Am J Cancer Res 2026;16(7):2740-2762
www.ajcr.us /ISSN:2156-6976/ajcr0173623

Original Article

Integration of serological markers with O-RADS ultrasound risk stratification achieves high diagnostic accuracy for early ovarian cancer: development and validation of an interpretable model

Tingting Hua[1*], Xinchun Wu[2*], Weihui Liu[3], Zhiying Jia[1], Lianhua Zhang[1], Dilibaier Abuduniyazi[4], Yan Ma[1], Li Li[1], Xiaoyan Zhu[5], Muzaipaer Muhetaer[5], Abudushalamu Abulaiti[5], Fucheng Ma[1]

[1]Department of Ultrasound Diagnosis, The 3rd Affiliated Teaching Hospital of Xinjiang Medical University (Affiliated Cancer Hospital), Urumqi 830010, Xinjiang Uygur Autonomous Region, China; [2]Medical Imaging Center, The Fifth Affiliated Hospital of Xinjiang Medical University, Urumqi 830026, Xinjiang Uygur Autonomous Region, China; [3]Department of Ultrasound, Midong District Traditional Chinese Medicine Hospital, Urumqi 831400, Xinjiang Uygur Autonomous Region, China; [4]Department of Ultrasound, Xinjiang Uygur Medicine Hospital (Second People's Hospital of Xinjiang Uygur Autonomous Region), Urumqi 830001, Xinjiang Uygur Autonomous Region, China; [5]Department of Ultrasound, Shule County People's Hospital, Kashgar Prefecture 844000, Xinjiang Uygur Autonomous Region, China. [*]Equal contributors and co-first authors.

Received April 27, 2026; Accepted June 16, 2026; Epub July 15, 2026; Published July 30, 2026

Abstract: This study aimed to enhance preoperative diagnostic accuracy for early ovarian cancer (EOC) by constructing and validating an interpretable predictive model integrating serological markers with O-RADS ultrasound risk stratification. A total of 674 patients with adnexal masses (270 EOC, 404 benign) were enrolled and randomly assigned to training (n=438) and validation (n=236) sets. Multivariate logistic regression identified cancer antigen 125 (CA125; P=0.002), human epididymal protein 4 (HE4; P=0.001), carcinoembryonic antigen (CEA; P=0.005), maximum tumor diameter (P=0.005), and Ovarian-Adnexal Reporting and Data System (O-RADS) classification (P<0.001) as independent predictors. Among individual indices, O-RADS performed best (AUC=0.854), followed by CA125 (AUC=0.783), HE4 (AUC=0.744), CEA (AUC=0.717), and maximum tumor diameter (MTD; AUC=0.696). The comprehensive model achieved an AUC of 0.969 (95% CI: 0.953-0.985) in the training set and 0.973 (95% CI: 0.952-0.994) in the validation set, with sensitivity of 0.907/0.897, specificity of 0.914/0.980, and accuracy of 0.911/0.949, significantly outperforming both the marker combination model (AUC=0.918) and individual indices (DeLong test, all P<0.05). Robust calibration was confirmed (training set: Brier score=0.062, Hosmer-Lemeshow P=0.126; validation set: Brier score=0.056, Hosmer-Lemeshow P=0.578), and decision curve analysis demonstrated positive net benefits across a wide threshold probability range. SHAP analysis identified O-RADS classification, CA125, and HE4 as the most influential predictors, with notable interactions among them. Bootstrap resampling and 5-fold cross-validation confirmed model stability. The comprehensive model demonstrated excellent diagnostic performance, robust calibration, and clinical utility, providing a reliable interpretable tool for preoperative EOC risk stratification.

Keywords: Early ovarian cancer, O-RADS, CA125, HE4, logistic regression, SHAP, predictive models

Introduction

Ovarian cancer (OC) is among the most lethal gynecologic malignancies, characterized by an insidious onset and rapid progression. According to GLOBOCAN 2022 estimates, approximately 324,000 new cases and 207,000 OC-related deaths occur annually worldwide, with incidence and mortality rates continuing to rise [1, 2]. The deep-seated pelvic location of the ovaries, coupled with the absence of specific early symptoms, results in over 70% of patients being diagnosed at advanced stages (FIGO III-IV), when curative intervention is no longer feasible [3, 4]. The prognosis of OC is intimately linked to the stage at initial diagno-

https://doi.org/10.62347/PGIS8026

sis: patients with early-stage disease (FIGO I-II) who receive timely and optimal treatment have a 5-year survival rate exceeding 85-90%, whereas this rate drops to below 30% for those with advanced-stage disease [4, 5]. Therefore, improving the accurate identification of early ovarian cancer (EOC) is a critical determinant of patient outcomes and a major priority in gynecologic oncology research.

The primary preoperative tools for adnexal mass evaluation include serum tumor markers and imaging modalities. Among serological markers, carbohydrate antigen 125 (CA125) remains the most widely used biomarker in OC diagnosis. However, its clinical utility is limited by suboptimal specificity, as elevated levels are frequently observed in benign gynecologic conditions (e.g., endometriosis, pelvic inflammatory disease, uterine leiomyomas) and physiological states (e.g., menstruation, pregnancy), resulting in false-positive findings and unnecessary surgical interventions [5]. Human epididymal protein 4 (HE4) has emerged as a complementary marker with superior specificity for distinguishing benign from malignant adnexal masses and is less influenced by benign gynecologic conditions than CA125. Nevertheless, HE4 alone lacks sufficient sensitivity for standalone screening [6]. Carcinoembryonic antigen (CEA), though primarily associated with gastrointestinal malignancies, may also be elevated in certain OC subtypes (e.g., mucinous carcinoma); however, its overall diagnostic value as an independent marker is limited [7]. To overcome the deficiencies of single markers, combined algorithms such as the Risk of Ovarian Malignancy Algorithm (ROMA) and the Risk of Malignancy Index (RMI) have been developed. While these algorithms improve diagnostic accuracy compared with individual markers, their performance varies across different populations, menopausal statuses, and histological subtypes, and they remain incompletely validated in EOC-specific settings [8].

In the imaging domain, transvaginal ultrasound (US) is the first-line imaging modality for adnexal mass characterization. The Ovarian-Adnexal Reporting and Data System (O-RADS), developed by the American College of Radiology, provides a standardized lexicon and evidence-based risk stratification system that categorizes adnexal masses into five categories according to morphological features, including internal components, papillary projections, solid areas, and vascularity [9]. The updated O-RADS US v2022 further refined the classification criteria to improve diagnostic consistency and specificity [10]. However, a notable diagnostic challenge persists in the “gray zone” (particularly O-RADS categories 3-4), where atypical imaging features of early-stage malignancies may overlap substantially with those of benign lesions such as endometriomas, teratomas, or tubo-ovarian abscesses, resulting in considerable diagnostic uncertainty [11, 12]. Relying on O-RADS alone, therefore, may not adequately capture all EOC cases.

The biological rationale for combining serological markers with imaging stratification lies in their complementary information dimensions: serological markers reflect the tumor’s functional biological behavior (e.g., proliferation, secretion, and metabolic alterations), whereas O-RADS US stratification captures its morphological and architectural characteristics. These two data modalities are mechanistically independent yet clinically synergistic [13]. In recent years, multimodal data fusion strategies have become a prominent research direction in oncology diagnostics. Several studies have demonstrated that predictive models or nomograms integrating serological and imaging information outperform individual markers or imaging scores alone in ovarian tumor risk assessment [11, 12, 14]. However, several important gaps remain in the current literature. First, most existing models are developed for all-stage OC rather than specifically targeting EOC, where the diagnostic challenge is greatest and the clinical benefit of accurate detection is most profound. Second, many high-performing predictive models function as “black boxes”, lacking interpretability - that is, they cannot elucidate the specific contribution of each predictor variable to an individual patient’s risk estimation, which limits their transparency, credibility, and ultimately their clinical adoption [15]. Third, the interaction effects between key predictors (e.g., how elevated CA125 modifies the predictive contribution of O-RADS classification, and vice versa) have rarely been systematically investigated, despite their potential importance for understanding the underlying decision-making mechanisms of the model.

In the era of precision medicine, there is a growing demand for predictive tools that are not only

accurate and robust but also interpretable, transparent, and actionable at the individual patient level. Model interpretability - the ability to decompose and quantify each predictor's specific contribution to a given prediction - is essential for building clinician trust and facilitating shared decision-making [15]. Shapley Additive Explanations (SHAP), a game-theory-based interpretability framework, has recently gained attention in medical machine learning due to its ability to provide both global (population-level) and local (individual-level) explanations of model predictions, including feature importance rankings, individual prediction decomposition, and pairwise feature interaction analysis.

Against this background, the present study was designed to address the above gaps by: (1) Focusing specifically on EOC (FIGO I-II) rather than all-stage OC; (2) Systematically integrating three serological markers (CA125, HE4, and CEA) with O-RADS US risk stratification and key tumor structural characteristics (maximum tumor diameter); (3) Employing SHAP to provide comprehensive model interpretability, including feature importance quantification, individual-level prediction explanation, and interaction effect analysis; and (4) Rigorously evaluating model performance through multi-dimensional assessment (discrimination, calibration, and clinical net benefit) in both training and independent validation cohorts. We hypothesized that the comprehensive model would demonstrate superior diagnostic performance compared with individual markers or the serological marker combination alone, while maintaining good calibration and clinical utility across diverse clinical subgroups.

Materials and methods

Study design

Using a single-centered, retrospective design, this study included patients with pathologically diagnosed adnexal masses who were surgically treated in our institution (July 2023-March 2026). This retrospective study was approved by the Institutional Ethics Committee of the Third Affiliated Teaching Hospital of Xinjiang Medical University (Affiliated Cancer Hospital). The requirement for written informed consent was waived by the Ethics Committee due to the retrospective nature of the study and the use of anonymized clinical data. The pathological results were used to dichotomize patients into an EOC group (FIGO stages I-II) and a benign lesion group.

Sample size computation was based on the previously reported malignancy incidence of 21.7%-28.3% among the population with adnexal mass [16, 17]. Also, the difference among different research groups was considered by taking a median value of P=0.20 in this study. Using the formula $N = Z^2 \times [P \times (1\text{-}P)]/E^2$, where α=0.05 (Z=1.96) and the allowable error E=0.05, the calculated minimum sample size was 246 cases. Finally, 674 patients, which met the requirements of sample size, were enrolled. Using random sampling, they were allocated to a training set (n=438) for model construction and a validation set (n=236) for independent performance evaluation according to a 7:3 ratio.

Clinical data and measurements

Eligibility requirements for study enrollment: (1) Aged ≥18 years; (2) Underwent elective surgical treatment for adnexal mass(es) in our institution between July 2023 and March 2026; (3) Had complete postoperative histopathological diagnosis confirming either primary early ovarian cancer (FIGO I-II) or benign adnexal lesions; (4) Completed serological marker detection (CA125, HE4, CEA) and standardized transvaginal ultrasound examination with O-RADS classification within 2 weeks prior to surgery; (5) Had complete and retrievable clinical records. Exclusion grounds: (1) Previous or concurrent other malignancies; (2) Receipt of any preoperative anti-tumor therapy (chemotherapy, radiotherapy, targeted therapy, or hormonal therapy); (3) FIGO stage III-IV ovarian cancer, borderline ovarian tumors, or metastatic tumors to the ovary; (4) Recurrent adnexal masses or emergency surgery; (5) Pregnancy, breastfeeding, or severe renal insufficiency (confounding tumor marker levels) [6, 8]; (6) Incomplete key clinical, serological, or imaging data (Figure S1).

Patients' clinical data, including age, body mass index (BMI), menopausal status, family tumor history, and hypertension/diabetes history, were consulted and recorded. The serological markers assessed covered CA125, HE4, and CEA. O-RADS classification, MTD, intratumoral solid component, papillary projections,

seroperitoneum, and bilateral lesions were the imaging parameters evaluated.

All continuous variables were tested for distribution before modeling. Variables with a skewed distribution were log-transformed or analyzed in the form of medians. Categorical variables were processed using dummy variable encoding. Only patients with complete clinical, serological, and imaging data were included in the final analysis. Cases with missing key variables were excluded prior to analysis, and thus no imputation methods were applied.

The quantification of serological markers employed the enzyme-linked immunosorbent assay (ELISA) technique using kits manufactured by Shanghai Enzyme-linked Biotechnology Co., Ltd. (Catalog Numbers: CA125-ELISA-2023, HE4-ELISA-2023, CEA-ELISA-2023; Batch Number: 202306-202602). The instructions were strictly followed during the measurement. All blood samples were collected on an empty stomach in the early morning. Following a 10-minute centrifugation at 3000 rpm, the serum was separated and stored at -80°C for testing.

US examinations were completed by an experienced sonographer (>5 years' experience) using a high-resolution transvaginal US machine (GE Healthcare, Voluson E8). All cases were graded and evaluated according to the O-RADS standard, and any disagreements (if any) were reviewed and agreed upon by two senior physicians.

Model development and outcomes

In this study, EOC (yes/no) was used as the principal outcome variable (binary outcome). During training, all the candidate variables were analyzed by univariate Logistic regression to screen for outcome-associated significant correlates ($P<0.05$). Subsequently, the screened variables were incorporated into the multivariate Logistic regression model, followed by final model variable determination via stepwise selection.

Given that some ultrasonic features have been integrated into the O-RADS classification system (e.g., solid components, papillary projections, seroperitoneum), variables with overlapping concepts were removed during modeling to avoid information duplication and model redundancy. Only those with independent clinical significance were retained.

Multicollinearity evaluation used the Variance Inflation Factor (VIF). Obvious collinearity was indicated by VIF >5, entailing variable removal.

Based on the final identified variables, two types of models were constructed: (1) A combination model of serological markers (CA125, HE4, CEA), referred to as the marker combination model; and (2) A comprehensive model (integrating serological markers with MTD and O-RADS classification).

The secondary endpoints included the diagnostic efficacy of individual markers and models, assessed through the area under the curve (AUC), sensitivity, specificity, accuracy, F1 value, and Youden index, as well as the performance of the models across different clinical subgroups.

Statistical analysis

Statistical data analyses were completed by R software (version 4.5.1) and SPSS (version 27.0). Measurement data underwent normality testing with the Shapiro-Wilk test. Those of Gaussian distribution were represented by the mean ± standard deviation and assessed for between-group differences using independent samples t-tests. If violating the Gaussian distribution, measurement data were shown as the median (quartile) and analyzed by Mann-Whitney U tests. Categorical variables were summarized as frequencies and corresponding percentages, and between-group comparisons were performed using the χ^2 test or Fisher's exact test, as appropriate. Correlations between variables were assessed by Spearman's correlation, and multicollinearity was evaluated by VIF, with a VIF >5 indicating the existence of potential collinearity issues. To avoid model over-fitting and variable redundancy, variables were screened by combining clinical significance and statistical results before multivariate analysis. In the training set, univariate Logistic regression analysis was first performed for initial screening of the candidate variables ($P<0.05$); those with significance were then incorporated into the multivariate Logistic regression model, using the stepwise regres-

sion method based on the Akaike Information Criterion (AIC) to determine the predictors for final modeling. Regression model coefficients were used to construct the prediction probabilities. The diagnostic efficiency of the models was evaluated by receiver operating characteristic (ROC) curves and AUCs, and the 95% confidence intervals (CIs) were calculated. AUC differences among different models were examined by the DeLong test. Optimal cutoff points were determined by the maximum Youden index. The models' discriminative capacities were evaluated by the AUC and concordance-statistic (C-statistic), with calibration evaluated by the calibration curve, Brier score, and Hosmer-Lemeshow (H-L) goodness-of-fit test. Practicability assessment used the decision curve analysis (DCA) by determining net benefits across threshold probabilities. External (validation set) and internal validation (1,000 bootstrapping iterations and 5-fold cross-validation) for stability and generalization assessments were also conducted. SHAP-based interpretability evaluation through feature importance ranking (average absolute SHAP value), individual prediction contribution, and variable interaction analyses was carried out. Sensitivity analysis evaluated whether O-RADS-specific US features provided additional predictive value. This study also built an extended model by adding intratumoral solid components and papillary projections to the main model. AUCs determined model discrimination, the DeLong test examined inter-model differences, the H-L test and Brier score assessed calibration, and the AIC and Bayesian information criterion (BIC) checked fitting. VIF calculation was conducted for potential multicollinearity assessment. All tests (two-tailed) took the $P<0.05$ significance criterion.

Results

Clinical and ultrasonic feature comparison

674 eligible patients were included, assigned based on postoperative pathology to EOC (270 cases) and benign lesion (404 cases) groups. FIGO stage I was predominant among EOC cases, followed by stage II; pathologically, serous carcinoma was the most prevalent, then came mucinous, endometrioid, and other types. The study cohorts were balanced in marital status, education, residence, and history of hypertension and diabetes ($P>0.05$; **Table 1**). Concerning clinical features, the EOC group was greater in terms of age ($P<0.001$), BMI ($P=0.027$), postmenopausal woman proportion ($P<0.001$), and family tumor history positivity rate relative to their benign lesion counterparts ($P=0.002$). From the perspective of serological markers, CA125, HE4, and CEA presented higher levels in EOC patients than in those with benign lesions ($P<0.001$; **Table 1**). As far as ultrasonic features are concerned, the O-RADS classification was higher in patients with EOC, with a larger proportion of patients falling into the category 4 or 5 ($P<0.001$); EOC cases also exhibited larger MTD, as well as greater rates of intratumoral solid component, papillary projections, seroperitoneum, and bilateral lesions, when compared to the benign lesion group (all $P<0.001$; **Table 1**).

Baseline data comparison (training set vs. validation set)

To verify dataset splitting rationality, this study conducted a balance test on the baseline characteristics of the training (n=438) and validation (n=236) sets. Statistical significance was absent regarding general demographic characteristics, like marital status, educational background, residence, menopause status, hypertension/diabetes history, and family tumor history ($P>0.05$). In terms of clinical and serological indices, age, BMI, CA125, HE4, CEA, and MTD also differed non-significantly ($P>0.05$). Inter-group comparability was also confirmed in ultrasonic features, including the O-RADS classification distribution and the incidences of intratumoral solid component, papillary projections, seroperitoneum, and bilateral lesions ($P>0.05$). Notably, all variables showed a standardized mean difference (SMD) below 0.1, suggesting well-balanced baseline characteristics between the training and validation sets and supporting dataset division rationality (**Table 2**).

Clinical and ultrasonic feature comparison (EOC vs. benign lesion groups in the training set)

In the training set (n=438), EOC cases (n=172) showed equivalent distributions in marital status, educational background, residence, hypertension/diabetes history, and BMI to benign lesion patients (n=266; $P>0.05$; **Table 3**), but were greater in age ($P<0.001$), the proportion

Table 1. Comparison of clinical features and ultrasonic parameters (EOC group vs. benign lesion group)

Variable	Total	EOC group (n=270)	Benign lesion group (n=404)	Statistic	P
Marital status				0.065	0.798
Married	577 (85.61%)	230 (85.19%)	347 (85.89%)		
Single/divorced/widowed	97 (14.39%)	40 (14.81%)	57 (14.11%)		
Education				3.238	0.198
Junior high school or below	182 (27.00%)	83 (30.74%)	99 (24.50%)		
Senior high school/technical secondary school	239 (35.46%)	92 (34.07%)	147 (36.39%)		
Junior college or above	253 (37.54%)	95 (35.19%)	158 (39.11%)		
Residence				0.325	0.569
Urban	438 (64.99%)	172 (63.70%)	266 (65.84%)		
Rural	236 (35.01%)	98 (36.30%)	138 (34.16%)		
Menopause status				38.043	<0.001
Premenopause	372 (55.19%)	110 (40.74%)	262 (64.85%)		
Postmenopause	302 (44.81%)	160 (59.26%)	142 (35.15%)		
Hypertension				1.473	0.225
Yes	185 (27.45%)	81 (30.00%)	104 (25.74%)		
No	489 (72.55%)	189 (70.00%)	300 (74.26%)		
Diabetes				1.980	0.159
Yes	92 (13.65%)	43 (15.93%)	49 (12.13%)		
No	582 (86.35%)	227 (84.07%)	355 (87.87%)		
Family tumor history				9.157	0.002
Yes	116 (17.21%)	61 (22.59%)	55 (13.61%)		
No	558 (82.79%)	209 (77.41%)	349 (86.39%)		
O-RADS ultrasound risk stratification				267.991	<0.001
Category 2	107 (15.88%)	3 (1.11%)	104 (25.74%)		
Category 3	185 (27.45%)	30 (11.11%)	155 (38.37%)		
Category 4	250 (37.09%)	113 (41.85%)	137 (33.91%)		
Category 5	132 (19.58%)	124 (45.93%)	8 (1.98%)		
Intratumoral solid component				89.529	<0.001
Yes	300 (44.51%)	180 (66.67%)	120 (29.70%)		
No	374 (55.49%)	90 (33.33%)	284 (70.30%)		
Papillary projections				73.739	<0.001
Yes	175 (25.96%)	118 (43.70%)	57 (14.11%)		
No	499 (74.04%)	152 (56.30%)	347 (85.89%)		
Seroperitoneum				36.689	<0.001
Yes	69 (10.24%)	51 (18.89%)	18 (4.46%)		
No	605 (89.76%)	219 (81.11%)	386 (95.54%)		
Bilateral lesions				63.065	<0.001
Yes	134 (19.88%)	94 (34.81%)	40 (9.90%)		
No	540 (80.12%)	176 (65.19%)	364 (90.10%)		
FIGO staging					
I	162 (60.00%)	162 (60.00%)			
II	108 (40.00%)	108 (40.00%)			
Pathological type					
Serous carcinoma	150 (55.56%)	150 (55.56%)			
Mucinous carcinoma	53 (19.63%)	53 (19.63%)			
Endometrioid carcinoma	40 (14.81%)	40 (14.81%)			
Other	27 (10.00%)	27 (10.00%)			
Age	47.76 [40.05, 55.32]	51.67 [44.99, 58.90]	45.11 [38.29, 52.11]	7.108	<0.001
BMI	23.42±3.02	23.73±3.11	23.21±2.94	-2.214	0.027
CA125	40.34 [21.87, 68.76]	81.84 [37.26, 150.21]	32.25 [17.67, 44.19]	12.815	<0.001
HE4	77.83 [53.95, 122.87]	140.75 [62.63, 224.92]	69.38 [51.78, 89.08]	10.438	<0.001
CEA	3.00 [1.55, 5.18]	5.05 [2.35, 8.31]	2.44 [1.35, 3.96]	9.632	<0.001
Maximum tumor diameter	6.06±2.78	7.28±2.90	5.25±2.37	-9.942	<0.001

Note: EOC, Early Ovarian Cancer; O-RADS, Ovarian-Adnexal Reporting and Data System; FIGO, International Federation of Gynecology and Obstetrics; BMI, Body Mass Index; CA125, Cancer Antigen 125; HE4, Human Epididymis Protein 4; CEA, Carcinoembryonic Antigen.

Table 2. Baseline feature comparison (training set vs. validation set)

Variable	Training set (n=438)	Validation set (n=236)	Statistic	P	SMD
Marital status			0.832	0.362	0.075
Married	371 (84.7%)	206 (87.3%)			
Single/divorced/widowed	67 (15.3%)	30 (12.7%)			
Education			0.281	0.869	0.040
Junior high school or below	121 (27.6%)	61 (25.8%)			
Senior high school/technical secondary school	153 (34.9%)	86 (36.4%)			
Junior college or above	164 (37.4%)	89 (37.7%)			
Residence			0.012	0.914	0.009
Urban	284 (64.8%)	154 (65.3%)			
Rural	154 (35.2%)	82 (34.7%)			
Menopause status			0.042	0.838	0.017
Premenopause	243 (55.5%)	129 (54.7%)			
Postmenopause	195 (44.5%)	107 (45.3%)			
Hypertension			1.093	0.296	0.085
Yes	126 (28.8%)	59 (25.0%)			
No	312 (71.2%)	177 (75.0%)			
Diabetes			0.271	0.603	0.042
Yes	62 (14.2%)	30 (12.7%)			
No	376 (85.8%)	206 (87.3%)			
Family tumor history			0.313	0.576	0.046
Yes	78 (17.8%)	38 (16.1%)			
No	360 (82.2%)	198 (83.9%)			
O-RADS ultrasound risk stratification			1.469	0.689	0.090
Category 2	70 (16.0%)	37 (15.7%)			
Category 3	114 (26.0%)	71 (30.1%)			
Category 4	168 (38.4%)	82 (34.7%)			
Category 5	86 (19.6%)	46 (19.5%)			
Intratumoral solid component			0.245	0.621	0.040
Yes	198 (45.2%)	102 (43.2%)			
No	240 (54.8%)	134 (56.8%)			
Papillary projections			1.111	0.292	0.085
Yes	108 (24.7%)	67 (28.4%)			
No	330 (75.3%)	169 (71.6%)			
Seroperitoneum			1.228	0.268	0.091
Yes	49 (11.2%)	20 (8.5%)			
No	389 (88.8%)	216 (91.5%)			
Bilateral lesions			1.056	0.304	0.082
Yes	82 (18.7%)	52 (22.0%)			
No	356 (81.3%)	184 (78.0%)			
Age	48.03±10.78	47.64±13.16	0.407	0.684	0.032
BMI	23.47±2.93	23.32±3.19	0.607	0.544	0.048
CA125	40.64 [22.71, 67.95]	39.55 [21.15, 70.62]	0.185	0.853	0.034
HE4	76.66 [54.42, 125.77]	82.00 [53.79, 115.04]	0.046	0.963	0.091
CEA	3.08 [1.49, 5.22]	2.97 [1.63, 5.12]	0.022	0.982	0.016
Maximum tumor diameter	6.10 [4.03, 8.05]	5.58 [4.07, 7.58]	0.931	0.352	0.071

Note: O-RADS, Ovarian-Adnexal Reporting and Data System; BMI, Body Mass Index; CA125, Cancer Antigen 125; HE4, Human Epididymis Protein 4; CEA, Carcinoembryonic Antigen; SMD, Standardized Mean Difference.

Table 3. Comparison of clinical characteristics and ultrasonic parameters in the training set (EOC group vs. benign lesion group)

Variable	Total (n=438)	EOC group (n=172)	Benign lesion group (n=266)	Statistic	P
Marital status				0.127	0.722
Married	371 (84.70%)	147 (85.47%)	224 (84.21%)		
Single/divorced/widowed	67 (15.30%)	25 (14.53%)	42 (15.79%)		
Education				2.694	0.26
Junior high school or below	121 (27.63%)	55 (31.98%)	66 (24.81%)		
Senior high school/technical secondary school	153 (34.93%)	56 (32.56%)	97 (36.47%)		
unior college or above	164 (37.44%)	61 (35.47%)	103 (38.72%)		
Residence				0.098	0.755
Urban	284 (64.84%)	110 (63.95%)	174 (65.41%)		
Rural	154 (35.16%)	62 (36.05%)	92 (34.59%)		
Menopause status				25.054	<0.001
Premenopause	243 (55.48%)	70 (40.70%)	173 (65.04%)		
Postmenopause	195 (44.52%)	102 (59.30%)	93 (34.96%)		
Hypertension				1.986	0.159
Yes	126 (28.77%)	56 (32.56%)	70 (26.32%)		
No	312 (71.23%)	116 (67.44%)	196 (73.68%)		
Diabetes				2.518	0.113
Yes	62 (14.16%)	30 (17.44%)	32 (12.03%)		
No	376 (85.84%)	142 (82.56%)	234 (87.97%)		
Family tumor history				4.582	0.032
Yes	78 (17.81%)	39 (22.67%)	39 (14.66%)		
No	360 (82.19%)	133 (77.33%)	227 (85.34%)		
O-RADS ultrasound risk stratification				167.421	<0.001
Category 2	70 (15.98%)	1 (0.58%)	69 (25.94%)		
Category 3	114 (26.03%)	19 (11.05%)	95 (35.71%)		
Category 4	168 (38.36%)	73 (42.44%)	95 (35.71%)		
Category 5	86 (19.63%)	79 (45.93%)	7 (2.63%)		
Intratumoral solid component				65.752	<0.001
Yes	198 (45.21%)	119 (69.19%)	79 (29.70%)		
No	240 (54.79%)	53 (30.81%)	187 (70.30%)		
Papillary projections				36.432	<0.001
Yes	108 (24.66%)	69 (40.12%)	39 (14.66%)		
No	330 (75.34%)	103 (59.88%)	227 (85.34%)		
Seroperitoneum				23.926	<0.001
Yes	49 (11.19%)	35 (20.35%)	14 (5.26%)		
No	389 (88.81%)	137 (79.65%)	252 (94.74%)		
Bilateral lesions				35.634	<0.001
Yes	82 (18.72%)	56 (32.56%)	26 (9.77%)		
No	356 (81.28%)	116 (67.44%)	240 (90.23%)		
Age	48.03±10.78	51.56±10.11	45.49±10.55	-6.051	<0.001
BMI	23.47±2.93	23.64±3.07	23.35±2.82	-1.014	0.311
CA125 (U/mL)	40.64 [22.71, 67.95]	79.43 [37.48, 144.16]	32.64 [17.90, 44.01]	10.113	<0.001
HE4 (pmol/L)	76.66 [54.42, 125.77]	143.83 [63.27, 247.00]	67.93 [49.97, 85.82]	8.730	<0.001
CEA (ng/mL)	3.08 [1.49, 5.22]	4.93 [2.35, 7.88]	2.43 [1.31, 4.07]	7.760	<0.001
Maximum tumor diameter (cm)	6.10 [4.03, 8.05]	7.26 [5.25, 9.44]	5.33 [3.49, 6.79]	7.011	<0.001

Note: EOC, Early Ovarian Cancer; O-RADS, Ovarian-Adnexal Reporting and Data System; BMI, Body Mass Index; CA125, Cancer Antigen 125; HE4, Human Epididymis Protein 4; CEA, Carcinoembryonic Antigen.

of postmenopausal women (P<0.001), family tumor history positivity rate (P=0.032), and CA125, HE4, and CEA concentrations (P<0.001; **Table 3**). Regarding ultrasonic features, the inter-group comparison showed higher O-RADS classification and greater proportions of categories 4 and 5 lesions in the EOC group versus the benign lesion group (P<0.001); larger MTD,

as well as greater numbers of patients with intratumoral solid component, papillary projections, seroperitoneum, and bilateral lesions, were also identified in the EOC group (all P<0.001; **Table 3**).

Multicollinearity diagnostics of predictors in the training set

Using Spearman correlation and variance inflation factor (VIF) analyses, this study assessed multicollinearity among 12 predictors in the training set (n=438). According to the correlation analysis, O-RADS classification was moderately correlated with intratumoral solid component (r=0.636) and papillary projections (r=0.513), and CA125 was moderately correlated with HE4 (r=0.456) (**Figure 1A**). All VIF values were below 5, with O-RADS classification showing the highest VIF value (3.079), indicating no severe multicollinearity among the predictors and supporting their inclusion in the subsequent modeling analysis (**Figure 1B**).

Univariate and multivariate Logistic regression analysis of predictor variables of EOC in the training set

Before multivariate analysis, intratumoral solid component and papillary projections were excluded because these ultrasound features had already been incorporated into the O-RADS classification and their simultaneous inclusion could have resulted in conceptual overlap. Consequently, nine variables were included in the logistic regression analysis. Univariate analysis identified age, CA125, HE4, CEA, MTD, menopausal status, family tumor history, O-RADS classification, and bilateral lesions as significant correlates of EOC (all P<0.05; **Figure 2**). After incorporating these nine variables into the multivariate logistic regression model, CA125 (P=0.002), HE4 (P=0.001), CEA (P=0.005), MTD (P=0.005), and O-RADS classification (P<0.001) were identified as independent predictors of EOC, whereas age (P=0.723), menopausal status (P=0.879), family tumor history (P=0.268), and bilateral lesions (P=0.555) did not reach statistical significance (**Figure 2**).

Comparison of diagnostic efficiency of single indexes and the joint models for EOC

Based on the multivariate logistic regression findings in the training set, two predictive models were constructed: a serological marker combination model comprising CA125, HE4, and CEA, and a comprehensive model further incorporating MTD and O-RADS classification. ROC curves were used to evaluate the diagnostic performance of the individual predictors and the two models. Among the individual predictors, O-RADS classification showed the highest diagnostic performance (AUC=0.854), followed by CA125 (AUC=0.783), HE4 (AUC=0.744), CEA (AUC=0.717), and MTD (AUC=0.696). The marker combination model achieved an AUC of 0.918, whereas the comprehensive model showed the highest diagnostic performance, with an AUC of 0.969 (**Figure 3A-G**). DeLong tests showed that the AUC of the comprehensive model was significantly higher than those of all individual predictors and the marker combination model (all P<0.001). The marker combination model also significantly outperformed all individual predictors, including O-RADS classification (all P<0.05). Among CA125, HE4, CEA, and MTD, only the difference between CA125 and MTD was statistically significant (P=0.012), whereas the remaining pairwise differences were not statistically significant (**Figure 3H**).

Comprehensive performance evaluation of the comprehensive model across datasets

This study systematically evaluated the comprehensive model across the training and validation sets from three dimensions: discriminative power, calibration performance, and clinical net benefits, to gain a comprehensive understanding of its diagnostic performance and clinical practical utility.

The ROC curve-based discriminant analysis showed that the AUC of the comprehensive model in the training set was 0.969 (95% CI: 0.953-0.985), and the sensitivity, specificity, accuracy, F1 score, and Youden index were 0.907, 0.914, 0.911, 0.895, and 0.821, respectively, at the corresponding optimal cut-points; during validation, the model had an AUC of 0.973 (95% CI: 0.952-0.994), with 0.897 sensitivity, 0.980 specificity, and 0.949 accuracy, as well as F1 score of 0.929 and a Youden index of 0.876. The highly consistent AUC across the datasets, plus a concordance (C) statistic exceeding 0.96, supported the model's excellent discrimination with no obvious over-fitting issues (**Figure 4A** and **4B**).

In terms of calibration performance, the training set showed a Brier score of 0.0624, an

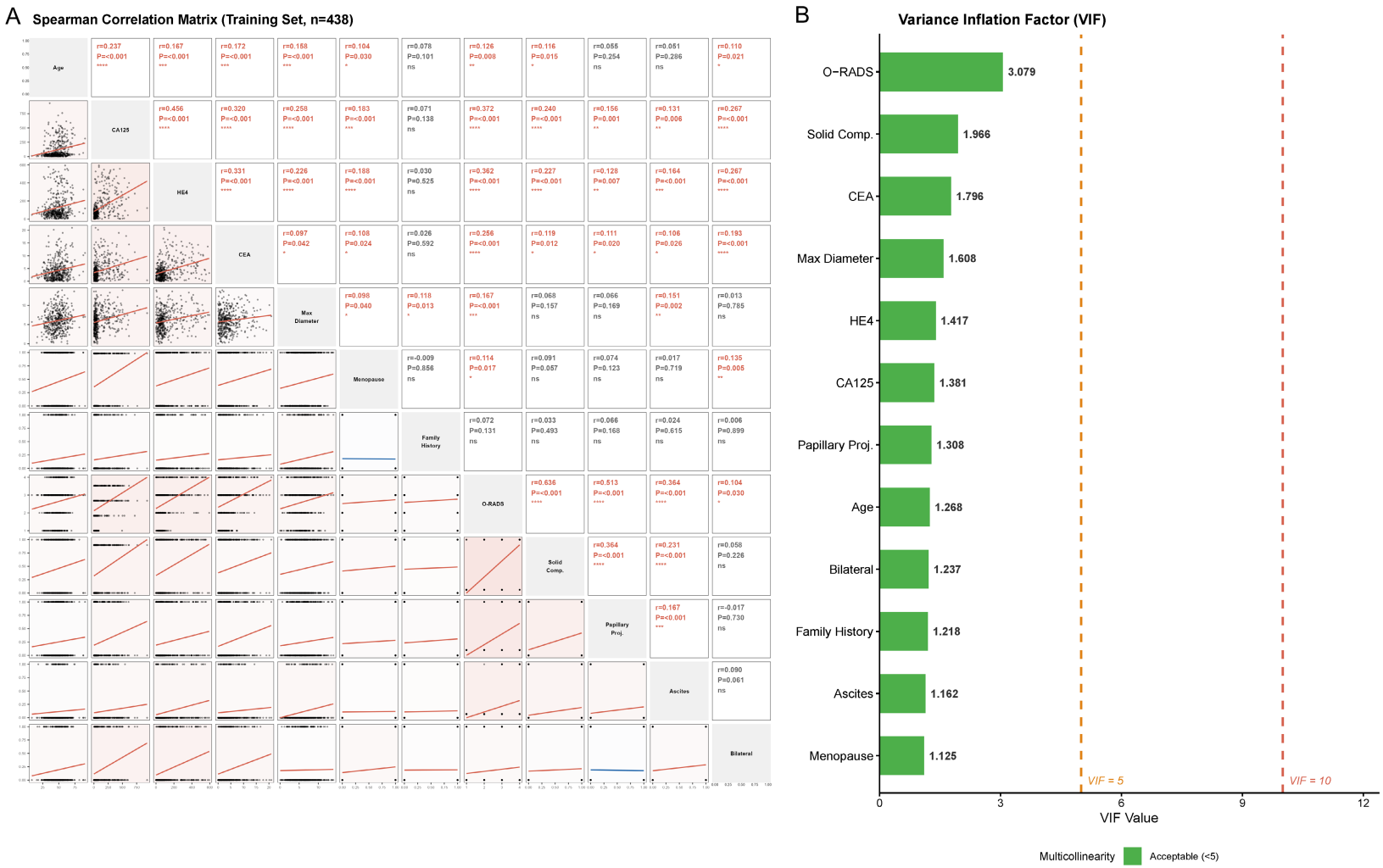

Figure 1. Spearman correlation matrix and VIF analysis of predictor variables in the training set. A. Spearman correlation matrix of 12 predictor variables in the training set (n=438); B. VIF values of various predictor variables. Note: VIF, Variance Inflation Factor; O-RADS, Ovarian-Adnexal Reporting and Data System; CA125, Cancer Antigen 125; HE4, Human Epididymis Protein 4; CEA, Carcinoembryonic Antigen.

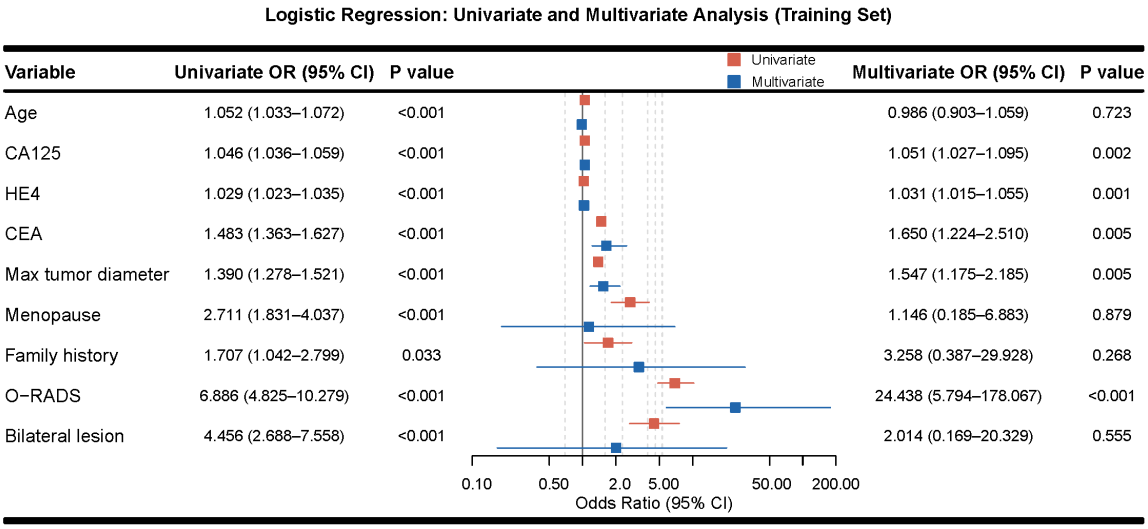

Figure 2. Forest plot of univariate and multivariate Logistic regression analysis of predictor variables of early ovarian cancer in the training set. The plot shows the results of univariate (orange) and multivariate (blue) Logistic regression analysis of 9 predictor variables in the training set (n=438), with the odds ratios (ORs) on the horizontal axis and the error bars representing 95% CIs. Note: OR, Odds Ratio; CI, Confidence Interval; CA125, Cancer Antigen 125; HE4, Human Epididymis Protein 4; CEA, Carcinoembryonic Antigen; O-RADS, Ovarian-Adnexal Reporting and Data System.

expected-to-observed (E/O) ratio of 1.000, a calibration intercept of 0.000, a calibration slope of 1.000, and an H-L test *P* value of 0.126. In the validation set, the corresponding values were a Brier score of 0.0560, an E/O ratio of 1.091, a calibration intercept of -0.532, a calibration slope of 1.099, and an H-L test *P* value of 0.578. The H-L test *P* values in both datasets were greater than 0.05, and the E/O ratios were close to 1, indicating satisfactory agreement between the predicted probabilities and the observed outcomes (**Figure 4C** and **4D**).

The clinical net benefits were evaluated by the DCA. In the training set, the comprehensive model provided continuous positive net benefits within 0.10-0.50 threshold probability. The model also maintained positive net benefits in the validation set in a wide threshold probability range, with the net profit achieved at each threshold highly consistent with the training set. The above substantiates the high practicability of the comprehensive model in clinical decision-making scenarios with different risk preferences (**Figure 4E** and **4F**).

SHAP analysis of the comprehensive model

To deeply explain the decision-making mechanism of the comprehensive model, this study adopted the SHAP approach to systematically interprets it from three aspects: feature importance, individual predictive contribution, and feature interactions.

From the feature importance perspective, the bar chart of SHAP feature importance showed that the contributions of the five predictor variables to the model prediction, from high to low, are as follows: O-RADS classification (average |SHAP|=0.150, attribution proportion: 30.2%), CA125 (average |SHAP|=0.122, attribution proportion: 24.7%), HE4 (average |SHAP|=0.105, attribution proportion: 21.2%), CEA (average |SHAP|=0.073, attribution proportion: 14.7%), and MTD (average |SHAP|=0.045, attribution proportion: 9.2%), establishing O-RADS classification as the predictor variable that contributes the most in the comprehensive model (**Figure 5A**). The SHAP beeswarm plot further demonstrated a strong positive association between the feature values of all five variables and their corresponding SHAP values (Spearman's r >0.93, P<0.001), indicating that higher feature values contributed more strongly to increased predicted probability of EOC (**Figure 5B**).

Concerning individual predictive contribution, the waterfall charts presented in **Figure 6** shows the decomposition of individual SHAP

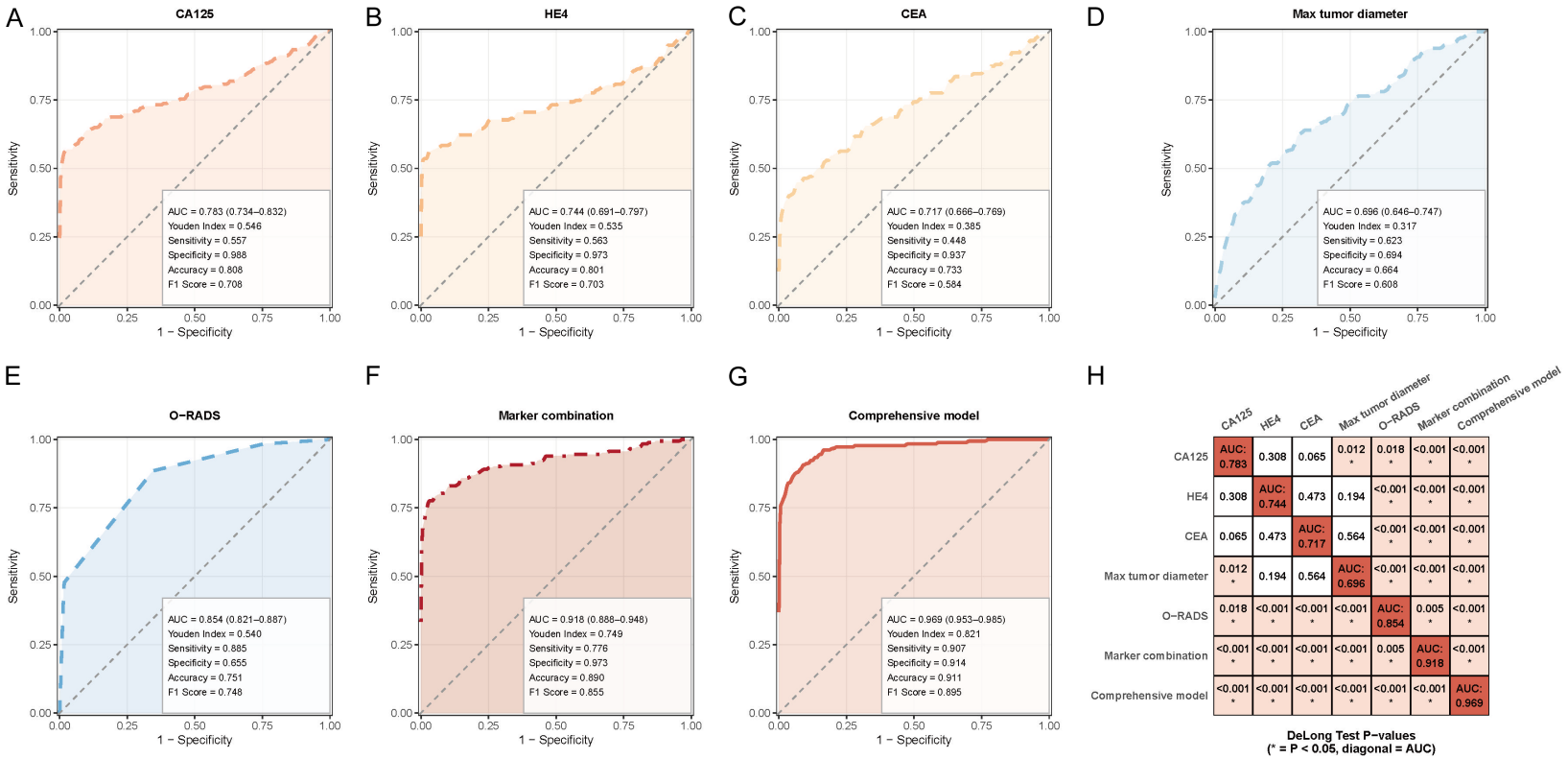

Figure 3. ROC curve analysis and DeLong test of the diagnostic efficacy analysis of single markers and joint models for early ovarian cancer. A-E. ROC curves of CA125, HE4, CEA, maximum tumor diameter, and O-RADS classification as individual markers; F. ROC curve of the (serological) marker combination model; G. ROC curve of the comprehensive model; H. DeLong test *P*-value heat map of the AUCs of indicators and models. Note: ROC, Receiver Operating Characteristic; CA125, Cancer Antigen 125; HE4, Human Epididymis Protein 4; CEA, Carcinoembryonic Antigen; O-RADS, Ovarian-Adnexal Reporting and Data System; AUC, Area Under the Curve; CI, Confidence Interval.

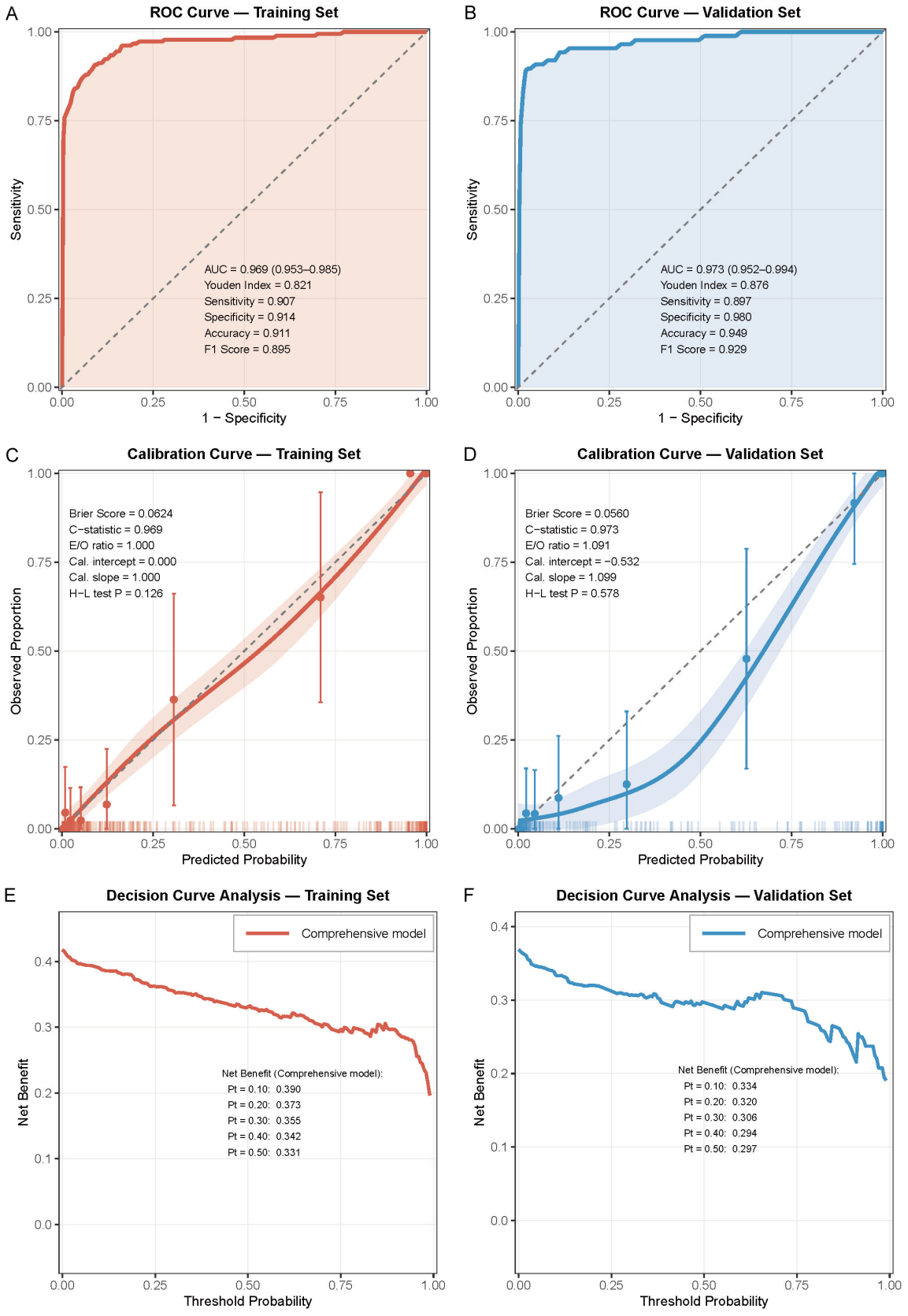

Figure 4. ROC curve, calibration curve, and decision curve analyses of the comprehensive model across the training and validation sets. A. Training set ROC curve; B. Validation set ROC curve; C. Training set calibration; D. Validation

set calibration curve; E. Training set DCA; F. Validation set DCA. Note: ROC, Receiver Operating Characteristic; DCA, Decision Curve Analysis; AUC, Area Under the Curve; CI, Confidence Interval; E/O, Expected/Observed ratio; H-L test, Hosmer-Lemeshow test; C-statistic, Concordance Statistic.

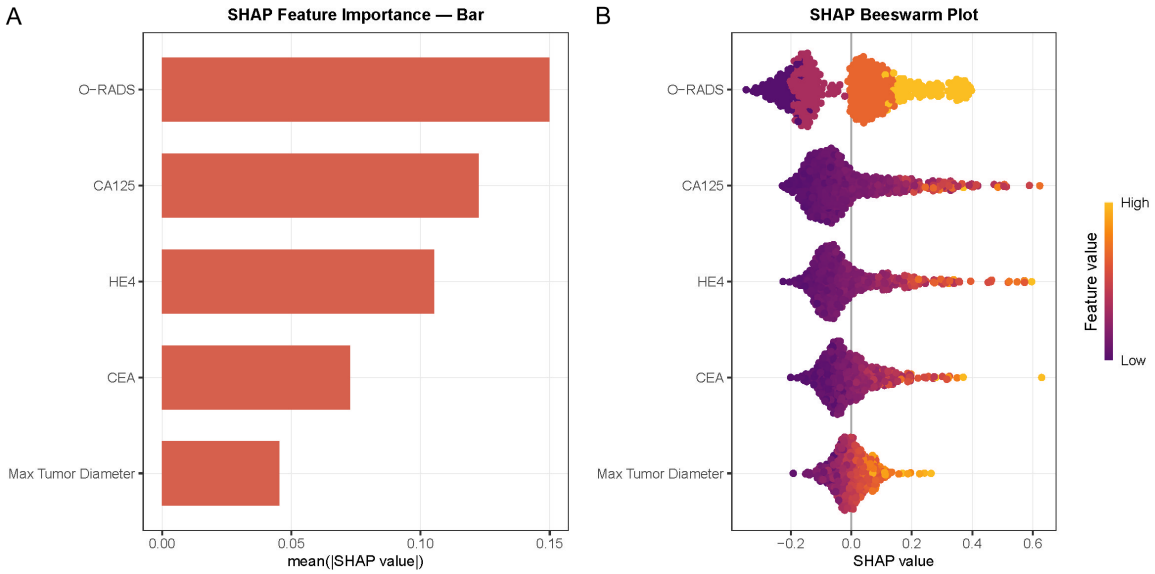

Figure 5. SHAP feature importance bar chart and beeswarm plot of the comprehensive model. A. SHAP feature importance bar chart of various predictor variables, with the average absolute SHAP value shown on the horizontal axis; B. SHAP beeswarm plot, where each point represents a sample, the color indicates the eigenvalue, and the horizontal axis shows the SHAP value. Note: SHAP, Shapley Additive Explanations; O-RADS, Ovarian-Adnexal Reporting and Data System; CA125, Cancer Antigen 125; HE4, Human Epididymis Protein 4; CEA, Carcinoembryonic Antigen.

contribution of 2 malignant and 2 benign representative samples. In malignant samples, all the predictor variables contributed positively, driving the model's prediction probability to shift in a malignant direction. In benign lesions, variables such as O-RADS classification, HE4, and CA125 all showed significant negative contributions, which together pull the prediction probability of the model down to a benign range. The above results show that the model exhibits high variable specificity and clinical interpretability in its predictions for different individuals (**Figure 6A-D**).

In terms of feature interactions, the SHAP dependence plots revealed that the CA125-SHAP value increased as the CA125 level rose, with the positive effect further enhanced at high HE4 levels, suggesting the presence of a synergistic interaction between CA125 and HE4 (**Figure 7A**). A similar pattern was observed for HE4, namely, HE4 level elevations led to increases in HE4-SHAP values, with high CA125 levels contributing to an amplifying effect (**Figure 7B**). The CEA-SHAP value increased with increasing CEA contents, and the positive effect was more prominent when the O-RADS classification was higher (**Figure 7C**). The MTD-SHAP value also rose in parallel with the increase in MTD, and O-RADS classification had a relatively independent influence on it (**Figure 7D**). The SHAP value of O-RADS classification increased in a stepwise manner as the classification level rose, and the positive effect of Category 5 lesions was particularly significant when the HE4 level was high (**Figure 7E**).

Subgroup analysis, FIGO staging performance, and threshold sensitivity analysis of the comprehensive model

To further verify the stability and applicability of the comprehensive model in different clinical subgroups, this study carried out subgroup analysis from three dimensions: age, menopausal status, and O-RADS classification. Meanwhile, the FIGO staging performance and model performance under different diagnostic thresholds were systematically evaluated.

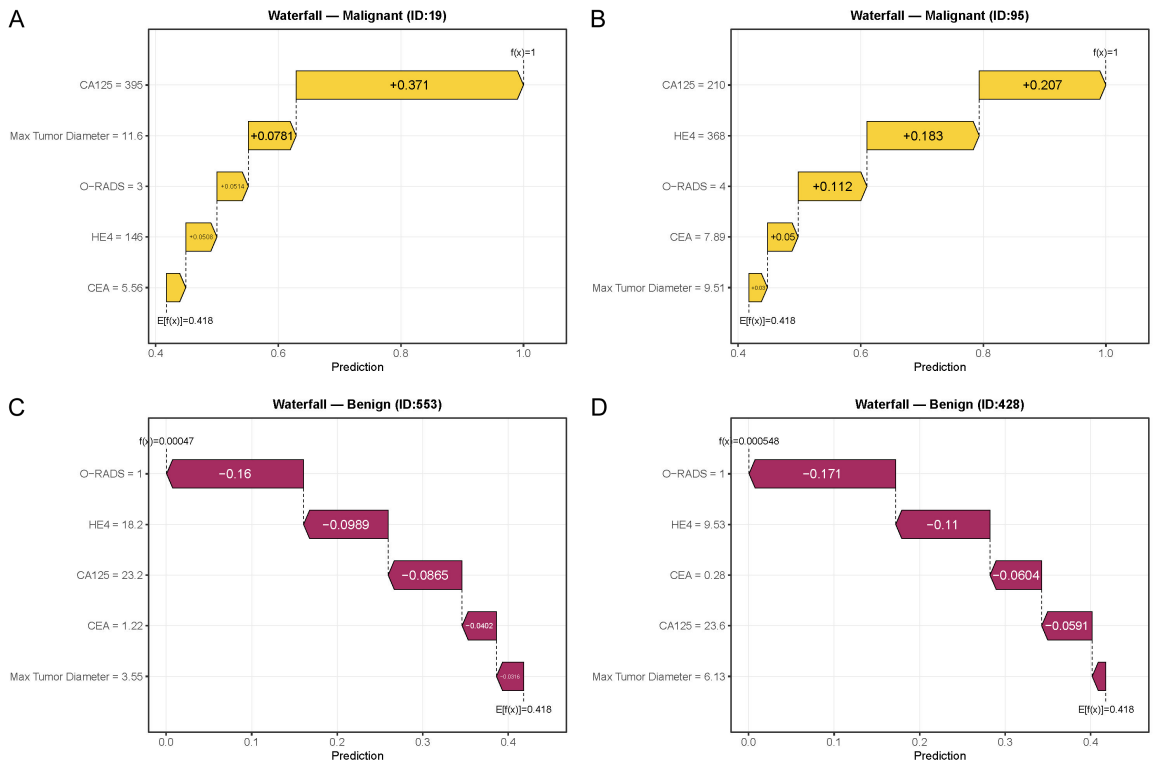

Figure 6. SHAP waterfall chart of representative samples of the comprehensive model. A, B. SHAP waterfall charts of two malignant samples; C, D. SHAP waterfall charts of two benign samples. The contribution of each variable is indicated by positive (yellow) or negative (purple) arrows. Note: SHAP, Shapley Additive Explanations; O-RADS, Ovarian-Adnexal Reporting and Data System; CA125, Cancer Antigen 125; HE4, Human Epididymis Protein 4; CEA, Carcinoembryonic Antigen.

When stratified by age, the model showed an AUC of 0.932 for patients aged under 50 years and 0.944 for those aged 50 or above in the training set; the corresponding AUC for the <50- and ≥50-year-old subgroups in the validation set was 0.926 and 0.931, respectively, with highly overlapped 95% CIs among the subgroups. It suggests that the model has stable diagnostic efficiency across age brackets (**Figure 8A**).

Subgroup analysis based on menopausal status revealed an AUC of 0.946 for the premenopausal subgroup and 0.936 for the postmenopausal subgroup, with the corresponding AUC in the validation set being 0.926 and 0.934, respectively. The consistency in diagnostic efficacy between the subgroups across datasets shows that menopausal status has no significant effect on model performance (**Figure 8B**).

Stratification according to O-RADS classification revealed an AUC of 0.773 and 0.878 for O-RADS category 1-2, and 0.884 and 0.867 for category 3, in the training and validation cohorts, respectively. It suggests that the comprehensive model still holds certain auxiliary differential value in low-grade O-RADS lesions with atypical ultrasonic features. For the O-RADS category 4 subgroup, however, we failed to calculate the effective AUC due to insufficient benign sample size (**Figure 8C**).

Regarding FIGO stage among malignant cases, the median predicted probability was higher for stage I than for stage II in the training set (0.936 vs. 0.873), whereas the median predicted probability was slightly lower for stage I than for stage II in the validation set (0.908 vs. 0.929) (**Figure 8D** and **8E**). Nevertheless, the predicted-probability distributions showed substantial overlap between the two FIGO stages, and the Mann-Whitney U tests showed no statistically significant difference in either the training set (P=0.210) or the validation set (P=0.340). These findings were expected because the model was developed to predict the probability of malignancy rather than FIGO

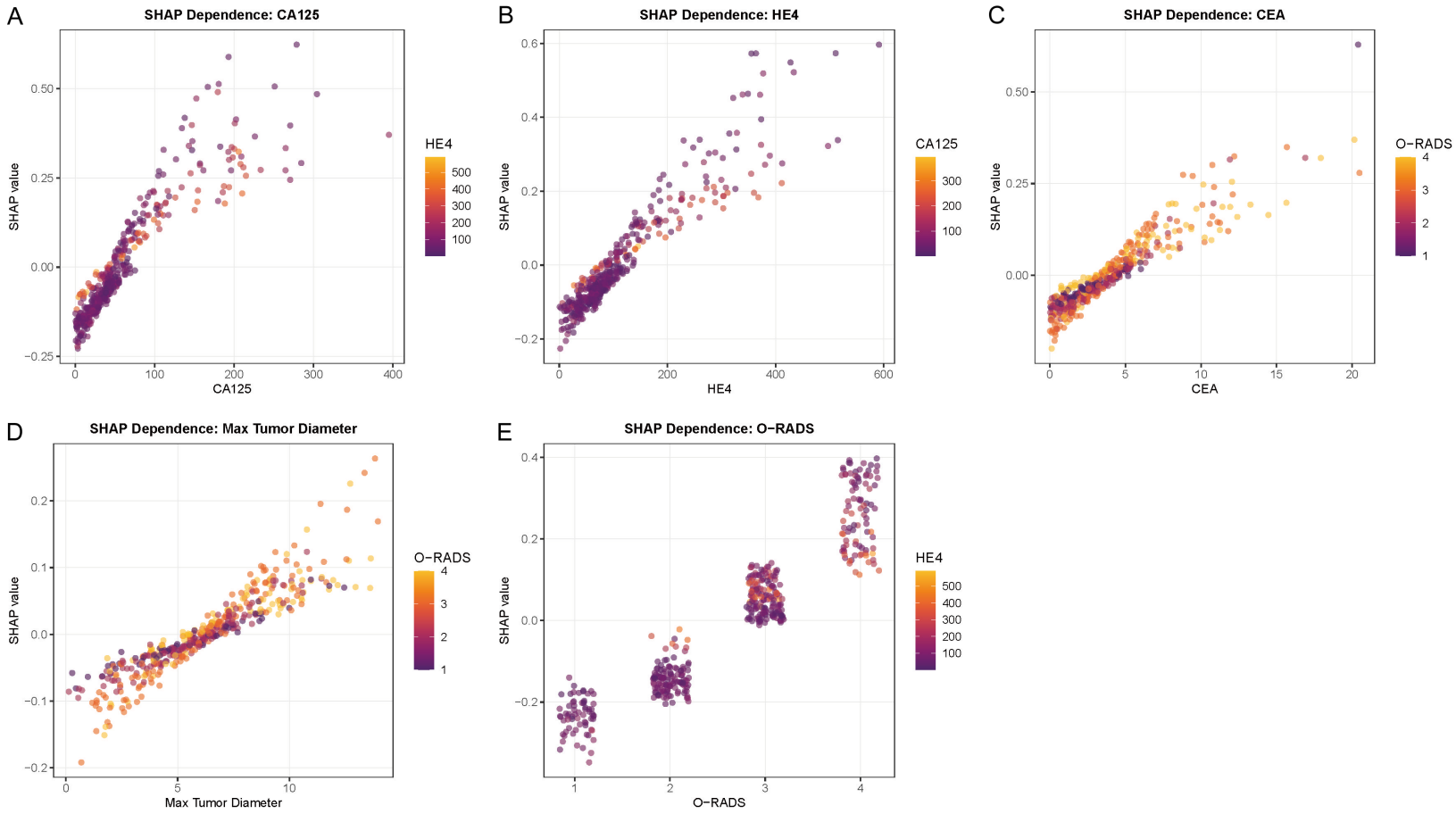

Figure 7. SHAP dependence plots of various predictor variables in the comprehensive model. A-E. SHAP dependence plots of CA125, HE4, CEA, maximum tumor diameter, and O-RADS classification. The horizontal axis indicates original feature values, the vertical axis shows the corresponding SHAP value, and the marker color represents the value of interactive features. Note: SHAP, Shapley Additive Explanations; O-RADS, Ovarian-Adnexal Reporting and Data System; CA125, Cancer Antigen 125; HE4, Human Epididymis Protein 4; CEA, Carcinoembryonic Antigen.

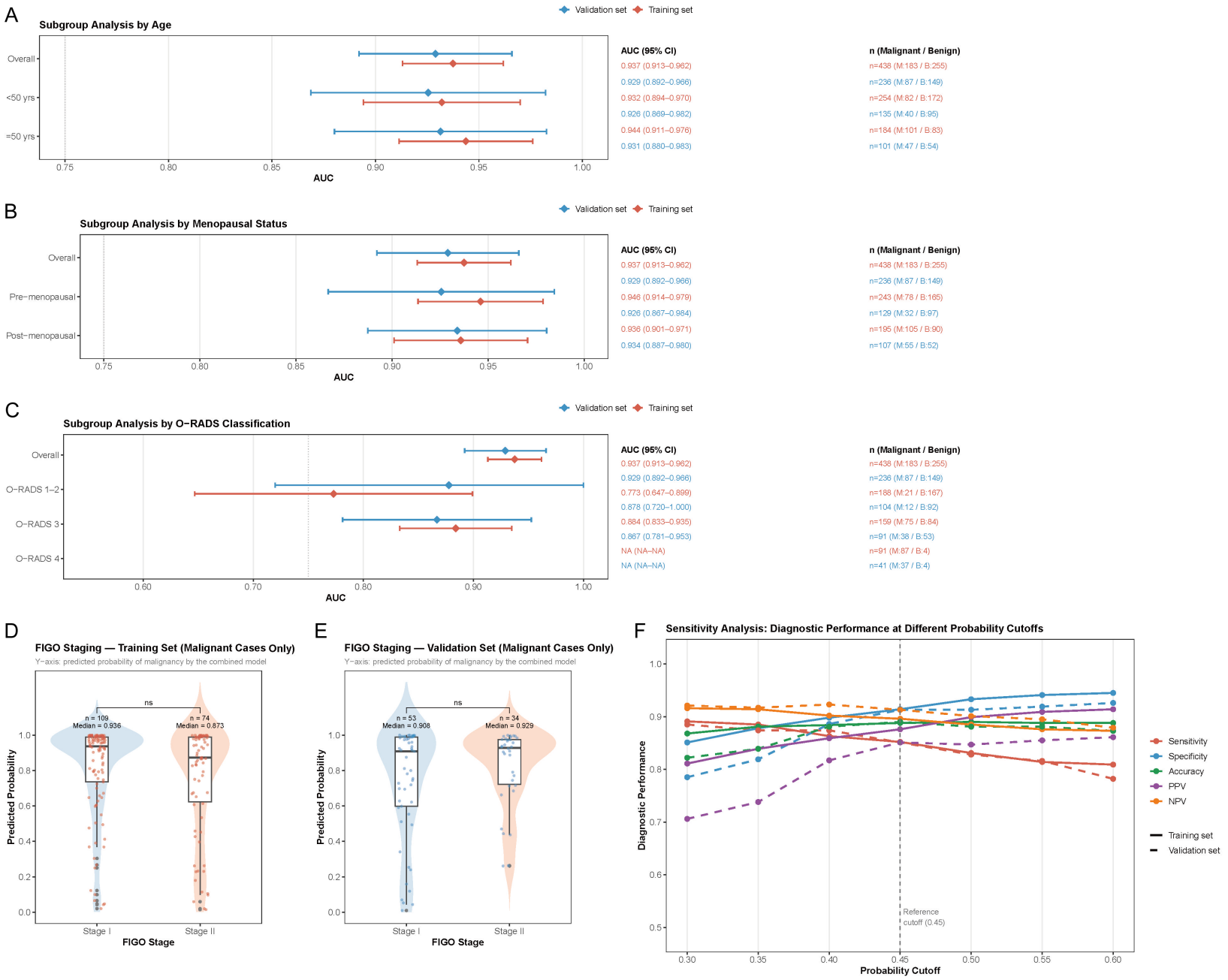
A
Subgroup Analysis by Age
Validation set
Training set
Overall
<50 yrs
=50 yrs
AUC (95% CI)
0.937 (0.913–0.962)
0.929 (0.892–0.966)
0.932 (0.894–0.970)
0.926 (0.869–0.982)
0.944 (0.911–0.976)
0.931 (0.880–0.983)
n (Malignant / Benign)
n=438 (M:183 / B:255)
n=236 (M:87 / B:149)
n=254 (M:82 / B:172)
n=135 (M:40 / B:95)
n=184 (M:101 / B:83)
n=101 (M:47 / B:54)
AUC
B
Subgroup Analysis by Menopausal Status
Overall
Pre−menopausal
Post−menopausal
0.937 (0.913–0.962)
0.929 (0.892–0.966)
0.946 (0.914–0.979)
0.926 (0.867–0.984)
0.936 (0.901–0.971)
0.934 (0.887–0.980)
n=438 (M:183 / B:255)
n=236 (M:87 / B:149)
n=243 (M:78 / B:165)
n=129 (M:32 / B:97)
n=195 (M:105 / B:90)
n=107 (M:55 / B:52)
C
Subgroup Analysis by O−RADS Classification
Overall
O−RADS 1–2
O−RADS 3
O−RADS 4
0.937 (0.913–0.962)
0.929 (0.892–0.966)
0.773 (0.647–0.899)
0.878 (0.720–1.000)
0.884 (0.833–0.935)
0.867 (0.781–0.953)
NA (NA–NA)
NA (NA–NA)
n=438 (M:183 / B:255)
n=236 (M:87 / B:149)
n=188 (M:21 / B:167)
n=104 (M:12 / B:92)
n=159 (M:75 / B:84)
n=91 (M:38 / B:53)
n=91 (M:87 / B:4)
n=41 (M:37 / B:4)
D
FIGO Staging — Training Set (Malignant Cases Only)
Y−axis: predicted probability of malignancy by the combined model
ns
n = 109
Median = 0.936
n = 74
Median = 0.873
Predicted Probability
Stage I
Stage II
FIGO Stage
E
FIGO Staging — Validation Set (Malignant Cases Only)
Y−axis: predicted probability of malignancy by the combined model
ns
n = 53
Median = 0.908
n = 34
Median = 0.929
Predicted Probability
Stage I
Stage II
FIGO Stage
F
Sensitivity Analysis: Diagnostic Performance at Different Probability Cutoffs
Diagnostic Performance
Reference
cutoff (0.45)
Probability Cutoff
Sensitivity
Specificity
Accuracy
PPV
NPV
Training set
Validation set

Figure 8. Subgroup analysis, FIGO stage prediction probability distribution, and diagnostic threshold sensitivity analysis of the comprehensive model. A. AUC forest plot of different age subgroups; B. AUC forest plot of subgroups grouped by menopausal status; C. AUC forest plot subgrouped by O-RADS classification; D. Violin plot of model prediction probability distribution of different FIGO stages in malignant cases in the training set; E. Violin plot of model prediction probability distribution of different FIGO stages in malignant cases in the validation set; F. Trend chart of the diagnostic performance of indicators under different diagnostic probability cutoffs in the training and validation sets. Note: FIGO, International Federation of Gynecology and Obstetrics; AUC, Area Under the Curve; O-RADS, Ovarian-Adnexal Reporting and Data System; CI, Confidence Interval; PPV, Positive Predictive Value; NPV, Negative Predictive Value.

stage. Therefore, the model-predicted probability should not be interpreted as an indicator of disease stage.

In the threshold sensitivity test, a gradual decline in sensitivity (0.891→0.809), an increase in specificity (0.851→0.945), a rise in positive predictive value (0.811→0.914), and a decline in negative predictive value (0.916→0.873) were observed when the threshold probability gradually increased from 0.30 to 0.60. The validation set also presented a similar monotonous trend, with the corresponding indicators being highly consistent across the training and validation sets under each threshold. The above supports the ability of the comprehensive model to maintain strong diagnostic performance in a wide threshold range. Clinical users can flexibly weigh the sensitivity and specificity according to actual needs (**Figure 8F**).

In addition, internal verification using 1000-iteration Bootstrapping showed that the adjusted AUC in the training set was 0.942, with a low optimism bias (-0.004); 5-fold cross-validation yielded an average AUC of 0.927 and an inter-fold standard deviation of 0.027, further validating the model's robust internal stability and generalization.

Sensitivity analysis of O-RADS and its component ultrasound feature

To assess whether intratumoral solid components and papillary projections provided additional predictive value beyond the O-RADS classification, a sensitivity analysis was performed by incorporating these variables into the main model. The extended model showed no significant improvement in discrimination compared with the main model. In the training set, the AUC was 0.975 for both models (P=0.835), while in the validation set, the AUCs were 0.973 for the main model and 0.972 for the extended model (P=0.321). In addition, the extended model demonstrated slightly higher AIC and BIC values than the main model, indicating no improvement in overall model fit despite increased complexity. In multivariable analysis, neither intratumoral solid components (P=0.834) nor papillary projections (P=0.104) remained independently significant. Furthermore, an alternative model excluding O-RADS but including these two ultrasound features showed lower predictive performance (AUC: 0.959 in both training and validation sets), suggesting that O-RADS captures imaging information more effectively than its individual component features.

Supplementary results: threshold-specific performance

To facilitate clinical decision-making, we evaluated the comprehensive model's performance at five probability thresholds (0.10, 0.20, 0.30, 0.40, and 0.50) in both the training (n=438) and validation (n=236) sets. Table S1 presents the sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and Youden index for each threshold. A threshold of 0.30 provided a balanced trade-off, with a Youden index of 0.818 (training) and 0.814 (validation), sensitivity of 0.922-0.944, specificity of 0.870-0.895, PPV of 0.817-0.860, and NPV of 0.943-0.962. Lower thresholds (0.10-0.20) maximized sensitivity (≥0.93) at the cost of lower specificity, suitable for screening or triage. Higher thresholds (0.40-0.50) prioritized specificity (≥0.89) and PPV (≥0.84), appropriate for confirmatory diagnosis or surgical referral. These data support clinical context-based flexible threshold selection.

Discussion

By systematically integrating serological markers (CA125, HE4, CEA), O-RADS US risk stratification, and tumor structural characteristics

collected from 674 adnexal mass-affected patients, this study built a comprehensive EOC predictive model that was subsequently verified to have excellent diagnostic efficiency (AUC=0.969 during training and 0.973 upon validation), far outperforming single indices and the marker combination model. Its calibration, clinical net gains, and good stability in internal validation were also confirmed. Hence, it is challenging for single-marker diagnostic tools to accurately identify EOC (a relatively hidden clinical stage); in contrast, risk assessment performance can be markedly enhanced through multi-dimensional information fusion. Due to the absent atypical manifestations and classic imaging features, EOC can be hardly detected by a single indicator that often cannot achieve good sensitivity and specificity simultaneously [16, 18]. This may underlie the gradual shift to the research on multimodal joint strategies. Utilizing real-world data, this study substantiates this notion and supports the practicability of "image+serology+structural features" joint modeling in EOC diagnosis. From a biological standpoint, the model's superior performance reflects biological complementarity. Serological markers capture the tumor's functional behavior, including proliferation, secretion, and metabolic alterations, whereas O-RADS ultrasound stratification reflects its morphological and architectural characteristics, such as internal complexity, solid components, and vascularity [13]. These two data dimensions are mechanistically independent yet clinically synergistic, which explains why their fusion markedly outperforms any single-modality approach.

Assessing single index performance, O-RADS classification (AUC=0.854) significantly surpassed CA125, HE4, CEA, and MTD. The O-RADS system proposed by Andreotti et al. [9] realizes the standardized risk stratification of adnexal masses by integrating characteristics such as mass morphology, solid composition, papillary projections, and blood flow signals. Strachowski et al. [10] further optimized the classification criteria in the updated version and improved the specificity. The AUC of O-RADS has been indicated to be mostly between 0.80 and 0.90 in the general population; in the context of EOC or intermediate risk stratification (especially in categories 3-4), however, the discrimination ability of O-RADS fluctuates [17, 19, 20]. Echoing these findings, the present study found that although the proportion of patients with O-RADS categories 4-5 was significantly increased in the malignant group, there was an obvious overlap in categories 3-4, suggesting the limitation of imaging stratification in gray zone lesions. Additionally, this research identified a greater prevalence of intratumoral solid component, papillary projections, and seroperitoneum in the malignant group, characteristics highly correlated with the O-RADS score, which further explains their superiority among individual markers. Yet, relying solely on O-RADS cannot cover all early cases, especially in lesions with atypical imaging manifestations or inapparent structural complexity that limit its diagnostic ability.

When analyzing the performance of serological markers, CA125, HE4, and CEA were all found to be significantly up-regulated in the malignant group, and were established as independent predictors by multivariate Logistic regression. HE4 and CA125, in particular, contributed a greater proportion, which largely agrees with previous research conclusions. CA125 (MUC16) is a high-molecular-weight glycoprotein shed from the surface of ovarian epithelial cells; its elevation reflects tumor burden and peritoneal irritation, but it is readily elevated in benign gynecologic conditions such as endometriosis and pelvic inflammatory disease, limiting its specificity [6, 21]. In contrast, HE4, encoded by the WFDC2 gene and belonging to the whey acidic protein family, exhibits a more restricted expression pattern and is less influenced by benign gynecologic conditions, thereby conferring the specificity that CA125 alone lacks [21-23]. Singh et al. [24] meta-analyzed tumor markers, similarly noting HE4's high stability in malignant lesion differentiation. In this study, the marker combination model (AUC=0.918) was advantageous over individual serological indices and O-RADS in diagnostic efficacy, which echoes previous literature [11, 12] and supports the functional complementation between serological information and imaging stratification. Importantly, CEA, though rarely used in OC diagnosis, retained independent significance during multivariate modeling. This likely reflects its elevated expression in mucinous ovarian carcinomas, which constituted 19.6% of the EOC cases in this cohort, highlighting the value of retaining CEA for capturing

subtype-specific information that CA125 and HE4 may miss [7]. MTD maintained its significance in the comprehensive model despite its weak diagnostic performance as a standalone; it suggests that MTD is of certain value in reflecting tumor burden and growth. Tumor size is often used as an auxiliary variable in risk assessment in multivariate models; however, its predictive value is maximized when interpreted together with image complexity and serological abnormalities, as larger tumors with higher O-RADS scores and elevated markers likely reflect biologically more aggressive lesions [25, 26].

During model building and performance evaluation, this study focused on model discrimination, calibration, and practicability. The model's robust generalization was demonstrated by a highly consistent AUC and C-statistic >0.96 across datasets. The consistency between model-predicted probabilities and actual risks (low Brier scores and comparable H-L test *P*-values) was also observed. As indicated by literature, sole reliance on AUC reporting while ignoring calibration performance makes the models infeasible for clinical promotion [27-29]. This study conducted a multi-dimensional evaluation to further enhance model reliability: DCA evidenced the model's stable and positive net profits in the 0.10-0.50 threshold range, supporting its practicability in different risk decision scenarios. Through DCA and threshold-specific performance evaluation, 0.30 was used as the model's initial threshold for general preoperative EOC risk assessment, achieving a Youden index of 0.814 in the validation set (0.944 sensitivity; 0.870 specificity). To minimize the situation of missed diagnosis of malignant lesions during tumor screening and clinical triage, a lower judgment threshold of 0.10 to 0.20 can be selected, thereby increasing the sensitivity to 0.93 or above. For scenarios where the risk of surgery exploration or other confirmatory methods is high or the cost is high, a high threshold of 0.40 to 0.50 should be selected, thereby optimizing specificity and PPV to ≥0.89 and ≥0.84, respectively. In a word, shared decision-making by taking into account patient preferences, local disease prevalence, and healthcare resources determined the final threshold. Besides, a trend conforming to the general laws of diagnostic models, namely gradually decreased sensitivity and NPV, as well as increased specificity and PPV with the threshold increasing, was observed, showing favorable clinical adjustability of the model we proposed. The stable outputs in Bootstrap resampling and 5-fold cross-validation further proved the model's good internal consistency.

SHAP provides an important supplement. The results confirmed O-RADS as the top-performing contributor, followed by CA125, HE4, CEA, and MTD, with SHAP values significantly and positively linked to their original values. It supports the high agreement between the model decision direction and clinical cognition, namely, a greaer imaging grade is associated with higher serological marker levels, larger tumors, and greater malignant possibilities. Geysels et al. [15] reported that interpretable models help reveal true variable-risk associations. Furthermore, this study found obvious variable-variable interactions (e.g., potentiated contribution of CA125 and amplified prediction effects of O-RADS category 5 at high HE4 expression, enhanced role played by CEA in high O-RADS classification). It implies a possible imaging feature-serological marker synergy, rather than a simple linear superposition relationship. Such synergies likely reflect the biological reality that concurrent upregulation of multiple serological markers, combined with high imaging complexity, collectively signifies a more aggressive tumor phenotype - a pattern that single-modality assessment alone cannot fully capture [15]. When assessing individual markers, the SHAP waterfall chart showed that the dominant variables influencing the predicted results for different patients varied. The results indicate that the comprehensive model we built possesses high interpretability that supports its clinical utility.

The model's stability and applicability were further validated by subgroup analyses based on age, menopausal status, and O-RADS classification. With constantly high AUC values obtained and highly overlapped CIs between corresponding subgroups, the model's consistent performance across different populations was demonstrated. Literature [30, 31] has reported O-RADS's relatively limited diagnostic capability in low-risk populations. The comprehensive model built in this study, however, showed good diagnostic efficiency in O-RADS categories 1-3, indicating that it has important supplementary value in "atypical cases". In

addition, this study found that the model's prediction probability for FIGO stage I was slightly higher than that for stage II, suggesting that it has a certain sensitivity to earlier-stage lesions. But given the overlap between them, the model is better suited for identifying malignant risks rather than staging judgments.

Despite the highly reliable results obtained, this study still has several limitations. First, this was a single-center retrospective study, which may have introduced selection bias. In addition, the validation cohort was derived from the same population as the training set; therefore, the present validation should be considered internal rather than true external validation. Future multi-center prospective studies with independent datasets are needed to further assess the generalizability of the model. Second, only cases with complete clinical, serological, and imaging data were included in the final analysis, and cases with missing key variables were excluded, which may also have introduced potential selection bias. Third, the limited samples in certain pathological subtypes can possibly impact model performance in rare tumor subtypes. Fourth, additional molecular markers, dynamic monitoring indices, and deep US image features were not analyzed. Future studies may consider integrating multi-omics data and high-dimensional imaging information for predictive performance enhancement [31]. Fifth, this study employed logistic regression as the modeling framework, which assumes linear relationships between predictors and the log-odds of the outcome. While SHAP analysis was used for post-hoc interpretability, the model may not capture higher-order non-linear interactions among predictors that more flexible machine learning approaches could identify. Future studies may consider comparing logistic regression with such models in this clinical setting.

Conclusion

The comprehensive model shows excellent discrimination, good calibration, and stable clinical practicability in diagnosing EOC. O-RADS stands as the best-performing predictor; serological markers and tumor structure characteristics offer significant incremental information. Through SHAP, the mechanism underlying each variable's contribution to model predictions is revealed, which enhances model explainability, favoring its clinical use. The model, being a promising tool to assist risk stratification and decision-making, sheds light on multimodal intelligent diagnosis.

Acknowledgements

Kashgar Region Science and Technology Plan Project Contract, Project Number: KS2024030.

Disclosure of conflict of interest

None.

Address correspondence to: Fucheng Ma, Department of Ultrasound Diagnosis, The 3rd Affiliated Teaching Hospital of Xinjiang Medical University (Affiliated Cancer Hospital), No. 789 Suzhou East Street, New Urban District, Urumqi 830010, Xinjiang Uygur Autonomous Region, China. Tel: +86-0991-7968027; E-mail: mafuchengmfc@yeah.net

References

[1] Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I and Jemal A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2024; 74: 229-263.

[2] Siegel RL, Kratzer TB, Giaquinto AN, Sung H and Jemal A. Cancer statistics, 2025. CA Cancer J Clin 2025; 75: 10-45.

[3] Hong MK and Ding DC. Early diagnosis of ovarian cancer: a comprehensive review of the advances, challenges, and future directions. Diagnostics (Basel) 2025; 15: 406.

[4] Lheureux S, Gourley C, Vergote I and Oza AM. Epithelial ovarian cancer. Lancet 2019; 393: 1240-1253.

[5] Momenimovahed Z, Mazidimoradi A, Allahqoli L and Salehiniya H. The role of CA-125 in the management of ovarian cancer: a systematic review. Cancer Rep (Hoboken) 2025; 8: e70142.

[6] Moore RG, MacLaughlan S and Bast RC Jr. Current state of biomarker development for clinical application in epithelial ovarian cancer. Gynecol Oncol 2010; 116: 240-245.

[7] Davenport CF, Rai N, Sharma P, Deeks J, Berhane S, Mallett S, Saha P, Solanki R, Bayliss S, Snell K and Sundar S. Diagnostic models combining clinical information, ultrasound and biochemical markers for ovarian cancer: cochrane systematic review and meta-analysis. Cancers (Basel) 2022; 14: 3621.

[8] Fathi A, Heidari M, Rasouli J and Ghasemnejad-Berenji H. Diagnostic value of the Risk of Ovarian Malignancy Algorithm (ROMA) index in

the detection of ovarian cancer in postmenopausal women: a systematic review and meta-analysis. BMC Womens Health 2025; 25: 280.
[9] Andreotti RF, Timmerman D, Strachowski LM, Froyman W, Benacerraf BR, Bennett GL, Bourne T, Brown DL, Coleman BG, Frates MC, Goldstein SR, Hamper UM, Horrow MM, Hernanz-Schulman M, Reinhold C, Rose SL, Whitcomb BP, Wolfman WL and Glanc P. O-RADS US risk stratification and management system: a consensus guideline from the ACR ovarian-adnexal reporting and data system committee. Radiology 2020; 294: 168-185.
[10] Strachowski LM, Jha P, Phillips CH, Blanchette Porter MM, Froyman W, Glanc P, Guo Y, Patel MD, Reinhold C, Suh-Burgmann EJ, Timmerman D and Andreotti RF. O-RADS US v2022: an update from the American college of radiology's ovarian-adnexal reporting and data system US committee. Radiology 2023; 308: e230685.
[11] Wang R, Li X, Li S, Fang S, Zhao C, Yang H and Yang Z. Clinical value of O-RADS combined with serum CA125 and HE4 for the diagnosis of ovarian tumours. Acta Radiol 2023; 64: 821-828.
[12] Pan RK, Zhang SQ, Zhang XY, Xu T, Cui XW, Li R, Yu M and Zhang B. Clinical value of ACR O-RADS combined with CA125 in the risk stratification of adnexal masses. Front Oncol 2024; 14: 1369900.
[13] Xie W, Wang Y, Du Z, Chen Y, Ke X, Wu T, Wang Z and Tang L. A nomogram combining clinical features, O-RADS US, and radiomics based on ultrasound imaging for diagnosing ovarian cancer. Sci Rep 2025; 15: 19279.
[14] Wang X, Lv X, Wang J, Zou L, Chen Z, Zhao R, Zhao L, Zhao M, Zhang X, Zhang B, Zhang J, Zhu Y, Shi X, Gao Y, Liu M, Ai L, Wang L, Liu X and Yang H. A multimodal uncertainty-aware AI system optimizes ovarian cancer risk assessment workflow. NPJ Digit Med 2025; 8: 614.
[15] Geysels A, Garofalo G, Timmerman S, Ceusters J, Buonomo F, Fischerová D, Testa AC, Moro F, Sladkevicius P, Jokubkiene L, Van Holsbeke C, Kudla M, Czekierdowski A, Epstein E, Groszmann Y, Blaschko M, Bourne T, De Moor B, Valentin L, Van Calster B, Timmerman D and Froyman W. ADNEX-AI: automated extraction of ultrasound predictors for interpretable ovarian cancer risk stratification. NPJ Precis Oncol 2025; 10: 18.
[16] Pelayo M, Sancho-Sauco J, Sanchez-Zurdo J, Abarca-Martinez L, Borrero-Gonzalez C, Sainz-Bueno JA, Alcazar JL and Pelayo-Delgado I. Ultrasound features and ultrasound scores in the differentiation between benign and malignant adnexal masses. Diagnostics (Basel) 2023; 13: 2152.
[17] Jha P, Gupta A, Baran TM, Maturen KE, Patel-Lippmann K, Zafar HM, Kamaya A, Antil N, Barroilhet L and Sadowski EA. Diagnostic performance of the ovarian-adnexal reporting and data system (O-RADS) ultrasound risk score in women in the United States. JAMA Netw Open 2022; 5: e2216370.
[18] Law RA, Giles BM, Culp-Hill R, Radnaa E, Goldberg M, Nichols CM, Wong M, Hansen C, Hill C, Eurich K, Prendergast E, Behbakht K, Bitler BG, Jeter A, Fa VS, White JR, Elias K and McElhinny A. Clinical evaluation of a multiomic diagnostic model for early-stage ovarian cancer detection. Diagnostics (Basel) 2025; 15: 2225.
[19] Timmerman S, Valentin L, Ceusters J, Testa AC, Landolfo C, Sladkevicius P, Van Holsbeke C, Domali E, Fruscio R, Epstein E, Franchi D, Kudla MJ, Chiappa V, Alcazar JL, Leone FPG, Buonomo F, Coccia ME, Guerriero S, Deo N, Jokubkiene L, Kaijser J, Scambia G, Andreotti R, Timmerman D, Bourne T, Van Calster B and Froyman W. External validation of the ovarian-adnexal reporting and data system (O-RADS) lexicon and the international ovarian tumor analysis 2-step strategy to stratify ovarian tumors into O-RADS risk groups. JAMA Oncol 2023; 9: 225-233.
[20] Cao L, Wei M, Liu Y, Fu J, Zhang H, Huang J, Pei X and Zhou J. Validation of American college of radiology ovarian-adnexal reporting and data system ultrasound (O-RADS US): analysis on 1054 adnexal masses. Gynecol Oncol 2021; 162: 107-112.
[21] Dochez V, Caillon H, Vaucel E, Dimet J, Winer N and Ducarme G. Biomarkers and algorithms for diagnosis of ovarian cancer: CA125, HE4, RMI and ROMA, a review. J Ovarian Res 2019; 12: 28.
[22] Zhang XY, Hong LL and Ling ZQ. MUC16/CA125 in cancer: new advances. Clin Chim Acta 2025; 565: 119981.
[23] Promchit K and Oranratanaphan S. Human epididymis protein 4 (HE4) and cancer antigen 125 (CA125) for prediction of optimal primary surgery in non-mucinous epithelial ovarian cancer. Asian Pac J Cancer Prev 2024; 25: 281-286.
[24] Singh AK, Yadav P, Shukla M, Samaiya S, Singh S and Sorout S. Comparative meta-analysis of carbohydrate antigen 125 (CA125), human epididymis protein 4 (HE4), and diagnostic indices (Risk of Malignancy Index (RMI) and risk of ovarian malignancy algorithm (ROMA)) for pre-operative detection of ovarian carcinoma. Cureus 2025; 17: e82415.
[25] Hu D, Ma D, Zhang ZJ, Zhang Y, Huang K and Li X. Prognosis comparison between small cell carcinoma of ovary and high-grade serous

ovarian cancer: a retrospective observational cohort study. Front Endocrinol (Lausanne) 2023; 14: 1103429.
[26] Feng Y. An integrated machine learning-based model for joint diagnosis of ovarian cancer with multiple test indicators. J Ovarian Res 2024; 17: 45.
[27] Noei Teymoordash S, Zendehdel H, Norouzi AR and Kashian M. Diagnostic accuracy of artificial intelligence algorithms to predict remove all macroscopic disease and survival rate after complete surgical cytoreduction in patients with ovarian cancer: a systematic review and meta-analysis. BMC Surg 2025; 25: 27.
[28] Binuya MAE, Engelhardt EG, Schats W, Schmidt MK and Steyerberg EW. Methodological guidance for the evaluation and updating of clinical prediction models: a systematic review. BMC Med Res Methodol 2022; 22: 316.
[29] van Geloven N, Giardiello D, Bonneville EF, Teece L, Ramspek CL, van Smeden M, Snell KIE, van Calster B, Pohar-Perme M, Riley RD, Putter H and Steyerberg E. Validation of prediction models in the presence of competing risks: a guide through modern methods. BMJ 2022; 377: e069249.
[30] Teng F, Xie H, Wei H, Che D, Wang H, Wu C, He X and Dong X. Diagnostic value of the ovarian adnexal reporting and data system ultrasound in ovarian masses: a 2-center study. Br J Radiol 2025; 98: 448-457.
[31] Almalki YE, Basha MAA, Nada MG, Metwally MI, Libda YI, Ebaid NY, Zaitoun MMA, Mahmoud NEM, Elsheikh AM, Radwan MHSS, Amin MI, Mohamed EM, Tantawy EF, Saber S, Mosallam W, Abdalla HM, Farag MAEM, Dawoud TM, Khater HM, Eldib DB, Altohamy JI, Abouelkheir RT, El Gendy WM, Alduraibi SK, Alshahrani MS, Ibrahim SA, Radwan AM, Obaya AA, Basha AMA and El-Maghraby AM. Ovarian-Adnexal Imaging-Reporting and Data System (O-RADS) ultrasound version 2019: a prospective validation and comparison to updated version (v2022) in pathologically confirmed adnexal masses. Eur Radiol 2025; 35: 3080-3095.

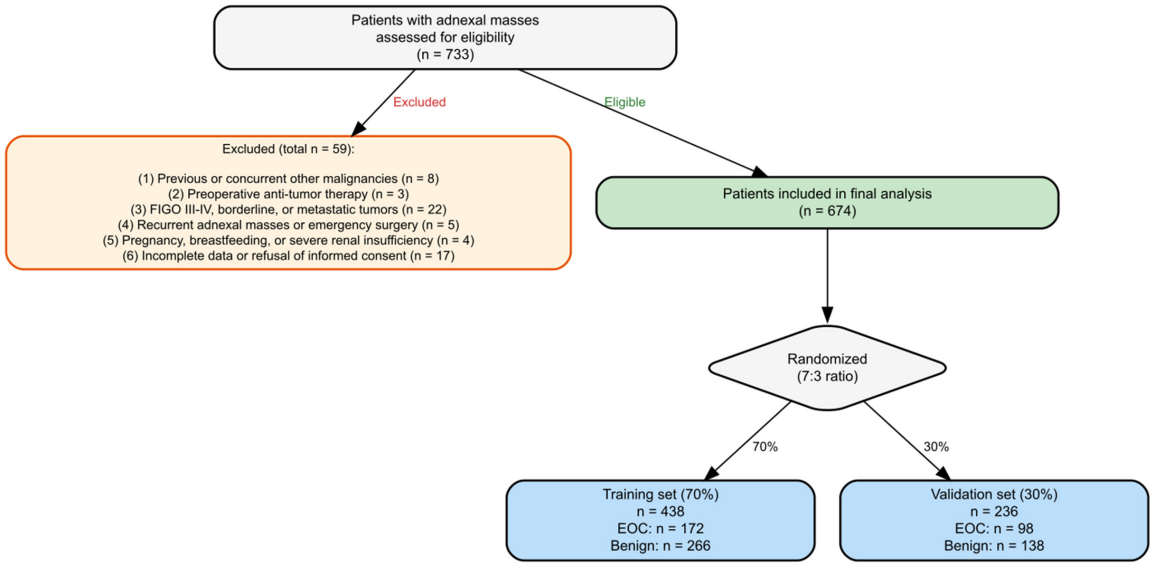

Figure S1. Research flowchart.

Table S1. Model performance at selected probability thresholds for early ovarian cancer prediction across datasets

Threshold	Dataset	Sensitivity	Specificity	PPV	NPV	Youden index
0.1	Training	0.967	0.748	0.728	0.97	0.715
0.1	Validation	0.967	0.788	0.737	0.975	0.754
0.2	Training	0.933	0.849	0.812	0.948	0.782
0.2	Validation	0.956	0.829	0.775	0.968	0.784
0.3	Training	0.922	0.895	0.86	0.943	0.818
0.3	Validation	0.944	0.87	0.817	0.962	0.814
0.4	Training	0.911	0.942	0.916	0.938	0.853
0.4	Validation	0.933	0.89	0.84	0.956	0.824
0.5	Training	0.9	0.957	0.936	0.932	0.857
0.5	Validation	0.911	0.938	0.901	0.945	0.849

Note: PPV, positive predictive value; NPV, negative predictive value. Threshold values represent predicted probability of early ovarian cancer. Training set: n=438 (180 malignant, 258 benign); Validation set: n=236 (90 malignant, 146 benign).