Additional Files

Additional File Table S1 Two-step LightGBM-based classification algorithm for training and prediction

Algorithm 1 Two-step LightGBM training

Input : Training data X = [Xt, Xg]
Xt ∈ ℝ^{n×4} (CA125, Age, HE4, CA72-4)
Xg ∈ ℝ^{n×G} (glycopeptides)
Labels y ∈ {0,1}^n
Cutoffs: L = 0.001, H = 0.999
PCA dimension m = 100

Output: Step-1 classifier f1, Step-2 classifier f2
Feature scalers S1, S2, PCA P, cutoffs (L,H)

1. Fit scaler S1 on Xt and obtain Zt = S1(Xt)
2. Train LightGBM f1 on (Zt, y)
3. Compute step-1 probabilities p1 = f1(Zt)
4. Define index set U = { j | L < p1(j) < H } (Indeterminate range)
5. Restrict data: XtU = Xt[U], XgU = Xg[U], yU = y[U]
6. Fit scaler S2 on log1p(XtU) and obtain Zt2 = S2(log1p(XtU))
7. Fit PCA P (m comps) on XgU and obtain Zg = P(XgU)
8. Concatenate features X2 = [Zt2, Zg]
9. Train LightGBM f2 on (X2, yU)
10. Save {f1, f2, S1, S2, P, L, H}

Algorithm 2 Two-step LightGBM prediction

Input : Unseen samples X = [Xt, Xg]
Trained f1, f2, S1, S2, P
Cutoffs L, H

Output: For each sample j:
step-1 prob p1(j), class c1(j) ∈ {0,1,2}
step-2 prob p2(j) (only if c1(j)=2)
final prob p*(j), class c*(j) ∈ {0,1}

1. Zt = S1(Xt)
2. p1 = f1(Zt) (step-1 probabilities)
3. For each j:
   if p1(j) ≥ H then c1(j) = 1 (Positive)
   else if p1(j) ≤ L then c1(j) = 0 (Negative)
   else c1(j) = 2 (Indeterminate range)
4. Let U = { j | c1(j) = 2 }
5. If U ≠ ∅ then
   XtU = Xt[U], XgU = Xg[U]
   Zt2 = S2(log1p(XtU))
   Zg = P(XgU)
   X2 = [Zt2, Zg]
   p2(U) = f2(X2) (step-2 probabilities)
6. Set final probabilities:
   p*(j) = p2(j) if j ∈ U, else p1(j)
7. Final class:
   c*(j) = 1 if p*(j) > 0.5, else 0
8. Return {p1, c1, p2, p*, c*}

Additional File Table S2 Area under the curve (AUC) and 95% confidence intervals of ROC analysis

(A) Comparison between EOC and benign gynecological diseases

	EOC vs All	EOC vs HE	EOC vs LE	EOC vs OCY	EOC vs EM
CA125	0.899 (0.879–0.916)	0.923 (0.906–0.939)	0.881 (0.850–0.909)	0.818 (0.768–0.865)	0.667 (0.611–0.722)
HE4	0.814 (0.787–0.839)	0.827 (0.800–0.853)	0.751 (0.709–0.791)	0.774 (0.737–0.808)	0.743 (0.696–0.782)
CA72-4	0.731 (0.704–0.757)	0.723 (0.696–0.748)	0.795 (0.751–0.840)	0.740 (0.683–0.795)	0.752 (0.699–0.799)
CSGSA EOC	0.962 (0.942–0.979)	0.971 (0.954–0.986)	0.959 (0.933–0.981)	0.918 (0.870–0.959)	0.879 (0.822–0.930)

(B) Comparison between each stage of EOC and non-EOC

	Stage I vs non-EOC	Stage II vs non-EOC	Stage III vs non-EOC	Stage IV vs non-EOC
CA125	0.838 (0.807–0.867)	0.930 (0.876–0.972)	0.961 (0.947–0.975)	0.993 (0.986–0.997)
HE4	0.716 (0.672–0.757)	0.884 (0.814–0.945)	0.911 (0.880–0.942)	0.940 (0.857–0.995)
CA72-4	0.672 (0.635–0.709)	0.722 (0.650–0.785)	0.798 (0.757–0.837)	0.852 (0.765–0.927)
CSGSA EOC	0.942 (0.907–0.971)	0.989 (0.974–0.999)	0.972 (0.941–0.997)	0.999 (0.994–1.000)

(C) Comparison between the histological subtypes of EOC and non-EOC patients

	CC vs non-EOC	MU vs non-EOC	EN vs non-EOC	SE vs non-EOC
CA125	0.859 (0.825–0.889)	0.809 (0.705–0.892)	0.894 (0.856–0.929)	0.963 (0.944–0.978)
HE4	0.702 (0.647–0.752)	0.673 (0.545–0.787)	0.845 (0.796–0.889)	0.936 (0.906–0.964)
CA72-4	0.664 (0.624–0.707)	0.811 (0.714–0.891)	0.677 (0.626–0.730)	0.821 (0.782–0.857)
CSGSA EOC	0.950 (0.902–0.985)	0.984 (0.967–0.996)	0.958 (0.922–0.985)	0.969 (0.937–0.992)

(D) Comparison between clear cell carcinoma and endometriosis

	CC vs EM
CA125	0.522 (0.441–0.600)
HE4	0.597 (0.524–0.667)
CA72-4	0.689 (0.616–0.759)
CSGSA CC	0.856 (0.771–0.927)

* Values in parentheses indicate 95% confidence intervals.

Additional File Table S3 Effect of different indeterminate threshold ranges on the performance of the two-step LightGBM framework

Indeterminate Threshold Range	Indeterminate Rate	AUC	Sensitivity	Specificity	Projected PPV
0.001–0.999	64%	0.962	65.0%	99.9%	18.7%
0.01–0.99	36%	0.940	32.0%	99.9%	8.0%
0.05–0.95	19%	0.903	35.7%	99.9%	8.9%

The threshold range indicates the lower and upper probability cutoffs used to define the indeterminate group in the first-stage classifier. The indeterminate rate represents the proportion of samples classified as indeterminate in the first stage and subsequently evaluated in the second-stage glycopeptide-based model. Sensitivity, specificity, and AUC were calculated using the independent hold-out test set. Projected PPV was estimated using Bayes' theorem under the population prevalence assumptions described in the Methods section and therefore represents a theoretical projection rather than a directly observed value.

Additional File Table S4 Effect of the number of retained principal components (PCs) on model performance

Number of PCs	AUC
30	0.950
50	0.956
100	0.962
150	0.959
200	0.962

Principal component analysis was applied to the enriched glycopeptide dataset prior to model development. AUC values were calculated using the independent hold-out test set after retaining the indicated number of PCs. The analysis was performed to assess the robustness of model performance to the number of retained PCs. Model performance remained relatively stable across a broad range of retained components, supporting the use of 100 PCs in the final model.

Additional File Table S5 Diagnostic performance of the proposed two-step LightGBM framework compared with a conventional biomarker-based logistic regression model

Model	AUC	Sensitivity	Specificity	Projected PPV (%)
Two-step LightGBM	0.962 (0.942–0.979)	65.0%	0.999	18.7%
Logistic Regression	0.889 (0.851–0.921)	45.0%	0.999	10.9%

The logistic regression model was constructed using age and the serum tumor markers CA125, HE4, and CA72-4. The two-step LightGBM framework incorporated age, serum tumor markers, and enriched glycopeptide-derived features. AUC values are presented with 95% confidence intervals. Sensitivity, specificity, and projected positive predictive value (PPV) were calculated using the independent hold-out test set. Projected PPV was estimated using Bayes' theorem under the population prevalence assumptions described in the Methods section. This analysis was performed to assess whether the proposed framework provides additional diagnostic value beyond conventional biomarker-based prediction.

Additional File Table S6 Diagnostic performance of the two-step LightGBM model for epithelial ovarian cancer detection without CA72-4

Table S6A Confusion matrix

		CSGSA		Sum
		Positive	**Negative**	
EOC		123	57	180
non-EOC	EM	2	21	23
	HE	1	315	316
	LE	0	26	26
	OCY	1	21	22
Sum		127	440	567

Table S6B Performance

Metric	**Observed (%)**	**95% CI**	**Prevalence-adjusted (%)***
Sensitivity	68.3	61.2–74.7	–
Specificity	99.0	97.4–99.6	–
PPV	96.9	94.1–99.5	4.4†
NPV	87.0	83.6–89.9	99.99†

* Prevalence-adjusted values were calculated using Bayes' theorem based on the reported real-world prevalence of epithelial ovarian cancer and relevant benign gynecological conditions (see Methods).
† The adjusted PPV and NPV represent theoretical estimates under the assumed population prevalence and were not derived from the hold-out test dataset.

Additional File Table S7 Sensitivity analysis of the projected positive predictive value under different assumed prevalences of epithelial ovarian cancer

Assumed EOC prevalence (per 100,000 women)	Projected PPV
10	7.4%
20	13.0%
30	18.7%
40	23.0%
50	27.5%
100	42.8%

The projected positive predictive value (PPV) was estimated using Bayes' theorem based on the sensitivity and specificity of the final two-step LightGBM model. Different prevalence values were applied to evaluate the robustness of the projected PPV under alternative population screening scenarios. The prevalence-adjusted PPV values represent theoretical estimates and were not directly observed in the study cohort. These results demonstrate that the projected PPV is highly dependent on the assumed disease prevalence and should therefore be interpreted within the context of the target screening population.

Additional File Table S8 Sensitivity of the projected positive predictive value to small changes in specificity

Specificity	Sensitivity	Projected PPV
99.55%	67.2%	4.4%
99.59%	65.6%	4.9%
99.89%	65.0%	18.7%

The projected positive predictive value (PPV) was estimated using Bayes' theorem assuming an epithelial ovarian cancer prevalence of 30 cases per 100,000 women. Sensitivity and specificity values correspond to alternative model configurations evaluated during model development. The analysis was performed to illustrate the impact of small changes in specificity on projected PPV under low-prevalence screening conditions. As expected, modest increases in specificity resulted in substantial increases in projected PPV because of the extremely low prevalence of ovarian cancer in the target screening population.

Additional File Table S9 Ethnic composition of the training and hold-out cohorts

Cohort	Group	Asian	Caucasian	Total
Training	EOC	307	66	373
Training	non-EOC	684	73	757
Hold-out	EOC	137	43	180
Hold-out	non-EOC	351	36	387

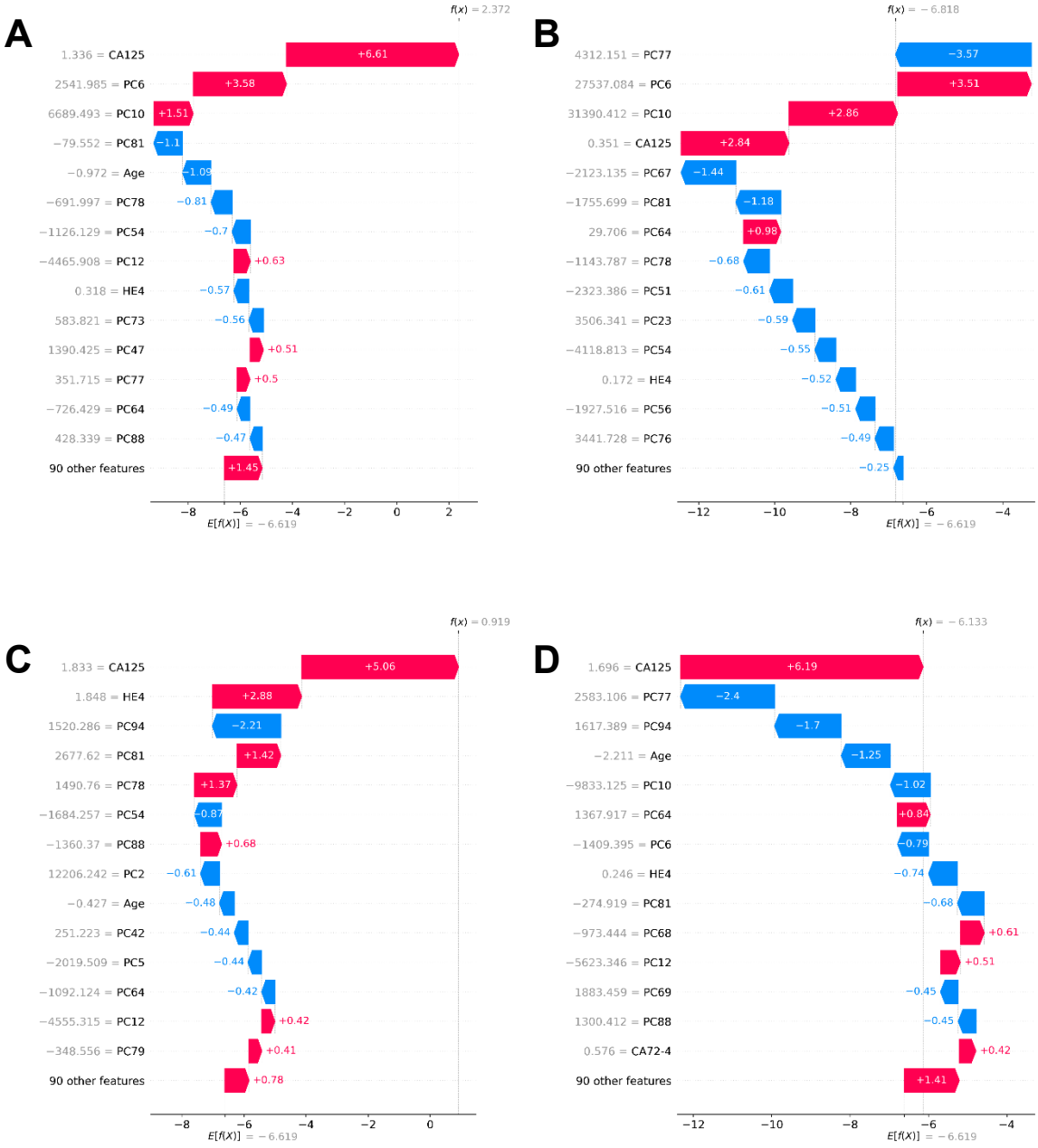
The study population included both Asian and Caucasian participants. Samples were randomly assigned to the training and hold-out cohorts, resulting in representation of both ethnic groups in each dataset. The present study was not designed or statistically powered to evaluate diagnostic performance separately by ethnicity. Therefore, additional validation in larger and more ethnically diverse prospective cohorts is required to determine the generalizability of the proposed framework across populations.

Additional File Table S10 Comparison of specificity at a fixed sensitivity of 90% for models with and without CA72-4

Model	Specificity	95% CI
With CA72-4	0.951	0.836–0.978
Without CA72-4	0.935	0.815–0.957

Specificity was evaluated at a fixed sensitivity of 90%.
Difference in specificity: 0.0155 (95% CI: -0.0611 to 0.0769)
Bootstrap: $p = 0.2496$
Exact McNemar: $p = 0.2101$

Specificity was evaluated at a fixed sensitivity of 90% to assess the contribution of CA72-4 in the high-specificity region relevant to ovarian cancer screening. Confidence intervals (95% CIs) were estimated via bootstrap resampling. The difference in specificity between models was evaluated using both bootstrap resampling and the exact McNemar test. Although the model including CA72-4 achieved a numerically higher specificity, the difference did not reach statistical significance, suggesting that the contribution of CA72-4 should be interpreted cautiously.



Additional File Fig. S1 SHAP waterfall plots for representative cases in the hold-out test cohort. **(A)** False-positive case (EM0002). **(B)** True-negative case (EM0006). **(C)** True-positive case (EOC0015). **(D)** False-negative case (EOC0003).

SHAP (SHapley Additive exPlanations) waterfall plots illustrate the contribution of individual features to the final model output (logit scale). The baseline value (E[f(X)]) represents the expected model output; red bars indicate features that increase the predicted probability of epithelial ovarian cancer (EOC), whereas blue bars indicate features that decrease it. The principal components represent orthogonal linear combinations of enriched glycopeptides and do not correspond to individual glycopeptide species. Only the top contributing features are shown, and the remaining features are aggregated. Biological interpretation was performed through downstream analyses of glycopeptides contributing strongly to key principal components.