11. Rohlfing T, Zahr NM, Sullivan EV, Pfefferbaum A: The SRI24 multichannel atlas of normal adult human brain structure. Hum Brain Mapp. 2010, 31:798-819. 10.1002/hbm.20906
12. Isensee F, Schell M, Pflueger I, et al.: Automated brain extraction of multisequence MRI using artificial neural networks. Hum Brain Mapp. 2019, 40:4952-64. 10.1002/hbm.24750
13. Koo TK, Li MY: A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016, 15:155-63. 10.1016/j.jcm.2016.02.012
14. Bettinelli A, Marturano F, Avanzo M, et al.: A novel benchmarking approach to assess the agreement among radiomic tools. Radiology. 2022, 303:533-41. 10.1148/radiol.211604
15. Ambroise C, McLachlan GJ: Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci U S A. 2002, 99:6562-6. 10.1073/pnas.102102699
16. Varma S, Simon R: Bias in error estimation when using cross-validation for model selection . BMC Bioinformatics. 2006, 7:91. 10.1186/1471-2105-7-91
17. Collins GS, Dhiman P, Ma J, et al.: Evaluation of clinical prediction models (part 1): from development to external validation. BMJ. 2024, 384:e074819. 10.1136/bmj-2023-074819
18. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW: Calibration: the Achilles heel of predictive analytics. BMC Med. 2019, 17:230. 10.1186/s12916-019-1466-7
19. Vickers AJ, Elkin EB: Decision curve analysis: a novel method for evaluating prediction models . Med Decis Making. 2006, 26:565-74. 10.1177/0272989X06295361
20. Lundberg SM, Erion G, Chen H, et al.: From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020, 2:56-67. 10.1038/s42256-019-0138-9
21. Liu T, Wang H, Guo Y, et al.: Tabular prior-data fitted network in real-world CT radiomics: benign vs. malignant renal tumor classification. Quant Imaging Med Surg. 2025, 15:10847-61. 10.21037/qims-2025-1132
22. Wu J, Tao G, Xie S, et al.: Prediction of three-year all-cause mortality in patients with heart failure and atrial fibrillation using the CatBoost model. BMC Cardiovasc Disord. 2025, 25:466. 10.1186/s12872-025-04928-w
23. Huang S, Han C, Shi Z, et al.: Postoperative CT habitat fusion model predicts prognosis in non-small cell lung cancer after neoadjuvant immunochemotherapy. NPJ Digit Med. 2026, 5:23-5. 10.1038/s41746-026-03415-6
24. Lundberg SM, Lee SI: A unified approach to interpreting model predictions . Adv Neural Inf Process Syst. 2017, 6:26-8.
25. Hosmer DW, Lemeshow S: Goodness of fit tests for the multiple logistic regression model . Commun Stat Theory Methods. 1980, 9:1043-69. 10.1080/03610928008827941
26. Soyama N, Kuratsu J, Ushio Y: Correlation between magnetic resonance images and histology in meningiomas: T2-weighted images indicate collagen contents in tissues. Neurol Med Chir (Tokyo). 1995, 35:438-41. 10.2176/nmc.35.438
27. Yoneoka Y, Fujii Y, Takahashi H, Nakada T: Pre-operative histopathological evaluation of meningiomas by 3 0T T2R MRI. Acta Neurochir (Wien). 2002, 144:953-7. 10.1007/s00701-002-1005-7
28. Yamaguchi N, Kawase T, Sagoh M, Ohira T, Shiga H, Toya S: Prediction of consistency of meningiomas with preoperative magnetic resonance imaging. Surg Neurol. 1997, 48:579-83. 10.1016/S0090-3019(96)00439-9
29. Saito T, Rehmsmeier M: The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015, 10:e0118432. 10.1371/journal.pone.0118432
30. Karabacak M, Patil S, Feng R, Shrivastava RK, Margetis K: A large scale multi institutional study for radiomics driven machine learning for meningioma grading. Sci Rep. 2024, 14:26191. 10.1038/s41598-024-78311-8
31. 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. 10.1093/jamia/ocac093

32. Aas K, Jullum M, Løland A: Explaining individual predictions when features are dependent: more accurate approximations to Shapley values. Artif Intell. 2021, 298:103502. 10.1016/j.artint.2021.103502