Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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
bless_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 Chen[1], Jinge Zhang[1], Xin He[2], Yili Gu[1], Ke Zou[3], Jun Xie[2], Yu-Xuan Zhang[4], Xijiao Liu[1*]

[1]Department of Radiology, West China Hospital of Sichuan University, Chengdu, Sichuan, China.
[2]Information Technology Center, West China Hospital of Sichuan University, Chengdu, China.
[3]Chengdu Finsiot Technology Co., Ltd., Chengdu, China.
[4]School of Information Science and Technology, Southwest Jiaotong University, Chengdu, China.

*Corresponding author(s). E-mail(s): bless_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

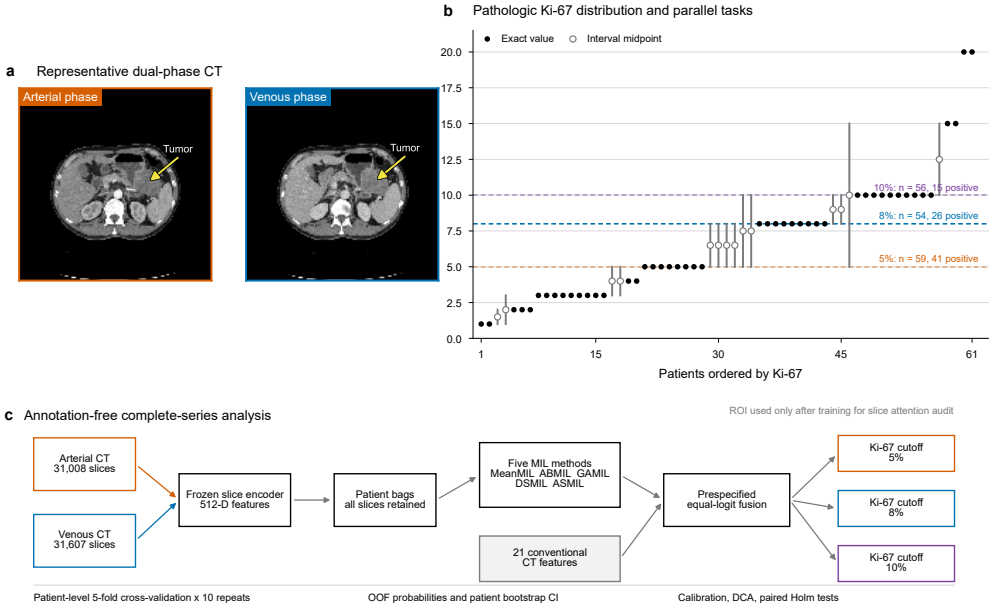

Fig. 1 Clinical overview and study design. **a**, Representative deidentified arterial-phase and venous-phase CT images from a patient with gastric GIST. Arrows indicate the tumor. **b**, Pathologic Ki-67 results for all 61 patients, with interval reports distinguished and cutoffs marked at 5%, 8%, and 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.

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

2 Materials and methods

2.1 Study population and outcomes

2.1.1 Study design and patients

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

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

2.1.2 Pathologic Ki-67 assessment

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

2.2 CT data and features

2.2.1 CT acquisition and conventional features

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

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

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

2.2.2 Image preprocessing and feature extraction

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

After this preprocessing, each axial slice was combined with its immediately preceding and following slices to form a three-channel input. The first or last slice was 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 $p \in \{\mathrm{A}, \mathrm{V}\}$ 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

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

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

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

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

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

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

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

2.3.4 Fusion strategy

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

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

This rule did not require fitting an additional fusion model.

2.3.5 Training and evaluation

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

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

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

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

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

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

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

2.4 Attention audit

Attention weights from ABMIL and GAMIL were compared between slices with and without tumor in 12 cutoff, phase, and model combinations. ROIs were mapped to slices by DICOM coordinates after all predictions were frozen. Attention mass, slice-level AUROC and AUPRC, top-one hit rate, and top-K overlap were calculated relative to their random or uniform references. Confidence intervals and tests were calculated with 10,000 patient-level bootstrap samples. Holm correction was 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

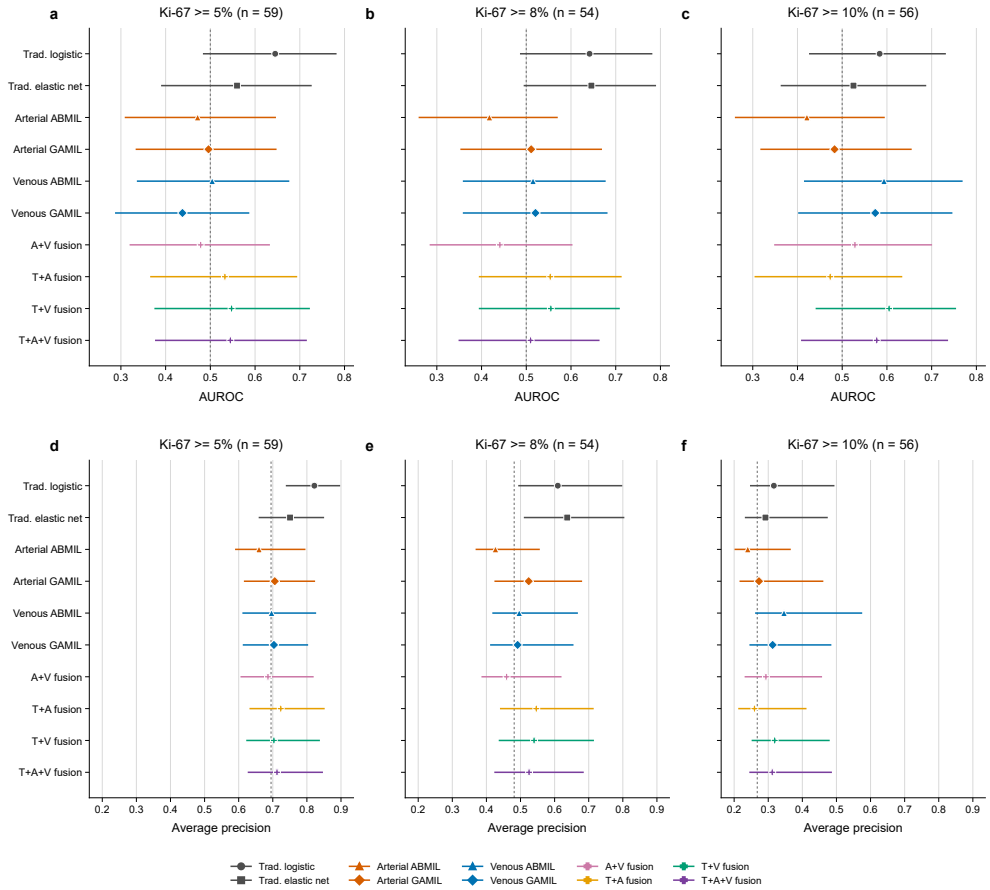

Fig. 2 Discrimination across models. Point estimates and 95% patient-stratified bootstrap confidence intervals are shown for AUROC (**a**–**c**) and AUPRC (**d**–**f**) across the three Ki-67 cutoffs. Dashed lines indicate an AUROC of 0.5 or the cutoff-specific positive prevalence for AUPRC. None of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

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

3.6 Comparison of MIL methods

The MIL method with the highest point estimate varied by phase, cutoff, and metric. Only two of 120 paired comparisons of five MIL methods (10 pairs, 2 phases, 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	ρ(T,A)	ρ(T,V)	ρ(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 without 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

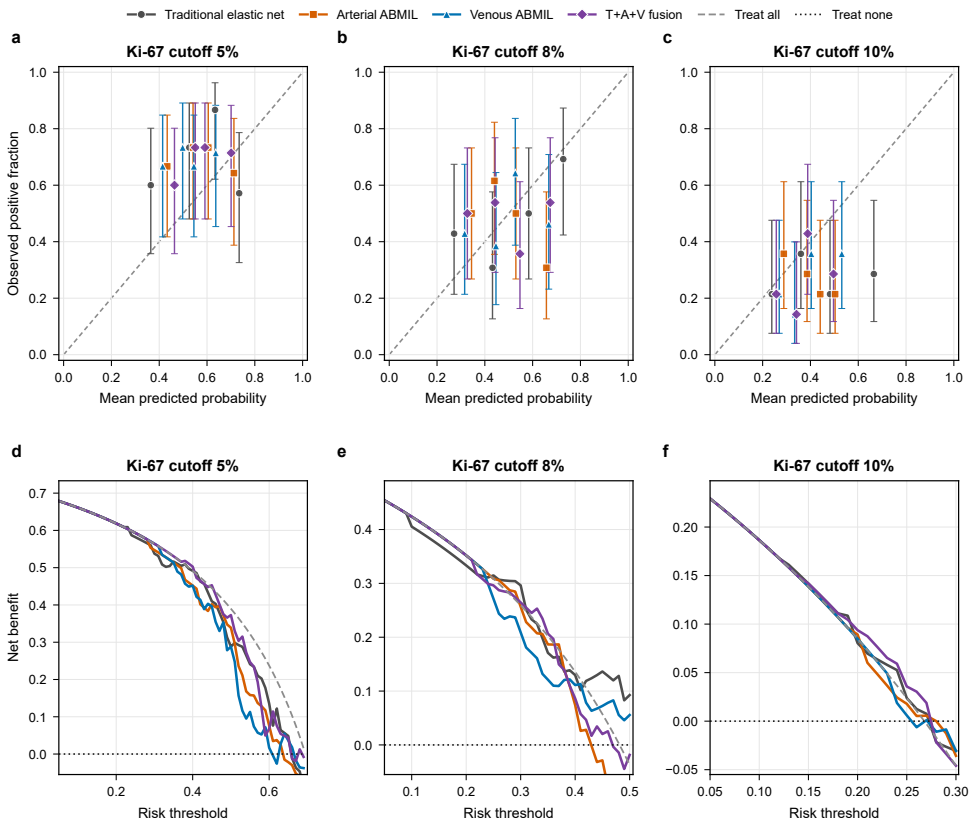

Fig. 3 Calibration and decision-curve analysis. Calibration panels (**a**–**c**) show four equal-frequency patient groups for conventional elastic net, arterial ABMIL, venous ABMIL, and three-component fusion. Symbols indicate observed positive fractions at the corresponding mean predicted probabilities, 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.

AUROC confidence interval included 0.5. The small cohort and wide AUROC confidence intervals preclude a firm conclusion that the approaches have equivalent performance.

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

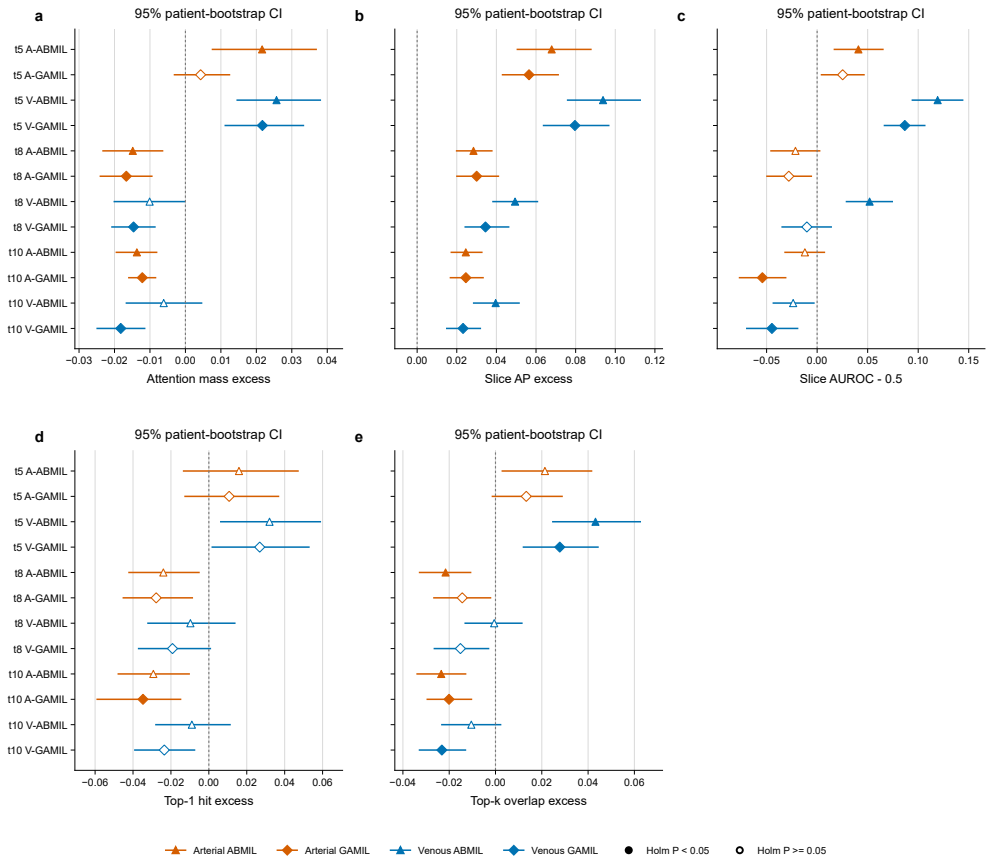

Fig. 4 Slice-level attention associations with tumor-bearing slices. Out-of-fold mean effects and 95% patient-level bootstrap confidence intervals are shown for five confirmatory attention measures in 12 single-phase MIL models. These measures comprise attention mass excess over uniform allocation (**a**), 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 $P \geq 0.05$.

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

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

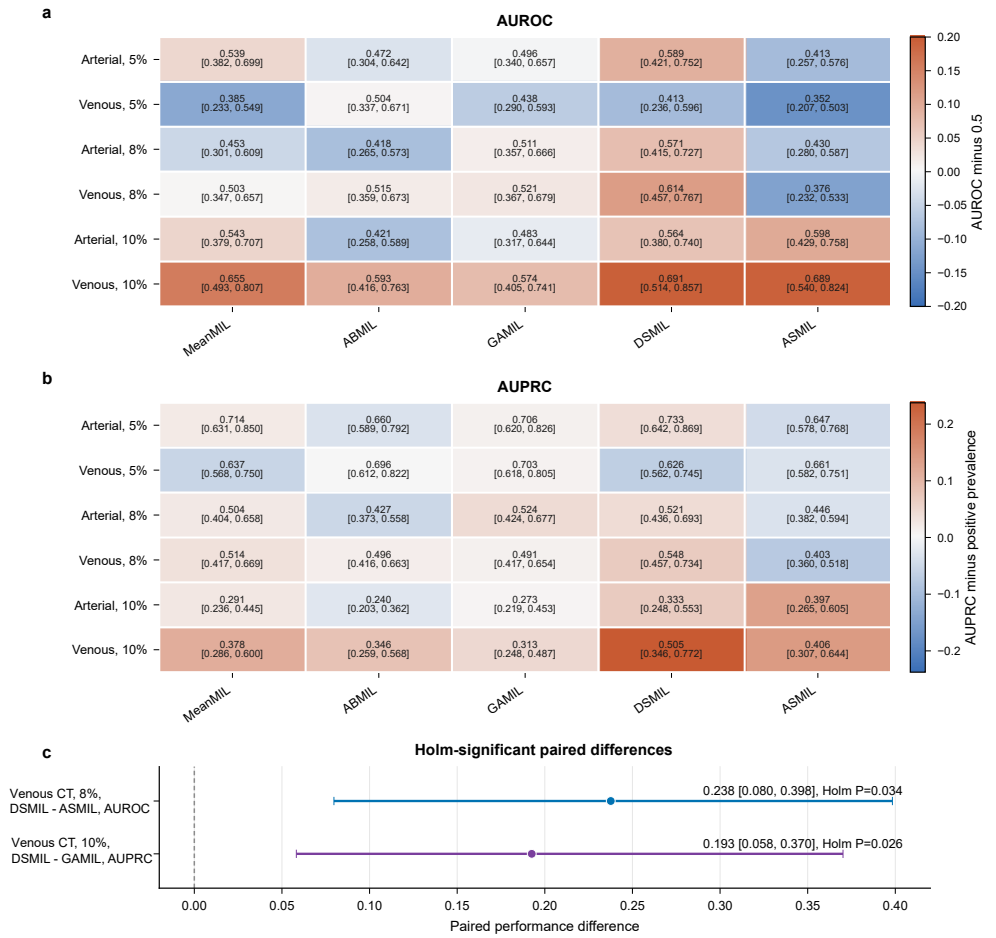

Fig. 5 Exploratory comparison of five multi-instance learning methods. Heatmaps show AUROC (**a**) and AUPRC (**b**) for arterial and venous phase CT at Ki-67 cutoffs of 5%, 8%, and 10%. Each cell reports the point estimate and 95% patient-stratified bootstrap confidence interval. Color encodes AUROC 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 pretrained on natural images rather than abdominal CT. Within each series, adjacent

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

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

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

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

5 Conclusions

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

6 Abbreviations

ABMIL Attention-based multi-instance learning

ASMIL Attention-stabilized multiple instance learning

AUPRC Area under the precision-recall curve

AUROC Area under the receiver operating characteristic curve

CI Confidence interval

CT Computed tomography

DCA Decision-curve analysis

DICOM Digital Imaging and Communications in Medicine

DSMIL Dual-stream multi-instance learning

GAMIL Gated-attention multi-instance learning

GIST Gastrointestinal stromal tumor

HU Hounsfield unit

IQR Interquartile range

MIL Multi-instance learning

OOF Out-of-fold

ROI Region of interest

SD Standard deviation

Acknowledgements

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

Author contributions

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

Funding

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

Data availability

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

Declarations

Ethics approval and consent to participate

The study was conducted in accordance with the Declaration of Helsinki. The Biomedical Ethics Review Committee of West China Hospital, Sichuan University approved the study (2022 Review No. 449 and 2025 Review No. 86). The requirement for written informed consent was waived by the committee because retrospective clinical data were used. All data were deidentified before analysis.

Consent for publication

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

Competing interests

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

[6] Deng J, Dong W, Socher R, Li LJ, Li K, Li FF (2009) ImageNet: A large-scale hierarchical image database. In: Proceedings of the Conference on Computer Vision and Pattern Recognition, pp 248–255, https://doi.org/10.1109/CVPR.2009.5206848

[7] Dietterich TG, Lathrop RH, Lozano-Pérez T (1997) Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence 89(1):31–71. https://doi.org/10.1016/S0004-3702(96)00034-3

[8] Efron B, Tibshirani RJ (1994) An Introduction to the Bootstrap. Chapman and Hall/CRC, New York, https://doi.org/10.1201/9780429246593

[9] He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the Conference on Computer Vision and Pattern Recognition, pp 770–778, https://doi.org/10.1109/CVPR.2016.90

[10] Holm S (1979) A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6:65–70. URL https://www.jstor.org/stable/4615733

[11] Ilse M, Tomczak JM, Welling M (2018) Attention-based deep multiple instance learning. In: Proceedings of the International Conference on Machine Learning, pp 2127–2136, URL https://proceedings.mlr.press/v80/ilse18a.html

[12] Joensuu H (2008) Risk stratification of patients diagnosed with gastrointestinal stromal tumor. Human Pathology 39(10):1411–1419. https://doi.org/10.1016/j.humpath.2008.06.025

[13] Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, Sanduleanu S, Larue RTHM, Even AJG, Jochems A, et al (2017) Radiomics: The bridge between medical imaging and personalized medicine. Nature Reviews Clinical Oncology 14:749–762. https://doi.org/10.1038/nrclinonc.2017.141

[14] Li B, Li Y, Eliceiri KW (2021) Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Proceedings of the Conference on Computer Vision and Pattern Recognition, pp 14318–14328, https://doi.org/10.1109/CVPR46437.2021.01409

[15] Li J, Wang AR, Chen XD, Pan H, Li SQ (2022) Ki67 for evaluating the prognosis of gastrointestinal stromal tumors: A systematic review and meta-analysis. Oncology Letters 23:189. https://doi.org/10.3892/ol.2022.13309

[16] Liu WH, Li M, Ren GQ, Tang ZY, Shan XH, Yang BQ (2025) Radiomics model based on dual-energy CT venous phase parameters to predict Ki-67 levels in gastrointestinal stromal tumors. Frontiers in Oncology 15:1502062. https://doi.org/10.3389/fonc.2025.1502062

[17] Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F (2021) Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering 5:555–570. https://doi.org/10.1038/s41551-020-00682-w

[18] Miettinen M, Lasota J (2006) Gastrointestinal stromal tumors: Pathology and prognosis at different sites. Seminars in Diagnostic Pathology 23(2):70–83. https://doi.org/10.1053/j.semdp.2006.09.001

[19] Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, Beam AL, Van Calster B, Celi LA, Denaxas S, et al (2025) PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ 388:e082505. https://doi.org/10.1136/bmj-2024-082505

[20] Papadakos SP, Argyrou A, Karniadakis I, Theocharopoulos C, Katsaros I, Machairas N, Vlachogiannakos J, Theocharis S (2025) Can CT radiomics predict the Ki-67 index of gastrointestinal stromal tumors (GISTs)? A systematic review and meta-analysis. Cancers 17:2855. https://doi.org/10.3390/cancers17172855

[21] Saito T, Rehmsmeier M (2015) The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE 10(3):e0118432. https://doi.org/10.1371/journal.pone.0118432

[22] Tejani AS, Klontzas ME, Gatti AA, Mongan JT, Moy L, Park SH, Kahn CEJr. (2024) Checklist for artificial intelligence in medical imaging (CLAIM): 2024 update. Radiology: Artificial Intelligence 6(4):e240300. https://doi.org/10.1148/ryai.240300

[23] Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW (2019) Calibration: The Achilles heel of predictive analytics. BMC Medicine 17:230. https://doi.org/10.1186/s12916-019-1466-7

[24] Vickers AJ, Elkin EB (2006) Decision curve analysis: A novel method for evaluating prediction models. Medical Decision Making 26(6):565–574. https://doi.org/10.1177/0272989X06295361

[25] Xie Z, Suo S, Zhang W, Zhang Q, Dai Y, Song Y, Li X, Zhou Y (2024) Prediction of high Ki-67 proliferation index of gastrointestinal stromal tumors based on CT at non-contrast-enhanced and different contrast-enhanced phases. European Radiology 34:2223–2232. https://doi.org/10.1007/s00330-023-10249-3

[26] Yang CW, Liu XJ, Zhao L, Che F, Yin Y, Chen HJ, Zhang B, Wu M, Song B (2021) Preoperative prediction of gastrointestinal stromal tumors with high Ki-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.
480 org/10.1002/mp.17276

481 [28] Ye L, Hamidi SM, Chi Z, Li G, Pilanci M, Ogawa T, Haseyama M, Pla-
482 taniotis KN (2026) ASMIL: Attention-stabilized multiple instance learning for
483 whole-slide imaging. In: Proceedings of the International Conference on Learning
484 Representations, URL https://openreview.net/forum?id=CYmjrbQRyM

485 [29] Youden WJ (1950) Index for rating diagnostic tests. Cancer 3(1):32–35. https:
486 //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://
491 doi.org/10.1148/radiol.2020191145