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.

WITHDRAWN: Predicting Malignancy in Solid Adnexal Masses: An Externally Validated Machine Learning Model Integrating Conventional and Contrast-Enhanced Ultrasound

Lixia Chen
chenlixia1976@126.com

Wuwu Zheng
Tingting Chi
Hui Li
Xiaona Cai

Research Article

Keywords: Adnexal mass, Ovarian cancer, Risk, Ultrasound, Contrast-enhanced ultrasound

Posted Date: April 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9064480/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

EDITORIAL NOTE:

The full text of this preprint has been withdrawn by the authors since they no longer stand by the findings/conclusions of this work. Therefore, the authors do not wish this work to be cited as a reference. Questions should be directed to the corresponding author.

Abstract

Background Traditional diagnostic models, or contrast-enhanced ultrasound-related models, are generally developed for broad applications across different adnexal masses. These broad models lack the precision needed to describe specific subtypes, such as solid ones.

Objective This study aimed to develop a machine learning model using conventional and contrast-enhanced ultrasound features to stratify malignancy in solid adnexal masses and to evaluate its performance relative to updated O-RADS on an independent external test set.

Methods A total of 277 solid adnexal masses were analysed in the development set. Missing data were addressed through multiple imputation, generating 20 datasets (m = 20). Feature selection was conducted using bootstrap-enhanced least absolute shrinkage and selection operator (LASSO) regression to retain the nine most stable predictors. Model performance and discrimination were assessed on three datasets. In the external test set, our model showed improved discrimination compared to updated O-RADS classifications, as determined by the DeLong test. Decision curve analysis (DCA) further demonstrated its greater clinical applicability.

Results In internal validation, the model showed robust discrimination (AUC: 0.926; 95% CI: 0.861–0.992) and good calibration (Brier score: 0.115). These results were further confirmed in the external test set (AUC: 0.949, 95% CI: 0.906–0.992) and with good calibration (Brier score: 0.101). The novel model outperformed the updated O-RADS categories in the external test set for O-RADS = 4 (AUC: 0.949 vs 0.656, DeLong test, $p < 0.001$) and O-RADS = 5 (AUC: 0.949 vs 0.593, DeLong test, $p < 0.001$). Decision curve analysis indicated that the model was suitable across a range of clinical trial thresholds.providing flexible thresholds, and building a visualised nomogram,

Conclusions The new model, integrating conventional ultrasound and CEUS features, can enhance diagnostic accuracy for solid adnexal masses, surpassing the updated O-RADS categorisation and providing increased practical value for clinical assessments.

Introduction

Ovarian cancer epidemiology is characterised by a paradoxical shift, including rising early-onset (40 years) incidence in some regions[1] and projected steeper increases in low- and middle-HDI countries by 2040[2], against a backdrop of declining global rates. This, coupled with the fact that more than 80% of cases are late-stage diagnoses with a pronounced survival disparity (90% in stage I vs 10–40% in stage III/IV)[3], highlights the imperative for accessible early diagnostic strategies.

Accurate differential diagnosis of adnexal masses, particularly early malignant lesions, remains a significant challenge in gynaecological imaging. Classic diagnostic models such as Simple Rules (SR), Simple Rules Risk (SR), and Assessing Different Neoplasias provide good guidelines for adnexal masses [4, 5], but fail to balance diagnostic sensitivity and specificity [6].O-RADS shows excellent sensitivity but often poor specificity, even after recent updates [7, 8]. Conventiional colour Doppler flow imaging (CDFI) fails to detect low-velocity blood flow within tumours, especially in early stages or hypovascular conditions, which can partially explain its poor specificity. Contrast-enhanced ultrasound (CEUS) has been introduced to address this problem. Injecting a microbubble contrast agent intravenously allows CEUS to visualise microvascular perfusion in real time, compensating for CDFI's inability to detect slow and small vessels. Hence, the new strategy is as follows: integrating qualitative or quantitative CEUS parameters with established ultrasound models to improve accuracy. Shi Y et al [9] exploited CEUS-derived microvascular perfusion features to improve the O-RADS classification and its limited specificity; The combination of IOTA SR and CEUS enhanced diagnostic accuracy for uncertain adnexal masses: malignancy suspicion was based on the presence of two or more of the following CEUS features (uneven enhancement, increased enhancement, abnormal vessel shape, or rapid washout) [10]. These studies upgraded or downgraded risk assessments by integrating CEUS with established models. Furthermore, the following models have been designed to increase specificity : Wu et al [11] developed a model using CA125, acoustic shadow and the mass-to-uterine peak intensity (PI) ratio;A more comprehensive weighted scoring system was later integrated with O-RADS US v 2022, CEUS features, and CA125 levels [12].

However, there are still two key limitations in the current research: First, the existing diagnostic models are designed for wide applicability, covering a range of adnexal masses from purely cystic to completely solid. Because different types of adnexal masses exhibit distinct pathophysiological and ultrasound characteristics, "broad-spectrum" models are less effective at differentiating specific subtypes, particularly solid adnexal masses (SAMs). Second, current research integrates CEUS and CA125 to construct models. Yet, combining ultrasound with CA125 is limited by its low sensitivity in early-stage disease [13] and its poor utility for sex cord-stromal tumours [14].

To address the limitations of current broad-spectrum diagnostic methods for adnexal masses, we seek to develop a machine learning model that combineds conventional ultrasound and CEUS features to evaluate the early risk of SAMs. We aim to accurately categorise early malignancy risk solely by considering direct sonographic characteristics, providing a practical tool with a focused approach that balances sensitivity and specificity for this particular subtype.

Methods

Study population

This was a single-center, pilot study. The study flowchart is shown in Fig. 1.

The prediction model was developed using retrospective data from the SAM's database (January 2008 to September 2024). The analysis of these de-identified data was granted a waiver of informed consent by the Ethics Committee. Its generalizability was assessed through external validation in a prospective cohort from October 2024 to January 2026, with informed consent from all participants. This study followed the principles of the Declaration of Helsinki and was approved by the Clinical Research Ethics Committee of the First Hospital of Wenzhou Medical University (KY2025-R204) .

The inclusion criteria were as follows: (1) age ≥ 18 years; (2) availability of complete imaging documentation, including conventional ultrasound, CDFI, and CEUS findings; (3) SAMs with at least 80% of the solid component in orthogonal sections, in accordance with the O-RADS criteria; (4) histopathological confirmation available; (5) provision of informed consent by the patients. Exclusion criteria were as follows: (1) absence of histopathological verification; (2) solid components showing non-enhancement; (3) non-adnexal primary tumours confirmed pathologically, acute pelvic inflammatory disease, metastatic adnexal tumours, or adnexal torsion; (4) a history of bilateral adnexectomy.

Ultrasound examination

We performed exams on various ultrasound machines (Mindray Resona 8/9, Philips iU22, Esaote MyLab 90). Transvaginal and abdominal imaging employed specialised probes, such as frequency ranges, available **in Supplementary File S1.** After initial grayscale and colour Doppler scans of the pelvis, a 1.5–2.4 mL bolus of SonoVue was administered intravenously and flushed with 5 mL of saline. CEUS imaging was then performed at a low mechanical index (MI < 0.1), with dynamic images recorded continuously for ≥ 3 min for subsequent analysis.

Image analysis

The conventional ultrasound features of SAMs are assessed as follows: MaxD (categorised as ≥ 10 cm or < 10 cm); boundary (indistinct or well-defined); hyper echo; echo uniformity (homogeneous or heterogeneous); morphology (regular or irregular); and the presence or absence of calcification, microcyst, and acoustic shadow. Additionally, the colour score (CS) was categorised into CS1 to CS4 on the basis of the O-RADS system, and then divided into two groups: CS1-2 and CS3-4.

The CEUS characteristics of SAMs were assessed as follows: peak intensity (PI) relative to the uterine myometrium was categorised as hyper- or iso-enhancement, and hypo-enhancement; enhancement homogeneity (homogeneous or heterogeneous); and the presence or absence of ring enhancement and perfusion defects. Additionally, the arrival time (AT) of the SAM was visually compared with the uterine myometrium and categorised into two types: Type 1 (enhancement simultaneous with or earlier than the myometrium) and Type 2 (enhancement later than the myometrium). TICs (time-intensity of curve) for the masses and uterine myometrium were derived from a small region of interest (hyper enhanced region of mass vs outer myometrium, <0.5 cm^2, positioned to avoid major vessels) and categorised into two types: Type I (early or synchronous wash-in with early washout) and Type II (all other patterns) (Fig. 2).

To maintain image consistency, all CEUS examinations and measurements are performed by the same ultrasound physician, with more than 25 years of experience in gynaecological ultrasound and 18 years of experience in CEUS. Image analysis was performed by two experienced ultrasound physicians unaware of the patient's clinical data, including pathological results. In cases of conflicting analyses, parties should discuss to reach a consensus. If disagreements continue, a third senior expert will serve as an arbitrator.

Statistical method

Ultrasound features analysis

Mean ± standard deviation (SD) and frequency counts were used to present continuous and categorical data, respectively. Group comparisons were conducted using the chi-square test.

Model Development

The study cohort was split into a development set (patients from Jan 2008 to Sep 2024) and an external test set (patients from Oct 2024 to Jan 2026). The development set was split randomly into the training and internal validation sets (8:2). This split was performed using double-stratified sampling (stratified by both pathological subtype and outcome status), implemented via the 'caret' package in R (version 4.4.3; set. seed(123)). Missing data in the training set were handled using multiple imputation by chained equations (m = 20, maxit = 10) with 20 imputed datasets.

For each imputed dataset, 1,000 bootstrap samples were generated via stratified sampling. Variable selection was conducted on each sample using LASSO regression with five-fold cross-validation and the lambda-1se rule for sparsity.Variables were retained as candidate predictors if (1) appeared in at least 80% of the imputed datasets and (2) were selected with a frequency greater than 50% of each dataset. Collinearity was measured using the variance inflation factor (VIF), and the events-per-variable ratio (EPV) is used to evaluate the sample size. A logistic regression model was constructed using the final set of predictors, and a nomogram was developed.

Model Performance Assessment

Model performance was evaluated in terms of discrimination (area under the receiver operating characteristic curve, AUC), calibration (calibration curves and Brier scores), and clinical utility (decision curve analysis, DCA). The optimal probability threshold was determined from the internal validation set through three methods: sensitivity ≥ 90%, maximising the Youden index, and a fixed threshold of 0.5. Select the optimal threshold with high stability and good generalisation ability, based on comprehensive diagnostic performance and net benefit. Platt scaling was applied to the external test set using five-fold cross-validation to correct miscalibration while avoiding overfitting. The model's performance was then evaluated against the updated O-RADS classification, considering malignancy as O-RADS ≥ 4 or O-RADS = 5.

A two-tailed p-value < 0.05 indicated statistical significance. Analyses were conducted using SPSS (v26; IBM Corp.) and R (v4.4.3; R Foundation).

Results

Pathological results

The development set comprised 269 patients with 277 SAMs from an initial pool of 450 patients with 491 adnexal masses, and the external test set consisted of 96 patients with 104 SAMs selected from 174 patients with 190 adnexal masses. In the development set, there were 124 postmenopausal and 145 premenopausal women, with a mean age of 50.2 ± 13.8 years (range, 22–84 years). The external test set included 58 postmenopausal and 38 premenopausal women, with a mean age of 54.0 ± 12.0 years (range, 28–82 years). The new prediction model utilised solely imaging features, excluding clinical parameters and laboratory tumour markers.

The distribution of pathological findings is summarised in Table 1. In the benign group, thecoma-fibroma and thecomas constituted 43.9% (72/164) and 23.2% (38/164) of the development set, respectively, and in the external test set, thecoma-fibromas and fibromas represented 47.3% (26/55) and 7.3% (4/55), respectively. Serous carcinoma and adult granulosa cell tumours were the most common malignancies. In the development set, serous carcinoma accounted for 38.9% (44/113), and in the external test set, it accounted for 44.9% (22/49); Adult granulosa cell tumours accounted for 16.8% (19/113) and 8.2% (4/49) in the respective sets. Pathological staging in the development set included 44 stage I, 12 stage II, 19 stage III, 12 stage IV, 2 unclear stage, and 24 borderline tumors. The external test set comprised 13 stage I, 3 stage II, 14 stage III, 5 stage IV, and 14 borderline tumors. Notably, the combined proportion of early-stage (FIGO I-II) malignant cancers and borderline tumours was 70.8% (80/113) in the development set and 61.2% (30/49) in the external testing set.

Table 1
The pathological subtypes

Pathological type	development set (n = 277)	external test set (n = 104)
Benign tumor		
Brenner tumor	5	1
Leidig tumor	3	NA
Mature teratoma	3	NA
Mature teratoma with foreign body reaction	1	NA
Mature teratoma with thyroid follicular component	1	NA
Adenofibroma	10	5
Steroid cell tumor	1	NA
Ovarian struma	9	1
Ovarian polypoid endometriotic nodule	1	NA
Thecoma	38	4
Thecoma with Focal Cellular Hyperplasia	NA	1
Cellular Thecoma	NA	1
Luteinized thecoma	2	NA
Thecomatous nodules	1	NA
Thecoma-fibroma	70	26
Thecoma-fibroma with cellular areas	2	NA
Tubal adenoma-like tumor	1	NA
Fibroma	13	8
Fibroma with Luteinization	NA	1
Sclerosing stromal tumor	2	NA
Papillary Cystadenoma, Mucinous Type	NA	1
Seromucinous cystadenoma	1	NA
Fibroma with Luteinization	1	NA
Ovarian ligament leiomyoma	NA	1
ovarian corpus albicans formation with peripheral vascular proliferation	NA	2
Mild Hyperplasia of Ovarian Theca Cells	NA	1
Hyperplasia of Thecomatous and Fibrous Tissue	NA	1
Tubal adenomyoma	NA	1
Total	164	55
Bordline tumor		
Cellular Fibroma	6	4
Serous Borderline Tumor	12	7
Mucinous Borderline Tumor	1	1
Granulosa-Theca Cell Tumor	4	1
Sertoli-Leydig Cell Tumor, Moderately	1	NA

Pathological type	development set (n = 277)	external test set (n = 104)
Sertoli-Leydig Cell Tumor, Tubular Structures	NA	1
Total	24	14
Malignant tumor		
Serous Carcinoma	44	22
Adult Granulosa Cell Tumor	19	4
Endometrioid Carcinoma of the Ovary	6	1
Clear Cell Carcinoma	5	3
Dysgerminoma	3	NA
Ovarian Adenocarcinoma	3	NA
Embryonal Carcinoma	2	NA
Malignant Brenner Tumor	1	1
Ovarian Small Cell Carcinoma Hypercalcemic Type	1	1
Epithelioid Malignant Tumor	1	NA
Yolk Sac Tumor	1	NA
Adenosarcoma	1	NA
Carcinoid	1	NA
Squamous Cell Carcinoma	1	NA
Ovarian Aggressive B-Cell Lymphoma	NA	2
Malignant Mesothelioma, Epithelioid Type, with Focal Necrosis	NA	1
Total	89	35

Model Development

The conventional ultrasound and CEUS features in the datasets were analysed (Table 2) and showed no significant differences in hyper echo, calcification, boundary, and ring enhancement between the benign and malignant groups. Due to their clinical importance, they were included in all modelling without interaction terms. The training and validation data sets had consistent feature distributions and proved comparable. The external test set features were summarised in Supplementary File S2.

Table 2 ultrasound characteristics

	The development	Tumor Type,N(%)				Cohort ,N(%)			
Variables	set (training and validation n = 277), N%	benign n = 164	Malignant n = 113	p	Missing date(%)	Training set n = 224	Validation set n = 53	p	Test set n = 104
MaxD				<0.001	0(0)			0.797	
<10cm	184(66.4%)	131(79.9%)	53(46.9%)			76(33.9%)	17(32.1%)		82(78.8%)
≥ 10cm	93(33.6%)	33(20.1%)	60(53.1)			148(66.1%)	36(67.9%)		22(21.2%)
Hyper echo				0.072	0(0)			0.655	
Present	22(7.9%)	17(10.4%)	5(4.4%)			17(7.6%)	5(9.4%)		4(3.8%)
Absent	255(92.1%)	147(89.6%)	108(95.6%)			207(92.4%)	48(90.6%)		100(96.2%)
Boundary				0.059	0(0)			0.064	
well-defined	245(88.3%)	150(91.5%)	95(84.1%)			202(90.2%)	43(81.1%)		84(80.8%)
indistinct	32(11.6%)	14(8.5%)	18(15.9%)			22(9.8%)	10(18.9%)		20(19.2%)
Uniformity				0.001	0(0)			0.497	
Homogenous	100(36.1%)	72(43.9%)	28(24.8%)			83(37.1%)	17(32.1%)		16(15.4%)
heterogeneous	177(63.9%)	92(56.1%)	85(75.2%)			141(62.9%)	36(67.9%)		88(84.6%)
Calcification				0.674	0(0)			0.814	
Present	34(12.3%)	19(11.6%)	15(13.3%)			28(12.5%)	6 (11.3%)		21(20.2%)
Absent	243(87.7%)	145(88.4%)	98(86/7%)			196(87.5%)	47(88.7%)		83(79.8%)
Microcyst				<0.001	0(0)			0.376	
Present	23(8.3%)	5(3.0%)	18(15.9%)			17(7.6%)	6(11.3%)		5(4.8%)
Absent	254(91.7%)	159(97.0%)	95(84.1%)			207(92.4%)	47(88.7%)		99(95.2%)
Acoustic Shadow				<0.001	0(0)			0.393	
Present	106(38.3%)	98(59.8%)	8(7.1%)			83(37.1%)	23(43.4%)		39(37.5%)
Absent	171(61.7%)	66(40.2%)	105(92.9%)			141(62.9%)	30(56.6%)		65(62.5%)
Morphology				<0.001	0(0)			0.623	
regular	211(76.2%)	145(88.4%)	66(58.4%)			172(76.8%)	39(73.6%)		84(80.8%)
irregular	66(23.8%)	19(11.6%)	47(41.6%)			52(23.2%)	14(26.4%)		20(19.2%)
Color Score				<0.001	0(0)			0.236	
cs1-2	201(72.6%)	143(87.2%)	58(51.3%)			166(74.1%)	35(66.0%)		77(74.0%)
cs3-4	76(27.4%)	21(12.8%)	55(48.7%)			58(25.9%)	18(34.0%)		27(26.0%)
Perfusion defects				<0.001	0(0)			0.391	
Present	36(13.0%)	6(3.7%)	30(26.5%)			31(13.8%)	5(9.4%)		11(10.6%)
Absent	241(87.0%)	158(96.3%)	83(73.5%)			193(86.2%)	48(90.6%)		93(89.4%)
Ring enhancement				0.783	0(0)			0.875	
Present	76(27.4%)	46(28.0%)	30(26.5%)			61(27.2%)	15(28.3%)		24(23.1%)
Absent	201(72.6%)	118(72.0%)	83(73.5%)			163(72.8%)	38(71.7%)		80(76.9%)

	The development	Tumor Type,N(%)				Cohort ,N(%)			
Variables	set (training and validation n = 277), N%	benign n = 164	Malignant n = 113	p	Missing date(%)	Training set n = 224	Validation set n = 53	p	Test set n = 104
Peak uniformity				<0.001	0(0)			0.562	
homogenous	109(39.4%)	85(51.8%)	24(21.2%)			90(40.2%)	19(35.8%)		49(47.1%)
heterogeneous	168(60.6%)	79(48.2%)	89(78.8%)			134(59.8%)	34(64.2%)		55(52.9%)
Peak intensity				< 0.001	0(0)			0.652	
hypo	123(44.4%)	111(67.7%)	12(10.6%)			98(43.8%)	25(47.2%)		53(51.0%)
Iso/hyper	154(55.6%)	53(32.3%)	101(89.4%)			126(56.3%)	28(52.8%)		51(49.0%)
AT type*				<0.001	72(26.0%)			0.454	
Type1	91(32.9%)	35(21.3%)	56(49.6%)			114(50.9%)	30(56.6%)		41(39.4%)
Type2	114(41.2%)	94(57.3%)	20(17.7%)			110(49.1%)	23(43.4%)		63(60.6%)
TIC type*				<0.001	72(26.0%)			0.711	
type	71 (25.6%)	25(15.2%)	46(40.7%)			95(42.4%)	21(39.6%)		35(33.7%)
type	134 (48.4%)	104(63.4%)	30(26.5%)			129(57.6%)	32(60.4%)		69(66.3%)

Table supl 2 ultrasound features of external rest set

Variable	group		P-value
	benign	malignant	
Hyper echo			
Present	2	2	1.000
Absent	53	47	
MaxD			
<10	47	36	0.129
≥ 10	8	13	
Microcyst			
Present	0	5	0.015
Absence	55	44	
Acoustic Shadow			
Present	36	2	0.000
Absent	19	47	
Morphology			
Regular	46	39	0.594
irregular	9	10	
Color Score			
CS1-2	52	25	0.000
CS3-4	3	24	
Perfusion defects			
Present	1	11	0.001
Absent	54	38	
Ring enhancement			
Present	9	15	0.085
Absent	46	34	
Peak Intensity			
hyper/iso	6	45	0.000
hypo	49	4	

The development set was divided into a training set (n = 224) and a validation set (n = 53) at an 8:2 ratio, stratified by pathological subgroups and positive outcome status for balance. An external test set of 104 cases was utilised for model external validation. This study employed multiple imputation to address missing data in the AT and TIC variables, generating 20 imputed training datasets. Then, we assessed the imputation stability. The standard deviations of 0.022 for TIC type and 0.013 for AT type ($p < 0.05$) indicated consistent and reliable imputation results across datasets. All other complete variables did not require imputation.

Following the Bootstrap-LASSO procedure with stability filtering, we identified nine major predictor variables (MPVs): MaxD, hyper echo, acoustic shadow, morphology, microcyst, CS, PI, perfusion defects, and ring enhancement. Then, we got the variable selection ranking (Fig. 3).

The final logistic model was fitted to complete cases without Rubin's pooling. VIFs for all predictor variables ranged from 1.05 to 1.97 (mean: 1.41), suggesting no multicollinearity. EPV ratio was 10.2 (92 events/9 variables), meeting recommended statistical power thresholds. We used 9 MPVs to construct the final model Log-odds(Malignancy) = -1.566 -1.864×Hyper echo + 0.519×MaxD + 1.189×CS + 1.179×Morphology + 0.946×Microcyst − 2.110×Acoustic shadow + 1.996×PI -1.296×Ring enhancement + 0.940×Perfusion defects, and constructed a nomogram (Fig. 4).

Table 3 shows the results of the multivariate logistic regression analysis for factors associated with malignancy in SAMs. Among the nine variables included in the analysis, six remained independent predictors of malignancy: PI, morphology, microcyst, perfusion defects, ring enhancement, acoustic shadow. The strongest predictor was PI (OR = 7.360, 95% CI:2.527–21.437, $p < 0.001$), indicating that patients with hyper- or iso-enhancement were more than seven times as likely to have malignant masses as those with hypo-enhancement. In contrast, the remaining three variables, including hyper echo, MaxD, and colour score did not retain statistical significance in the multivariate model (all $p > 0.05$).

Table 3
Multivariate Logistic Regression Analysis for Malignancy of solid adnexal masses

Variable	β	OR	0R 95%CI	P-value
Hyperechoic	-1.864	0.155	0.033–0.733	0.019
MaxD	0.519	1.681	0.593–4.763	0.329
Microcyst	0.946	2.575	0.578–11.460	0.215
Acoustic Shadow	-2.110	0.121	0.036–0.412	0.001
Morphology	1.179	3.251	1.049–10.073	0.041
Color Score	1.189	3.283	1.211–8.901	0.020
Perfusion defects	0.940	2.561	0.673–9.741	0.168
Ring enhancement	-1.296	0.274	0.104–0.723	0.009
Peak intensity	1.996	7.360	2.527–21.437	< 0.001

Model Performance and Validation

The model showed excellent discrimination with AUCs of 0.922 (95% CI: 0.890–0.954), 0.926 (95% CI: 0.861–0.992), and 0.949 (95% CI: 0.906–0.992) across the training, internal validation, and external test sets, respectively (Fig. 5). The Delong test revealed no significant differences in AUC among the datasets at various thresholds: training vs validation (Z = -0.114, $p = 0.910$), training vs external test (Z = -0.974, $p = 0.331$), and internal validation vs external test (Z = -0.565, $p = 0.574$). The Brier scores for the training, internal validation, and external test sets were 0.114, 0.115, 0.101, respectively, indicating consistent performance across datasets and no evidence of overfitting.

Classification thresholds for 0.297, 0.415, and 0.5 were established based on sensitivity (≥ 90%), maximising the Youden index, and a fixed threshold, with performance assessed across three datasets (Table 4).

Table 4
Model performance across datasets

Dataset	Optimal threshold	Sensitivity	Specificity	PPV	NPV	Accuracy	Brier score	Net Benefit
Train	0.5	0.837	0.818	0.762	0.878	0.826	0.114	0.237(0.152–0.313)
Train	0.415(0.062–0.719)	0.859	0.811	0.760	0.892	0.830	0.114	0.272(0.196–0.346)
Train	0.297(0.062–0.581)	0.880	0.795	0.750	0.905	0.830	0.114	0.310(0.238–0.378)
Validation	0.5	0.810	0.812	0.739	0.867	0.811	0.115	0.208(0.038–0.359)
Validation	0.415(0.062–0.719)	0.857	0.812	0.750	0.897	0.830	0.115	0.262(0.093–0.413)
Validation	0.297(0.062–0.581)	0.905	0.750	0.704	0.923	0.811	0.115	0.2937(0.139–0.432)
Test	0.5	0.816	0.945	0.930	0.852	0.885	0.101	0.356(0.250–0.462)
Test	0.415(0.062–0.719)	0.816	0.945	0.930	0.852	0.885	0.101	0.361(0.261–0.470)
Test	0.297(0.062–0.581)	0.837	0.891	0.872	0.860	0.865	0.101	0.370(0.269–0.474)
Test(platt)	0.5	0.837	0.891	0.872	0.860	0.865	0.082	0.337(0.221–0.452)
Test(platt)	0.297(0.062–0.581)	0.959	0.873	0.870	0.960	0.913	0.082	0.424(0.318–0.526)

The 0.415 and 0.5 thresholds were similar in robustness ($p < 0.05$), whereas 0.415 had wider confidence intervals. The fixed 0.5 thresholds have sensitivities and specificities of 0.810 and 0.812 in internal validation, and 0.816 and 0.945 in external test sets. For high sensitivity, 0.297

has sensitivities and specificities of 0.905 and 0.750 in the internal validation set, and 0.837 and 0.891 in the external test set. For enhanced sensitivity, we recommend utilising 0.297. For balance, we suggest a stable 0.5.

The calibration curves for the training and internal validation sets were well-balanced, indicating good agreement between predicted and observed risks. The external validation set showed slight differences in intercepts and a risk underestimation, as shown by an intercept of 1.083 and a slope of 1.421. The Hosmer-Lemeshow test showed good calibration for the training (χ^2=7.583, df = 6, p = 0.270) and internal validation sets (χ^2=1.862, df = 3, p = 0.601), but significant misfit in the external test set (χ^2=13.959, df = 6, p = 0.030) (Fig. 6).

Platt scaling improved diagnostic results on an external test set, with a calibration intercept of -0.028 (95% CI: -0.668–0.610) and a slope of 0.944 (95% CI: 0.598–1.289) (Fig. 7), with strong agreement between predicted and observed outcomes.

Despite a significant HL test result, calibration metrics indicate satisfactory calibration. The Brier score improved from 0.101 to 0.082, reflecting reduced prediction error, while sensitivities and specificities at thresholds of 0.297 and 0.5 were 0.959 and 0.873, and 0.837 and 0.891 respectively. The new model showed better net benefit than 'treat-all' and 'treat-none' strategies across all thresholds, indicating potential for improved patient outcomes (Fig. 8).

In the external test set, our new model outperformed the updated O-RADS classification in diagnostic accuracy (Table 5), showing an AUC of 0.949 (95% CI: 0.906–0.992). The AUC values for O-RADS 4 and O-RADS 5 were 0.656 (95% CI: 0.573–0.739) and 0.593 (95% CI: 0.507–0.697), respectively (Fig. 9).

Table 5
Diagnostic performance of the updated O-RADS versus the original O-RADS in the external test set

	Threshold	Sensitivity	Specificity	PPV	NPV	ACC	AUC(95%CI)	Brier score
O-RADS ≥ 4		**0.857**	**0.455**	**0.583**	**0.781**	**0.644**	**0.656(0.573–0.739)**	**0.221**
O-RADS = 5		**0.367**	**0.818**	**0.643**	**0.592**	**0.606**	**0.593(0.507–0.678)**	**0.238**
New model	**0.297**	**0.837**	**0.891**	**0.872**	**0.860**	**0.865**	**0.949(0.906–0.992)**	**0.101**
	0.500	**0.816**	**0.946**	**0.930**	**0.853**	**0.885**	**0.949(0.906–0.992)**	**0.101**
New model(platt)	**0.297**	**0.959**	**0.873**	**0.870**	**0.960**	**0.913**	**0.943 (0.893–0.993)**	**0.082**
	0.500	**0.837**	**0.891**	**0.872**	**0.860**	**0.865**	**0.943 (0.893–0.993)**	**0.082**

The Delong test indicated significant differences between the new model and O-RADS ≥ 4 (z = -7.070, p < 0.001), as well as between the new model and O-RADS = 5 (z = -7.625, p < 0.001). No significant difference was observed between O-RADS ≥ 4 and O-RADS = 5 (z = 1.295, p = 0.195). The decision curve analysis showed that our new model consistently provided higher net benefits than the updated O-RADS classifications across a wide range of clinically relevant threshold probabilities (Fig. 10).

Discussion

The heterogeneous spectrum of adnexal masses—ranging from purely cystic to completely solid, each with distinct characteristics—poses a challenge for diagnosis[15]. Traditional diagnostic models or CEUS-related research are designed for broad applicability and therefore address a wide spectrum of adnexal masses. Accordingly, these models lack the precision needed to accurately interpret specific types, particularly SAMs.

We developed and validated a machine learning model using solely imaging features of conventional ultrasound and CEUS features to classify the risk of malignancy in SAM. The model was designed to be accurate for early detection. We excluded late-stage indicators like ascites, lymph nodes, and peritoneal involvement, as well as non-enhanced SAMs. Serum tumour markers were not included. The high proportion of early-stage cancer and borderline tumours in our data (> 60%) also favour the construction of models with enhanced sensitivity for early detection.

The model was well discriminated on both internal validation and external test sets (AUC > 0.90), with high sensitivity and good specificity at the chosen threshold. To ensure the reliability of the probability outputs, we corrected the calibration offset observed in the external test set with Platt scaling. After correction, the model was well-calibrated (a calibration slope close to 1) and, together with a low Brier score, its risk predictions were reliable in absolute terms.

To implement this model practically, we developed a multi-level application scheme. First, we created a highly sensitive "exclusion threshold" (0.297)to minimise the chance of missed diagnoses during screening or in high-risk scenarios. Second, a "triage threshold" 0.5is set to

optimise sensitivity and specificity for clinical diagnosis and treatment decisions, as summarized in Table 4. This dual decision framework allows the tool to adapt to different risk preferences across scenarios.

To facilitate individualised clinical interpretation, the final model was presented as an intuitive nomogram that integrates the nine MPVs, thereby translating the complex algorithmic output into a visual personal risk assessment tool. MaxD ≥ 10 cm, higher CS (CS3-4), irregular morphology, PI (hyper- or iso-enhancement), and the presence of microcysts and perfusion defects were independently associated with an increased risk of malignancy. Conversely, hyper echo, acoustic shadows, and ring enhancement showed significant protective effects. The MPVs incorporated into our final model may reflect their established pathological and clinical relevance, supporting the biological plausibility of the model.

Malignant SAMs are typically larger, vascularised, and irregular in morphology, consistent with established criteria [16]. In the updated O-RADS classification system, any solid mass with irregular morphology, regardless of size or CS, was classified as category 5, indicating a high risk of malignancy. Our model reduced over-classification of benign irregular masses as O-RADS grade 5, correctly reclassifying them as low-risk, for instance, irregularly shaped thecoma-fibromas. This advantage was confirmed in both the development and external validation sets, showing potential to complement current classification systems.

The presence of acoustic shadow, a key benign predictor, is consistent with the O-RADS classification [17]. Fibrous-rich benign pathologies (e.g., thecoma-fibromas, fibromas) attenuate sound, leading to acoustic shadows [18]. In our development set, benign masses (59.8%, 98/164) showed a notably higher incidence of acoustic shadows than malignant masses (7.1%, 8/113). The rarity of this predictor in the malignant group of the external test set (only two malignant cases) confirms its reliability as a robust protective indicator.

The hyper echo of a SAMs, characterised by the 'pearl sign' in struma ovarii, is a valuable benign sonographic marker [19, 20]. Consistent with these findings, the hyper echoic masses in our study often exhibit cs3-4 and PI (hyper- or iso-enhancement) on CEUS, closely mimicking malignancy and heightening the risk of misdiagnosis. Consequently, recognising the hyper echo of SAMs on conventional ultrasound assessments is critical to accurately reclassify these masses as benign.

The presence of microcystic features suggests an associated micropapillary structure in certain cancers [21]. This finding necessitates careful analysis, as it can also be present in benign solid masses, such as adenofibromas[22]. But adenofibromas typically exhibit acoustic shadows and hypo-enhancement. Therefore, this model, which integrates multiple ultrasound characteristics, is favourable for improving diagnostic accuracy.

Despite a notable disparity in CS distribution between benign and malignant groups, over half of the malignant tumours had CS1-2 in both the development and external test sets, accounting for 51.3% (58/113) and 51.0% (25/49), respectively. Conventional ultrasound may underestimate the vascularity of certain SAMs. CEUS primarily addresses this limitation by demonstrating PI (hyper- or iso-enhancement), thereby revealing true hypervascularity. This imaging sign aligns with the pathological hallmark of malignant tumours: a rich microvascular supply.This finding aligns with previous studies and confirms the PI as a key predictor [23]. Our study further showed that PI (hyper- or iso-enhancement) (OR = 7.360) compared with PI (hypo-enhancement) confers a more than sevenfold greater risk of malignancy.

Ring enhancement is also one of the crucial indicators for distinguishing benign from malignant tumours. Benign tumours typically show complete ring enhancement owing to intact capsules, whereas malignant tumours often show disrupted capsules, leading to interrupted or absent ring enhancement. However, we observed ring enhancement in 26.5% (30/113) of malignancies in the development set and 30.6% (15/49) in the external test set, mainly in early-stage (Stage I) or borderline tumours. Therefore, interpreting ring enhancement requires consideration of tumour stage to prevent overlooking early-stage malignancies or borderline tumours.

Perfusion defect, a predictor of malignancy, appears on grayscale ultrasound as an anechoic or solid-like area; however, CEUS reveals a larger non-enhancing region corresponding to the anechoic area, or demonstrates non-enhancement in the solid-appearing area. This feature is consistent with the pathology of necrotic hemorrhage in malignant tumours.

Furthermore, we excluded two variables from our study: AT type, assessed visually and TIC type. This decision contrasts with some previous studies [24], which suggested that integrating subjective visual assessment of CEUS and TIC analysis with the O-RADS US scoring system significantly enhances diagnostic performance while achieving comparable accuracy. Excluding AT and TIC types does not diminish their theoretical significance; rather, we are grounded in practical challenges and data quality issues observed in clinical environments. In some instances, technical constraints may hinder the execution of AT visual assessment and TIC analysis. These include: 1) the physical separation of a large or distant mass from the uterine wall, preventing their simultaneous inclusion within a single region of interest for comparative analysis; 2) deep-seated lesions where significant signal attenuation degrades TIC curve fidelity; and 3) the absence of a uterine myometrial reference following hysterectomy. These factors result in a high, non-random rate of data omission. Although we attempted to address missing data through interpolation, the resulting datasets remained unstable. Thus, the divergence is between our study's conclusions and other literature.

In summary, by focusing on early imaging features, calibrating the model, providing flexible thresholds, and building a visualised nomogram, we are able to provide not only a high-performing classifier, but also a tool for reliably quantifying risk and supporting multi-scenario clinical decision making.

Limitations

The limitations of this study are (1) a single-centre design, (2) uneven sample distribution, and a time gap between development and external validation, which may lead to baseline risk underestimation. After scaling the model, the calibration was adequate, supporting its potential clinical application. Nonetheless, validation in larger cohorts is needed to confirm generalizability; and (3) classification of borderline tumours as malignant, which potentially oversimplified their indolent biology and affected surgical planning [25]; 4). Our model is highly targeted, focusing exclusively on SAMs, which limits its scope of application. Future research will concentrate on: (1) multicenter validation; (2) creating a diagnostic model for fertility preservation in younger patients with borderline tumors; (3) combining radiomics, tumor markers, and sonographic features for a model.

Conclusions

We developed a novel machine learning model integrating direct characteristics of conventional ultrasound and CEUS features to predict the malignant risk of SAMs, specifically tailored for primary imaging characteristics during initial outpatient evaluations.Our model performed better diagnostically, with higher sensitivity and specificity, than the updated O-RADS classification system for early detection of malignant SAMs. It can help clinicians with rapid risk stratification, prevent unnecessary surgeries for benign conditions, and provide special attention to malignant cases. But multicenter studies are needed to validate its clinical impact and cost-effectiveness before widespread use.

Abbreviations

CEUS Contrast-enhanced ultrasound

SAMs Solid Adnexal Masses

LASSO Least Absolute Shrinkage and Selection Operator

DCA Decision curve analysis

O-RADS Ovarian-Adnexal Reporting and Data System

SR Simple rules

PI Peak intensity

TIC Time intensity curve

AT Arrival time

CS Colour score

MPVs Major predictor variables

CA125 Cancer Antigen 125

MI Low mechanical index

HL test Hosmer-Lemeshow test

VIF Variance inflation factor

EPV Events per variable

Declarations

Acknowledgements

No

Funding Declaration

Medical and health project of Wenzhou Science and Technology Bureau（Y20220463(Self-funded)

title: Application of CEUS in the differential diagnosis of ovarian tumor）

The research was supported through from the First Affiliated Hospital of Wenzhou Medical University. No specific grant from external funding agencies was received.

Declaration of Competing Interest

The authors declare no competing interests.

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Consent for publication

Not applicable

Human Ethics and Consent to Participate declarations

This study followed the principles of the Declaration of Helsinki and was approved by the Clinical Research Ethics Committee of the First Hospital of Wenzhou Medical University. (KY2025-R204) .

Authors' contributions

Lixia Chen *: Supervision, data curation, writing, review & editing (lead);

Wuwu Zheng: Formal analysis and ultrasound evaluation;

Tingting Chi: ultrasound evaluation;

Hui Li: data curation;

Xiaona Cai: Patient follow-up,data curation;

References

1. Huang J, Chan WC, Ngai CH, et al. Worldwide Burden, Risk Factors, and Temporal Trends of Ovarian Cancer: A Global Study. Cancers (Basel). 2022;14(9):2230. 10.3390/cancers14092230. Published 2022 Apr 29.
2. Cabasag CJ, Fagan PJ, Ferlay J, et al. Ovarian cancer today and tomorrow: A global assessment by world region and Human Development Index using GLOBOCAN 2020. Int J Cancer. 2022;151(9):1535–41. 10.1002/ijc.34002.
3. Caruso G, Weroha SJ, Cliby W, Ovarian Cancer. Rev JAMA. 2025;334(14):1278–91. 10.1001/jama.2025.9495.
4. Timmerman D, Testa AC, Bourne T, et al. Simple ultrasound-based rules for the diagnosis of ovarian cancer. Ultrasound Obstet Gynecol. 2008;31(6):681–90. 10.1002/uog.5365.
5. Hiett AK, Sonek JD, Guy M, Reid TJ. Performance of IOTA Simple Rules, Simple Rules risk assessment, ADNEX model and O-RADS in differentiating between benign and malignant adnexal lesions in North American women. Ultrasound Obstet Gynecol. 2022;59(5):668–76. 10.1002/uog.24777.
6. Sundar S, Agarwal R, Davenport C, et al. Risk-prediction models in postmenopausal patients with symptoms of suspected ovarian cancer in the UK (ROCkeTS): a multicentre, prospective diagnostic accuracy study. Lancet Oncol. 2024;25(10):1371–86. 10.1016/S1470-2045(24)00406-6.
7. Andreotti RF, Timmerman D, Strachowski LM, et al. Radiology. 2020;294(1):168–85. 10.1148/radiol.2019191150. O-RADS US Risk Stratification and Management System: A Consensus Guideline from the ACR Ovarian-Adnexal Reporting and Data System Committee.
8. Strachowski LM, Jha P, Phillips CH, et al. Radiology. 2023;308(3):e230685. 10.1148/radiol.230685. O-RADS US v2022: An Update from the American College of Radiology's Ovarian-Adnexal Reporting and Data System US Committee.
9. Shi Y, Li H, Wu X, Li X, Yang M. O-RADS combined with contrast-enhanced ultrasound in risk stratification of adnexal masses. J Ovarian Res. 2023;16(1):153. Published 2023 Aug 3. 10.1186/s13048-023-01243-w

10. Tian C, Han YW, Shi ZJ, et al. Diagnostic value of the International Ovarian Tumor Analysis simple rules combined with contrast-enhanced ultrasound for adnexal masses. Int J Gynecol Cancer. 2025;35(2):100049. 10.1016/j.ijgc.2024.100049.
11. Wu M, Zhang M, Qu E, et al. A modified CEUS risk stratification model for adnexal masses with solid components: prospective multicenter study and risk adjustment. Eur Radiol. 2024;34(9):5978–88. 10.1007/s00330-024-10639-1.
12. Jiang Z, Pu W, Luo X et al. Integrating O-RADS US v2022, CEUS, and CA125 to enhance the diagnostic differentiation of ovarian masses: development of the OCC-US model. Cancer Imaging. 2025;25(1):96. Published 2025 Jul 30. 10.1186/s40644-025-00918-5
13. Young Han C, Bedia JS, Yang WL, et al. Autoantibodies, antigen-autoantibody complexes and antigens complement CA125 for early detection of ovarian cancer. Br J Cancer. 2024;130(5):861–8. 10.1038/s41416-023-02560-z.
14. Nasioudis D, Wilson E, Mastroyannis SA, Latif NA. Prognostic significance of elevated pre-treatment serum CA-125 levels in patients with stage I ovarian sex cord-stromal tumors. Eur J Obstet Gynecol Reprod Biol. 2019;238:86–9. 10.1016/j.ejogrb.2019.05.002.
15. Timmerman D, Valentin L, Bourne TH, et al. Terms, definitions and measurements to describe the sonographic features of adnexal tumors:a consensus opinion from the International Ovarian Tumor Analysis (IOTA) Group. Ultrasound Obstet Gynecol. 2000;16(5):500–5. 10.1046/j.1469-0705.2000.00287.x.
16. Timmerman D, Van Calster B, Testa A, et al. Predicting the risk of malignancy in adnexal masses based on the Simple Rules from the International Ovarian Tumor Analysis group. Am J Obstet Gynecol. 2016;214(4):424–37. 10.1016/j.ajog.2016.01.007.
17. Luo H, Lin Z, Wu L, et al. Application of O-RADS ultrasound lexicon-based logistic regression analysis model in the diagnosis of solid component-containing ovarian malignancies. Biomed Res Int. 2022;2022:7187334.
18. Paladini D, Testa A, Van Holsbeke C, Mancari R, Timmerman D, Valentin L. Imaging in gynecological disease (5): clinical and ultrasound characteristics in fibroma and fibrothecoma of the ovary. Ultrasound Obstet Gynecol. 2009;34:188–95.
19. Savelli L, Testa AC, Timmerman D, Paladini D, Ljungberg O, Valentin L. Imaging of gynecological disease (4): clinical and ultrasound characteristics of struma ovarii. Ultrasound Obstet Gynecol. 2008;32(2):210–9. 10.1002/uog.5396.
20. Birbas E, Kanavos T, Gkrozou F, et al. Ovarian masses in children and adolescents: a review of the literature with emphasis on the diagnostic approach. Children. 2023;10:1114.
21. Liu D, Lyu G, Lai H, Li L, Gan Y, Yang S. Can the ultrasound microcystic pattern accurately predict borderline ovarian tumors? J Ovarian Res. 2023;16(1):162. 10.1186/s13048-023-01253-8. Published 2023 Aug 11.
22. Zheng X, Lyu G, Gan Y, et al. Microcystic pattern and shadowing are independent predictors of ovarian borderline tumors and cystadenofibromas in ultrasound. Eur Radiol. 2021;31:45–54.
23. Yuan K, Huang YJ, Mao MY, et al. Contrast-enhanced US to Improve Diagnostic Performance of O-RADS US Risk Stratification System for Malignancy. Radiology. 2023;308(2):e223003. 10.1148/radiol.223003.
24. Integrating Contrast-enhanced US to O-RADS US for Classification of Adnexal Lesions with Solid Components. Time-intensity Curve Analysis versus Visual Assessment. Radiol Imaging Cancer. 2024;6(6):e240024. 10.1148/rycan.240024.
25. Morice P, Scambia G, Abu-Rustum NR, et al. Fertility-sparing treatment and follow-up in patients with cervical cancer, ovarian cancer, and borderline ovarian tumours: guidelines from ESGO, ESHRE, and ESGE. Lancet Oncol. 2024;25:e602–10.

Figures

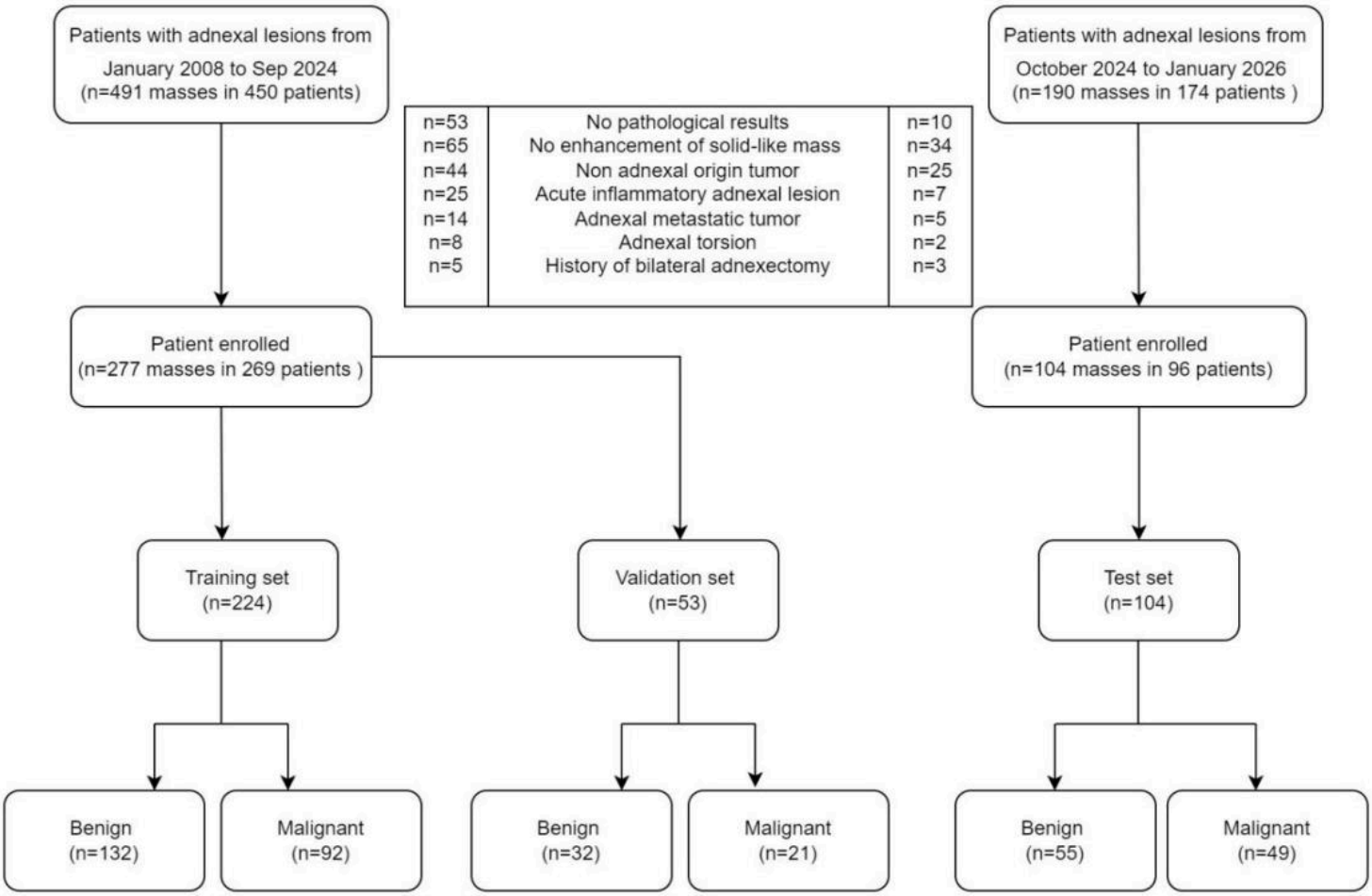

Figure 1

Patient selection process: the flowchart depicts the consecutive steps of screening, eligibility assessment, and exclusion from the initial population to the final analytical cohorts

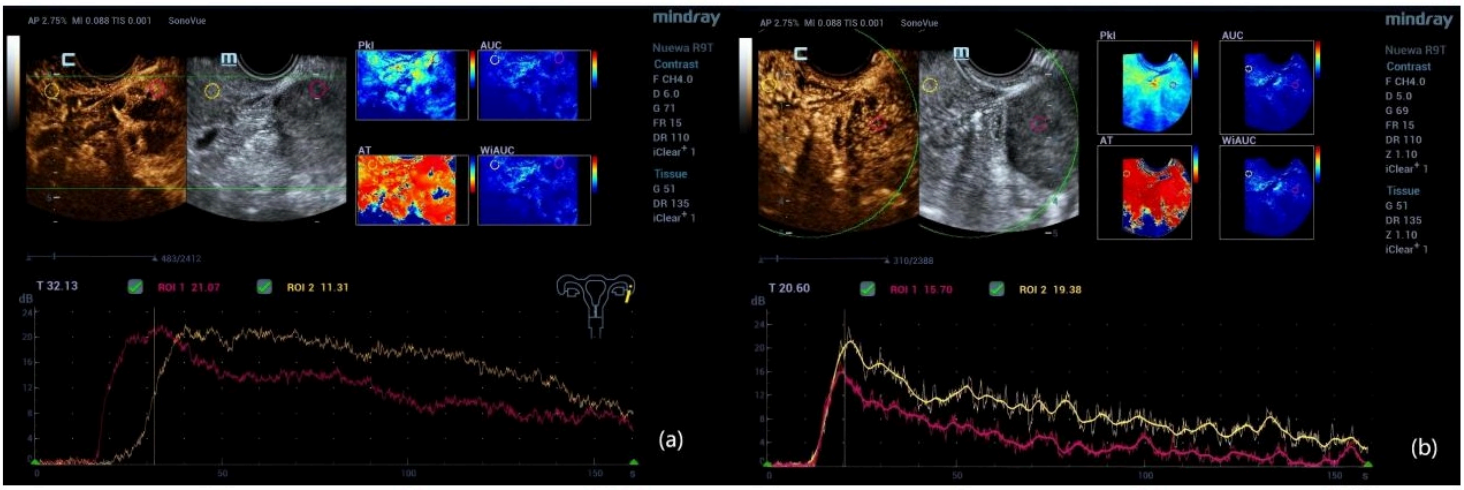

Figure 2

TICs of adnexal solid mass (RED)relative to uterine myometrium:

(a): early wash-in and washout, the pathology confirmed an adnexal high-grade serous carcinoma;

(b): synchronous wash-in and washout, the pathology confirmed a thecofibroma.

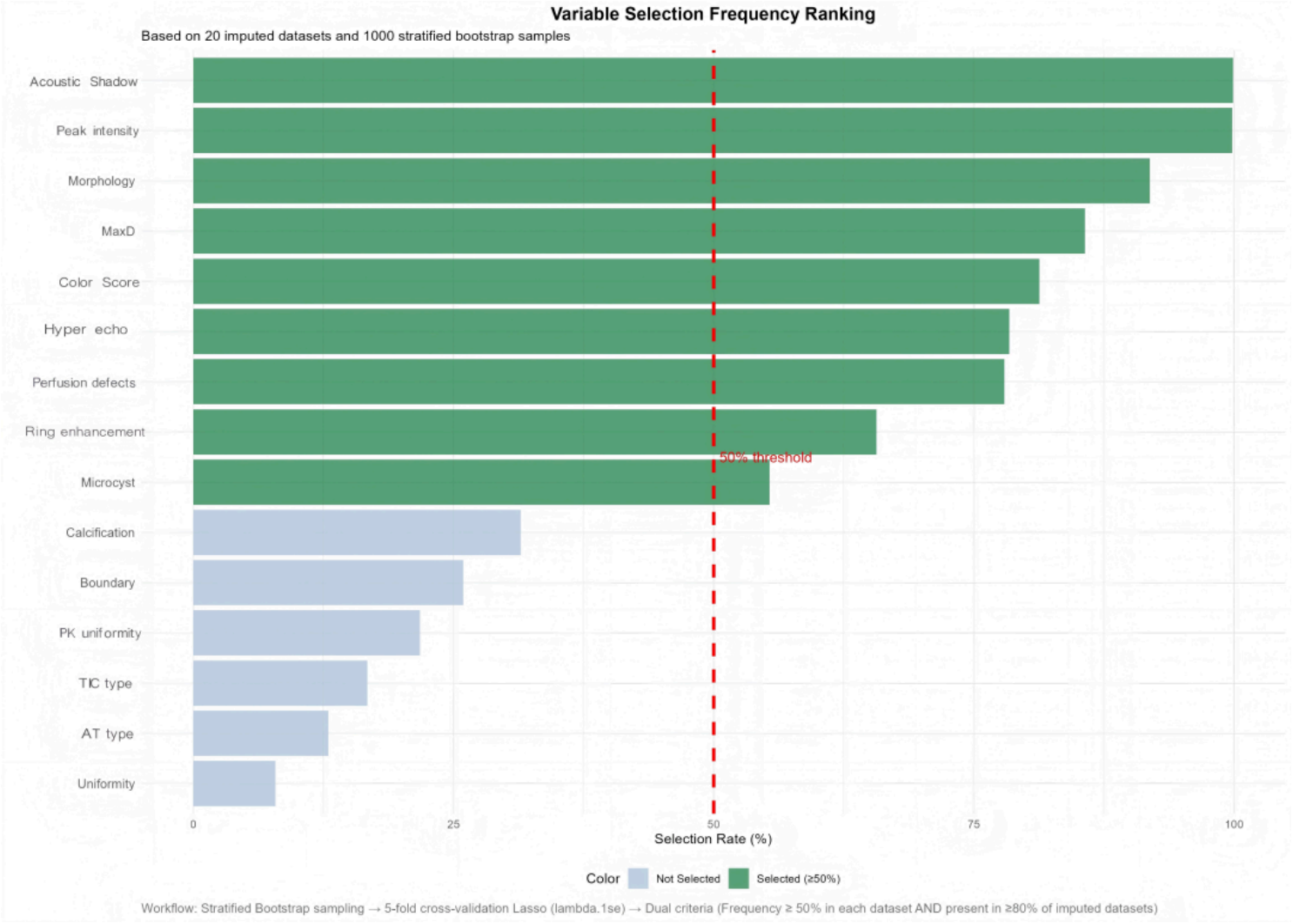

Figure 3

Waterfall plot of variable importance contributions

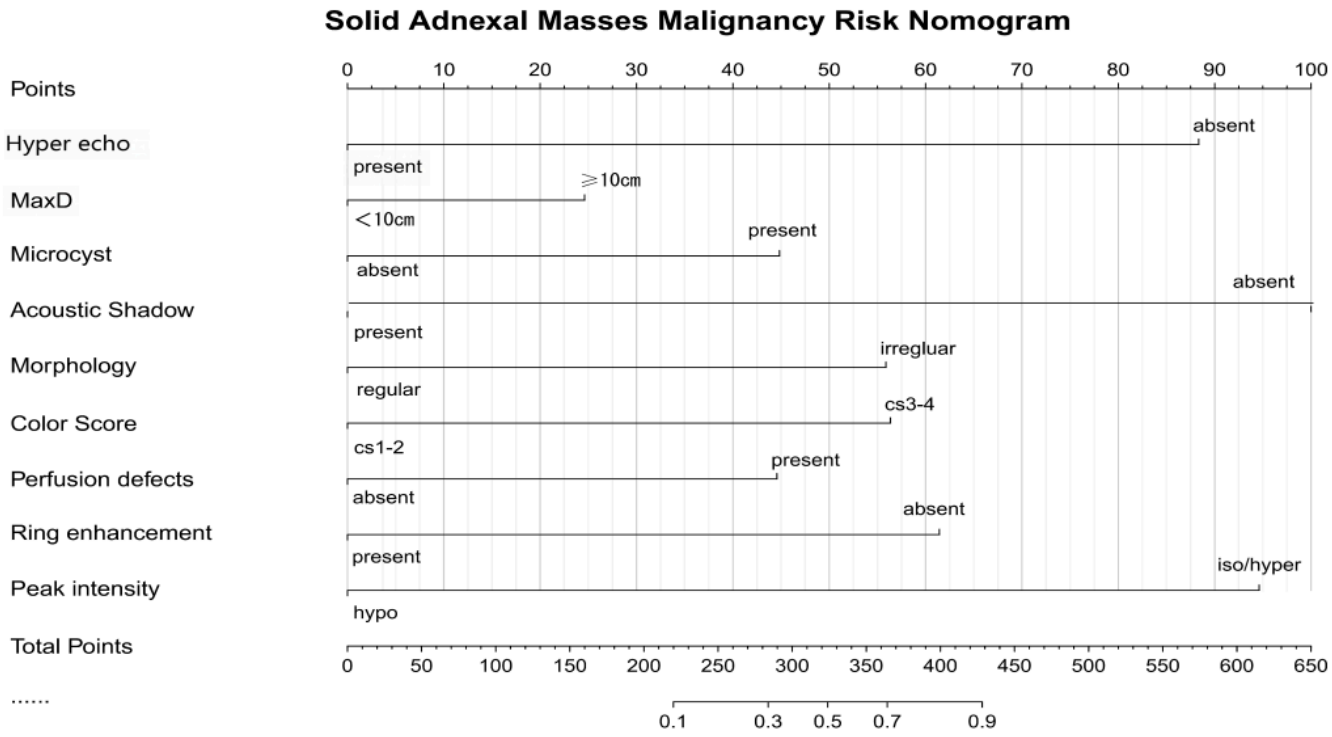

Figure 4

Malignant risk nomogram of solid adnexal masses

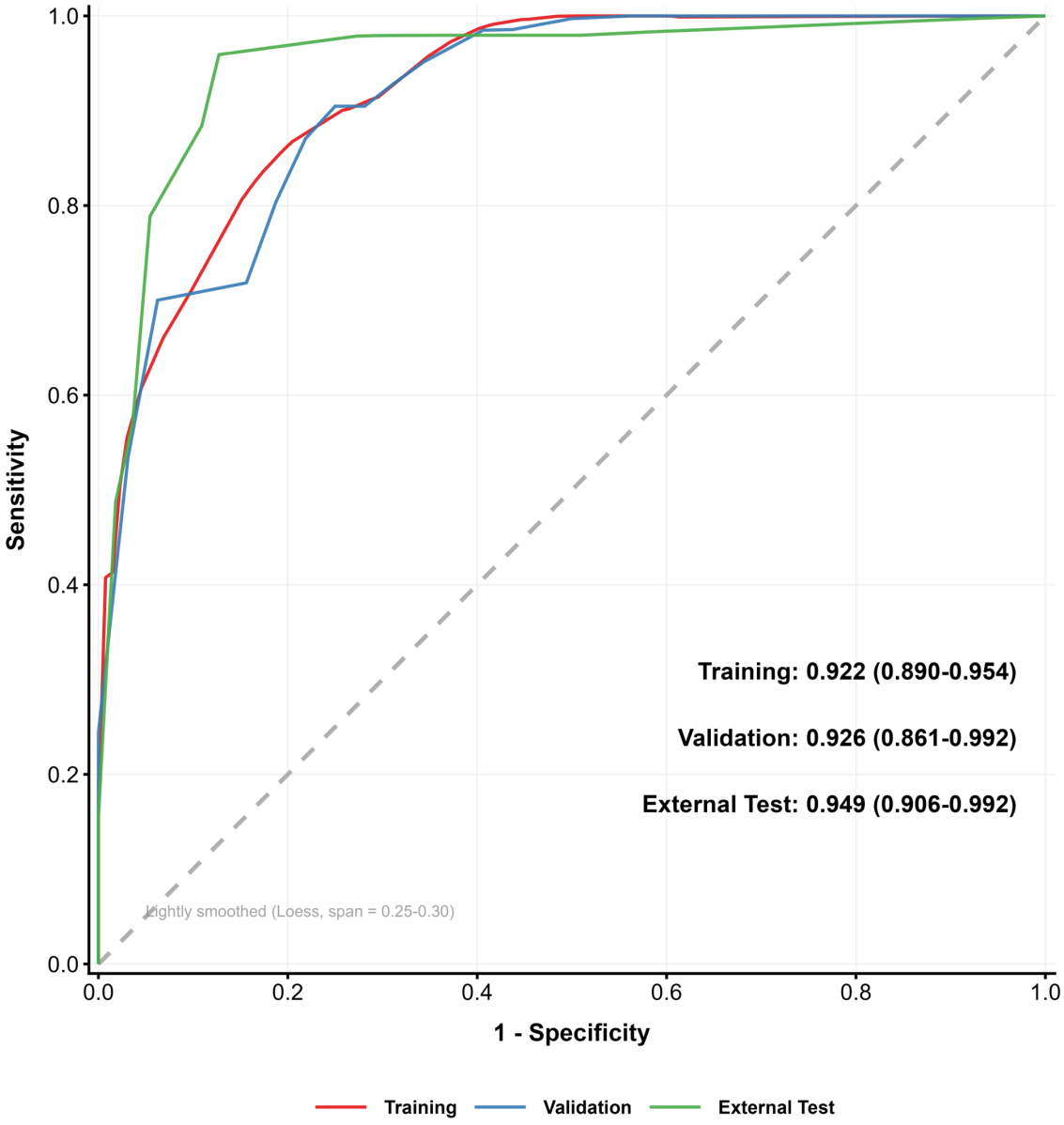

Figure 5

The ROCs of training, validation, external test sets

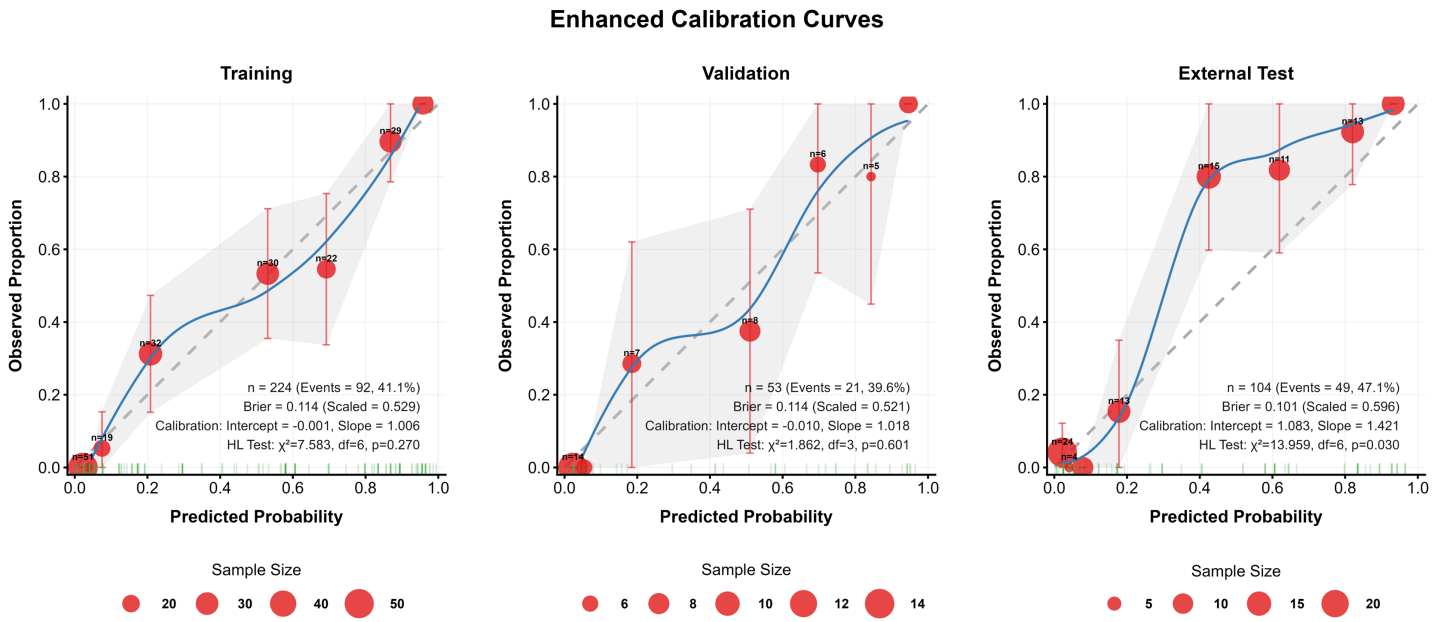

Figure 6

The Calibration Curve of train, validation, external test set

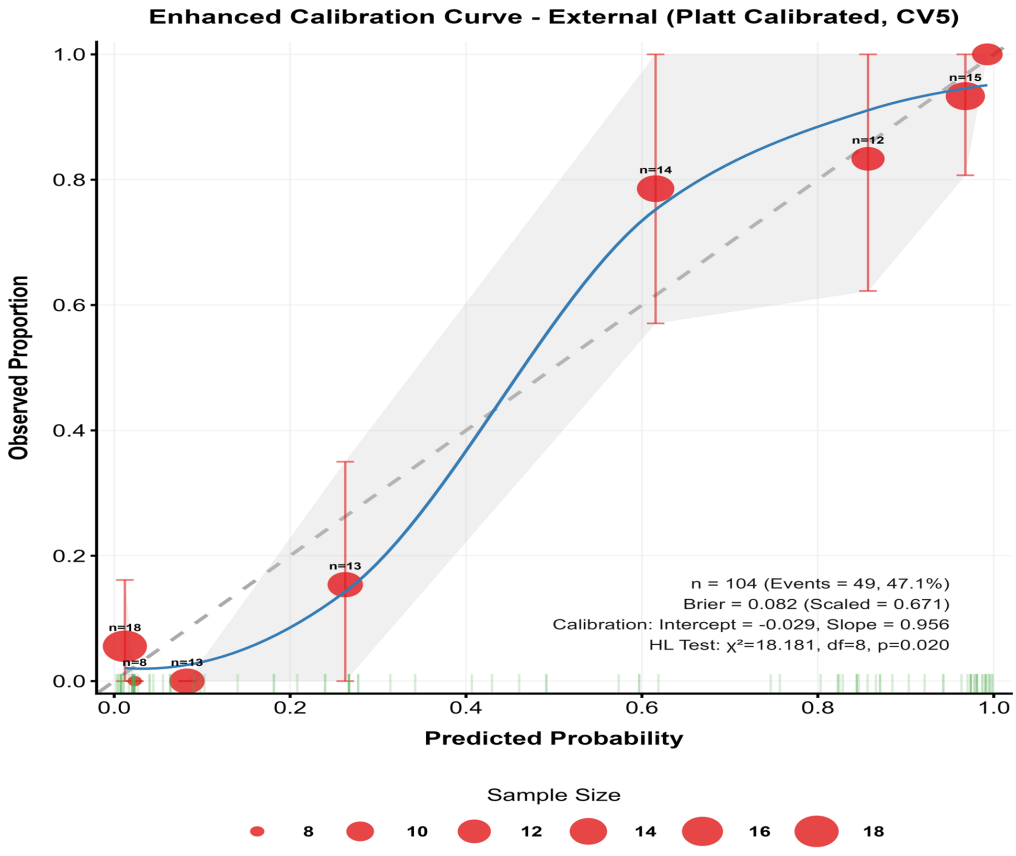

Figure 7

The Calibration Curve of external test set (Platt calibrated, CV5)

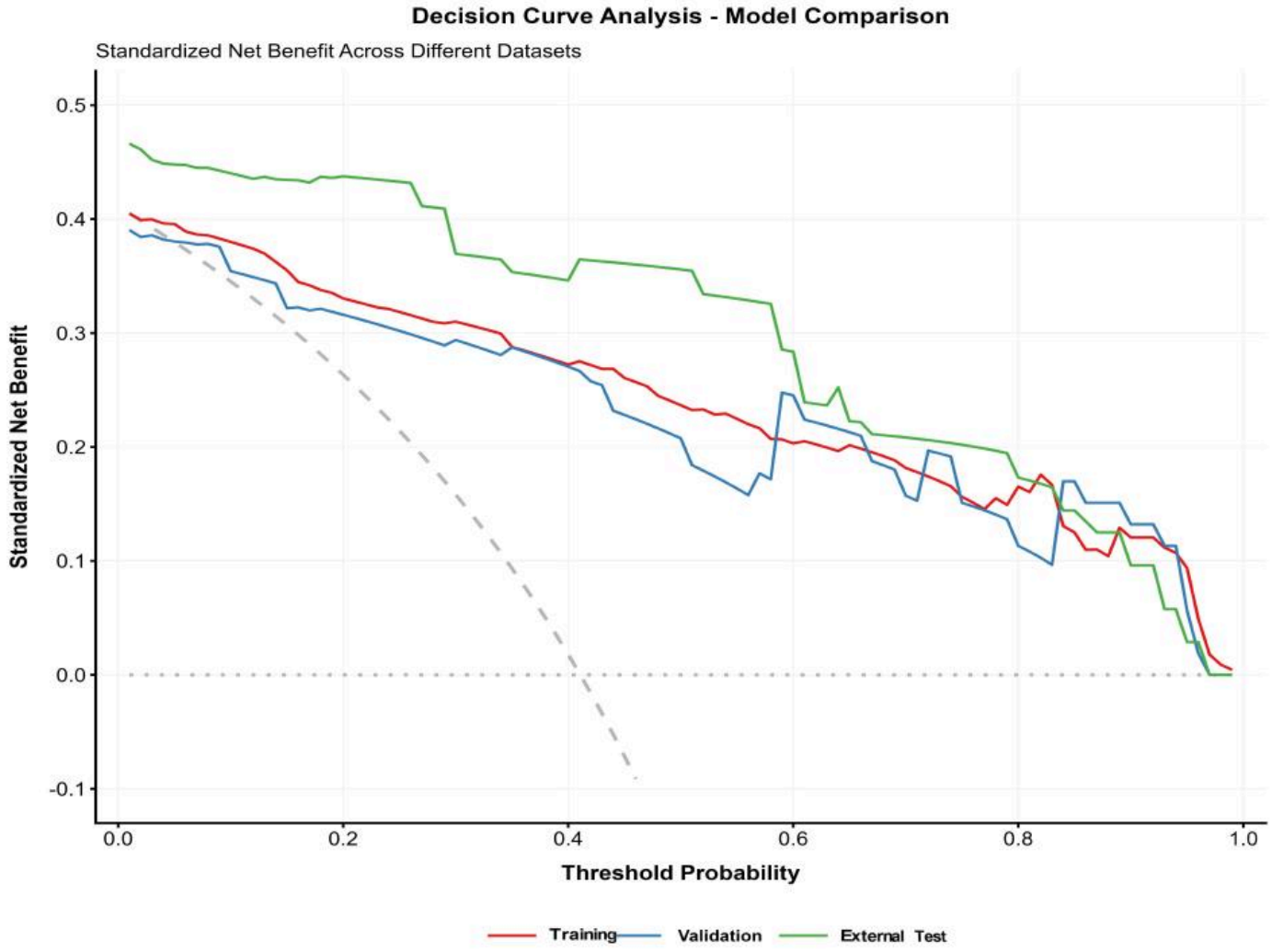

Figure 8

The DCA of train, validation, external test set

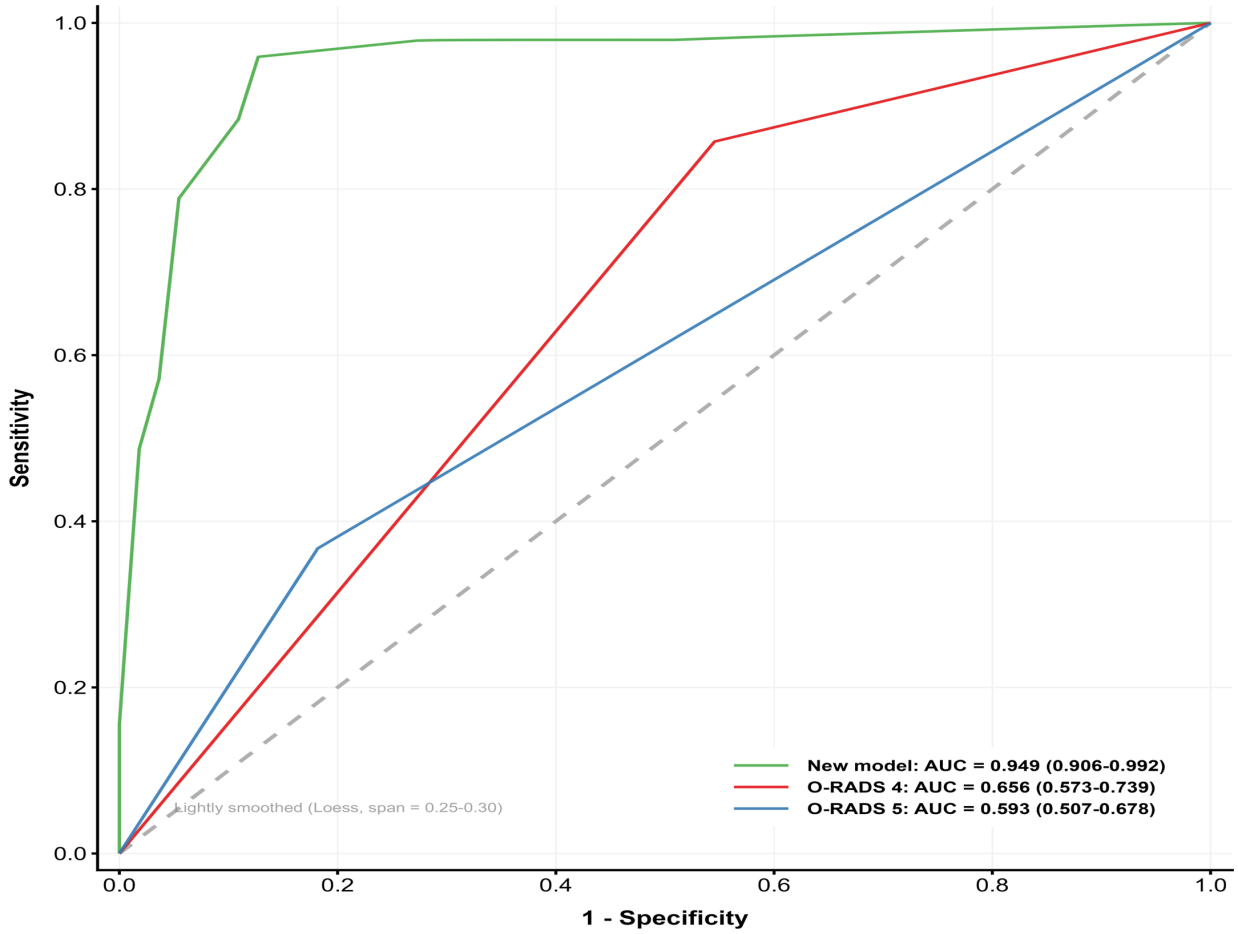

Figure 9

ROC of new model，O-RADS≥4，and O-RADS=5

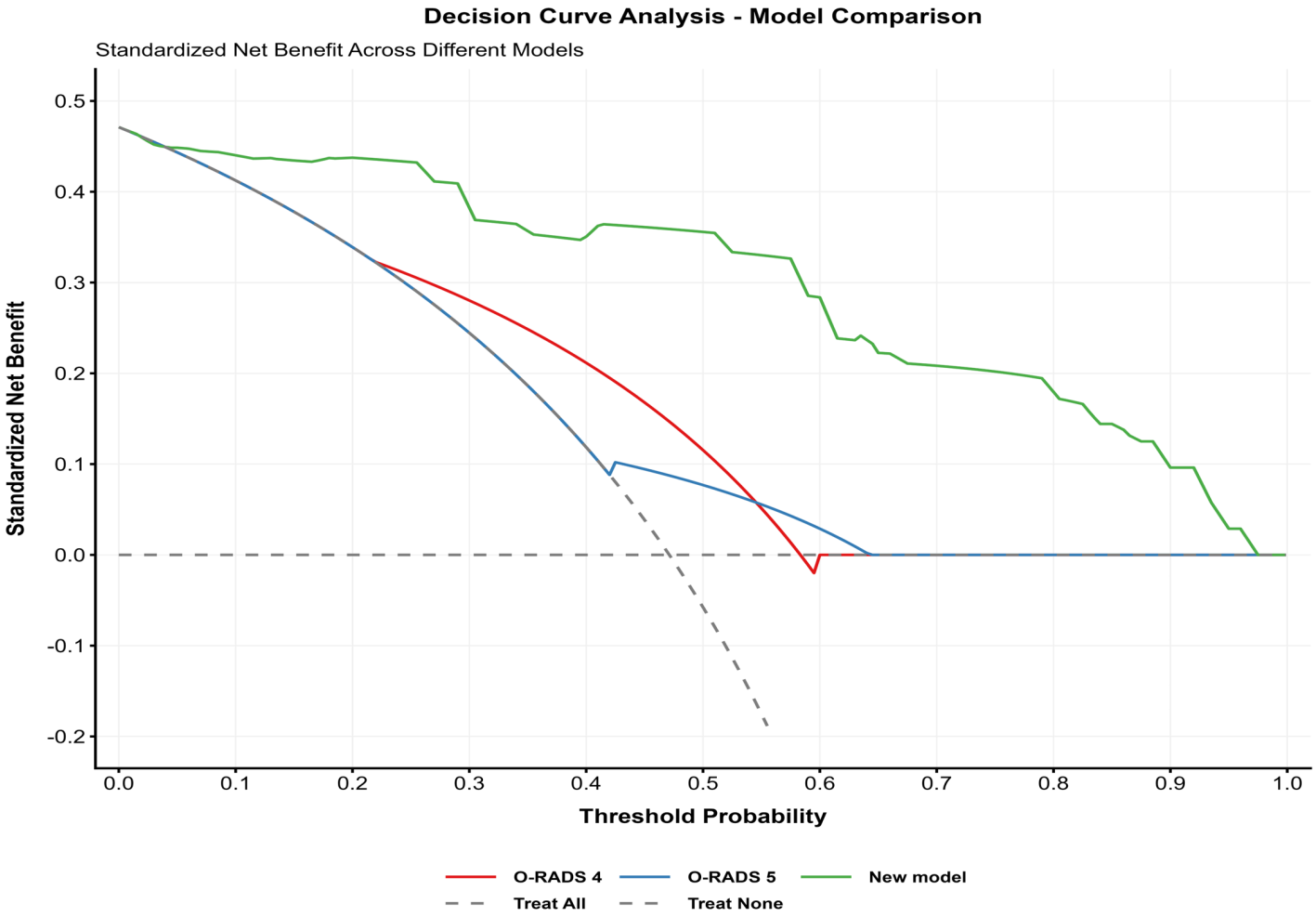

Figure 10

The DCA curves of the new model compared with the updated O-RADS classification (malignant for O-RADS ≥4 or O-RADS =5)