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.

Machine Learning Classification of Benign and Malignant Ovarian Tumours Using Clinical, Biochemical and Tumour-Marker Data

Subarna Akter Liza
Daffodil International University
Sadman Sadik Khan
sadman15-13696@diu.edu.bd

Daffodil International University
Nusrat Khan
Daffodil International University
Aminul Haque
Daffodil International University

Article

Keywords:

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11484670/v1

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

Additional Declarations: No competing interests reported.

Machine Learning Classification of Benign and Malignant Ovarian Tumours Using Clinical, Biochemical and Tumour-Marker Data

Subarna Akter Liza[1], Sadman Sadik Khan[1]*, Nusrat Khan[1], Aminul Haque[1]

[1] Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh

*Corresponding author: Sadman Sadik Khan, Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh. Email: sadman15-13696@diu.edu.bd

Abstract

Preoperative distinction between benign and malignant ovarian tumours guides referral and surgical planning, but machine-learning (ML) studies on routine laboratory data may overestimate performance when preprocessing is fitted before validation. We developed and internally validated a leakage-safe ML model using a public cohort of 349 patients (171 ovarian cancers, 178 benign tumours) with 49 clinical, haematological, biochemical and tumour-marker predictors. After a single stratified 70/30 split, nested cross-validation compared seven classifiers, with missingness filtering, imputation, minimum-redundancy maximum-relevance selection of 12 features, scaling and hyperparameter search fitted inside training folds only. Random forest (RF) achieved the highest outer-fold ROC-AUC (0.923 ± 0.038). On the untouched test set (n = 105), RF reached a ROC-AUC of 0.941 (95% CI 0.891–0.978), accuracy of 0.886, sensitivity of 0.804, specificity of 0.963 and Brier score of 0.093, with a positive net benefit across threshold probabilities of 0.03–0.99. RF outperformed age plus menopause and CA125 alone (Holm-adjusted DeLong p = 0.033 for both). SHAP, permutation importance and LIME identified HE4, CA125 and age as the main predictors, and tumour markers alone matched the full feature set. Under a leakage-safe protocol, routine blood tests allow accurate and interpretable classification of ovarian tumours.

Introduction

Ovarian cancer is the most lethal gynaecological malignancy, and most patients are diagnosed at an advanced stage[1,2]. A more common clinical problem is the adnexal mass of

uncertain nature. Most of these masses are benign. Even so, the choice between conservative management, local surgery and referral to a gynaecological oncologist depends on an accurate preoperative estimate of malignancy risk[3].

Serum CA125 is the most widely used marker. However, it rises in endometriosis, inflammation and other benign conditions, and it is often normal in premenopausal women with cancer. HE4 improves specificity. The ROMA algorithm combines HE4, CA125 and menopausal status[4]. The risk of malignancy index (RMI) combines CA125 with ultrasound findings and menopausal status[5]. Ultrasound-based models such as IOTA ADNEX perform well in expert hands[6]. These tools use only a few variables, and some require specialist imaging.

Routine haematology and blood chemistry are inexpensive and nearly always available before surgery. Lu et al. combined such tests with tumour markers in a decision-tree model and separated benign from malignant tumours in a Chinese cohort of 349 patients[7]. Later studies on the same public data reported accuracies close to 0.90 with ensembles and explanation methods[8]. When feature selection, imputation or tuning is performed on the full dataset before splitting, information from the test data leaks into model building. Such leakage inflates test results and is a recognised cause of irreproducible ML research[9–11]. Small samples make the problem worse, because selecting among many models and feature sets adds further optimism[12]. In addition, ML often fails to outperform logistic regression in clinical prediction[13].

The TRIPOD+AI statement and the PROBAST+AI tool set explicit standards for reporting prediction models and assessing their risk of bias[14,15]. Both require calibration, uncertainty estimates, comparison with simple baselines and a clear separation between development and validation.

Two clinical questions are often confused. Population screening asks whether asymptomatic women have ovarian cancer. In the UKCTOCS trial, multimodal screening with serial CA125 did not reduce ovarian-cancer mortality[16]. Preoperative classification asks whether a known adnexal mass is benign or malignant. This decision determines conservative management, local surgery or referral to a gynaecological oncologist[3]. The present study, like the cohort it uses, addresses only the second question.

CA125 is elevated in most advanced epithelial cancers. It is also raised by endometriosis, fibroids, pelvic inflammation and menstruation, which lowers its specificity before

menopause[17]. HE4 is less affected by these benign conditions. ROMA combines HE4, CA125 and menopausal status in a logistic model to stratify women with a pelvic mass into high- and low-risk groups[4]. The RMI combines CA125, an ultrasound score and menopausal status[5]. Reviews report that marker combinations generally outperform single markers. Their performance, however, varies with menopausal status, cancer stage and the benign case mix[3,17]. These indices are simple and interpretable, which makes them strong baselines for ML models.

The International Ovarian Tumor Analysis (IOTA) group developed simple ultrasound rules and the multiclass ADNEX model. ADNEX combines patient age, CA125, centre type and ultrasound features[6,18]. In meta-analyses, these approaches outperform marker-only indices when used by trained examiners[3]. Their dependence on expert ultrasound limits their use where such expertise is scarce. This motivates models built from blood tests alone.

Kawakami et al. trained gradient boosting, random forest and support vector machine models on 32 routine preoperative blood biomarkers[19]. Their models discriminated epithelial ovarian cancer from benign tumours and predicted stage and prognosis. Lu et al. released the 349-patient cohort used here[7,8]. The data were collected at the Third Affiliated Hospital of Soochow University between July 2011 and July 2018. They combine blood routine, general chemistry and tumour-marker results. Lu et al. built an interpretable decision-tree model in which tumour markers were among the most informative variables[7]. Ahamad et al. applied several ML classifiers to the same data with an 80/20 split and reported an accuracy of 0.88 and an AUC of 0.87[8]. These studies differ in feature selection, splitting and validation. Calibration and decision-curve analysis are rarely reported.

Data leakage occurs when information from the evaluation data influences model building. Typical examples are feature selection, imputation or scaling fitted on the full dataset[9,10]. Kapoor and Narayanan found leakage in hundreds of ML-based studies across fields and identified pre-split feature selection as a common type[10]. Selecting models and hyperparameters on the same folds used for evaluation adds further optimism; nested cross-validation avoids this bias[11,12]. With small samples, simple train–test splits give unstable estimates. The bias from improper validation can exceed the differences between algorithms[20]. Moreover, when strong predictors exist, flexible ML models often fail to outperform logistic regression[13].

SHAP provides additive feature attributions that can be summarised globally[21]. LIME fits local surrogate models around single predictions[22]. Both describe model behaviour, not causal mechanisms. Local explanations of individual patients cannot support population-level conclusions[23,24]. TRIPOD+AI specifies what ML prediction studies should report, including missing-data handling, hyperparameter tuning, calibration, uncertainty and fairness[14]. PROBAST+AI assesses risk of bias and applicability. It penalises analyses that ignore overfitting or lack calibration[15]. Calibration[25] and decision-curve analysis[26,27] are now expected for models intended to inform clinical decisions.

Table 1 summarises these approaches and their limitations. Three gaps follow from this literature. First, to our knowledge, performance on this cohort has not been estimated with nested cross-validation and fold-wise preprocessing. Second, the incremental value of ML over a simple model of age, menopause, CA125 and HE4 has not been tested with correlated-ROC methods. Third, calibration, net benefit, subgroup behaviour and the role of missing data have received little attention. The present study addresses all three gaps.

Table 1. Approaches to preoperative classification of ovarian masses.

Approach	Main inputs	Strengths	Limitations	Refs
Single tumour markers (CA125, HE4)	Serum assay	Cheap, widely available, standardised	Limited specificity (CA125) or sensitivity; menopause-dependent	[4,17]
Marker-based indices (ROMA)	HE4, CA125, menopause	Simple logistic model, interpretable	Few variables; performance varies by stage and case mix	[4,17]
Ultrasound indices and models (RMI, IOTA rules, ADNEX)	Ultrasound features, CA125, age	Best-validated, multicentre evidence	Need trained examiners and imaging	[5,6,18]
ML on routine blood tests	Haematology, chemistry, tumour markers	Uses routine data; can model interactions	Small single-centre cohorts; frequent leakage; little calibration or baseline comparison	[7,8,19]
This study	49 routine variables, leakage-safe nested CV	Calibration, DeLong vs clinical baselines, ablation, decision curves, SHAP and LIME	Single cohort; no imaging; internal validation only	—

We therefore re-examined the 349-patient cohort with a fully leakage-safe protocol and addressed four research questions:

1. RQ1. Which classifier gives the best internally validated discrimination when every preprocessing step is learned inside the training folds?

2. RQ2. How well does the selected model perform on an untouched independent test set, in terms of discrimination, calibration and clinically relevant error rates?
3. RQ3. Does the ML model add value beyond age, menopausal status, CA125 and HE4, and which laboratory panels carry the signal?
4. RQ4. Which predictors drive the model's predictions globally and for individual patients?

This study makes five contributions. First, it provides a leakage-controlled analysis of a widely used ovarian tumour cohort. The test set is separated before any processing, and every learned step is refitted inside nested cross-validation folds (Figure 1). Second, it compares seven conventional and ensemble classifiers on identical, paired folds. Discrimination, calibration and clinically relevant error rates are reported with confidence intervals and formal statistical tests. Third, it measures the incremental value of ML over simple clinical-biomarker models with the DeLong test and decision-curve analysis. A feature-group ablation identifies the laboratory panels that carry the predictive signal. Fourth, it combines global (SHAP, permutation importance) and local (LIME) explanations. It also shows that missingness of several tumour markers is associated with the outcome, which matters when models are transferred between hospitals. Fifth, it provides a fully specified and reproducible pipeline, so that the frozen model can be validated externally.

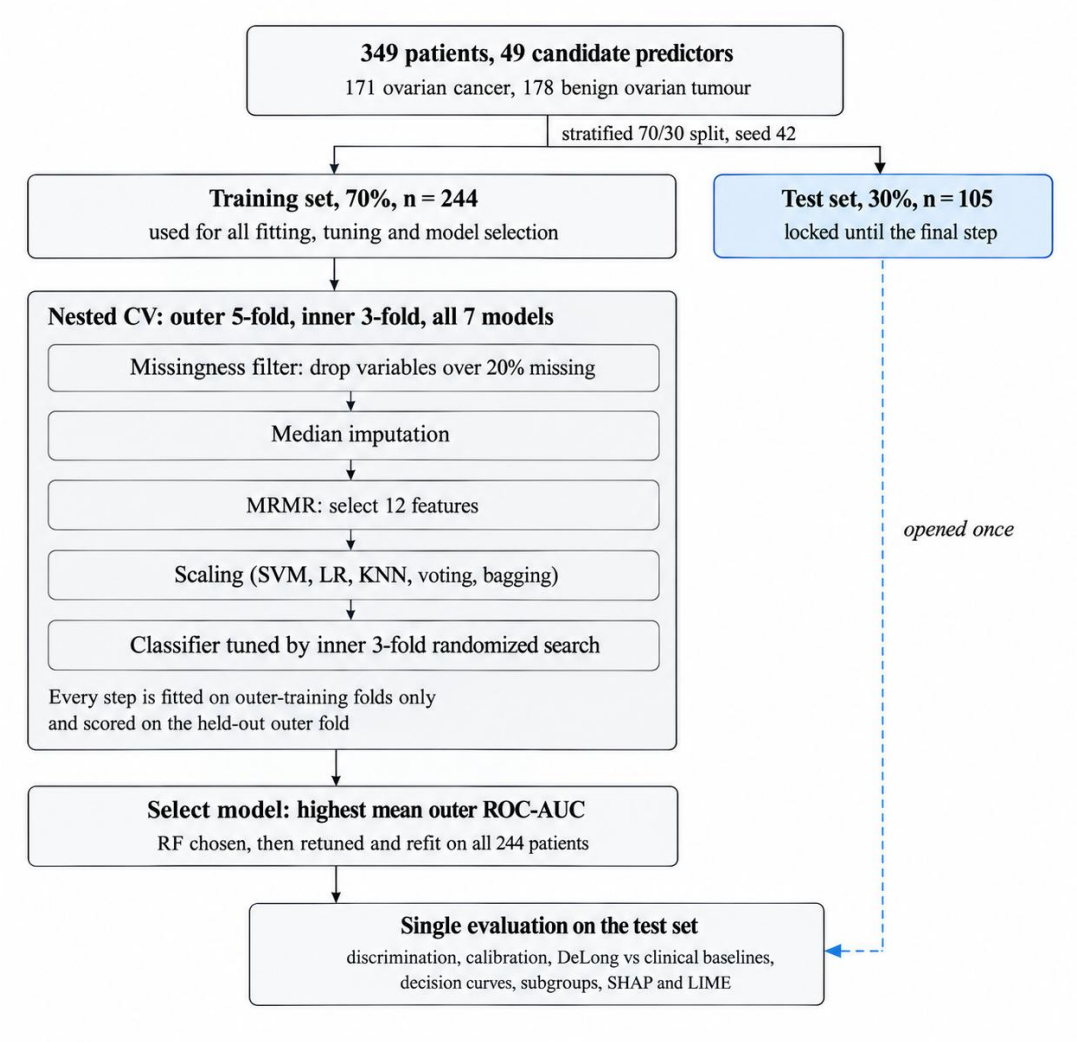

Figure 1. Leakage-safe study pipeline. Every learned step sits inside the nested cross-validation on the training set; the test set is opened once, for the final model.

Results

Cohort characteristics

Patients with cancer were older than those with benign tumours (median 53 vs 36 years). They were also more often postmenopausal (56.1% vs 12.9%). After FDR correction, 27 of 49 variables differed between groups (Table 2). HE4 and CA125 showed the largest effects (rank-biserial r = 0.79 and 0.67), followed by age, menopause, NEU and albumin. CEA and CA19-9 did not differ significantly.

Table 2. Characteristics of the 349 patients (15 variables with the smallest p-values; full table in Supplementary Table S1).

Variable	Benign (n = 178)	Cancer (n = 171)	Effect size	FDR-adjusted p
HE4 (pmol/L)	43.8 [38.8–51.4]	140.9 [61.0–388.4]	r = 0.79	< 0.001
CA125 (U/mL)	22.7 [13.9–48.7]	241.5 [48.3–753.6]	r = 0.67	< 0.001
Age (years)	36 [27–46]	53 [45–63]	r = 0.60	< 0.001

Variable	Benign (n = 178)	Cancer (n = 171)	Effect size	FDR-adjusted p
Postmenopausal, n (%)	23 (12.9)	96 (56.1)	V = 0.45	< 0.001
NEU (%)	59.5 [52.8–65.7]	70.6 [63.5–78.2]	r = 0.55	< 0.001
Albumin (g/L)	43.4 [41.4–45.9]	39.2 [34.7–44.2]	r = −0.42	< 0.001
LYM%	29.8 [23.4–36.4]	23.4 [15.6–28.3]	r = −0.39	< 0.001
ALP (U/L)	65 [55–77]	77 [64–97]	r = 0.38	< 0.001
LYM# (10^9/L)	1.67 [1.31–2.02]	1.34 [1.08–1.67]	r = −0.32	< 0.001
IBIL (μmol/L)	6.0 [4.8–7.6]	5.0 [3.6–6.4]	r = −0.28	< 0.001
CA72-4 (U/mL)	1.74 [0.79–3.38]	4.83 [2.00–12.04]	r = 0.50	< 0.001
TBIL (μmol/L)	9.2 [7.3–11.5]	7.6 [5.7–9.6]	r = −0.28	< 0.001
PLT (10^9/L)	224 [192–267]	265 [205–325]	r = 0.27	< 0.001
PCT (%)	0.23 [0.19–0.27]	0.26 [0.21–0.32]	r = 0.25	< 0.001
AST (U/L)	16 [13–20]	19 [14.5–24]	r = 0.24	< 0.001

Median [IQR]; Mann–Whitney U with rank-biserial r (positive = higher in cancer) or χ^2 with Cramér's V; Benjamini–Hochberg adjustment over 49 tests.

Missing data

Missingness was strongly related to the outcome. HE4, CEA, AFP and CA19-9 were missing only in cancer patients (11.7–14.0% vs 0%; all $p < 0.001$). NEU was missing in 50.0% of benign and 1.2% of cancer patients ($p < 0.001$). CA72-4 missingness did not differ between groups (64.6% vs 73.1%; $p = 0.11$). The training-fold filter removed CA72-4 and NEU in every fold.

Model comparison in nested cross-validation

RF had the highest mean outer-fold ROC-AUC (0.923 ± 0.038), followed by XGBoost (0.921 ± 0.027) and soft voting (0.916 ± 0.032) (Table 3). Distance- and margin-based models performed worse, and KNN had low sensitivity (0.517). The Friedman test showed a difference across the seven models ($\chi^2 = 24.1$, $p < 0.001$). No pairwise Wilcoxon test remained significant after Holm correction, as expected with five folds. The three leading tree-based and ensemble models were statistically indistinguishable.

Table 3. Outer 5-fold nested CV on the training set (n = 244), mean ± SD.

Model	ROC-AUC	PR-AUC	Sensitivity	Specificity	MCC	Brier
RF	0.923 ± 0.038	0.935 ± 0.029	0.817	0.887	0.708	0.110
XGBoost	0.921 ± 0.027	0.939 ± 0.017	0.817	0.887	0.708	0.108
Soft voting	0.916 ± 0.032	0.934 ± 0.030	0.825	0.839	0.667	0.127
Bagging (SVM)	0.885 ± 0.038	0.900 ± 0.024	0.800	0.855	0.660	0.153
SVM	0.863 ± 0.045	0.880	0.675	0.856	0.545	0.153
Logistic regression	0.862 ± 0.049	0.886	0.742	0.830	0.577	0.155
KNN	0.854 ± 0.038	0.860	0.517	0.928	0.497	0.172

SDs for all metrics and t-based 95% CIs are given in Supplementary Table S2. Figure 2 shows the pooled out-of-fold ROC and PR curves for all seven models.

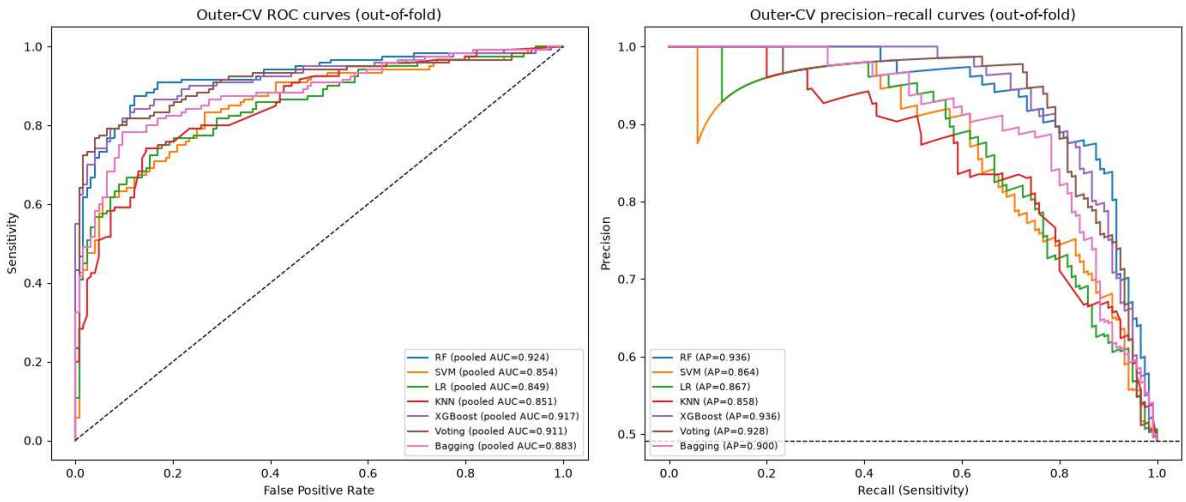

Figure 2. Pooled out-of-fold ROC (left) and precision–recall (right) curves for the seven candidate models from the outer 5-fold cross-validation on the training set. The test set was not used.

Final model and selected features

After retuning on the full training set, the RF used 600 trees, a maximum depth of 5 and all features per split (max_features = None). The other settings were min_samples_split = 2 and min_samples_leaf = 1 (inner-CV ROC-AUC 0.930). The final 12 MRMR features, in order of selection, were HE4, Na, CEA, CA125, AFP, MCH, age, PDW, TP, CA19-9, LYM# and AST. HE4, CA125, CEA, AFP and Na were chosen in all 35 outer-fold fits (Supplementary Table S4).

Independent test-set performance

On the 105 test patients, RF reached a ROC-AUC of 0.941 and a PR-AUC of 0.952 (Table 4; Figure 3). It missed 10 of 51 cancers and misclassified 2 of 54 benign tumours as cancer (Figure 4). Test results lay within or slightly above the outer-CV confidence intervals. The cross-validated and test estimates therefore agree.

Table 4. Final RF model: outer-CV estimate vs independent test.

Metric	Outer CV, mean ± SD	Outer CV 95% CI	Independent test	Test 95% CI (bootstrap)
ROC-AUC	0.923 ± 0.038	0.876–0.970	0.941	0.891–0.978
PR-AUC	0.935 ± 0.029	0.899–0.971	0.952	0.906–0.982
Accuracy	0.852 ± 0.046	0.795–0.910	0.886	0.819–0.943
Sensitivity	0.817 ± 0.023	0.788–0.845	0.804	0.680–0.911
Specificity	0.887 ± 0.077	0.791–0.983	0.963	0.904–1.000
PPV	0.879 ± 0.073	0.789–0.970	0.953	0.881–1.000

Metric	Outer CV, mean ± SD	Outer CV 95% CI	Independent test	Test 95% CI (bootstrap)
NPV	0.832 ± 0.027	0.799–0.865	0.839	0.740–0.930
F1	0.846 ± 0.043	0.793–0.899	0.872	0.786–0.939
MCC	0.708 ± 0.094	0.590–0.825	0.779	0.654–0.888
Brier score	0.110 ± 0.024	0.081–0.139	0.093	0.060–0.132

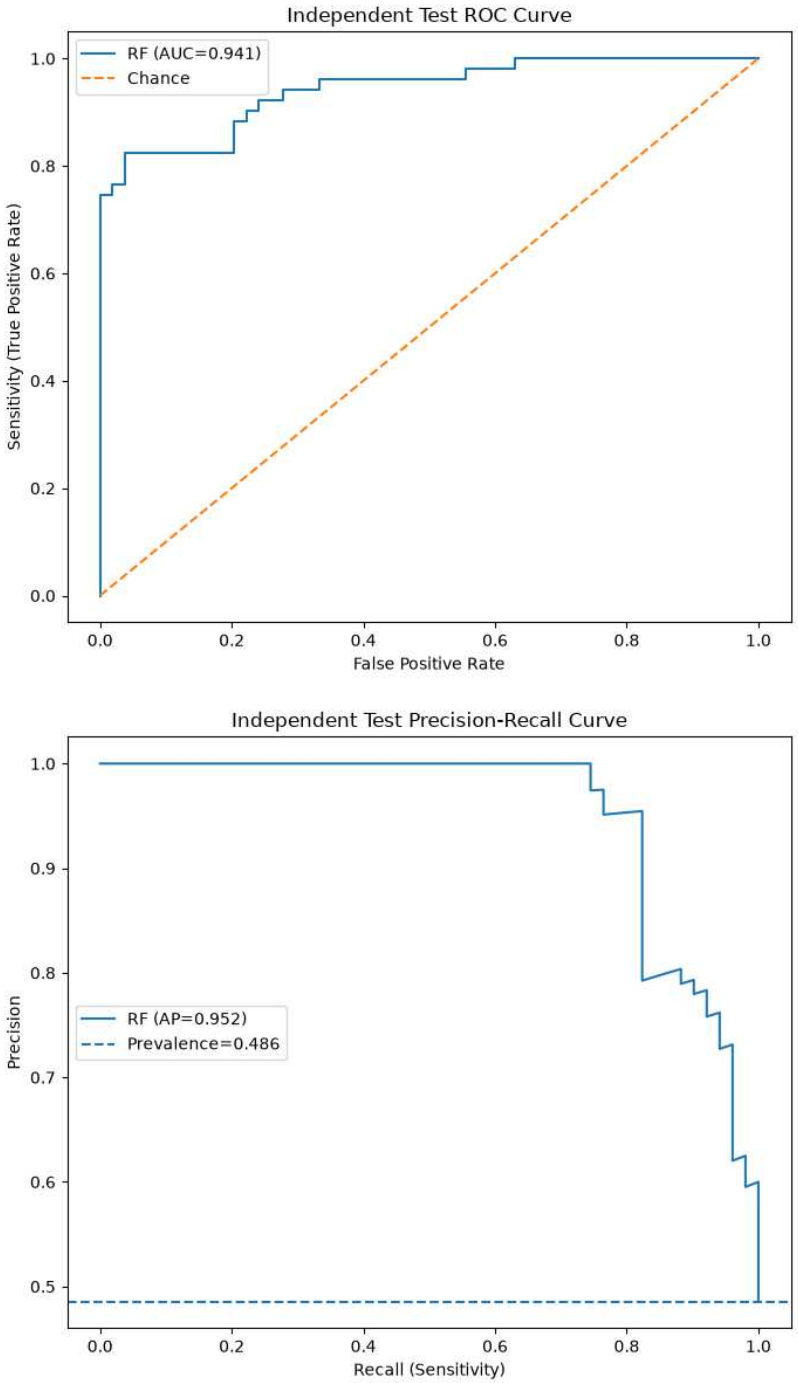

Figure 3. Independent test-set ROC curve (top; AUC 0.941) and precision–recall curve (bottom; average precision 0.952) of the final RF model. The dashed lines show chance and outcome prevalence (0.486).

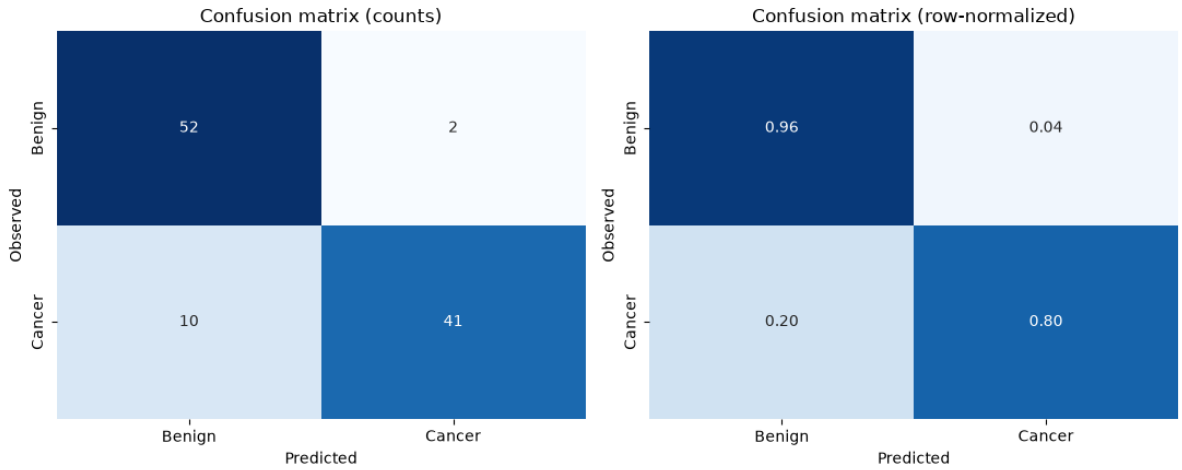

Figure 4. Confusion matrices of the final RF model on the test set at the 0.50 threshold: counts (left) and row-normalised proportions (right). The diagonal of the normalised matrix gives specificity (0.963) and sensitivity (0.804).

Calibration

The test Brier score was 0.093. Calibration-in-the-large was 0.29, and the calibration slope was 1.32 (ideal values 0 and 1). A slope above 1 indicates somewhat conservative predictions. Predicted probabilities clustered towards the middle more than the observed risks justify. This is a known property of averaged tree ensembles (Figure 5).

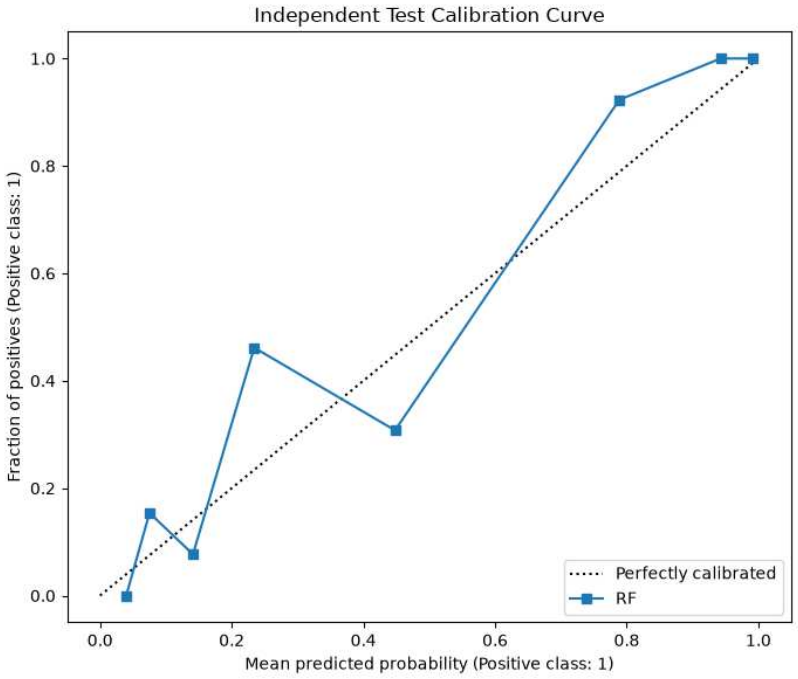

Figure 5. Calibration curve of the final RF model on the independent test set (8 quantile bins). The dashed diagonal is perfect calibration. Brier score 0.093; calibration intercept 0.29; slope 1.32.

Clinical baselines and incremental value

Among the simple models, age + menopause reached a test ROC-AUC of 0.826, CA125 alone 0.838 and HE4 alone 0.882 (Table 5). Age + menopause + CA125 + HE4 reached 0.898. Adding CA125 to age and menopause improved discrimination (ΔAUC 0.066; Holm-adjusted DeLong p = 0.033). Adding HE4 gave little further gain (ΔAUC 0.007; p = 0.26).

RF improved on age + menopause (ΔAUC 0.115; p = 0.033) and on CA125 alone (ΔAUC 0.103; p = 0.033). Its advantages over HE4 alone (0.059) and CA125 + HE4 (0.057) were significant before, but not after, Holm correction (adjusted p = 0.10). The difference from age + menopause + CA125 + HE4 was not significant (ΔAUC 0.043, 95% CI −0.008 to 0.095; adjusted p = 0.26). However, RF had higher sensitivity at the 0.50 threshold (0.804 vs 0.745) and a lower Brier score (0.093 vs 0.124).

Table 5. Independent test performance of clinical baselines and the RF model (n = 105).

Model	ROC-AUC	Sensitivity	Specificity	Brier	ΔAUC, RF vs model (95% CI)	DeLong p (Holm)
Age + menopause	0.826	0.588	0.833	0.173	0.115 (0.036 to 0.194)	0.033
CA125	0.838	0.608	0.944	0.171	0.103 (0.035 to 0.176)	0.033
HE4	0.882	0.588	0.981	0.163	0.059 (0.011 to 0.113)	0.104
CA125 + HE4	0.885	0.647	0.944	0.148	0.057 (0.009 to 0.108)	0.104
Age + menopause + CA125	0.892	0.725	0.889	0.132	—	—
Age + menopause + CA125 + HE4	0.898	0.745	0.889	0.124	0.043 (−0.008 to 0.095)	0.263
All tumour markers	0.896	0.647	0.926	0.139	0.045 (−0.005 to 0.103)	0.263
RF (12 MRMR features)	0.941	0.804	0.963	0.093	reference	—

Baselines are logistic regressions fitted on the training set. ΔAUC CIs are paired bootstrap (2,000 resamples); p-values are DeLong tests with Holm correction over nine comparisons.

Which laboratory panels carry the signal

In the nested-CV ablation, tumour markers alone (five usable markers) reached a ROC-AUC of 0.923 (Figure 6). This was essentially equal to all 49 features (0.915) and to tumour markers combined with blood routine (0.925) or chemistry (0.921). Blood routine alone (0.719) and chemistry alone (0.822) were clearly weaker, and their combination did not help

(0.796). Tumour markers therefore carry most of the discriminative signal. A compact marker-based feature set is sufficient for high discrimination.

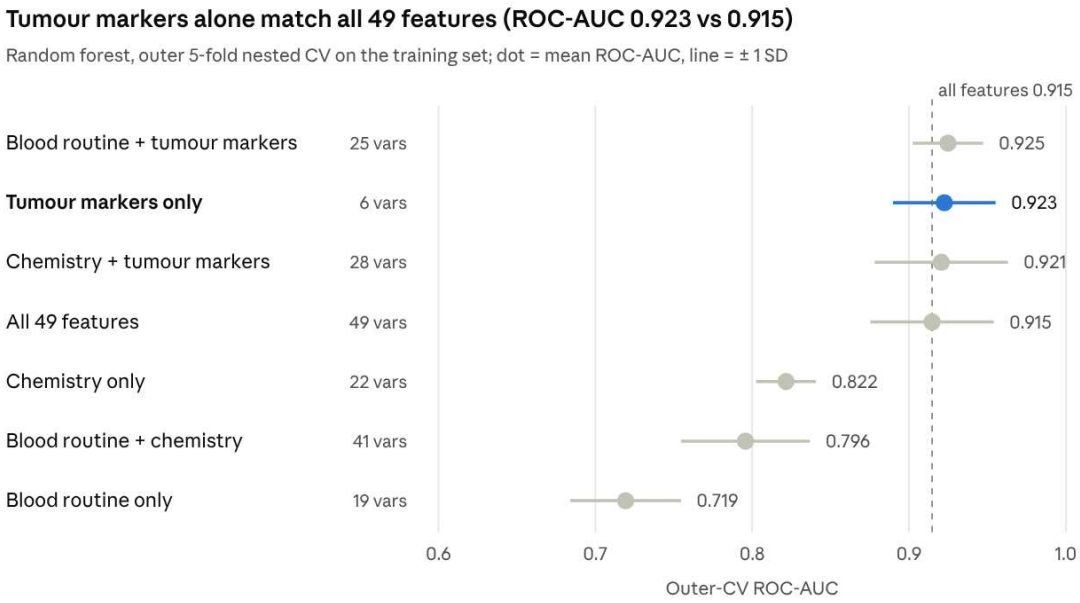

Figure 6. Feature-group ablation. Every set containing the tumour markers reaches ROC-AUC 0.915–0.925; sets without them stay at 0.72–0.82.

Global and local explanations

HE4 dominated the SHAP explanation (mean |SHAP| 0.248), followed by CA125 (0.077), age (0.053) and CEA (0.039) (Figure 7). Permutation importance gave the same order for the top two features. Shuffling HE4 reduced the test ROC-AUC by 0.241, and shuffling CA125 by 0.042; AST, CEA and age followed (Figure 8). In the dependence plots, predicted cancer risk rose steeply with HE4 and CA125 and more gradually with age. Haematological and chemistry variables contributed little individually.

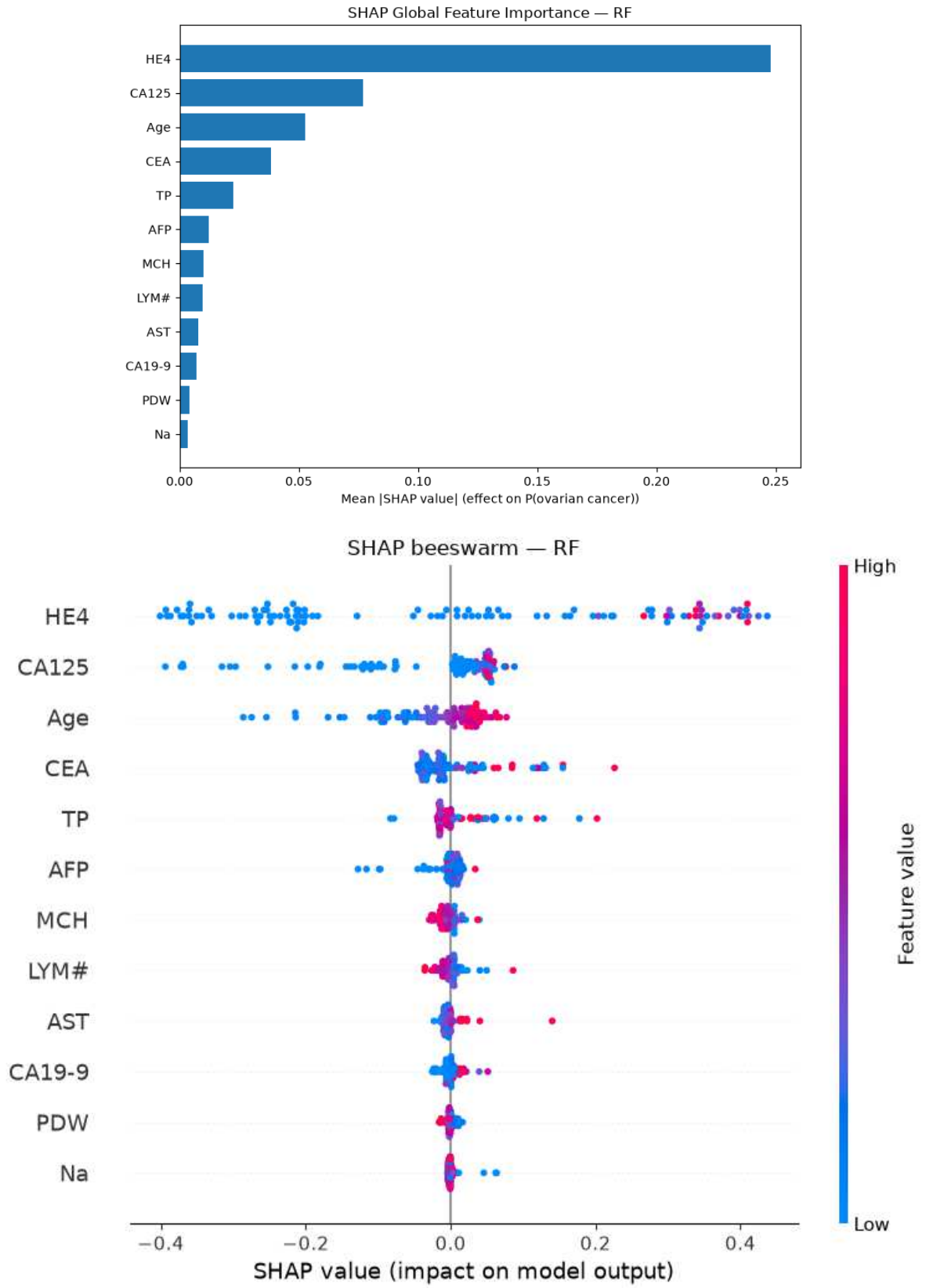

Figure 7. SHAP explanation of the final RF model on the test set: mean absolute SHAP value per feature (top) and beeswarm plot of individual SHAP values coloured by feature value (bottom). Positive values push towards ovarian cancer.

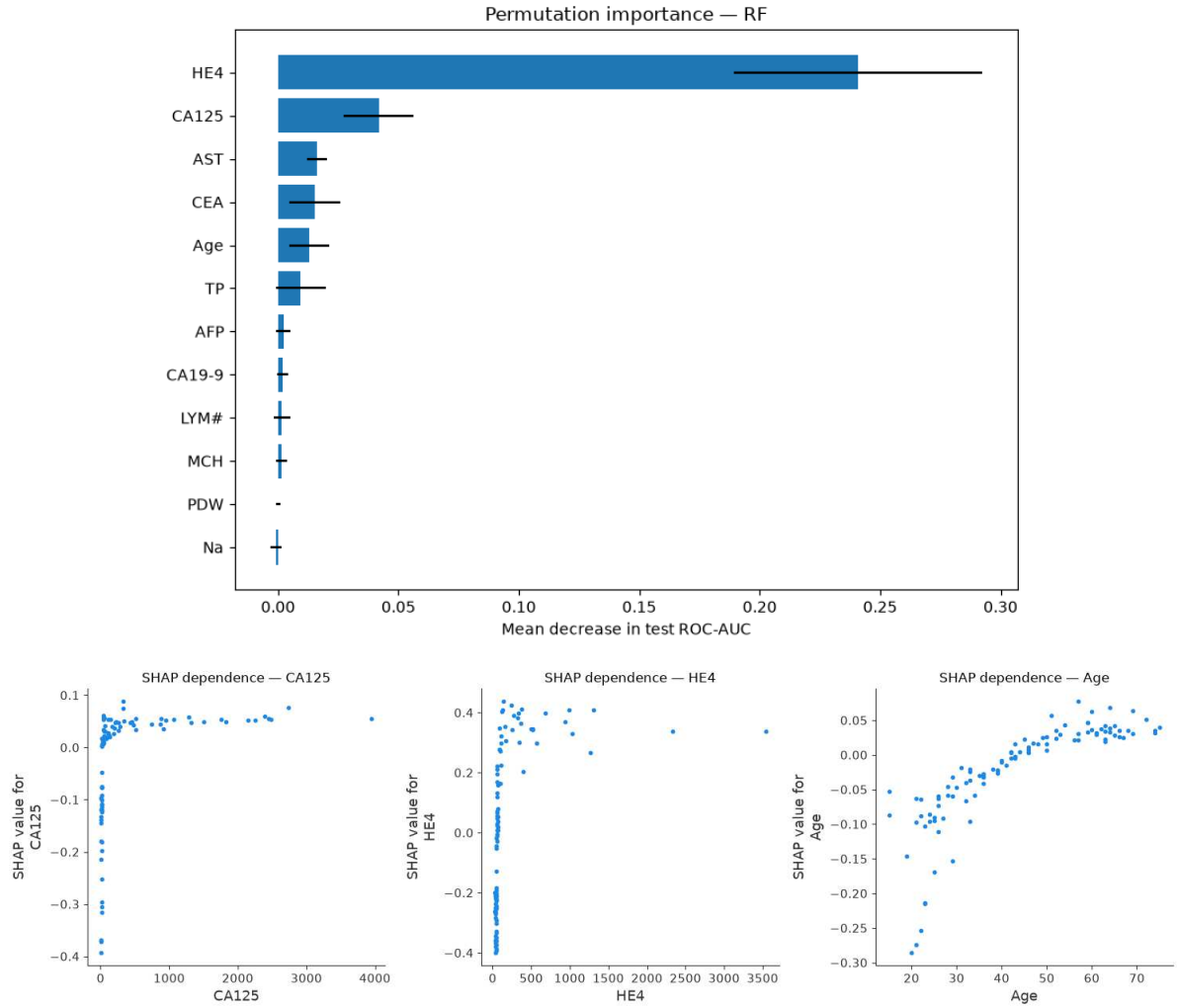

Figure 8. Permutation importance (mean decrease in test ROC-AUC over 50 shuffles, ± SD; top) and SHAP dependence plots for CA125, HE4 and age (bottom).

LIME explanations for three test patients agreed with the global picture (Table 6). For a correctly classified benign case (predicted probability 0.07), HE4 ≤ 42.7 and CA125 ≤ 21.1 contributed most towards benign. Two correctly classified cancers both had a predicted probability of 0.99. In both, HE4 > 125 contributed most towards cancer, followed by CA125 > 204 or CEA > 2.0 and age > 57. A high predicted probability reflects the model's confidence for that patient. It is not evidence of clinical reliability.

Table 6. LIME local explanations for three representative test patients (four largest contributions).

Patient	Observed	P(cancer)	Contributions toward cancer (+) or benign (−)
Case A	Benign	0.07	HE4 ≤ 42.7 (−0.28); CA125 ≤ 21.1 (−0.22); 0.88 < CEA ≤ 1.29 (−0.09); age ≤ 34 (−0.09)
Case B	Cancer	0.99	HE4 > 125.1 (+0.38); CA125 > 204.2 (+0.10); 0.88 < CEA ≤ 1.29 (−0.08); age > 57 (+0.06)

Patient	Observed	P(cancer)	Contributions toward cancer (+) or benign (−)
Case C	Cancer	0.99	HE4 > 125.1 (+0.39); CEA > 2.0 (+0.15); CA125 > 204.2 (+0.10); TP ≤ 67.7 (+0.07)

Weights are LIME surrogate coefficients for P(cancer); full explanations in Supplementary Table S8.

Decision-curve analysis

RF gave a higher net benefit than treat-all and treat-none across thresholds from 0.03 to 0.99 (Figure 9; Table 7). Its net benefit equalled or exceeded that of the age + menopause + CA125 + HE4 model at 94% of thresholds. For example, it was 0.375 vs 0.330 at 0.40, and 0.371 vs 0.305 at 0.50. The two curves crossed near 0.30.

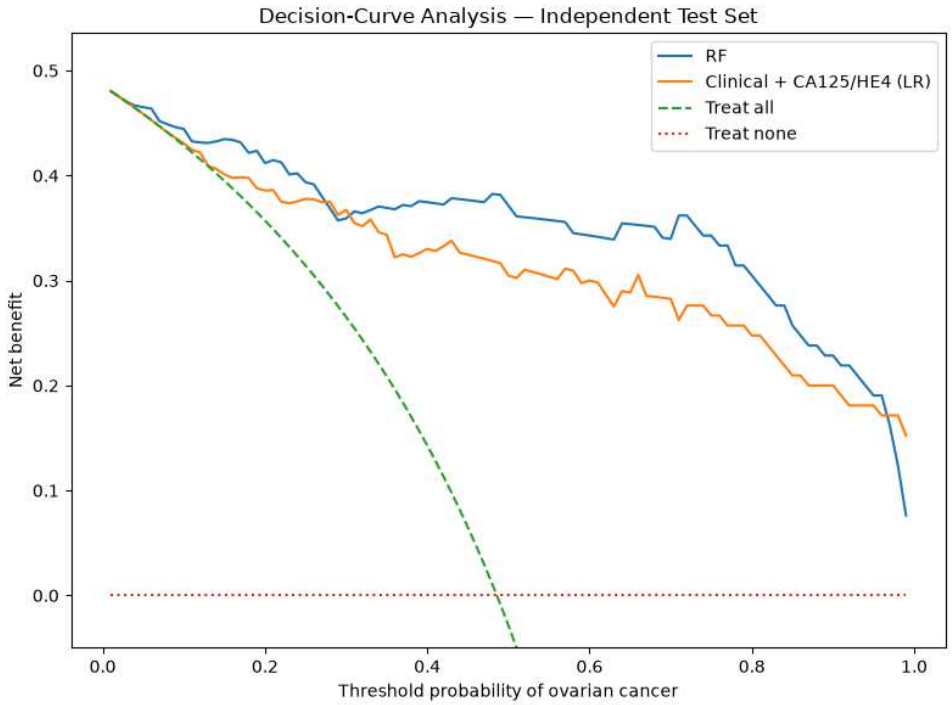

Figure 9. Decision-curve analysis on the independent test set: net benefit of the RF model, the age + menopause + CA125 + HE4 logistic model, treat-all and treat-none across threshold probabilities.

Table 7. Net benefit on the independent test set at selected threshold probabilities.

Threshold	RF	Age + menopause + CA125 + HE4	Treat all	Treat none
0.10	0.444	0.431	0.429	0
0.20	0.412	0.386	0.357	0
0.30	0.359	0.367	0.265	0
0.40	0.375	0.330	0.143	0
0.50	0.371	0.305	−0.029	0
0.60	0.343	0.300	−0.286	0

Threshold	RF	Age + menopause + CA125 + HE4	Treat all	Treat none
0.70	0.340	0.283	−0.714	0

Net benefit = TP/n − FP/n × pt/(1 − pt). Values for all thresholds 0.01–0.99 are in Supplementary Data S1.

Subgroup analysis

Discrimination was higher in patients aged ≥ 50 years (ROC-AUC 0.978, n = 39) than in those aged < 50 years (0.887, n = 66) (Table 8). It was also higher in postmenopausal (0.964, n = 34) than in premenopausal patients (0.900, n = 71). Sensitivity was lower in younger (0.667) and premenopausal patients (0.696), a group in which CA125 is known to be less informative. Subgroup results are descriptive.

Table 8. Test-set performance of the RF model by subgroup (descriptive).

Subgroup	N	Cancers	Accuracy	Sensitivity	Specificity	ROC-AUC
Age < 50 years	66	21	0.879	0.667	0.978	0.887
Age ≥ 50 years	39	30	0.897	0.900	0.889	0.978
Premenopausal	71	23	0.887	0.696	0.979	0.900
Postmenopausal	34	28	0.882	0.893	0.833	0.964

Threshold 0.50. Subgroups were not used for model development; no fairness claim is made.

Discussion

Principal findings

With every preprocessing step confined to training folds, RF, XGBoost and soft voting discriminated benign from malignant ovarian tumours equally well (outer-CV ROC-AUC 0.92). The selected RF model kept this performance on the untouched test set (ROC-AUC 0.941). Specificity was high (0.963), the Brier score was low (0.093), and cross-validated and test estimates agreed closely. RF significantly outperformed age plus menopause and CA125 alone. HE4, CA125 and age carried most of the signal, and tumour markers alone matched the full 49-variable set. The gain over a four-variable logistic model (ΔAUC 0.043) did not reach statistical significance. A compact, interpretable marker-based model is therefore the practical target for external validation.

Comparison with previous work

Lu et al. built an interpretable decision-tree model on this cohort, with tumour markers among the most informative variables[7]. Ahamad et al. reported an accuracy of 0.88 and an AUC of 0.87 on the same data with an 80/20 split[8]. Our test accuracy (0.886) is similar, and our ROC-AUC (0.941) is higher. Both were obtained without any information from the test set entering feature selection, imputation or tuning. The studies differ in splits, feature sets and validation schemes, so this comparison is indirect. The modest gain of ML over a few strong markers agrees with systematic evidence comparing ML with logistic regression[13].

Interpretation of the predictors

HE4 was the dominant predictor in every analysis. It had the largest univariable effect and the highest SHAP and permutation importance, and it was selected in every fold. This agrees with its established role in ROMA[4]. CA125 added information, particularly in combination with age and menopause. Age partly reflects menopausal status; MRMR selected age rather than the binary menopause variable. Albumin, lymphocyte indices, platelets and bilirubin differed between groups. These differences may reflect systemic inflammation and nutritional status in malignancy. However, these variables added little once tumour markers were available. Sodium was selected in every fold but contributed almost nothing to predictions. MRMR favours weakly relevant features that are uncorrelated with those already chosen, so selection frequency should not be read as clinical importance.

Missingness as a data-collection signal

Missing HE4, CEA, AFP and CA19-9 values occurred only in cancer patients. Missing NEU values occurred mainly in benign patients. This pattern most likely reflects test-ordering practice rather than biology. A model that learned “missing HE4 means cancer” would not transfer to another hospital. In our pipeline, missing values were replaced by training-fold medians and no missingness indicators were used. The model therefore cannot exploit these patterns directly.

Clinical implications

The model is intended for patients who already have an ovarian mass. It supports benign-versus-malignant classification and is not a population-screening tool. Its high specificity (0.963) and positive predictive value (0.953) support its use in confirming suspected malignancy before oncological referral. Sensitivity was 0.804 at the prespecified 0.50

threshold. It can be increased by lowering the threshold, guided by the decision curve. The curve showed a positive net benefit over treat-all and treat-none strategies across thresholds of 0.03–0.99. Because the model needs only routine laboratory tests, it could support decisions where specialist ultrasound is unavailable.

Strengths

The main strength is the leakage-safe design. A single stratified split was made before any processing, and every preprocessing step was refitted within nested cross-validation folds. The selection rule was prespecified, and the test set was evaluated once. Following TRIPOD+AI, we report calibration, bootstrap confidence intervals, decision curves and clinical baselines with DeLong tests. We also provide a panel ablation and both global and local explanations. The fixed 12-feature MRMR selection keeps the model compact. The fully specified pipeline (fixed seed, version-pinned environment) makes every reported number reproducible.

Limitations

1. The data come from a single centre and were analysed retrospectively. Nested cross-validation and an untouched test set limit optimism. Performance in other populations and laboratories should still be confirmed by external validation.
2. Disease stage and histological subtype are not included in the public dataset. Performance could therefore not be reported separately for early-stage or borderline tumours.
3. The dataset contains no imaging data. A direct comparison with ultrasound-based scores such as RMI and IOTA ADNEX was therefore not possible[5,6].
4. The cohort size (349 patients) is typical of single-centre biomarker studies. It limits the power of pairwise model comparisons, which we report with multiplicity correction.

Future work and external validation plan

The frozen final pipeline, including preprocessing, the 12 features, the hyperparameters and the 0.50 threshold, will be applied without retuning to an independent cohort from another institution or time period. That study should include ultrasound-based scores for direct comparison with RMI, ROMA and ADNEX. It should also record stage and histology and use a sample size based on published criteria for validation studies[28]. Calibration should be

reported, with recalibration if needed. A prospective evaluation of clinical utility should follow only if external performance holds.

Conclusions

Routine clinical and laboratory data can classify ovarian tumours as benign or malignant with high accuracy when the model is developed under a strictly leakage-safe protocol. In this study, every learned step was refitted inside nested cross-validation folds. These steps were missingness filtering, median imputation, MRMR selection of 12 features, scaling and hyperparameter tuning. The independent test set was used only once, for the final model.

Among seven classifiers, random forest achieved the highest cross-validated ROC-AUC (0.923 ± 0.038). XGBoost and soft voting performed almost identically. On the untouched test set, the random forest reached a ROC-AUC of 0.941, a PR-AUC of 0.952, an accuracy of 0.886, a specificity of 0.963 and a Brier score of 0.093. These results agreed closely with the cross-validated estimates. The model gave a positive net benefit across threshold probabilities of 0.03–0.99. It significantly outperformed both age plus menopausal status and CA125 alone.

Global (SHAP, permutation importance) and local (LIME) explanations consistently identified HE4, CA125 and age as the main predictors. This is in line with established clinical knowledge. The feature-group ablation showed that tumour markers alone reproduced the performance of all 49 variables. A compact, interpretable marker-based model is therefore feasible, and it requires only routine blood tests.

Missingness of several tumour markers was associated with the outcome. This should be considered whenever such models are transferred between hospitals. The fully specified, reproducible pipeline makes the frozen model ready for external evaluation. The next step towards clinical use is multicentre validation that includes ultrasound-based reference models, disease stage and histology.

Methods

Study design and reporting

This retrospective study develops and internally validates a diagnostic prediction model. It is reported according to TRIPOD+AI[14], and risk of bias was assessed with PROBAST+AI[15] (Supplementary Table S10). The data split, validation scheme, number of selected features,

primary metric and model-selection rule were specified in advance. The test set played no role in preprocessing, feature selection, tuning or model selection.

Data source and ethics

We used the publicly available, de-identified dataset released by Lu et al. as supplementary material (file "Supplementary data_Biomarker")[7,29]. It includes 349 patients with pathologically confirmed ovarian tumours treated at the Third Affiliated Hospital of Soochow University, China, between July 2011 and July 2018[7,8]. The file contains no direct identifiers. This was a secondary analysis of anonymised, public data without patient contact. Ethical approval and informed consent were therefore not required for this secondary analysis. All methods were carried out in accordance with relevant guidelines and regulations.

Participants and outcome

All 349 patients were analysed: 171 with ovarian cancer and 178 with benign ovarian tumours. The binary outcome was the final pathological diagnosis (1 = ovarian cancer, 0 = benign tumour). In the data file, this variable is coded TYPE, and TYPE = 0 denotes cancer, consistent with the class counts reported in[7]. Disease stage is not available. The model therefore addresses benign versus malignant classification of a known ovarian mass, not screening or early-stage detection.

Candidate predictors

Forty-nine predictors were available, all measured before treatment. They comprised two clinical variables (age, menopausal status) and 19 blood-routine variables, such as lymphocyte, monocyte, eosinophil and basophil counts, haemoglobin and platelet indices. They also included 22 general-chemistry variables, such as albumin, globulin, bilirubin fractions, liver enzymes, electrolytes, urea, creatinine and uric acid. The remaining six were tumour markers (AFP, CA125, CA19-9, CA72-4, CEA, HE4). Values reported beyond a laboratory limit were set to that limit (AFP > 1210 → 1210; CA125 > 5000 → 5000; CA19-9 > 1000 → 1000; CA19-9 < 0.6 → 0.5). The patient identifier was removed.

Missing data

CA72-4 (68.8% missing) and NEU (26.1% missing) exceeded a prespecified 20% threshold. All other variables had at most 6.9% missing values. The missingness rule and median imputation were learned inside each training fold. They were then applied unchanged to the

corresponding validation or test data. We also tested whether missingness was related to the outcome (χ^2 test of the missing indicator).

Data splitting and validation

The data were split once into a training set (70%, n = 244) and an independent test set (30%, n = 105). The split was stratified by outcome with a fixed seed (42). The test set was not used for preprocessing, feature selection, tuning or model choice.

Nested cross-validation was performed on the training set. The outer loop (stratified 5-fold) estimated performance, and the inner loop (stratified 3-fold) tuned hyperparameters. All models used the same outer folds, so their results are paired. Within every training fold, the full pipeline was refitted in this order: missingness filter, median imputation, MRMR selection of 12 features, standard scaling (scale-sensitive models only) and classifier (Figure 1).

Feature selection

MRMR[30] selected a fixed number of 12 features in every fold. Relevance was the mutual information between each feature and the outcome. Redundancy was the mean absolute Pearson correlation with the features already chosen. At each step, the feature with the highest relevance minus redundancy was added. Both quantities were computed from the training fold only. No other feature-selection method was used in the final pipeline.

Models and hyperparameter optimisation

Seven classifiers were compared: random forest (RF)[31], support vector machine (SVM; RBF or linear kernel)[32], logistic regression (LR), k-nearest neighbours (KNN), XGBoost[33], a soft-voting ensemble (LR + SVM + RF) and bagged SVMs. All models except XGBoost were implemented in scikit-learn[34]. RF, SVM, LR and the ensembles used balanced class weights. SVM, LR, KNN, voting and bagging received standardised inputs, whereas tree-based models did not.

Hyperparameters were tuned by randomized search[35], with inner-CV ROC-AUC as the objective. We sampled 8 configurations per outer fold and 12 for the final refit. Search spaces are listed in Supplementary Table S3. The SVM penalty range (C = 0.1–100) excluded very small values. At such values the decision function becomes nearly flat, and Platt-scaled probabilities become unstable.

Model selection and final model

The primary selection criterion was the mean outer-fold ROC-AUC. Mean PR-AUC and then the lower Brier score served as tie-breakers. The selected model type was retuned on the whole training set by inner 3-fold CV, refitted and applied once to the test set. The classification threshold was fixed at 0.50 in advance and was not tuned on test data.

Performance measures

We report accuracy, sensitivity, specificity, positive predictive value (PPV, precision), negative predictive value (NPV), F1, ROC-AUC, PR-AUC (average precision)[36], Matthews correlation coefficient (MCC) and Brier score. Ovarian cancer was the positive class:

$$Sensitivity = TP / (TP + FN); \quad Specificity = TN / (TN + FP); \quad PPV = TP / (TP + FP); \quad NPV = TN / (TN + FN)$$

Outer-CV results are given as mean ± SD, with t-based 95% CIs over the five folds. Test-set 95% CIs were obtained from 2,000 bootstrap resamples (percentile method)[37]. Calibration was assessed with a quantile calibration curve and the Brier score. We also estimated calibration-in-the-large (intercept with the linear predictor as offset) and the calibration slope[25,38].

Statistical analysis

Patient characteristics are given as median [IQR] or n (%). Groups were compared with the Mann–Whitney U or χ^2 test. Effect sizes were the rank-biserial r or Cramér's V, and the Benjamini–Hochberg procedure controlled the false-discovery rate across 49 tests[39]. Models were compared on paired outer-fold ROC-AUCs with the Friedman test[40,41]. Pairwise Wilcoxon signed-rank tests with Holm correction followed[42]. With five folds, the smallest attainable two-sided Wilcoxon p-value is 0.0625, so pairwise tests are exploratory. On the test set, the ROC-AUCs of the final and baseline models were compared with the DeLong test for correlated curves[43,44]. Holm correction and paired bootstrap confidence intervals for ΔAUC were applied.

Clinical baselines and incremental value

Logistic regression baselines (C = 1) were fitted on the training set and evaluated on the test set. They used the same training-only missingness filter, imputation and scaling. The baselines were: (i) age + menopause; (ii) CA125 alone; (iii) HE4 alone; (iv) CA125 + HE4;

(v) age + menopause + CA125; (vi) age + menopause + CA125 + HE4; and (vii) all tumour markers. Models (i), (v) and (vi) form an incremental chain.

Feature-group ablation

We tested whether combining laboratory panels improves discrimination. The selected model type was re-evaluated with the same nested CV on seven feature sets: blood routine, chemistry, tumour markers, each pair of panels and all features. Sets with fewer than 12 candidates used all available features. Only outer-CV results are reported, so the test set was not reused.

Interpretability

Global explanations used SHAP TreeExplainer values on the test set[21]. Permutation importance was measured as the decrease in test ROC-AUC over 50 shuffles. SHAP dependence plots were drawn for CA125, HE4 and age. Local explanations used LIME[22] for three representative test patients. These explanations describe model behaviour, not causal effects.

Decision-curve and subgroup analyses

Net benefit was computed on the test set across threshold probabilities of 0.01–0.99. It was compared with treat-all and treat-none strategies[26,27]. Descriptive subgroup results are reported by age (< 50 vs ≥ 50 years) and menopausal status.

Software and reproducibility

Analyses used Python 3.12.14, NumPy 2.5.3, pandas 3.0.6, SciPy 1.18.1, scikit-learn 1.9.1, XGBoost 3.4.1, SHAP 0.52.0 and LIME 0.2.0.1, with random seed 42. The notebook, environment record and all output tables are available (see Code availability).

Data availability

The dataset analysed in this study was released by Lu et al.[7] and is publicly available from Mendeley Data ("Data for: Using machine learning to predict ovarian cancer", version 11, CC BY-NC 3.0 licence) at https://doi.org/10.17632/th7fztbrv9.11[29]. The original article is available at https://doi.org/10.1016/j.ijmedinf.2020.104195. All derived output tables are provided as Supplementary Information.

Code availability

The complete analysis code (Jupyter notebook) and environment record are available from the corresponding author upon reasonable request.

References

1. Bray, F. et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **74**, 229–263 (2024).
2. Siegel, R. L., Giaquinto, A. N. & Jemal, A. Cancer statistics, 2024. *CA Cancer J. Clin.* **74**, 12–49 (2024).
3. Kaijser, J. et al. Presurgical diagnosis of adnexal tumours using mathematical models and scoring systems: a systematic review and meta-analysis. *Hum. Reprod. Update* **20**, 449–462 (2014).
4. Moore, R. G. et al. A novel multiple marker bioassay utilizing HE4 and CA125 for the prediction of ovarian cancer in patients with a pelvic mass. *Gynecol. Oncol.* **112**, 40–46 (2009).
5. Jacobs, I. et al. A risk of malignancy index incorporating CA 125, ultrasound and menopausal status for the accurate preoperative diagnosis of ovarian cancer. *Br. J. Obstet. Gynaecol.* **97**, 922–929 (1990).
6. Van Calster, B. et al. Evaluating the risk of ovarian cancer before surgery using the ADNEX model to differentiate between benign, borderline, early and advanced stage invasive, and secondary metastatic tumours: prospective multicentre diagnostic study. *BMJ* **349**, g5920 (2014).
7. Lu, M. et al. Using machine learning to predict ovarian cancer. *Int. J. Med. Inform.* **141**, 104195 (2020).
8. Ahamad, M. M. et al. Early-stage detection of ovarian cancer based on clinical data using machine learning approaches. *J. Pers. Med.* **12**, 1211 (2022).
9. Kaufman, S., Rosset, S., Perlich, C. & Stitelman, O. Leakage in data mining: formulation, detection, and avoidance. *ACM Trans. Knowl. Discov. Data* **6**, 15 (2012).
10. Kapoor, S. & Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* **4**, 100804 (2023).
11. Varma, S. & Simon, R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* **7**, 91 (2006).
12. Cawley, G. C. & Talbot, N. L. C. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.* **11**, 2079–2107 (2010).
13. Christodoulou, E. et al. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J. Clin. Epidemiol.* **110**, 12–22 (2019).
14. Collins, G. S. et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **385**, e078378 (2024).

15. Moons, K. G. M. et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ* **388**, e082505 (2025).
16. Menon, U. et al. Ovarian cancer population screening and mortality after long-term follow-up in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised controlled trial. *Lancet* **397**, 2182–2193 (2021).
17. Dochez, V. et al. Biomarkers and algorithms for diagnosis of ovarian cancer: CA125, HE4, RMI and ROMA, a review. *J. Ovarian Res.* **12**, 28 (2019).
18. Timmerman, D. et al. Simple ultrasound-based rules for the diagnosis of ovarian cancer. *Ultrasound Obstet. Gynecol.* **31**, 681–690 (2008).
19. Kawakami, E. et al. Application of artificial intelligence for preoperative diagnostic and prognostic prediction in epithelial ovarian cancer based on blood biomarkers. *Clin. Cancer Res.* **25**, 3006–3015 (2019).
20. Vabalas, A., Gowen, E., Poliakoff, E. & Casson, A. J. Machine learning algorithm validation with a limited sample size. *PLoS ONE* **14**, e0224365 (2019).
21. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* 4765–4774 (2017).
22. Ribeiro, M. T., Singh, S. & Guestrin, C. "Why should I trust you?" Explaining the predictions of any classifier. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1135–1144 (2016).
23. Ghassemi, M., Oakden-Rayner, L. & Beam, A. L. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit. Health* **3**, e745–e750 (2021).
24. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
25. Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L. & Steyerberg, E. W. Calibration: the Achilles heel of predictive analytics. *BMC Med.* **17**, 230 (2019).
26. Vickers, A. J. & Elkin, E. B. Decision curve analysis: a novel method for evaluating prediction models. *Med. Decis. Making* **26**, 565–574 (2006).
27. Vickers, A. J., Van Calster, B. & Steyerberg, E. W. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* **352**, i6 (2016).
28. Riley, R. D. et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Stat. Med.* **40**, 4230–4251 (2021).
29. Lu, M. et al. Data for: Using machine learning to predict ovarian cancer. *Mendeley Data* https://doi.org/10.17632/th7fztbrv9.11 (2020).
30. Peng, H., Long, F. & Ding, C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**, 1226–1238 (2005).
31. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
32. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
33. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (2016).

34. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
35. Bergstra, J. & Bengio, Y. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **13**, 281–305 (2012).
36. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **10**, e0118432 (2015).
37. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap* (Chapman & Hall, 1993).
38. Steyerberg, E. W. et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* **21**, 128–138 (2010).
39. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).
40. Friedman, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Stat. Assoc.* **32**, 675–701 (1937).
41. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **7**, 1–30 (2006).
42. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6**, 65–70 (1979).
43. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845 (1988).
44. Sun, X. & Xu, W. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Process. Lett.* **21**, 1389–1393 (2014).

Acknowledgements

We thank Lu et al. for releasing the original dataset.

Author contributions

S.A.L. developed the software and wrote the original draft. S.S.K. developed the software, prepared the visualisations and reviewed and edited the manuscript. N.K. reviewed and edited the manuscript. A.H. supervised the study. All authors reviewed and approved the final manuscript.

Funding

This research received no external funding.

Competing interests

The authors declare no competing interests.

Supplementary information

Every supplementary item is generated by the analysis notebook (folder ovarian_ml_outputs/).

Item	Content	Source file
Table S1	All 49 variables: median [IQR], tests, effect sizes, FDR-adjusted p	table1_descriptive_BH_FDR.csv
Table S2	Outer-CV metrics, all models: mean, SD, 95% CI; per-fold values	outer_cv_95CI.csv, outer_cv_per_fold.csv
Table S3	Hyperparameter search spaces, per-fold selections and the full final configuration	hyperparameter_search_space_all_models.csv, outer_fold_hyperparameters.csv, final_hyperparameters.csv
Table S4	MRMR selection frequency across 35 outer-fold fits; final 12 features	mrmr_selection_frequency.csv, final_12_mrmr_features.csv
Table S5	Pairwise Wilcoxon tests with Holm correction; Friedman test	pairwise_wilcoxon_holm.csv, friedman_test.csv
Table S6	Missingness by outcome	missingness_vs_outcome.csv
Table S7	Feature-group ablation, per fold	feature_group_ablation_per_fold.csv
Table S8	SHAP, permutation importance and LIME tables	shap_global_importance.csv, permutation_importance.csv, lime_local_explanations.csv
Table S9	Out-of-fold predictions from the outer cross-validation, all models	outer_cv_out_of_fold_predictions.csv
Table S10	TRIPOD+AI checklist and PROBAST+AI self-assessment	Supplementary File S2 (official TRIPOD+AI and PROBAST+AI forms)
Figure S1	Full notebook report with every intermediate output	Ovarian_Cancer_Results_Report.html
Figure S2	SHAP beeswarm and dependence plots	figures/05_SHAP_global_2.png, figures/06_…_2.png
Data S1	Test-set predictions and decision-curve values	independent_test_predictions.csv, decision_curve_net_benefit.csv
Data S2	Reproducibility record (versions, seed, settings)	reproducibility.json