

Contour-free multi-instance learning versus conventional CT features for preoperative Ki-67 stratification in gastric gastrointestinal stromal tumors: a retrospective comparison

Bin Tian

West China Hospital of Sichuan University

Zeng Wang

West China Hospital of Sichuan University

Zhixia Chen

West China Hospital of Sichuan University

Jinge Zhang

West China Hospital of Sichuan University

Xin He

West China Hospital of Sichuan University

Yili Gu

West China Hospital of Sichuan University

Ke Zou

Chengdu Finsiot Technology Co., Ltd

Jun Xie

West China Hospital of Sichuan University

Yu-Xuan Zhang

Southwest Jiaotong University

Xijiao Liu

bleess_jiao@163.com

West China Hospital of Sichuan University

Research Article

Keywords: Gastrointestinal stromal tumor, Ki-67 labeling index, computed tomography, multi-instance learning, artificial intelligence

Posted Date: October 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10941201/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Contour-free multi-instance learning versus
conventional CT features for preoperative Ki-67
stratification in gastric gastrointestinal stromal
tumors: a retrospective comparison

Bin Tian^{1†}, Zeng Wang^{2†}, Zhixia Chen¹, Jing Zhang¹, Xin He²,
Yili Gu¹, Ke Zou³, Jun Xie², Yu-Xuan Zhang⁴, Xijiao Liu^{1*}

¹Department of Radiology, West China Hospital of Sichuan University,
Chengdu, Sichuan, China.

²Information Technology Center, West China Hospital of Sichuan
University, Chengdu, China.

³Chengdu Finsiot Technology Co., Ltd., Chengdu, China.

⁴School of Information Science and Technology, Southwest Jiaotong
University, Chengdu, China.

*Corresponding author(s). E-mail(s): bleess_jiao@163.com;

[†]These authors contributed equally to this work.

Abstract

Background: Manual tumor contouring adds a separate preprocessing step to CT-based prediction. We therefore evaluated whether complete-series multi-instance learning (MIL) could perform preoperative Ki-67 stratification from dual-phase CT without tumor contours and provide greater discrimination than conventional CT features.

Methods: This retrospective study included 61 patients. Ki-67 cutoffs of 5%, 8%, and 10% defined three cutoff-specific tasks. Patients whose reported Ki-67 interval spanned a cutoff were excluded only from the corresponding task. Twenty-one conventional variables were analyzed with two conventional models, and frozen

ResNet-18 features from the complete arterial and venous phase series were analyzed separately with five MIL methods. Four prespecified equal-logit fusion combinations were evaluated. Patient-level stratified fivefold cross-validation was repeated 10 times. Assessment comprised AUROC, AUPRC, calibration assessment, decision-curve analysis, paired bootstrap comparisons with Holm correction, and a post-training attention audit.

Results: The three tasks included 59, 54, and 56 patients, with 41, 26, and 15 positive cases. The highest AUROCs were 0.645 (95% CI, 0.485–0.781) at 5%, 0.646 (95% CI, 0.496–0.788) at 8%, and 0.605 (95% CI, 0.442–0.753) at 10%. The maximum AUPRCs at 5%, 8%, and 10% were 0.822, 0.637, and 0.346. Every AUROC confidence interval included 0.5, and no comparison among the prespecified individual models or between fusion and its components remained significant after Holm correction. In the attention audit, 34 of 60 tests remained significant after global Holm correction, with 21 estimates above and 13 below their prespecified references.

Conclusions: Complete-series MIL was applied separately to arterial and venous CT series for cutoff-specific Ki-67 stratification and did not require tumor contour inputs. It showed no clear discrimination advantage over conventional CT features in this cohort. External validation and direct assessment of annotation burden are required before clinical use.

Keywords: Gastrointestinal stromal tumor, Ki-67 labeling index, computed tomography, multi-instance learning, artificial intelligence

1 Introduction

Gastrointestinal stromal tumors (GISTs) show substantial variation in biological behavior, ranging from indolent tumors to aggressive disease with metastatic potential. Established risk systems integrate tumor size, mitotic count, primary site, and tumor rupture to estimate the risk of recurrence [3, 12, 18]. Although tumor size and primary site can be assessed before surgery, the mitotic count and confirmation of tumor rupture generally depend on intraoperative or pathologic findings. This limits complete preoperative risk assessment. The Ki-67 index is another measure of tumor proliferation, and higher values have been associated with higher risk categories and recurrence [15]. Its assessment also requires biopsy or resected tissue and may be affected by sampling heterogeneity. A preoperative imaging approach could therefore

provide a noninvasive estimate of tumor proliferative activity, although histopathology remains the reference standard.

Contrast-enhanced computed tomography (CT) is routinely used for GIST staging and follow-up. Tumor size, necrosis, and enhancement have been associated with a higher Ki-67 index [26]. Radiomics can quantify heterogeneity within a tumor, but the reproducibility of radiomic results is influenced by image processing and feature definitions [13, 30]. Published CT studies have used Ki-67 cutoffs of 5%, 8%, or 10% and different modeling approaches. These studies have used conventional imaging findings as predictors, applied radiomics or deep learning, or examined spectral parameters [1, 16, 25]. A recent meta-analysis found that diagnostic performance varied according to the cutoff [20]. Direct comparison is limited by differences in patient selection, tumor segmentation, CT acquisition, and validation.

CT models based on radiomics alone or combined radiomic and deep features have relied on manually delineated tumor regions of interest (ROIs) [1, 27]. Manual contouring adds a separate image-processing step and increases the analysis workload. Each patient has one pathologic Ki-67 label, whereas every phase-specific CT series contains multiple slices and only some depict the tumor. This arrangement of one patient-level label and multiple slices can be modeled with multi-instance learning (MIL), which treats the series as a bag and its slices as instances without requiring slice labels or tumor contours [7]. Attention pooling, dual-stream modeling, and attention stabilization represent different MIL designs [11, 14, 28], and MIL has been applied to whole-slide pathology images with slide-level labels and CT images with patient-level labels [2, 4, 17].

This study evaluated contour-free MIL applied separately to complete arterial and venous CT series for gastric GIST Ki-67 stratification at prespecified 5%, 8%, and 10% cutoffs. Two conventional CT models, five MIL methods, and four equal-logit fusion combinations were compared using the same patient-level repeated cross-validation

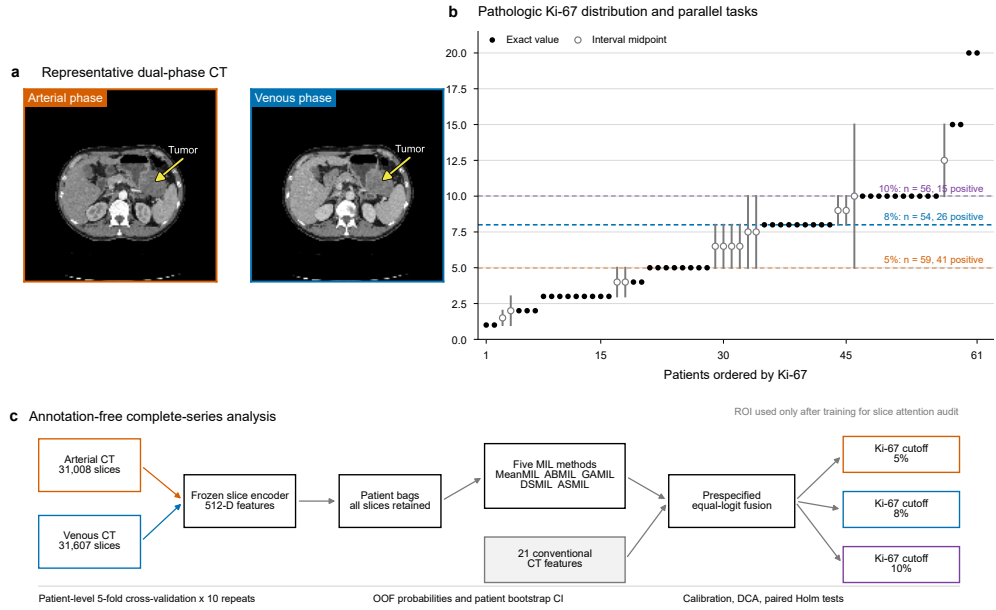


Fig. 1 Clinical overview and study design. **a**, Representative deidentified arterial-phase and venous-phase CT images from a patient with gastric GIST. Arrows indicate the tumor. **b**, Pathologic Ki-67 results for all 61 patients, with interval reports distinguished and cutoffs marked at 5%, 8%, and 10%. **c**, Complete arterial and venous CT series were encoded slice by slice and aggregated separately at the patient level using five MIL methods. Two conventional CT models and four equal-logit fusion combinations were evaluated using the same data partitions. After predictions were finalized, ROI geometry was used only for the attention audit, not as input to the prediction models.

39 splits. The main comparison examined whether MIL or fusion had greater discrimina-
 40 tion than the conventional models. ROIs were used only for the post-training attention
 41 audit. The prediction models did not generate tumor segmentation masks. The clinical
 42 data structure and analysis workflow are shown in Fig. 1.

43 2 Materials and methods

44 2.1 Study population and outcomes

45 2.1.1 Study design and patients

46 This retrospective single-center study included consecutive patients who underwent
 47 radical resection at a tertiary referral center between March 2022 and December 2024.

48 The Biomedical Ethics Review Committee of West China Hospital, Sichuan Univer-
49 sity approved the study (2022 Review No. 449 and 2025 Review No. 86) and waived
50 the requirement for written informed consent because retrospective clinical data were
51 used. All data were deidentified before analysis. Eligibility required postoperative con-
52 firmation of gastric GIST, contrast-enhanced spectral CT of the abdomen and pelvis
53 within 30 days before radical resection, and no tumor-directed treatment before CT.
54 Exclusion criteria were nondiagnostic image quality or tumor rupture. CT was per-
55 formed 2 to 23 days before surgery. The final cohort comprised 61 patients. No formal
56 sample size calculation was performed because the study included all eligible patients
57 available during the study period.

58 **2.1.2 Pathologic Ki-67 assessment**

59 Resected tumors underwent routine histopathologic examination and Ki-67 immuno-
60 histochemistry. The Ki-67-stained section was screened at low magnification, and
61 tumor cells were counted manually at high magnification in the area with the high-
62 est density of Ki-67-positive nuclei. The Ki-67 index was the percentage of counted
63 tumor cell nuclei that showed positive staining. When the report provided a single
64 percentage, that value was used directly. For interval reports, the documented lower
65 and upper bounds were retained rather than replaced by a midpoint. At each cut-
66 off, an interval was labeled positive when its lower bound was at or above the cutoff
67 and negative when its upper bound was below it. Intervals spanning the cutoff were
68 excluded only from that task.

69 **2.2 CT data and features**

70 **2.2.1 CT acquisition and conventional features**

71 CT was acquired with a dual-layer detector-based spectral CT system (HAWK CT,
72 Philips Healthcare). Iodinated contrast material was administered at 1.0 mL/kg and

73 3.0 to 4.5 mL/s. Late arterial phase images were obtained at 50 seconds and venous
74 phase images at 80 seconds. Tube voltage was 120 kVp in 60 patients and 140 kVp
75 in one. Automatic current modulation, pitch 1.386, rotation time 0.27 seconds, and a
76 B reconstruction kernel were used. Conventional images had a 512 by 512 matrix, 1.0
77 mm section thickness, and a 1.0 mm reconstruction interval.

78 Two readers with 10 and 3 years of abdominal imaging experience independently
79 reviewed the CT images while blinded to Ki-67. Disagreements were resolved by con-
80 sensus. The 21 conventional variables covered demographics, tumor location and size,
81 shape, margin, calcification, intratumoral gas, ulceration, growth and enhancement
82 patterns, enlarged feeding or draining vessels, necrosis, cystic change, adjacent-organ
83 invasion, mesenteric fat infiltration, enlarged lymph nodes, and distant metastasis. The
84 prediction models did not use ROI-derived spectral measurements or postoperative
85 pathologic variables.

86 **2.2.2 Image preprocessing and feature extraction**

87 Arterial and venous phase conventional CT series were processed separately using
88 the same pipeline. Within each series, DICOM slices were first ordered by physical
89 position and converted to Hounsfield units using the stored rescale slope and intercept.
90 Intensities were then clipped to an abdominal window with a center of 50 HU and a
91 width of 400 HU and mapped to the range from 0 to 255. To standardize the field
92 of view without locating the tumor, a body mask was generated using a threshold of
93 -500 HU. Five evenly spaced slices were used to define a square body bounding box
94 with a 5% margin, and pixels outside the body mask were set to zero to suppress the
95 table and background. The resulting crop was limited to the patient body and did not
96 use tumor location.

97 After this preprocessing, each axial slice was combined with its immediately pre-
98 ceding and following slices to form a three-channel input. The first or last slice was
99 repeated when an adjacent slice was unavailable. Each input was resized to 224 by 224

pixels with bilinear interpolation, scaled to the range from 0 to 1, and standardized using the ImageNet channel means and standard deviations. A ResNet-18 pretrained on ImageNet was used as a frozen encoder, and global average pooling produced one 512-dimensional feature vector for each axial position [6, 9]. Pathologic labels and ROIs were not used during feature extraction. All feature vectors from each phase were retained as the patient bag for MIL after automated checks of feature dimensions, finite values, slice counts, and slice order.

2.3 Model development and evaluation

ROIs were not inputs to any prediction model and were introduced only after predictions were frozen for the attention audit.

2.3.1 Task definition

The modeling task was to estimate from preoperative CT whether the postoperative pathologic Ki-67 index met or exceeded each cutoff. Each patient i constituted one independent sample. The conventional input was the 21-variable feature vector $\mathbf{x}_i$, whereas the image inputs were the complete arterial and venous phase CT series, modeled separately. Let $p \in \{A, V\}$ denote the CT phase and $N_i^{(p)}$ the number of ordered axial positions in that phase. After slice-level feature extraction, each position j was represented by $\mathbf{h}_{ij}^{(p)}$, and all vectors from phase p formed the patient bag

$$\mathcal{B}_i^{(p)} = \left\{ \mathbf{h}_{ij}^{(p)} \right\}_{j=1}^{N_i^{(p)}}, \quad \mathbf{h}_{ij}^{(p)} \in \mathbb{R}^{512}. \quad (1)$$

Supervision was provided only by the patient-level pathologic Ki-67 result. For an exact percentage v_i , the bounds were set to $L_i = U_i = v_i$. For an interval report, L_i and U_i were its documented lower and upper bounds. With $\mathcal{C} = \{5\%, 8\%, 10\%\}$

denoting the cutoff set, the label for cutoff $c \in \mathcal{C}$ was

$$y_i^{(c)} = \begin{cases} 1, & L_i \geq c, \\ 0, & U_i < c, \\ \text{undefined}, & L_i < c \leq U_i. \end{cases} \quad (2)$$

A patient with an undefined label was excluded only from the task defined by c and remained eligible for the other cutoff tasks. For each cutoff, conventional models received $\mathbf{x}_i$, whereas phase-specific MIL models received $\mathcal{B}_i^{(p)}$. Each model returned the patient-level probability that Ki-67 met or exceeded that cutoff.

2.3.2 Conventional CT models

The 21 variables were modeled using class-balanced logistic regression with either L_2 or elastic-net regularization. Continuous variables were median-imputed and standardized, whereas categorical variables were mode-imputed and one-hot encoded. These transformations and the classifier formed a single pipeline. Preprocessing parameters were estimated only from the patients used for fitting. For both regularization schemes, the candidate values of C were $\{0.03, 0.1, 0.3, 1, 3, 10\}$. Elastic net additionally used L_1 ratios of $\{0.1, 0.5, 0.9\}$.

2.3.3 Multi-instance learning models

Phase-specific MIL models reduced each complete patient bag to a patient representation. Attention-based models assigned a learned weight to each CT slice and combined

137 the weighted features for classification. The slice features were aggregated as

$$\begin{aligned}
 a_{ij}^{(p)} &= \frac{\exp\left[g\left(\mathbf{h}_{ij}^{(p)}\right)\right]}{\sum_{k=1}^{N_i^{(p)}} \exp\left[g\left(\mathbf{h}_{ik}^{(p)}\right)\right]}, \\
 \mathbf{z}_i^{(p)} &= \sum_{j=1}^{N_i^{(p)}} a_{ij}^{(p)} \mathbf{h}_{ij}^{(p)},
 \end{aligned}
 \tag{3}$$

138 where $g(\cdot)$ is learned, $a_{ij}^{(p)}$ is the normalized weight for slice j , and $\mathbf{z}_i^{(p)}$ is the resulting
 139 bag representation. Five methods were compared. MeanMIL averaged instance log-
 140 its. ABMIL used learned attention, and GAMIL added gated attention [11]. DSMIL
 141 combined critical-instance and bag streams [14]. ASMIL used aggregation tokens and
 142 online-anchor attention stabilization [28].

143 2.3.4 Fusion strategy

144 Fusion components were prespecified as conventional elastic net, arterial ABMIL, and
 145 venous ABMIL. The four combinations were arterial plus venous, conventional plus
 146 arterial, conventional plus venous, and all three components. For component set $\mathcal{M}$,
 147 probabilities and validation-derived Youden thresholds [29] were fused with equal logit
 148 weights as follows.

$$\begin{aligned}
 \hat{p}_{i,F}^{(c)} &= \sigma \left[\frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \text{logit} \left(\hat{p}_{i,m}^{(c)} \right) \right], \\
 q_F^{(c)} &= \sigma \left[\frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \text{logit} \left(q_m^{(c)} \right) \right].
 \end{aligned}
 \tag{4}$$

149 This rule did not require fitting an additional fusion model.

150 2.3.5 Training and evaluation

151 Evaluation used patient-level stratified fivefold cross-validation repeated 10 times.
 152 Common outer-fold assignments were generated for all three cutoff tasks by stratify-
 153 ing the midpoint of each reported Ki-67 result into four bands. The bands were below

154 5%, 5% to below 8%, 8% to below 10%, and 10% or greater. The midpoint was used
155 for stratification and not for task-label assignment. After the task-specific interval
156 exclusions defined in Section 2.3.1, each outer training set was further split into an
157 inner training partition containing 80% of the patients and an inner validation par-
158 tition with the remaining 20%, both stratified by the binary label. During candidate
159 selection, models were fitted on the inner training partition. Hyperparameters and
160 operating thresholds were selected on the inner validation partition. Every training,
161 validation, and outer test partition contained both classes, and all slices from a patient
162 remained in the same partition.

163 Within each outer split, conventional hyperparameters were chosen by validation
164 AUROC, with AUPRC resolving ties. These comprised C for L_2 logistic regression
165 and both C and the L_1 ratio for elastic net. MIL models used AdamW, class-weighted
166 binary cross-entropy, one patient bag per batch, and a maximum of 100 epochs. For
167 each of four joint learning-rate, dropout, and weight-decay configurations, the check-
168 point with the lowest validation loss was retained, and training stopped after 15 epochs
169 without improvement in validation loss. The four configurations were then ranked
170 by the validation loss of their retained checkpoints, with AUROC and then AUPRC
171 resolving ties.

172 After model selection, each component operating threshold was derived from its
173 inner validation predictions using the Youden index. The selected conventional pipeline
174 and MIL configuration were then refitted on the complete outer training set, compris-
175 ing the inner training and validation partitions, with MIL trained for the selected epoch
176 count. Outer test patients were used for prediction only after the model configuration
177 and operating threshold had been fixed.

178 Across the five outer folds, each eligible patient received one out-of-fold (OOF)
179 probability per repeat for each model. The 10 repeat-level probabilities were averaged
180 into a single patient-level prediction used for AUROC and AUPRC [21]. Secondary

181 assessment comprised sensitivity and specificity at the fold-specific validation-derived
182 thresholds, Brier score, calibration plots, and decision curves [23, 24]. Agreement
183 among the fusion components was assessed with Spearman correlations between
184 patient-level OOF probability means. Repeat-to-repeat stability was summarized by
185 the within-patient SD of OOF probabilities across repeats (Section 3.3).

186 All statistical resampling was performed at the patient level. Confidence intervals
187 for model performance used 2,000 patient-stratified bootstrap samples [8], whereas
188 AUROC and AUPRC differences used 10,000 synchronized paired bootstrap samples.
189 Holm correction was applied within prespecified comparison families [10]. For the six
190 conventional and single-phase MIL models, each cutoff and metric defined 15 pair-
191 wise comparisons. Fusion analyses used nine fusion versus component comparisons for
192 each cutoff and metric. The exploratory MIL analysis used 10 method pairs for each
193 cutoff, phase, and metric. Each of these sets was corrected separately. Uncorrected
194 rankings were treated as descriptive. Analyses used Python 3.11, PyTorch 2.8, and
195 scikit-learn 1.7 with deterministic seeds. TRIPOD+AI and the 2024 CLAIM update
196 were consulted during manuscript preparation [5, 22]. PROBAST+AI was used to
197 review potential sources of bias [19].

198 2.4 Attention audit

199 Attention weights from ABMIL and GAMIL were compared between slices with and
200 without tumor in 12 cutoff, phase, and model combinations. ROIs were mapped
201 to slices by DICOM coordinates after all predictions were frozen. Attention mass,
202 slice-level AUROC and AUPRC, top-one hit rate, and top- K overlap were calcu-
203 lated relative to their random or uniform references. Confidence intervals and tests
204 were calculated with 10,000 patient-level bootstrap samples. Holm correction was
205 applied across all 60 tests. One patient lacked an available ROI in both phases and

Table 1 Characteristics of the study cohort

Characteristic	Value	Characteristic	Value
Age, years	60 (52–67)	Gastric location	61 (100.0)
Male sex	35 (57.4)	Growth pattern, exophytic	26 (42.6)
Ki-67 midpoint, %	6.5 (3.0–10.0)	Growth pattern, endophytic	20 (32.8)
Interval-reported Ki-67	14 (23.0)	Growth pattern, mixed	14 (23.0)
Long diameter, mm	29 (21–50)	Growth pattern, intramural	1 (1.6)
Short diameter, mm	24 (16–40)	Heterogeneous enhancement	38 (62.3)
Well-defined margin	54 (88.5)	Enhancement degree, mild	11 (18.0)
Oval shape	46 (75.4)	Enhancement degree, moderate	31 (50.8)
Lobulated shape	8 (13.1)	Enhancement degree, marked	19 (31.1)
Irregular shape	5 (8.2)	Enlarged feeding or draining vessels	34 (55.7)
Regular shape	2 (3.3)	Necrosis	37 (60.7)
Calcification	12 (19.7)	Cystic change	3 (4.9)
Intratumoral gas	11 (18.0)	Adjacent organ invasion	1 (1.6)
Ulcer	13 (21.3)	Mesenteric fat infiltration	3 (4.9)
Enlarged lymph nodes	1 (1.6)	Distant metastasis	0 (0.0)

Data are median (IQR) or number (percentage). Equivalent category labels in the clinical data extraction sheet were harmonized before tabulation. IQR, interquartile range.

was excluded from the audit. Model development and prediction retained the full task-specific cohorts. This audit measured association rather than tumor localization.

3 Results

3.1 Cohort and task composition

The 61 patients had a median age of 60 years (IQR, 52–67). Of these, 35 (57.4%) were male. The median of the Ki-67 midpoints was 6.5% (IQR, 3.0%–10.0%), and 14 results (23.0%) were intervals. Complete series comprised 31,008 arterial and 31,607 venous slices. After task-specific interval exclusions, the 5%, 8%, and 10% tasks included 59, 54, and 56 patients, with 41, 26, and 15 positive cases. Cohort characteristics are reported in Table 1, and the study workflow is shown in Fig. 1.

3.2 Model discrimination

Every AUROC confidence interval included 0.5, and none of the 90 paired model comparisons was significant after Holm correction (Fig. 2). The models with the

highest point estimates differed by cutoff. At 5%, conventional logistic regression yielded an AUROC of 0.645 (95% CI, 0.485–0.781) and an AUPRC of 0.822 (95% CI, 0.740–0.895). At 8%, conventional elastic net had an AUROC of 0.646 (95% CI, 0.496–0.788) and an AUPRC of 0.637 (95% CI, 0.512–0.802). At 10%, conventional plus venous ABMIL had the highest AUROC, 0.605 (95% CI, 0.442–0.753), whereas venous ABMIL had the highest AUPRC, 0.346 (95% CI, 0.264–0.573). Threshold-dependent operating characteristics also varied across models. At the fold-specific validation-derived thresholds, sensitivity ranged from 0.000 to 0.732 and specificity from 0.167 to 0.927. Point estimates did not consistently favor either CT phase across the three cutoffs. Conventional plus venous ABMIL had the highest fusion AUROC at all cutoffs. None of 54 fusion-versus-component comparisons was significant after correction.

3.3 Component agreement and stability

Fusion-component correlations were low to moderate, with Spearman ρ ranging from -0.156 to 0.382 (Table 2). Median within-patient OOF probability SD ranged from 0.165 to 0.261 . Three-component fusion was less variable than either phase-specific ABMIL model but more variable than conventional elastic net.

3.4 Calibration and decision analysis

Calibration intervals overlapped substantially, and decision curves crossed frequently. No model had consistently greater net benefit across all cutoffs and risk thresholds (Fig. 3). The lowest Brier scores at 5%, 8%, and 10% were 0.217, 0.236, and 0.210.

3.5 Attention audit

Reference-adjusted attention results differed across cutoffs and measures (Fig. 4). The audit included 58, 53, and 55 patients at the three cutoffs. Of 60 tests, 34 remained significant after global Holm correction, with 21 estimates above and 13 below their references. All 12 significant results at 5% were above their references, whereas significant

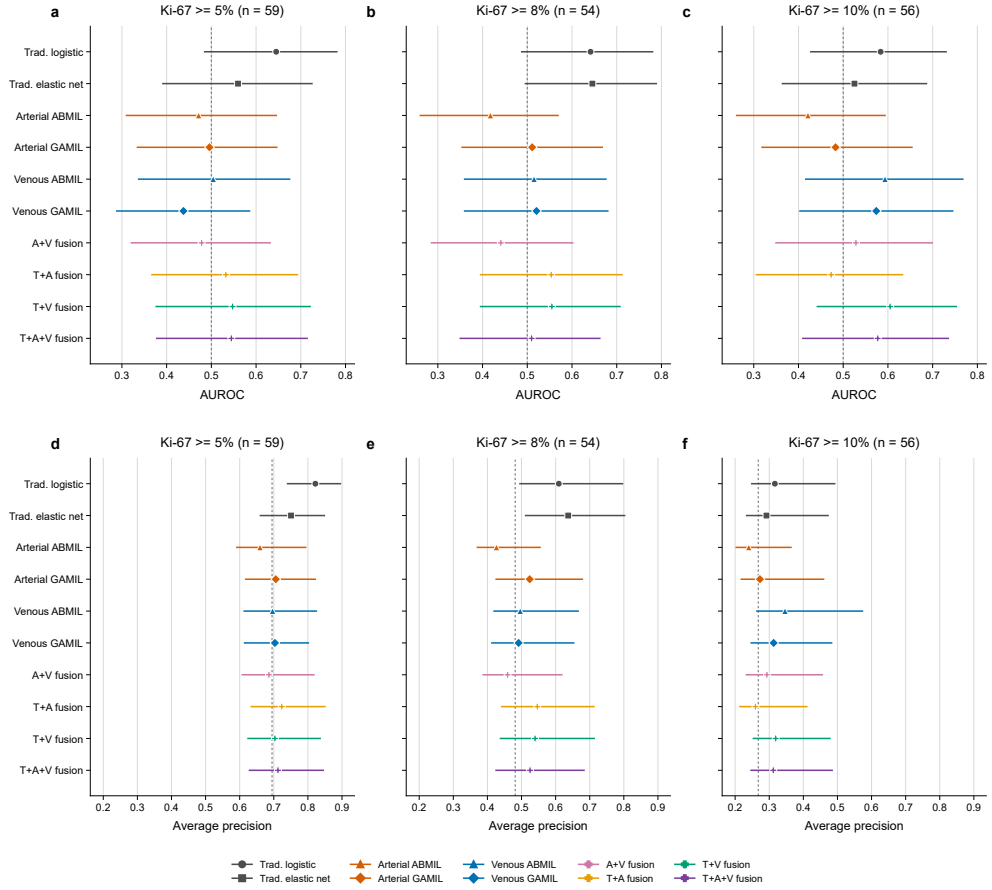


Fig. 2 Discrimination across models. Point estimates and 95% patient-stratified bootstrap confidence intervals are shown for AUROC (a–c) and AUPRC (d–f) across the three Ki-67 cutoffs. Dashed lines indicate an AUROC of 0.5 or the cutoff-specific positive prevalence for AUPRC. None of the 90 paired model comparisons or 54 fusion-versus-component comparisons was significant after Holm correction.

244 results at 8% and 10% included estimates on both sides of the reference. Slice-
 245 level AUPRC exceeded prevalence in all 12 cutoff, phase, and model combinations.
 246 Estimates for the other measures were not uniformly above their references.

247 3.6 Comparison of MIL methods

248 The MIL method with the highest point estimate varied by phase, cutoff, and met-
 249 ric. Only two of 120 paired comparisons of five MIL methods (10 pairs, 2 phases,
 250 3 cutoffs, and 2 metrics) remained significant (Fig. 5). Venous DSMIL exceeded

Table 2 Agreement and repeat-to-repeat variability among fusion components

a. Component agreement			
Ki-67 cutoff	$\rho(T,A)$	$\rho(T,V)$	$\rho(A,V)$
5%	-0.065	-0.106	0.100
8%	0.077	0.221	0.382
10%	-0.049	-0.156	0.164
b. Repeat-to-repeat prediction variability			
Ki-67 cutoff	Prediction model	Median SD (IQR)	
5%	T	0.165 (0.132–0.201)	
	A	0.221 (0.190–0.261)	
	V	0.228 (0.187–0.253)	
	T+A+V	0.186 (0.148–0.213)	
8%	T	0.168 (0.135–0.196)	
	A	0.231 (0.201–0.257)	
	V	0.261 (0.226–0.293)	
	T+A+V	0.186 (0.163–0.213)	
10%	T	0.166 (0.135–0.207)	
	A	0.207 (0.176–0.244)	
	V	0.245 (0.212–0.271)	
	T+A+V	0.175 (0.148–0.208)	

Spearman correlations were calculated from patient-level OOF probability means. SD was calculated for each patient across 10 OOF predictions and is reported as median (IQR) across patients. T denotes conventional elastic net. A denotes arterial ABMIL. V denotes venous ABMIL. OOF denotes out-of-fold. SD denotes standard deviation.

venous ASMIL for 8% AUROC by 0.238 (95% CI, 0.080–0.398, Holm $P = 0.034$).

It also exceeded venous GAMIL for 10% AUPRC by 0.193 (95% CI, 0.058–0.370,

Holm $P = 0.026$). After Holm correction, no tested comparison showed a statistically

significant advantage of an MIL or fusion model over a conventional CT model.

4 Discussion

This study evaluated complete-series MIL for cutoff-specific Ki-67 stratification with-

out tumor contour inputs. Patient-level predictions were generated from complete

arterial or venous CT series, in which only a subset of slices contained tumor. This

patient-bag formulation provides a way to investigate imaging biomarkers without

lesion localization as a preprocessing step. However, neither MIL nor fusion showed

a Holm-corrected discrimination advantage over conventional CT models, and every

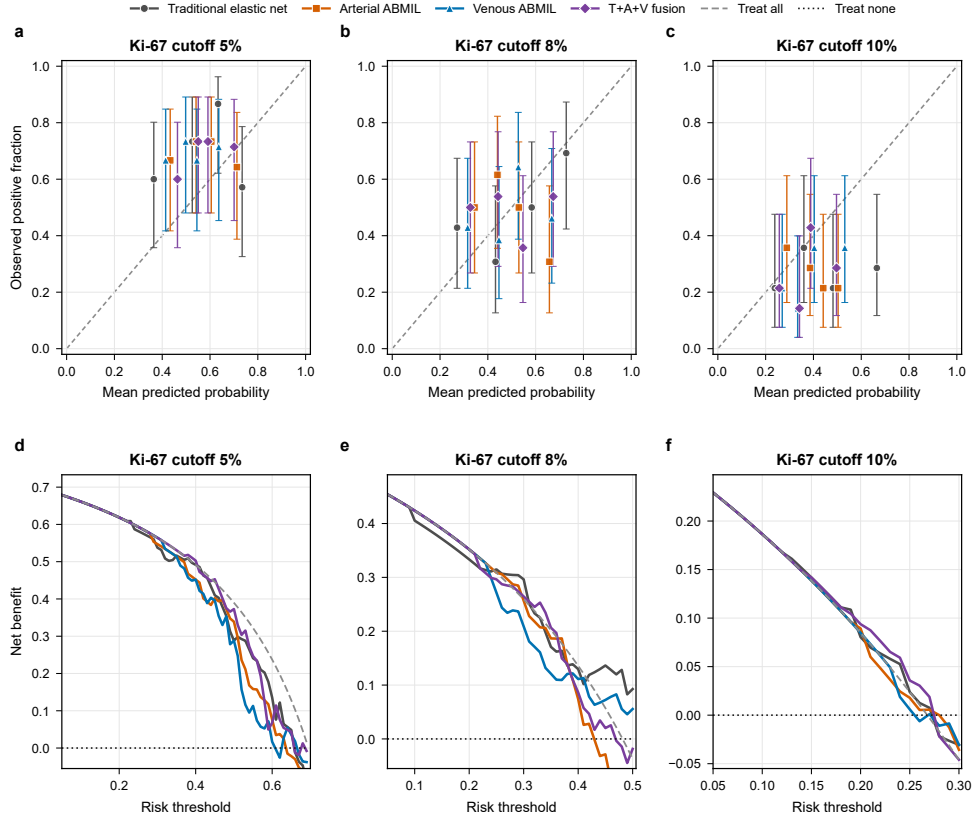


Fig. 3 Calibration and decision-curve analysis. Calibration panels (a–c) show four equal-frequency patient groups for conventional elastic net, arterial ABMIL, venous ABMIL, and three-component fusion. Symbols indicate observed positive fractions at the corresponding mean predicted probabilities, and vertical bars are Wilson 95% binomial confidence intervals. Points are not connected because each group contains a small number of patients. The dashed diagonal is the identity line. Decision panels (d–f) show net benefit for the same four models and the treat-all and treat-none strategies. The displayed threshold range for each task was defined from outcome prevalence to keep the treat-all net benefit above approximately -0.05 . The four displayed models were specified before the plots were examined.

262 AUROC confidence interval included 0.5. The small cohort and wide AUROC con-
 263 fidence intervals preclude a firm conclusion that the approaches have equivalent
 264 performance.

265 Prior contour-dependent radiomics studies address a different learning problem. A
 266 tumor ROI provides explicit lesion localization and excludes most non-tumor image
 267 regions before model fitting, whereas each bag in this study retained a complete arterial
 268 or venous series. MIL instead aggregated information across the full set of slices to

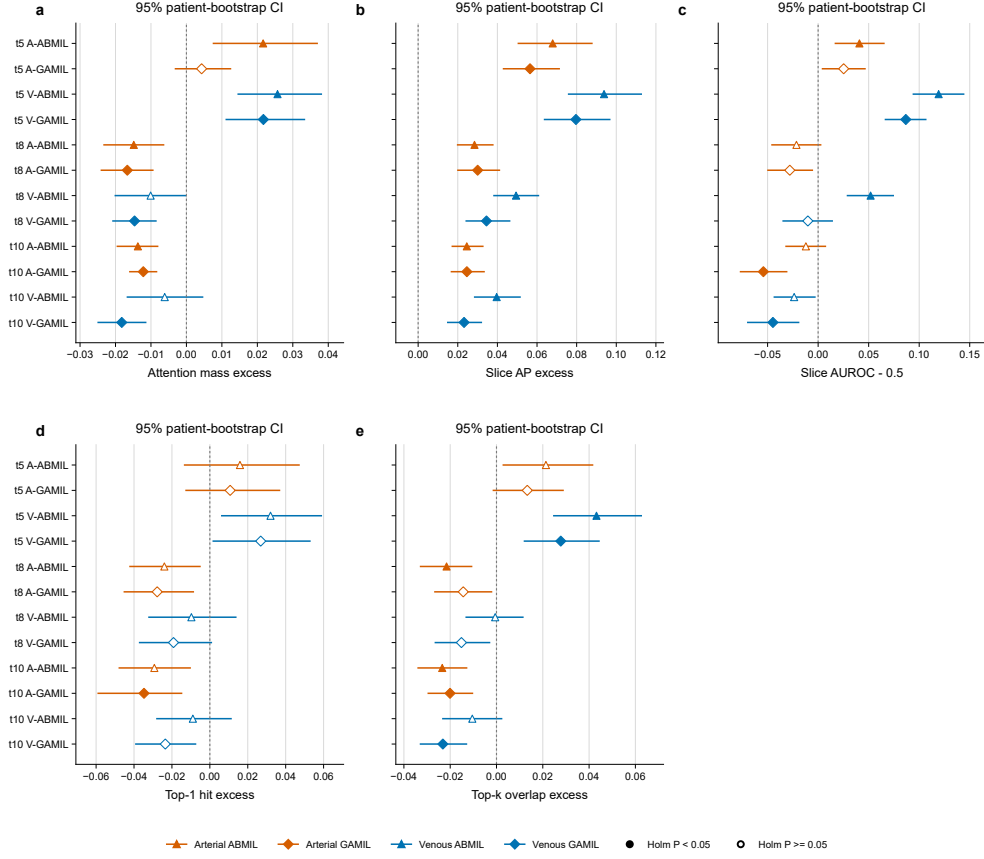


Fig. 4 Slice-level attention associations with tumor-bearing slices. Out-of-fold mean effects and 95% patient-level bootstrap confidence intervals are shown for five confirmatory attention measures in 12 single-phase MIL models. These measures comprise attention mass excess over uniform allocation (a), slice-level AUPRC excess over the prevalence of tumor-bearing slices (b), slice-level AUROC minus 0.5 (c), top-one hit excess over prevalence (d), and top- K overlap excess over prevalence (e). Filled symbols indicate global-family Holm $P < 0.05$, and open symbols indicate Holm $P \geq 0.05$.

269 generate a patient-level prediction. This distinction limits direct numerical comparison
 270 between the two settings. The 21 variables used by the conventional models included
 271 size, morphology, enhancement, and necrosis. Compared with the neural MIL models,
 272 the logistic models based on these clinically interpretable variables required fewer
 273 fitted parameters and may be easier to estimate from a small dataset.

274 Limited discrimination may reflect both sample size and image representation.
 275 Task-specific exclusions left 54 and 56 patients at the 8% and 10% cutoffs, and the 10%

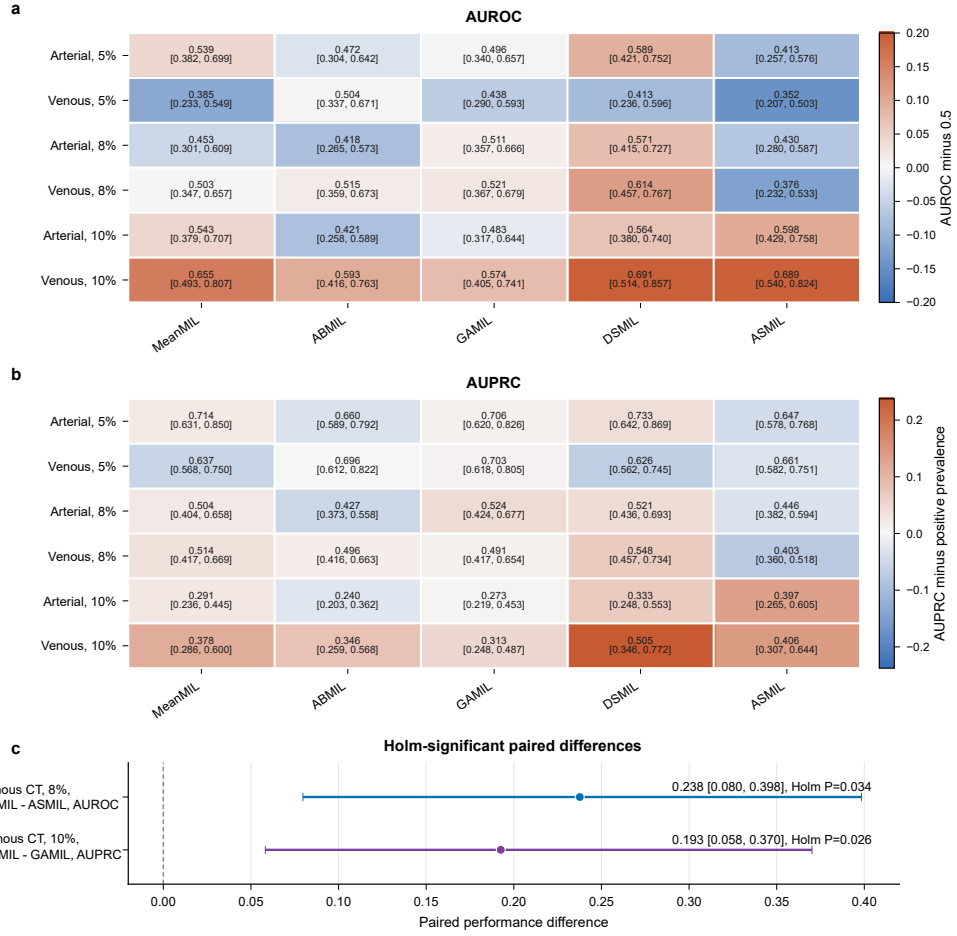


Fig. 5 Exploratory comparison of five multi-instance learning methods. Heatmaps show AUROC (**a**) and AUPRC (**b**) for arterial and venous phase CT at Ki-67 cutoffs of 5%, 8%, and 10%. Each cell reports the point estimate and 95% patient-stratified bootstrap confidence interval. Color encodes AUROC minus 0.5 or AUPRC minus the cutoff-specific positive prevalence, while the printed values remain on the original metric scale. Panel **c** shows the two paired differences that remained significant after Holm correction. The remaining 118 paired comparisons were not significant after correction.

task included only 15 positive cases. Although limited in size, the cohort combined surgically confirmed gastric GIST, complete preoperative arterial and venous CT series, postoperative Ki-67 results, and ROI information for the independent attention audit. These sample sizes provided limited data for learning attention weights and selecting among the five MIL models and their configurations. The frozen ResNet-18 was pre-trained on natural images rather than abdominal CT. Within each series, adjacent

282 slices are highly correlated, and useful signal may extend across the tumor, surround-
283 ing anatomy, and enhancement patterns. A slice encoder followed by bag pooling may
284 not fully represent relationships across adjacent slices or variations in enhancement.
285 The lack of spatial alignment between arterial and venous series may also have limited
286 the modeling of corresponding anatomy across the two phases.

287 The secondary analyses examined whether differences in patient rankings could
288 support fusion and whether learned attention was associated with tumor-bearing slices.
289 Low to moderate component correlations showed that the conventional, arterial, and
290 venous models did not produce identical patient rankings, but equal logit fusion did
291 not consistently improve on the component with the highest point estimate. Learned
292 fusion weights might account for unequal component reliability, although estimating
293 them from 61 patients would add another layer of model selection. All significant
294 attention estimates at the 5% cutoff were above their respective references, whereas
295 those at 8% and 10% included values above and below their respective references.
296 Attention was optimized for patient-level classification and received no localization
297 supervision. The observed variation does not support interpreting these weights as an
298 anatomic localization map.

299 Patient-level partitions, fold-confined preprocessing and model selection, and
300 validation-derived thresholds reduced opportunities for information leakage. Fusion
301 weights were fixed before the outer test results were examined. Synchronized boot-
302 strap samples preserved the pairing between model predictions, and Holm correction
303 controlled family-wise error across each comparison family. Patients with interval-
304 reported Ki-67 spanning a cutoff were not assigned an artificial binary label, and all
305 evaluated model families and comparisons were reported. The analysis remains limited
306 by its small, single-center retrospective cohort and the further reduction in sample size
307 after task-specific exclusions. The measured Ki-67 index may be affected by intratu-
308 moral heterogeneity and tissue sampling. Tube voltage was 120 kVp in 60 patients and

309 140 kVp in one. Spectral parameters were omitted because extracting them usually
310 requires lesion ROIs and would have changed the contour-free prediction task. We did
311 not compare the models with radiologists or directly measure preparation time, inter-
312 observer variability, or reader workload. The internal DCA alone does not establish
313 clinical utility.

314 5 Conclusions

315 Complete-series MIL generated cutoff-specific Ki-67 predictions from arterial and
316 venous CT without tumor contour inputs. The approach showed no Holm-corrected
317 discrimination advantage over conventional CT models in this cohort. The study
318 defines a patient-level framework for evaluating complete CT series without lesion
319 delineation. Its effect on annotation time and clinical workload was not measured.
320 Independent multicenter studies with larger cohorts should test this design using
321 CT-specific representations, spatial registration between arterial and venous phases,
322 radiologist comparison, and direct measurements of annotation burden before clinical
323 use is considered.

324 6 Abbreviations

325 *ABMIL* Attention-based multi-instance learning
326 *ASMIL* Attention-stabilized multiple instance learning
327 *AUPRC* Area under the precision-recall curve
328 *AUROC* Area under the receiver operating characteristic curve
329 *CI* Confidence interval
330 *CT* Computed tomography
331 *DCA* Decision-curve analysis
332 *DICOM* Digital Imaging and Communications in Medicine
333 *DSMIL* Dual-stream multi-instance learning

334 *GAMIL* Gated-attention multi-instance learning
335 *GIST* Gastrointestinal stromal tumor
336 *HU* Hounsfield unit
337 *IQR* Interquartile range
338 *MIL* Multi-instance learning
339 *OOF* Out-of-fold
340 *ROI* Region of interest
341 *SD* Standard deviation

342 **Acknowledgements**

343 Generative artificial intelligence was used only for English-language editing.

344 **Author contributions**

345 B.T. led the study design, methodology development, software implementation, formal
346 analysis, visualization, and preparation of the original manuscript. Z.W. contributed
347 to the study conception, methodology, data management, validation, funding acquisition,
348 and manuscript revision. Z.C. contributed to clinical data collection, data
349 curation, and investigation. J.Z. contributed to CT data collection, image preprocessing,
350 and data validation. X.H. contributed to data management, software validation,
351 and verification of the results. Y.G. contributed to image review, ROI annotation, and
352 data quality control. K.Z. contributed to statistical analysis and verification of the
353 results. J.X. contributed to the study conception, resource coordination, validation,
354 and manuscript revision. Y.X.Z. contributed to the literature review, visualization,
355 and manuscript revision. X.L. supervised the study and contributed to its conception,
356 project administration, resource coordination, funding acquisition, clinical
357 interpretation, and manuscript revision. All authors read and approved the final
358 manuscript.

359 **Funding**

360 This work was supported by the Chengdu City Science and Technology Project
361 (2026-YF11-00019-HZ), the Clinical Medicine Research Promotion Program of the
362 China Medical Foundation (2024CMFA10), and the Design and Development of an
363 Embedded Intelligent Agent for Imaging Order Entry Assistance (HX-H2509443).

364 **Data availability**

365 The datasets generated and analyzed during the current study are not publicly
366 available because they contain information derived from patient CT examinations. Dei-
367 dentified derived data and analysis code are available from the corresponding author
368 on reasonable request, subject to institutional and ethics approval.

369 **Declarations**

370 **Ethics approval and consent to participate**

371 The study was conducted in accordance with the Declaration of Helsinki. The Biomed-
372 ical Ethics Review Committee of West China Hospital, Sichuan University approved
373 the study (2022 Review No. 449 and 2025 Review No. 86). The requirement for writ-
374 ten informed consent was waived by the committee because retrospective clinical data
375 were used. All data were deidentified before analysis.

376 **Consent for publication**

377 Not applicable. All clinical data and CT images were deidentified, and no identifying
378 information is presented.

379 **Competing interests**

380 The authors declare that they have no competing interests.

References

- [1] Cai W, Guo K, Chen Y, Shi Y, Chen J (2024) Sub-regional CT radiomics for the prediction of Ki-67 proliferation index in gastrointestinal stromal tumors: A multi-center study. *Academic Radiology* 31:4974–4984. <https://doi.org/10.1016/j.acra.2024.06.036>
- [2] Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, Brogi E, Reuter VE, Klimstra DS, Fuchs TJ (2019) Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine* 25:1301–1309. <https://doi.org/10.1038/s41591-019-0508-1>
- [3] Casali PG, Blay JY, Abecassis N, Bajpai J, Bauer S, Biagini R, Bielack S, Bonvalot S, Boukovinas I, Bovee JVMG, et al (2022) Gastrointestinal stromal tumours: ESMO-EURACAN-GENTURIS clinical practice guidelines for diagnosis, treatment and follow-up. *Annals of Oncology* 33(1):20–33. <https://doi.org/10.1016/j.annonc.2021.09.005>
- [4] Chang R, Qi S, Wu Y, Song Q, Yue Y, Zhang X, Guan Y, Qian W (2022) Deep multiple instance learning for predicting chemotherapy response in non-small cell lung cancer using pretreatment CT images. *Scientific Reports* 12:19829. <https://doi.org/10.1038/s41598-022-24278-3>
- [5] Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, Ghassemi M, Liu X, Reitsma JB, van Smeden M, et al (2024) TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 385:e078378. <https://doi.org/10.1136/bmj-2023-078378>

- 405 [6] Deng J, Dong W, Socher R, Li LJ, Li K, Li FF (2009) ImageNet: A large-scale
406 hierarchical image database. In: Proceedings of the Conference on Computer
407 Vision and Pattern Recognition, pp 248–255, [https://doi.org/10.1109/CVPR.](https://doi.org/10.1109/CVPR.2009.5206848)
408 [2009.5206848](https://doi.org/10.1109/CVPR.2009.5206848)
- 409 [7] Dietterich TG, Lathrop RH, Lozano-Pérez T (1997) Solving the multiple instance
410 problem with axis-parallel rectangles. *Artificial Intelligence* 89(1):31–71. [https:](https://doi.org/10.1016/S0004-3702(96)00034-3)
411 [//doi.org/10.1016/S0004-3702\(96\)00034-3](https://doi.org/10.1016/S0004-3702(96)00034-3)
- 412 [8] Efron B, Tibshirani RJ (1994) *An Introduction to the Bootstrap*. Chapman and
413 Hall/CRC, New York, <https://doi.org/10.1201/9780429246593>
- 414 [9] He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition.
415 In: Proceedings of the Conference on Computer Vision and Pattern Recognition,
416 pp 770–778, <https://doi.org/10.1109/CVPR.2016.90>
- 417 [10] Holm S (1979) A simple sequentially rejective multiple test procedure. *Scandina-*
418 *vian Journal of Statistics* 6:65–70. URL <https://www.jstor.org/stable/4615733>
- 419 [11] Ilse M, Tomczak JM, Welling M (2018) Attention-based deep multiple instance
420 learning. In: Proceedings of the International Conference on Machine Learning,
421 pp 2127–2136, URL <https://proceedings.mlr.press/v80/ilse18a.html>
- 422 [12] Joensuu H (2008) Risk stratification of patients diagnosed with gastrointestinal
423 stromal tumor. *Human Pathology* 39(10):1411–1419. [https://doi.org/10.1016/j.](https://doi.org/10.1016/j.humpath.2008.06.025)
424 [humpath.2008.06.025](https://doi.org/10.1016/j.humpath.2008.06.025)
- 425 [13] Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren
426 J, Sanduleanu S, Larue RTHM, Even AJG, Jochems A, et al (2017) Radiomics:
427 The bridge between medical imaging and personalized medicine. *Nature Reviews*
428 *Clinical Oncology* 14:749–762. <https://doi.org/10.1038/nrclinonc.2017.141>

- 429 [14] Li B, Li Y, Eliceiri KW (2021) Dual-stream multiple instance learning network
430 for whole slide image classification with self-supervised contrastive learning. In:
431 Proceedings of the Conference on Computer Vision and Pattern Recognition, pp
432 14318–14328, <https://doi.org/10.1109/CVPR46437.2021.01409>
- 433 [15] Li J, Wang AR, Chen XD, Pan H, Li SQ (2022) Ki67 for evaluating the prog-
434 nosis of gastrointestinal stromal tumors: A systematic review and meta-analysis.
435 *Oncology Letters* 23:189. <https://doi.org/10.3892/ol.2022.13309>
- 436 [16] Liu WH, Li M, Ren GQ, Tang ZY, Shan XH, Yang BQ (2025) Radiomics model
437 based on dual-energy CT venous phase parameters to predict Ki-67 levels in
438 gastrointestinal stromal tumors. *Frontiers in Oncology* 15:1502062. <https://doi.org/10.3389/fonc.2025.1502062>
- 440 [17] Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F
441 (2021) Data-efficient and weakly supervised computational pathology on whole-
442 slide images. *Nature Biomedical Engineering* 5:555–570. <https://doi.org/10.1038/s41551-020-00682-w>
- 444 [18] Miettinen M, Lasota J (2006) Gastrointestinal stromal tumors: Pathology and
445 prognosis at different sites. *Seminars in Diagnostic Pathology* 23(2):70–83. <https://doi.org/10.1053/j.semdp.2006.09.001>
- 447 [19] Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P,
448 Beam AL, Van Calster B, Celi LA, Denaxas S, et al (2025) PROBAST+AI:
449 An updated quality, risk of bias, and applicability assessment tool for prediction
450 models using regression or artificial intelligence methods. *BMJ* 388:e082505. <https://doi.org/10.1136/bmj-2024-082505>

- 452 [20] Papadakos SP, Argyrou A, Karniadakis I, Theocharopoulos C, Katsaros I,
453 Machairas N, Vlachogiannakos J, Theocharis S (2025) Can CT radiomics predict
454 the Ki-67 index of gastrointestinal stromal tumors (GISTs)? A systematic review
455 and meta-analysis. *Cancers* 17:2855. <https://doi.org/10.3390/cancers17172855>
- 456 [21] Saito T, Rehmsmeier M (2015) The precision-recall plot is more informative than
457 the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS*
458 *ONE* 10(3):e0118432. <https://doi.org/10.1371/journal.pone.0118432>
- 459 [22] Tejani AS, Klontzas ME, Gatti AA, Mongan JT, Moy L, Park SH, Kahn CE Jr.
460 (2024) Checklist for artificial intelligence in medical imaging (CLAIM): 2024
461 update. *Radiology: Artificial Intelligence* 6(4):e240300. [https://doi.org/10.1148/](https://doi.org/10.1148/ryai.240300)
462 [ryai.240300](https://doi.org/10.1148/ryai.240300)
- 463 [23] Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW (2019)
464 Calibration: The Achilles heel of predictive analytics. *BMC Medicine* 17:230.
465 <https://doi.org/10.1186/s12916-019-1466-7>
- 466 [24] Vickers AJ, Elkin EB (2006) Decision curve analysis: A novel method for eval-
467 uating prediction models. *Medical Decision Making* 26(6):565–574. [https://doi.](https://doi.org/10.1177/0272989X06295361)
468 [org/10.1177/0272989X06295361](https://doi.org/10.1177/0272989X06295361)
- 469 [25] Xie Z, Suo S, Zhang W, Zhang Q, Dai Y, Song Y, Li X, Zhou Y (2024) Prediction
470 of high Ki-67 proliferation index of gastrointestinal stromal tumors based on
471 CT at non-contrast-enhanced and different contrast-enhanced phases. *European*
472 *Radiology* 34:2223–2232. <https://doi.org/10.1007/s00330-023-10249-3>
- 473 [26] Yang CW, Liu XJ, Zhao L, Che F, Yin Y, Chen HJ, Zhang B, Wu M, Song B
474 (2021) Preoperative prediction of gastrointestinal stromal tumors with high Ki-
475 67 proliferation index based on CT features. *Annals of Translational Medicine*

- 476 9(20):1556. <https://doi.org/10.21037/atm-21-4669>
- 477 [27] Yang P, Wu J, Liu M, Zheng Y, Zhao X, Mao Y (2024) Preoperative CT-based
 478 radiomics and deep learning model for predicting risk stratification of gastric
 479 gastrointestinal stromal tumors. *Medical Physics* 51(10):7257–7268. [https://doi.](https://doi.org/10.1002/mp.17276)
 480 [org/10.1002/mp.17276](https://doi.org/10.1002/mp.17276)
- 481 [28] Ye L, Hamidi SM, Chi Z, Li G, Pilanci M, Ogawa T, Haseyama M, Platanotis KN (2026) ASML: Attention-stabilized multiple instance learning for
 482 whole-slide imaging. In: *Proceedings of the International Conference on Learning*
 483 *Representations*, URL <https://openreview.net/forum?id=CYmjrbQRyM>
- 485 [29] Youden WJ (1950) Index for rating diagnostic tests. *Cancer* 3(1):32–35. [https:](https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3)
 486 [//doi.org/10.1002/1097-0142\(1950\)3:1<32::AID-CNCR2820030106>3.0.CO;2-3](https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3)
- 487 [30] Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte
 488 A, Ashrafinia S, Bakas S, Beukinga RJ, Boellaard R, et al (2020) The image
 489 biomarker standardization initiative: Standardized quantitative radiomics for
 490 high-throughput image-based phenotyping. *Radiology* 295(2):328–338. [https://](https://doi.org/10.1148/radiol.2020191145)
 491 doi.org/10.1148/radiol.2020191145