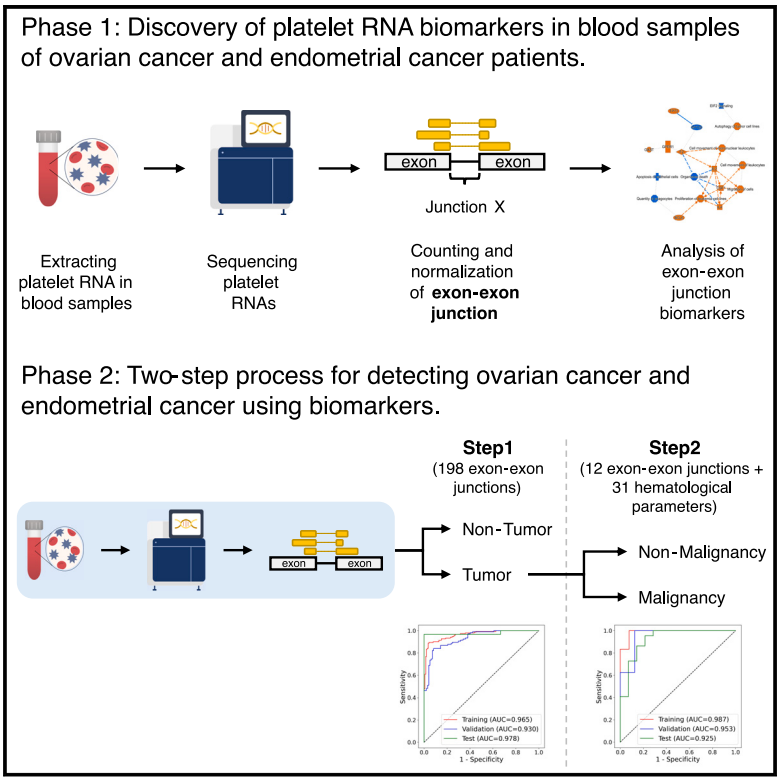
iScience

Article

Catalyzing early ovarian cancer detection: Platelet RNA-based precision screening

Graphical abstract



Highlights

- ISR in platelet transcriptome varies among cancer patients
- SII, PNI, and PLT/LYM ratio differ between benign and cancer groups
- Two-step model: Step 1 detects pelvic mass with >99% specificity
- Two-step model: Step 2 predicts tumor malignancy with >99% sensitivity

Authors

Eunyong Ahn, Se Ik Kim, Sungmin Park, ..., Cheol Lee, TaeJin Ahn, Yong-Sang Song

Correspondence

taejin.ahn@handong.edu (T.A.), yssong@snu.ac.kr (Y.-S.S.)

In brief

Health sciences; Medicine; Health informatics; Health technology

Ahn et al., 2025, iScience *28*, 112280
June 20, 2025
https://doi.org/10.1016/j.isci.2025.112280

CellPress

iScience

CellPress
OPEN ACCESS

Article

Catalyzing early ovarian cancer detection: Platelet RNA-based precision screening

Eunyong Ahn,[1,9] Se Ik Kim,[2,9] Sungmin Park,[1] Sarah Kim,[1] Hyejin Lee,[1] Yeochan Kim,[3] Sangick Park,[3] Suyeon Lee,[3] Dong Won Hwang,[2] Heeyeon Kim,[4,5] HyunA Jo,[4] Untack Cho,[4] Juwon Lee,[4] Cheol Lee,[6] TaeJin Ahn,[1,3,7,10,11,*] and Yong-Sang Song[8,10,*]

[1]Foretell My Health, Inc., 558 Handong-ro Buk-gu, Pohang 37554, Republic of Korea
[2]Department of Obstetrics and Gynecology, Seoul National University College of Medicine, Seoul 03080, South Korea
[3]School of Life Science, Handong Global University, Pohang 37554, Republic of Korea
[4]Cancer Research Institute, Seoul National University College of Medicine, Seoul 03080, Republic of Korea
[5]WCU Biomodulation, Department of Agricultural Biotechnology, Seoul National University, Seoul 08826, Republic of Korea
[6]Department of Pathology, Seoul National University College of Medicine, Seoul 03080, Republic of Korea
[7]Department of Advanced Convergence, Handong Global University, Pohang 37554, Republic of Korea
[8]Department of Obstetrics and Gynecology, Myongji Hospital, Gyeonggi-do 10475, Republic of Korea
[9]These authors contributed equally
[10]These authors contributed equally
[11]Lead contact
*Correspondence: taejin.ahn@handong.edu (T.A.), yssong@snu.ac.kr (Y.-S.S.)
https://doi.org/10.1016/j.isci.2025.112280

SUMMARY

Early detection of ovarian cancer is crucial for successful treatment, yet most cases are diagnosed at advanced stages due to a lack of effective screening. Recent advancements in RNA technology from platelets aid in early tumor detection. Here, we proposed our two-step method for assessing the existence of pelvic mass either located at ovaries or uterus with more than 99% specificity by utilizing exon-exon junction features with a sampling invariant normalization technique; then next our model finds the malignancy of detected mass with more than 99% negative predictive value for ovarian cancer to practically assist clinicians' further investigation via combined features of exon-exon junctions, and hematology parameters. We diverged from traditional methods by employing intron-spanning reads (ISR) counts rather than gene expression levels to use splice junctions as features in our models. If integrated with current screening methods, our algorithm holds promise for identifying ovarian or endometrial cancer in its early stages.

INTRODUCTION

Early detection of localized ovarian cancer holds the potential to ensure successful treatment for over 90% of affected women.[1] However, once the cancer metastasizes, the 5-year survival rate diminishes by less than 30%.[2] Unfortunately, due to the absence of disease-specific symptoms and effective screening tools, most cases of ovarian cancer are diagnosed at advanced stages.

To deal with this issue, the UK collaborative trial of ovarian cancer screening (UKCTOCS) was initiated to investigate whether population screening could reduce mortality rates associated with ovarian cancer.[3] Despite observing a decrease in the incidence of stage III or IV disease in the multimodal screening (MMS) group, comprised of longitudinal CA-125 and second-line transvaginal ultrasound scans, this reduction did not translate into a reduction in ovarian and tubal cancer deaths. Thus, they concluded that population-level general MMS screening for ovarian cancer cannot be recommended, necessitating the development of a screening strategy capable of detecting the disease earlier and in a larger proportion of women. Possible approach to high-risk women's ovarian cancer consists of screening intervals of 3–4 months and risk-reducing surgery, resulting in a significant reduction in the proportion of women diagnosed with advanced disease.[4,5] However, the extensive salpingo-oophorectomy could induce potential side effects associated with surgical removal, such as osteoporosis, cardiovascular disease, infection, or bleeding.[6,7] Therefore, we need to refine an early ovarian cancer detection method that considers the mechanisms behind the onset and advancement of ovarian cancer. This refinement is crucial for accurately distinguishing healthy individuals and effectively avoiding unnecessary extensive surgical procedures.

While inflammation and aging are speculated to play major roles in the initial stages of ovarian cancer development, gaps persist in understanding these mechanisms.[8] Chronic inflammation has long been associated with the development of various types of cancer, including ovarian cancer.[9] Inflammatory

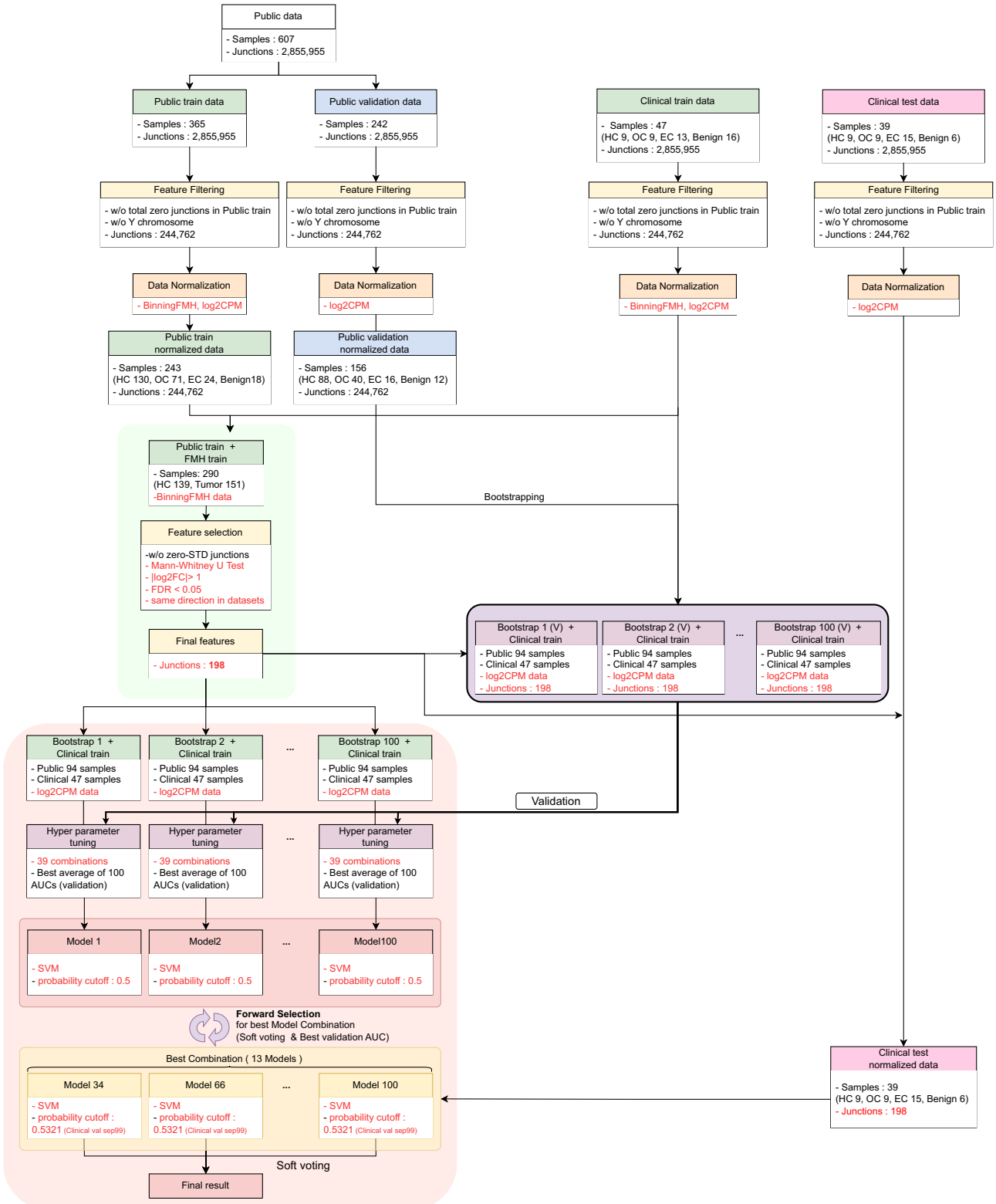


Figure 1. The flowchart of prediction model development

Table 1. Clinical characteristics of samples used for RNA sequencing analysis

Group	Stage	Age[a]	Weight	Height[a]	Albumin	CA-125
Healthy control	N.A.	40.5 (26.2, 61.8), $n = 18$	55.0 (52.1, 62.0), $n = 17$	160.0 (158.5, 161.0), $n = 17$	N.A.	N.A.
Benign	N.A.	47.0 (41.2, 52.8), $n = 22$	61.1 (56.4, 69.8), $n = 22$	161.9 (158.9, 163.0), $n = 22$	4.4 (4.3, 4.6), $n = 21$	30.0 (19.0, 59.8), $n = 11$
Cancer		54.0 (50.0, 61.0), $n = 46$	62.0 (55.9, 70.8), $n = 46$	157.7 (153.5, 161.0), $n = 46$	4.4 (4.2, 4.6), $n = 45$	46.8 (14.1, 214.7), $n = 42$
Ovarian cancer	I	52.5 (50.5, 62.0), $n = 6$	68.6 (60.0, 72.7), $n = 6$	159.8 (158.7, 160.2), $n = 6$	4.4 (4.2, 4.5), $n = 6$	172.0 (118.7, 196.8), $n = 6$
	II	56.0 (54.5, 57.5), $n = 2$	69.4 (65.2, 73.5), $n = 2$	157.2 (155.0, 159.3), $n = 2$	4.0 (3.7, 4.4), $n = 2$	764.0 (392.0, 1136.0), $n = 2$
	III	55.0 (51.0, 61.0), $n = 7$	61.4 (54.2, 70.0), $n = 7$	154.7 (153.4, 158.1), $n = 7$	4.2 (4.0, 4.3), $n = 7$	1332.0 (126.2, 1754.5), $n = 7$
	IV	51.0 (49.0, 51.5), $n = 3$	54.8 (49.0, 74.9), $n = 3$	159.1 (158.0, 160.9), $n = 3$	4.5 (4.2, 4.6), $n = 3$	578.0 (316.5, 880.5), $n = 3$
	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
Endometrial cancer	I	57.0 (54.0, 63.0), $n = 19$	60.0 (53.5, 65.8), $n = 19$	156.1 (152.2, 160.6), $n = 19$	4.5 (4.3, 4.7), $n = 19$	13.4 (5.8, 20.3), $n = 18$
	II	N.A.	N.A.	N.A.	N.A.	N.A.
	III	50.0 (50.0, 53.0), $n = 5$	66.6 (58.1, 71.3), $n = 5$	158.4 (156.8, 164.0), $n = 5$	4.4 (3.7, 4.6), $n = 5$	226.7 (175.2, 260.0), $n = 4$
	IV	N.A.	N.A.	N.A.	N.A.	N.A.
	N.A.	39.0 (32.5, 48.5), $n = 4$	65.9 (58.1, 75.2), $n = 4$	159.5 (156.6, 165.5), $n = 4$	4.6 (4.3, 4.8), $n = 3$	116.1 (73.0, 159.1), $n = 2$

Values represent the median, with the interquartile range [q25, q75] shown in brackets.
[a]p values from Wilcoxon test between Benign and Cancer are less than 0.05 after Bonferroni correction.

Table 2. Read composition of RNA-seq data

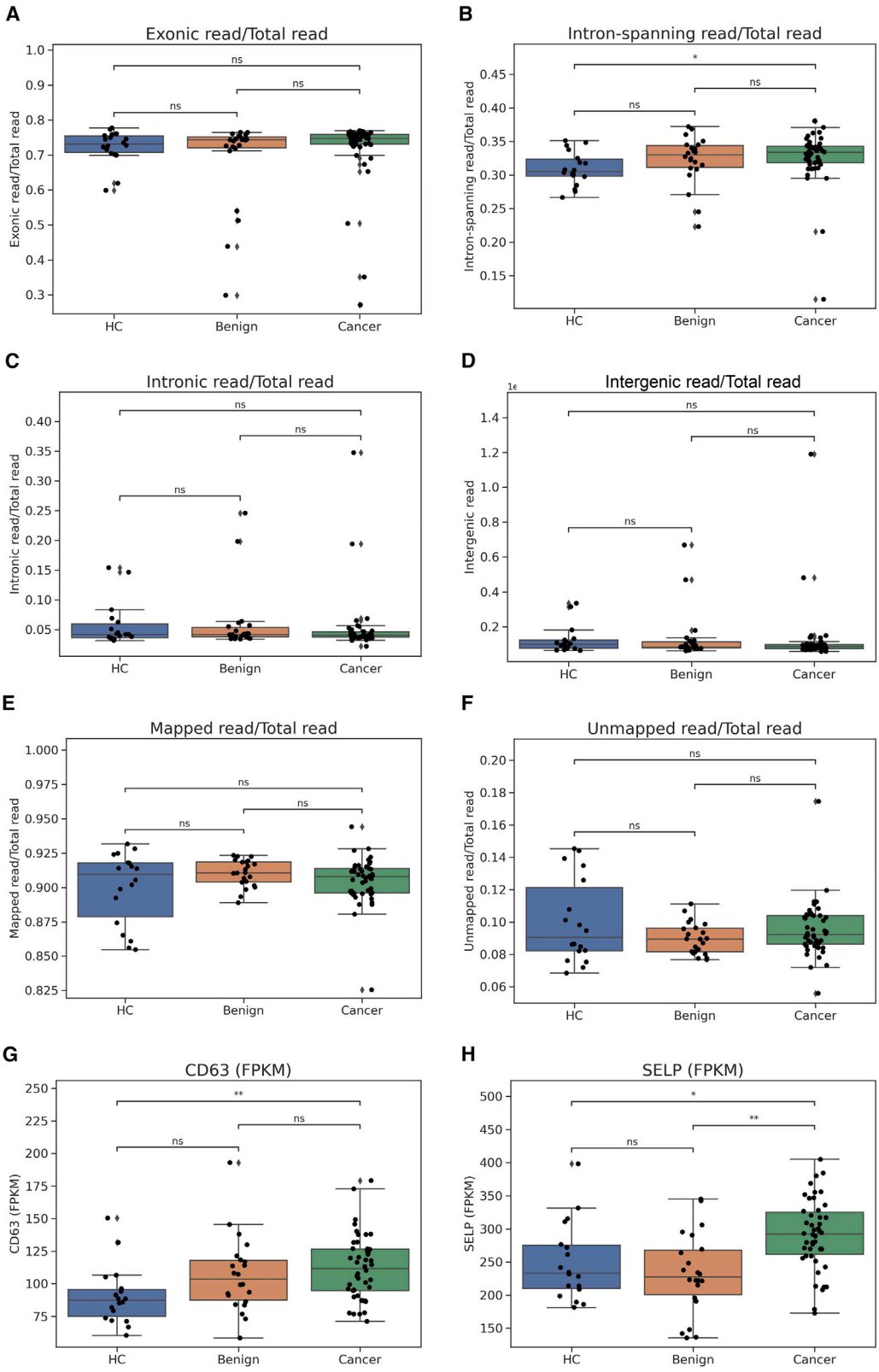
	Group		
	Healthy control ($n = 18$)	Benign ($n = 22$)	Cancer ($n = 46$)
exonic/total	0.73 (0.71, 0.75)	0.74 (0.72, 0.75)	0.75 (0.73, 0.76)
ISR/total	0.31 (0.3, 0.32)	0.33 (0.31, 0.34)	0.33 (0.32, 0.34)
intronic/total	0.04 (0.04, 0.06)	0.04 (0.04, 0.05)	0.04 (0.04, 0.05)
intergenic/total	0.02 (0.02, 0.03)	0.02 (0.02, 0.02)	0.02 (0.02, 0.02)
mapped/total	0.91 (0.88, 0.92)	0.91 (0.9, 0.92)	0.91 (0.9, 0.91)
unmapped/total	0.09 (0.08, 0.12)	0.09 (0.08, 0.1)	0.09 (0.09, 0.1)

Exonic: counts of exonic reads, ISR: counts of intron spanning reads, intronic: counts of intronic reads, intergenic: count of intergenic reads, mapped: count of mapped reads, unmapped: counts of unmapped read. p value from Wilcoxon test between Healthy control and Cancer is less than 0.05 after Bonferroni correction. Values represent the median, with the interquartile range (q25, q75) shown in brackets.

processes can create a microenvironment conducive to tumor initiation and progression.[10] Ovulation, which occurs regularly during a woman's reproductive years, involves repeated trauma to the ovarian epithelium.[11] This process can lead to inflammation and damage to the ovarian surface epithelium, creating opportunities for the initiation of cancerous changes. Thus, for ovarian cancer, chronic inflammation or repeated ovulation within the ovarian tissue may promote the accumulation of genetic mutations and the growth of abnormal cells.

Aging often coincides with hormonal changes, particularly a decline in estrogen levels and alterations in the balance of other hormones.[12] These hormonal changes can impact the ovarian microenvironment, potentially contributing to inflammation and the development of ovarian cancer.[8,13] Also, chronic inflammation over time in aging can induce DNA damage and impair DNA repair mechanisms within ovarian cells. Accumulated DNA damage may lead to the accumulation of mutations that promote malignant transformation. Also, aging is associated with changes in immune function, including alterations in the activity of immune cells and cytokine production.[14] Dysregulation of the immune system may contribute to chronic inflammation and impaired immune surveillance, allowing cancerous cells to evade detection and proliferate unchecked. Overall, the interplay between inflammation and aging creates a microenvironment within the ovaries that is conducive to the initiation and progression of ovarian cancer[10,15]; understanding these mechanisms is crucial for developing effective strategies for early detection, prevention, and treatment of ovarian cancer. Consequently, there is a pressing need to appreciate the initial pathophysiological changes attributed to "inflamm-aging", which cannot be reverted, initiating the disease.[8]

Platelets emerge as promising biomarkers for assessing "inflamm-aging" within epithelial ovarian tissue, given their involvement in the inflammatory cascade and their ability to reflect systemic inflammation.[16,17] Furthermore, platelets can interact directly or indirectly with ovarian tissue via the bloodstream, thus potentially reflecting local inflammatory processes.



(legend on next page)

Inflamm-aging, characterized by persistent low-grade inflammation accompanying aging,[18] is implicated in the pathogenesis of epithelial ovarian cancer, with platelets actively participating in the inflammatory response by releasing pro-inflammatory molecules.

Notable alterations in platelet count (PLT), structure, and levels of inflammatory markers have been observed in patients with epithelial ovarian cancer and other malignancies, suggesting their utility as diagnostic, prognostic, and monitoring tools.[19–21] Recent advancements in RNA technology from platelets aid in early tumor detection.[22] It is known that RNA from platelets can accurately detect ovarian cancer.[23] However, achieving accurate detection with high specificity is vital for population screening of diseases with low prevalence, such as ovarian cancer. Thus, here we proposed our two-step method for assessing the existence of gynecological tumors either located at ovaries or uterus with more than 99% specificity by utilizing exon-exon junction features with a sampling invariant normalization technique. Then next our model finds the malignancy of detected tumor with more than 99% negative predictive value to practically assist clinicians' further investigation via combined features of exon-exon junctions and hematology parameters.

RESULTS

We devised a two-stage approach to first anticipate the presence of pelvic masses and subsequently assess their malignancy using platelet transcriptome and hematology analysis. Although our primary objective was early ovarian cancer detection, we included uterine masses in our study. In clinical practice, once a pelvic mass is identified, further medical evaluation can easily differentiate between ovarian and uterine origins. This approach prevents overfitting in our machine learning model, as training solely on ovarian tumors could lead to limited generalizability across different types of gynecological masses.

For RNA sequencing analysis, age and cancer stage were matched when datasets were divided into train and test datasets 3 to 2 (Figure S1). The training and validation datasets included open-source data (n = 365 and n = 242, respectively, Figure 1), while our clinical data were utilized for training (n = 47) and testing (n = 39) for Model 1 and training (n = 46) and testing (n = 37) for Model 2. The clinical characteristics of the gynecological (ovarian or endometrial) cancer cases, benign tumor cases and healthy controls are presented in Table 1. Despite the age difference among the three groups, it is important to note that the risk of gynecological cancer typically rises post-menopause. Also, often in cancer cohort studies, the cancer patient group's age is generally higher than that of the healthy control group. Therefore, considering age-related disease prevalence, we do not believe that the performance of our model is influenced significantly by the differences in age across the groups. While CA-125 levels appeared to be generally higher in the cancer group, it's worth noting that these measurements were partially obtained from the benign group and did not demonstrate statistical significance.

Following preprocessing of Fastq files obtained from next-generation sequencing (NGS) analysis, the aligned reads were annotated utilizing the GRCh38 genomic reference. Upon comparing the composition of annotated reads, a notable alteration was observed solely in the composition of intron-spanning reads (ISR) within the cancer group not in that of exonic read, intronic read, intergenic read, mapped read, or unmapped reads (Tables 2 and S5 and Figure 2). The benign group also has an altered proportion of ISR levels compared to the control group, but a smaller sample size might be not enough to show statistical significance. An intron-spanning read is an RNA read that originates from a single region of a transcript, but it maps to two locations on the genome indicating the splice junction of two exons in the transcript. Given that platelets undergo alternative splicing in response to external disease-related signals,[24–26] it was anticipated that the alteration in ISR levels relative to total reads would be evident in the cancer group.

The expression levels of PTPRC, a leukocyte marker, remained unchanged in the tumor group, indicating that potential leukocyte contamination did not confound the observed differences in RNA profiles between the groups (Table S6). Additionally, expression levels of PF4, SELP, and CD63, which serve as markers of platelet activation, were assessed. Specifically, the FPKM (fragment per kilobase of transcript) level of SELP was significantly elevated in the cancer group compared to both the healthy control and benign groups. Overall, there was a discernible increase in CD63 expression levels observed from the healthy control to the benign and cancer groups, with the FPKM or TPM (transcripts per million) levels of CD63 in the cancer group being statistically higher than those in the healthy control group, as indicated by a p value of less than 0.05.

Platelet extraction from whole blood samples in EDTA-coated bottles was conducted, followed by hematology analysis of both

Figure 2. Analysis of read composition and RNA-seq data

Read composition of RNA sequencing data from Healthy control (HC), benign gynecological tumor (Benign), and cancer (Gynecological cancer) are shown. Data are presented as boxplots, where the box represents the interquartile range (IQR) from the first quartile (Q1) to the third quartile (Q3), and the horizontal line within the box indicates the median. The whiskers extend to data points within 1.5 times the IQR, while outliers are shown as diamond-shaped markers (◇). Each circular marker (●) represents an individual data point, providing a clear visualization of data distribution. Statistical significance was assessed using the Wilcoxon test. p values from Wilcoxon tests: * ≤0.05, ** ≤0.01, *** ≤0.001, **** ≤ 0.0001.

(A) The exonic read counts divided by total number of reads.

(B) The intron spanning read counts divided by total number of reads.

(C) The intronic read counts divided by total number of reads.

(D) The intergenic read counts divided by total number of reads.

(E) The mapped read counts divided by total number of reads.

(F) The unmapped read counts divided by total number of reads.

(G) The expression levels (FPKM) of CD63 gene.

(H) The expression levels (FPKM) of SELP gene.

Table 3. Hematology analysis results of whole blood samples

	Group		
	Healthy control	Benign	Cancer
WBC[a,c] (10^3/uL)	5.9 (4.83, 6.7) *n* = 42	4.9 (3.7, 6.3) *n* = 57	5.55 (4.5, 7.95) *n* = 78
RBC[b,c] (10^6/μL)	4.34 (4.18, 4.60) *n* = 42	3.94 (3.73, 4.31) *n* = 58	3.86 (3.54, 4.13) *n* = 78
PLT (10^3/μL)	249.5 (199, 273.25) *n* = 42	236.5 (193, 288) *n* = 58	255 (194.25, 312.25) *n* = 78
HGB[b,c] (g/dL)	13 (12.63, 13.67) *n* = 42	11.7 (10.9, 12.6) *n* = 58	11.85 (10.5, 12.5) *n* = 78
HCT[b,c] (%)	36.55 (35.9, 38.8) *n* = 42	33.65 (31.55, 35.78) *n* = 58	32.7 (29.7, 34.7) *n* = 78
MCV (fL)	85 (83.73, 86.95) *n* = 42	84.4 (82.2, 87.68) *n* = 58	85.2 (82.08, 87.18) *n* = 78
MCH (pg)	30.2 (29.23, 30.75) *n* = 42	30.05 (28.7, 31.13) *n* = 58	30.65 (28.95, 31.38) *n* = 78
MCHC (g/dL)	35.2 (34.7, 35.5) *n* = 42	35.2 (34.55, 36) *n* = 58	35.75 (34.73, 36.48) *n* = 78
LYM%[a,b]	33.4 (30.1, 40.1) *n* = 41	31.7 (27.9, 38.4) *n* = 57	21.75 (13.23, 28.7) *n* = 78
MXD%[a]	7 (6.2, 8.3) *n* = 41	7.3 (5.53, 10.35) *n* = 54	6.1 (4, 8.4) *n* = 77
NEUT%[a,b]	60 (53.2, 64.1) *n* = 41	61.55 (50.8, 66.05) *n* = 54	72.2 (63.7, 81.7) *n* = 77
LYM#[a,b,c] (10^3/uL)	1.9 (1.7, 2.2) *n* = 41	1.5 (1.2, 2.1) *n* = 57	1.2 (0.9, 1.78) *n* = 78
MXD# (10^3/μL)	0.4 (0.3, 0.5) *n* = 41	0.4 (0.3, 0.5) *n* = 54	0.3 (0.3, 0.5) *n* = 77
NEUT#[a] (10^3/uL)	3.1 (2.6, 4.3) *n* = 41	2.7 (1.9, 4.07) *n* = 54	3.9 (2.6, 5.9) *n* = 77
RDW-SD (fL)	41.4 (40.2, 43.4) *n* = 41	41.5 (40, 43.6) *n* = 57	40.55 (38.82, 42.6) *n* = 78
RDW-CV (%)	12.6 (12.1, 13.1) *n* = 41	12.75 (12.23, 13.48) *n* = 58	12.4 (11.8, 12.9) *n* = 78
PDW[c] (fL)	11.2 (10.4, 12.4) *n* = 41	12.2 (11.3, 13.3) *n* = 57	11.5 (10.5, 13) *n* = 77
MPV[a,c] (fL)	9.45 (9, 10.1) *n* = 42	10.2 (9.6, 10.6) *n* = 57	9.6 (9.1, 10.3) *n* = 77
P-LCR[a,c] (%)	21.6 (17, 24.9) *n* = 41	26.6 (21.4, 29.2) *n* = 57	22.2 (17.8, 27) *n* = 77
PCT (%)	0.24 (0.2, 0.26) *n* = 41	0.23 (0.2, 0.29) *n* = 57	0.25 (0.19, 0.32) *n* = 77
ResearchW[a,c] (10^3/μL)	5.89 (4.80, 6.74) *n* = 41	4.62 (3.64, 6.29) *n* = 57	5.55 (4.46, 7.95) *n* = 78
ResearchS[a,b,c] (10^3/μL)	1.91 (1.70, 2.21) *n* = 41	1.52 (1.19, 2.12) *n* = 57	1.21 (0.92, 1.74) *n* = 78
ResearchM (10^3/μL)	0.37 (0.33, 0.46) *n* = 41	0.38 (0.31, 0.49) *n* = 54	0.35 (0.25, 0.51) *n* = 77
ResearchL[a,c] (10^3/μL)	3.4 (2.65, 4.33) *n* = 41	2.68 (1.86, 3.83) *n* = 54	3.92 (2.63, 5.91) *n* = 77

The table displays each hematological parameter's median, 25%, and 75% quantiles. WBC: white blood cells, RBC: red blood cells, PLT: platelets, HGB: hemoglobin, HCT: the percentage of red blood cells in blood, MCV: Mean corpuscular volume, the average size of red blood cells, MCH: mean corpuscular hemoglobin, the average amount of hemoglobin within red blood cells, MCHC: mean corpuscular hemoglobin concentration, the average concentration of hemoglobin within red blood cells, LYM%: relative amounts of lymphocytes in WBC, MXD%: relative amounts of monocytes, eosinophils, and basophils in WBC. NEUT%: relative amounts of neutrophils in WBC, LYM: lymphocytes, MXD: monocytes, eosinophils, and basophils, NEUT: neutrophils, RDW-SD: the width of red cells size distribution histogram, RDW-CV: coefficient of variation of mean corpuscular volume, PDW: platelet distribution width, MPV: mean platelet volume, P-LCR: platelet larger cell ratio, the percentage of platelets that exceed the normal value of platelet volume of 12 fL in the total platelet count, PCT: plateletcrit, the volume occupied by platelets in the blood, ResearchW: number of cells between LD (lower detection line) and UD (upper detection line), ResearchS: number of cells between LD (lower detection line) and T1 (trough 1 line), ResearchM: number of cells between T1 (trough 1 line) and T2 (trough 2 line), ResearchS: number of cells between T2 (trough 2 line) and UD (upper detection line).

Values represent the median, with the interquartile range (q25, q75) shown in brackets.

[a]*p* values from Wilcoxon test between Benign and Cancer are less than 0.05 after Bonferroni correction.

[b]*p* values from Wilcoxon test between Healthy control and Cancer are less than 0.05 after Bonferroni correction.

[c]*p* values from Wilcoxon test between Healthy control and Benign are less than 0.05 after Bonferroni correction.

whole blood and platelet-rich plasma. Comparative analysis across three groups—healthy controls, benign, and cancer groups—revealed statistically significant variances in multiple hematological parameters. Notably, the healthy control group exhibited distinct disparities from the benign and cancer groups. Tables 3 and 4 (Figure 3) encapsulate the pivotal findings.

Due to the exclusive focus on hematology analysis for certain samples, the sample size for this analysis exceeds that of RNA sequencing. A reduction in white blood cell (WBC) and red blood cell (RBC) counts was observed in the cancer group. The median WBC count for the healthy control group was 5.9 × 10^3/μL, contrasting with 4.9 × 10^3/μL for the benign group and 5.55 × 10^3/μL for the cancer group. These discrepancies suggest potential alterations in immune status. Similarly, the median RBC count for the healthy control group was 4.34 × 10^6/μL, whereas the benign and cancer groups displayed counts of 3.94 × 10^6/μL and 3.86 × 10^6/μL, respectively, indicating disruptions in erythropoiesis within the pathological groups.

Furthermore, reductions in hemoglobin (HGB) and hematocrit levels were observed in both the benign and cancer groups, while mean corpuscular volume , mean corpuscular hemoglobin (MCH), and MCH concentration (MCHC) remained unchanged. The distribution of red blood cells remained unaffected by pathological conditions, although there were notable shifts in the

Table 4. Hematology analysis results of platelet-rich plasma samples

	Group		
	Healthy control	Benign	Cancer
PLT ($10^3/\mu L$)	407 (356.25, 476.5) n = 42	360 (242, 472.25) n = 58	373 (282, 456) n = 77
PDW[a] (fL)	10.8 (10.1, 11.7) n = 41	11.35 (10.6, 12.17) n = 58	10.6 (9.9, 11.6) n = 77
MPV[a,b] (fL)	9 (8.6, 9.58) n = 42	9.45 (8.9, 9.9) n = 58	8.9 (8.3, 9.4) n = 77
P-LCR[a,b] (%)	17.7 (14.9, 21.7) n = 41	20.85 (16.65, 24.63) n = 58	17.2 (13.8, 21.2) n = 77
PCT[c] (%)	0.37 (0.33, 0.44) n = 41	0.36 (0.25, 0.44) n = 58	0.33 (0.25, 0.42) n = 77

PLT: platelets, PDW: platelet distribution width, MPV: mean platelet volume, P-LCR: platelet larger cell ratio, the percentage of platelets that exceed the normal value of platelet volume of 12 fL in the total platelet count, PCT: plateletcrit, the volume occupied by platelets in the blood, ResearchW: number of cells between LD (lower detection line) and UD (upper detection line).

Values represent the median, with the interquartile range (q25, q75) shown in brackets.

[a]p values from Wilcoxon test between Benign and Cancer are less than 0.05 after Bonferroni correction.

[b]p values from Wilcoxon test between Healthy control and Benign are less than 0.05 after Bonferroni correction.

[c]p values from Wilcoxon test between Healthy control and Cancer are less than 0.05 after Bonferroni correction.

proportions of various WBC types. Specifically, lymphocyte percentage (LYM%) and mixed cell percentage (MXD%) decreased, while neutrophil percentage (NEUT%) increased significantly. The decline in lymphocyte count was observed progressively from the healthy control to the benign and cancer groups.

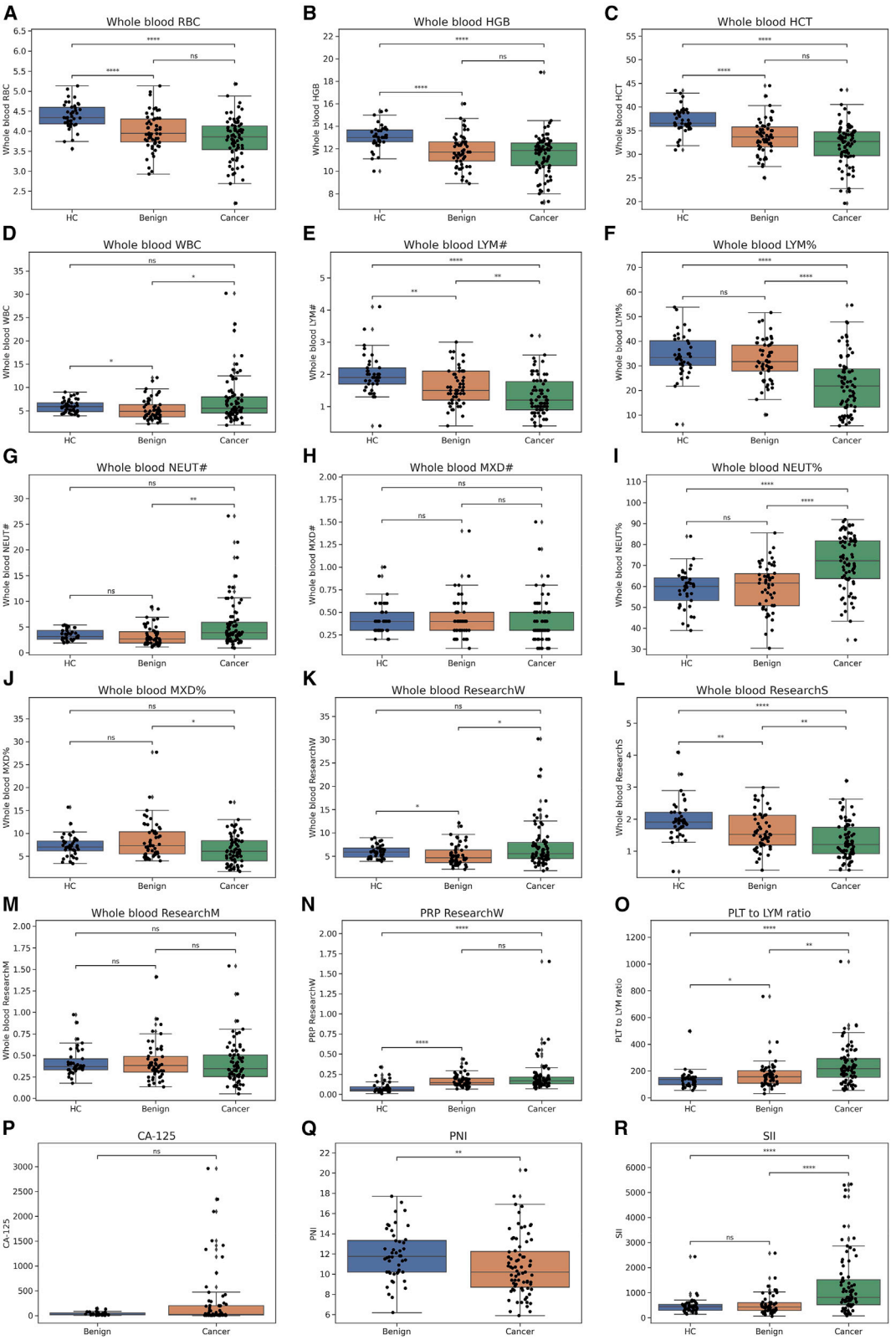
It is noteworthy that the difference in neutrophil count between the cancer and benign groups was statistically significant, whereas such distinctions were not evident in the other groups. Parameters such as ResearchW, ResearchS, ResearchM, and ResearchL, provided by Sysmex and indicative of WBC constitution, also exhibited significant disparities between the benign and cancer groups. The parameters ResearchW, ResearchS, ResearchM, and ResearchL, provided by the Sysmex hematology analyzer, are indicative of different WBC subpopulations. Specifically, ResearchW reflects the overall proportion of WBCs, ResearchS corresponds to small-sized WBC subpopulations mainly lymphocytes, ResearchM represents medium-sized WBC subpopulations mainly mixed cells, and ResearchL relates to large-sized WBC subpopulations, mainly neutrophils. These parameters are proprietary outputs from the Sysmex analyzer and are utilized to provide a detailed assessment of WBC distribution and characteristics. In platelet-rich plasma analysis, although there were slight variations, no significant differences were observed in the PLT between the groups. Platelet distribution width (PDW), which reflects the variation in platelet size, mean platelet volume , indicating the average size of platelets, platelet-large cell ratio (P-LCR), a measure of the proportion of large platelets, and plateletcrit (PCT), which represents the volume fraction of platelets in the blood, were slightly lower in the cancer group compared to the benign group.

We also measured the blood markers utilized for diagnosing cancer (Table 5). In advanced-stage patients, elevated CA-125 levels were observed; however, no significant disparity in CA-125 levels was noted between benign and cancerous cases. In our analysis, the mean CA-125 level was 281.496 (SD = 590.096) in the cancer group (n = 67) and 39.533 (SD = 37.168) in the benign group (n = 27). Although the mean CA-125 level was higher in the cancer group, the substantial variability in CA-125 levels, particularly in the cancer group, contributed to the lack of significant difference between the two groups. Conversely, hematological parameters such as systemic immune-inflammation index (SII), prognostic nutritional index (PNI), and PLT to LYM ratio exhibited notable variances between benign and cancer groups. Consequently, we further devised a model utilizing these hematological parameters to distinguish between benign and cancerous conditions as flowing paragraphs.

We performed Pearson's correlation analysis to identify potential associations or confounders among the molecular features described in the manuscript and the hematological measurements. The results of this analysis are presented as a heatmap (Figure S2), which highlights correlations across a range of features. Notably, several molecular features demonstrated associations with hematological parameters, indicating potential relationships between these variables. For example, we observed relatively high levels of association between whole blood WBC and whole blood NEUT#, as well as between whole blood LYM# and PNI, suggesting potential relevance of these correlations in the observed data patterns.

Additionally, we addressed potential age-related biases by randomly sampling half of the individuals in their 20s from the HC group and applying propensity score matching based on age to the benign and cancer groups. Statistical analyses confirmed no significant age differences between the matched groups (HC vs. Benign, p = 0.64; HC vs. Cancer, p = 0.13), ensuring that age is not a confounding factor in our analysis. The analysis results from age matched samples are provided in Tables S7–S9.

Given that ISR read counts in platelet transcriptome effectively capture the pathophysiological traits in cancer patients, we devised a machine learning model utilizing ISR counts from platelet RNA sequencing data to detect gynecological tumors. Prior to constructing the tumor prediction model, we undertook an investigation into normalization methods to identify the most suitable approach for platelet transcriptome data. Owing to the characteristics of platelet transcriptome and the scarcity of alternative splicing, numerous undetectable variants are recorded as 0, resulting in an extremely skewed dataset with a zero-inflated distribution. Methods such as rank-based methodology or variance stabilizing transformation aim to mitigate the variance of



(legend on next page)

zero-inflated data initially processed with log_2CPM, but they may obscure the effects of markers. In contrast, the binning FMH methodology preserves the quantitative attributes of the data preprocessed with log_2CPM, facilitating the identification of marker effects (Figure S3A). Furthermore, to prevent overfitting, we adopted an analysis approach that restricts feature selection to a minimal set at the statistical level (Figure S3B).

Thus, we selected 198 features using the binning FMH and developed the bootstrapped ensemble Model 1 to differentiate between gynecological tumor and non-tumor groups. The area under the curve (AUC) of our tumor prediction model for the training, validation, and test datasets were 0.965 (95% CI: 0.937, 0.981), 0.930 (95% CI: 0.886, 0.958), and 0.978 (95% CI: 0.874, 0.997), respectively (Figures 4 and S4 and Table S10). Particularly, noteworthy is the pre-defined cut-off value for the test dataset based on the validation set, yielding an accuracy, sensitivity, specificity, and balanced accuracy (the average of sensitivity and specificity) of 0.974, 0.967, 1, and 0.983, respectively. Additionally, we generated RNA sequencing data separately across 9 distinct batches (Figure S1), rather than as a single dataset. The high specificity achieved using the pre-defined cut-off value is attributed to a discrimination algorithm we developed, which leverages bootstrapping and ensembles public data combined with noise to mitigate overfitting of self-produced data and enhance model performance. Notably, without incorporating public data, the error rate escalates by 5-fold (see Figure S4).

Following the development of Model 1 for predicting gynecological tumors, we proceeded to devise Model 2 for predicting the malignancy of tumors identified by Model 1. In this phase, our primary focus was on enhancing the sensitivity of Model 2 to minimize the likelihood of overlooking cancer patients in our malignancy assessments. To achieve this objective, we expanded the scope of our model by incorporating 31 hematological parameters alongside the 9 combinatorial features (12 junctions) in ISR derived from RNA-sequencing data. The AUC for our ovarian cancer prediction model in the training and test datasets are 0.987 (95% CI: 0.845, 0.999) and 0.925 (95% CI: 0.793, 0.976), respectively (Figure 4). Notably, for Model 2, we achieved an accuracy, sensitivity, specificity, and balanced accuracy of 0.861 (95% CI: 0.713, 0.939), 0.909 (95% CI: 0.722, 0.975), 0.786 (95% CI: 0.524, 0.924), and 0.847 (95% CI: 0.713, 0.939), respectively (Table S11). Notably, only for ovarian cancer sensitivity, Model 2 achieved Sensitivity 1.0.

We evaluated the predictive performance of CA-125, both independently and in combination with our derived model (Model 2). Since the HC group lacks CA-125 values, both CA-125 and Model 2 were trained using only malignant and non-malignant tumors to ensure an accurate comparison of the performance. The CA-125 performance was assessed using logistic regression with a 5-fold cross-validation approach. For CA-125, the AUC in the training set was 0.581 (95% CI: 0.374–0.763), and in the test set, the AUC was 0.583 (95% CI: 0.401–0.745). This suggests that while CA-125 has moderate discriminative ability on its own, the performance is not optimal. Combining Model 2 with CA-125, trained on the same dataset, and showed significantly better performance. The AUC in the training set was 1.000 (95% CI: 0.845–1.000), and in the test set, it was 0.917 (95% CI: 0.758–0.975). To explore the potential benefit of combining CA-125 with Model 2, we employed a soft voting method, giving equal weight to the predicted probabilities of both models. This combined approach yielded improved performance metrics, indicating that the combination of CA-125 and Model 2 offers a more robust predictive tool compared to CA-125 alone. These findings underscore the importance of incorporating Model 2 in clinical settings to enhance the predictive accuracy when used alongside CA-125.

To appreciate the characteristics of selected 198 features (95 genes) in the Model1, we performed various functional analysis, including differential network analysis, ingenuity pathway

Figure 3. Hematology analysis results and CA-125 level

Hematological parameters and CA-125 levels in Healthy control (HC), Benign gynecological tumor (Benign), and Cancer (Gynecological cancer) groups are shown. *p* values from Wilcoxon tests: * $\leq$0.05, ** $\leq$0.01, *** $\leq$0.001, **** $\leq$0.0001.

(A) The red blood cell counts in whole blood (10^6/uL).

(B) The hemoglobin levels in whole blood (g/dL).

(C) The hematocrits in whole blood (%).

(D) The white blood cell count in whole blood (10^3/uL).

(E) The lymphocyte counts in whole blood (10^3/uL).

(F) The lymphocyte percentage among white blood cells in whole blood (%).

(G) The neutrophil counts in whole blood (10^3/uL).

(H) The mixed cell counts in whole blood (10^3/uL).

(I) The neutrophil percentage among white blood cells in whole blood (%).

(J) The mixed cell percentage among white blood cells in whole blood (%).

(K) The ResearchW in whole blood (10^3/uL).

(L) The ResearchS in whole blood (10^3/uL).

(M) The ResearchM in whole blood (10^3/uL).

(N) The ResearchW in PRP (10^3/uL).

(O) The platelet to lymphocyte ratio in whole blood.

(P) The CA-125 levels in whole blood.

(Q) The PNI levels in whole blood.

(R) The SII levels in whole blood. Data are presented as boxplots following the same conventions as described in Figure 2, where the box represents the interquartile range (IQR) from the first quartile (Q1) to the third quartile (Q3), with the central line indicating the median. The whiskers extend to data points within 1.5 times the IQR, while outliers are shown as diamond-shaped markers (◇). Each circular marker (●) represents an individual data point, providing a clear visualization of data distribution.

Table 5. Cancer-related blood makers

	Group		
	Healthy control	Benign	Cancer
CA-125 (units/mL)		22 (15, 54.25) *n* = 27	26 (13.1, 201.8) *n* = 67
SII[a,b]	442.75 (300.88, 530.77) *n* = 41	424.08 (292.87, 597.52) *n* = 54	816 (521.33, 1517.08) *n* = 77
PNI[a]		11.75 (10.2, 13.33) *n* = 48	10.2 (8.7, 12.25) *n* = 75
PLT to LYM ratio[a,b,c]	136.47 (97.27,150.53) *n* = 41	155 (107.41, 200.91) *n* = 57	215.9 (152.4, 292.38) *n* = 78
PDW to PLT ratio	0.05 (0.039, 0.056) *n* = 41	0.05 (0.04, 0.07) *n* = 57	0.04 (0.04, 0.06) *n* = 77
PLT to Albumin ratio		53.72 (43.81, 63.41) *n* = 49	58.81 (47.13, 70.1) *n* = 75
RDW to PLT ratio	0.17 (0.15, 0.21) *n* = 41	0.18 (0.14, 0.22) *n* = 57	0.16 (0.13, 0.21) *n* = 78
MPV to PLT ratio	0.039 (0.033, 0.048) *n* = 42	0.04 (0.03, 0.05) *n* = 57	0.04 (0.03, 0.05) *n* = 77

Systemic immune-inflammation index (SII) is calculated by (N×P)/L (N, P and L represent neutrophil counts, platelet counts and lymphocyte counts, respectively, $10^3/\mu L$). Prognostic nutritional index (PNI) is calculated by 10 × serum albumin (g/dL) + 0.005 × total lymphocyte count (per uL). Values represent the median, with the interquartile range [q25, q75] shown in brackets.

[a]*p* values from Wilcoxon test between Benign and Cancer are less than 0.05 after Bonferroni correction.

[b]*p* values from Wilcoxon test between Healthy control and Cancer are less than 0.05 after Bonferroni correction.

[c]*p* values from Wilcoxon test between Healthy control and Benign are less than 0.05 after Bonferroni correction.

analysis (IPA) analysis, and gene regulatory network (GRN) analysis. IPA was employed to ascertain the canonical pathways linked with 198 junction markers (comprising 95 genes) between the cancer and noncancer cohorts (Figure 5). The determination of canonical pathways was based on their respective *p* values. Among the pathways associated with the 198 junction markers (involving 95 genes) within the IPA reference dataset, 9 datasets, which are mainly related to transcription and translation, were found to be downregulated while 6 datasets, which are mainly related to immune response, were upregulated (Figure S5).

We also conducted GRN analysis to elucidate potential signaling mechanisms involved in gynecological cancer. Here, we constructed the GRN network using 95 genes from our 198 junction markers and retained the top 100 edges from cancer group. In order to concentrate on the mechanisms of oncogenesis, we separated the sample into cancer and non-cancer groups. In the cancer group, *UBQLN1* and *E2F1* emerged as a significant hub gene and a transcription factor, respectively, within the GRN networks (Figure S6).

DISCUSSION

Here, we described a two-step method that will practically aid physicians in their subsequent medical follow-up after using our algorithm. Although previous diagnosis guidelines for ovarian cancer could not exclusively find patients with ovarian cancer or tumors in their reproductive organs, our model uses a combination of exon-exon junction, transcriptome totality, and hematological data to determine the malignancy of the identified tumor with a negative predictive value of greater than 99% for ovarian cancer. Because usually these two types of cancers are easily differentiated and proximal, the metastasis between these two cancers is feasible. Thus, we developed an algorithm to predict ovarian or endometrial cancers together. Notably, we observed significant alterations in ISR in cancer samples, suggesting potential pathophysiological mechanisms related to quantitative regulation in splicing of platelet RNA. Hematological analysis also revealed significant differences between pathological and healthy groups. Notably, functional analyses identified key genes and pathways associated with gynecological cancer, highlighting the potential of platelet transcriptome analysis in advancing precision oncology.

Notably, we diverged from traditional methods by employing ISR counts rather than gene expression levels in our model development to use splice junctions as features in our models. Conventionally, gene expression levels in platelets have been quantified using ISR alone, excluding other subtype reads from sequencing files to prevent contamination from leukocyte DNA.[27] However, our methodology introduces a novel dimension by focusing on quantitative changes in ISR counts, enabling us to pinpoint specific splicing alterations with precision. Our findings, exemplified in Figure S7, demonstrate instances where disease-associated statistically significant quantitative alterations in ISR were detected (*p* value from Wilcoxon test <1.00e-10), whereas corresponding changes at the gene level were not observed (*p* value from Wilcoxon test >0.5). This underscores the enhanced sensitivity and specificity of our approach in capturing subtle molecular signatures indicative of cancer, potentially revolutionizing diagnostic paradigms in oncology.

In Model 2, pairwise differences were examined in RNA expression from tumor-educated platelets (TEPs) to construct a diagnostic model for ovarian cancer. The identified gene pairs—*MAX-DAPP1*, *MAX-MTPN*, *MTPN-PTGS1*, *KIF2A-TSPAN33*, *TSPAN33-MTPN*, and *ACTN1-MTPN*—demonstrated significant differences in expression patterns between ovarian cancer patients and controls. These differences reflect altered platelet RNA profiles influenced by systemic cancer-associated processes rather than direct molecular interactions or causative relationships. In cancer, tumor-derived signals are known to reprogram platelets through mechanisms such as RNA transfer, selective RNA packaging, and alternative splicing. These processes collectively alter platelet RNA profiles, which can then serve as a valuable source of biomarkers for cancer detection. For example, *MTPN* (myotrophin), implicated in cytoskeletal dynamics, appears in several key gene pairs, potentially reflecting tumor-induced changes in platelet structure and

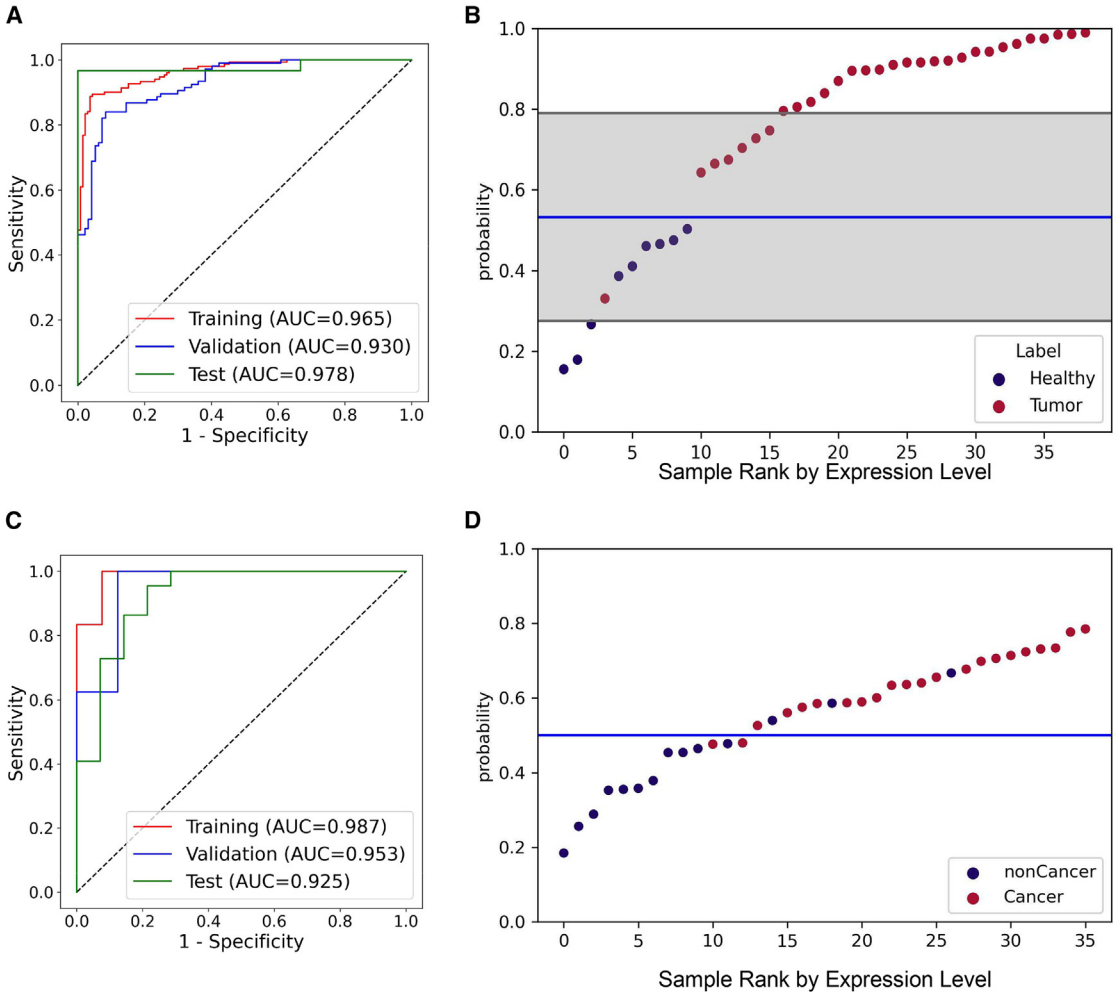


Figure 4. The performance results of prediction models

(A and B) Model 1 for gynecological tumor detection: (A) ROC curves of training, validation, and test datasets.

(B) Probability output is illustrated in a plot. Gray colored region is the region between the highest value of healthy from training set and the lowest value of tumor from training set. Blue line is cut-off value decided based on validation dataset.

(C and D) Model 2 for malignancy of tumor prediction: (C) ROC curves of training, and test datasets (D) Probability output is illustrated in a plot. Blue line is cut-off value decided based on training dataset.

function.[28] Similarly, the consistent involvement of *KIF2A* (kinesin family member 2A), paired with *TSPAN33* (tetraspanin 33), may indicate modifications in platelet intracellular transport mechanisms in response to tumor signals.[29,30]

In our investigation, distinctive alterations in platelet characteristics, along with broader changes in hematology parameters, were observed in cancer patients. While our study primarily focused on developing a classification model, it is conceivable that similar approaches could be extrapolated to other cancer types in the future. The identification of specific platelet transcriptome changes associated with cancer underscores the potential of our methodology to serve as a versatile tool for cancer diagnosis across various malignancies. Moving forward, exploring these possibilities could lead to the development of comprehensive diagnostic strategies with widespread applicability in oncology. Such advancements hold promise for enhancing early detection and personalized management approaches for cancer patients.

Surgery is the mainstay in the management of gynecologic malignancies. Especially for ovarian cancer, open surgery, rather than minimally invasive surgery, is the current standard. Herein, our newly developed model with high diagnostic performance or sensitivity may aid both physicians and patients in determining the surgical approach: open surgery or minimally invasive surgery. For example, if a woman is expected to have ovarian cancer rather than benign ovarian tumors, she might avoid minimally invasive surgery, which might result in tumor leakage or spillage and intraperitoneal dissemination. Instead, physicians and patients might choose laparotomy for such high-risk women. These findings underscore the transformative potential of platelet transcriptome analysis in advancing precision oncology.

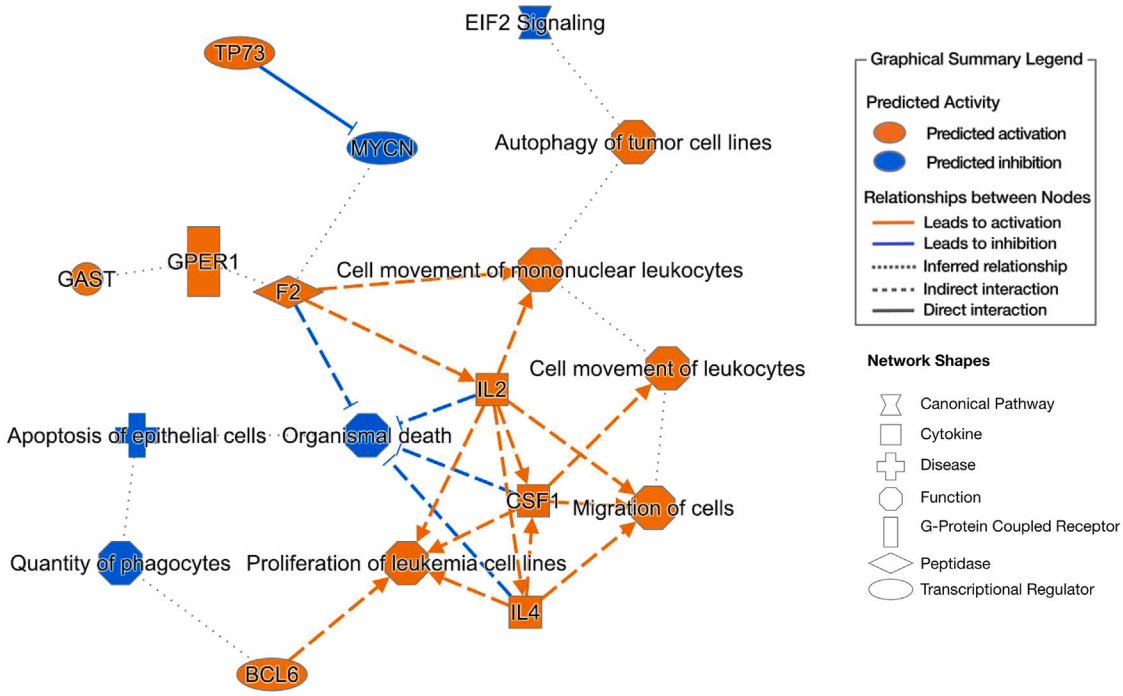


Figure 5. The functional analysis of genes in prediction models

Graphical summary showing the most significant transcription factors and regulators affected by 198 junction (95 genes) markers. The orange color goes with the direction of activation while the blue is the inhibition state.

Limitations of the study

Our findings suggest that the platelet transcriptome can serve as an early biomarker for cancer. However, this study has certain limitations. Notably, it does not provide direct evidence demonstrating that platelets influence the etiology of cancer progression. For example, inflammation and menopause are significant risk factors for ovarian cancer and are linked to changes in platelet physiology. While these factors imply a potential connection between platelets and the etiology of ovarian cancer, our research did not include experimental evidence to substantiate this association. Additionally, our machine learning model was constructed using a limited dataset. Our hematology analyses, though revealing certain trends, may not provide specific markers uniquely associated with the presence of ovarian cancer. As such, their broader applicability in a screening setting is limited, and further research is needed to identify more cancer-specific hematologic markers. Consequently, it may not accurately identify subtype-specific features necessary to distinguish particular types of cancer within the population. This limitation could also be influenced by variations in ethnicities or hereditary factors.

RESOURCE AVAILABILITY

Lead contact

Requests for further information should be directed to and will be fulfilled by the lead contact, TaeJin Ahn (taejin.ahn@handong.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The raw sequencing data in FASTQ format have been archived in the NCBI GEO database under the accession number GEO: GSE278917 and are publicly accessible as of the publication date.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

This work was supported by grants from Korea Health Industry Development Institute (HI16C2037), Korea Foundation for Cancer Research (CB-2022-A-2), Ministry of SMEs and Startups (RS-2023-00240924), and National Research Foundation of Korea (RS-2022-NR073612). The biospecimens for this study were provided by the Seoul National University Hospital Human Biobank, a member of the National Biobank of Korea, which is supported by the Ministry of Health and Welfare. All samples, derived from the National Biobank of Korea, were obtained with informed consent under institutional review board-approved protocols.

AUTHOR CONTRIBUTIONS

E.A., S.I.K., T.A., and Y.-S.S. designed and supervised the research. S.P. and S.K. organize the clinical study and lead computational methodology development. S.P., S.K., Y.K., S.P., and S.L. performed Bioinformatic analysis. H.L., H.K., H.J., U.C., and J.L. performed experiments. S.I.K., D.W.H., and Y.-S.S. recruited patients for clinical research. C.L. reviewed and confirmed all the cases in our study. E.A. wrote the first version of manuscript. All authors revised and approved the final version of manuscript.

DECLARATION OF INTERESTS

E.A., S.I.K., S.P., S.K., T.A., and Y.-S.S. are stockholders of Foretell My Health Ltd. We affirm that this conflict of interest does not affect the integrity,

objectivity, or validity of the research conducted. All efforts have been made to ensure that the research is conducted impartially and accurately. Patent applications related to some part of these findings in this study have been filed by the authors.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Clinical specimens
  - Ethics approval and consent to participate
- METHOD DETAILS
  - Hematology analysis and platelet RNA extraction
  - RNA-sequencing
  - Opensource dataset
  - Preprocessing of RNA-seq data
  - Developing machine learning-based gynecological tumor diagnosis model
  - Developing machine learning-based malignancy prediction model
  - Functional analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2025.112280.

Received: April 29, 2024
Revised: October 6, 2024
Accepted: March 19, 2025
Published: March 24, 2025

REFERENCES

1. Hayat, M.J., Howlader, N., Reichman, M.E., and Edwards, B.K. (2007). Cancer Statistics, Trends, and Multiple Primary Cancer Analyses from the Surveillance, Epidemiology, and End Results (SEER) Program. Oncologist *12*, 20–37. https://doi.org/10.1634/theoncologist.12-1-20.
2. Badgwell, D., and Bast, R.C. (2007). Early Detection of Ovarian Cancer (IOS Press).
3. Menon, U., Gentry-Maharaj, A., Burnell, M., Singh, N., Ryan, A., Karpinskyj, C., Carlino, G., Taylor, J., Massingham, S.K., Raikou, M., et al. (2021). 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. https://doi.org/10.1016/S0140-6736(21)00731-5.
4. Rosenthal, A.N., Fraser, L.S.M., Philpott, S., Manchanda, R., Burnell, M., Badman, P., Hadwin, R., Rizzuto, I., Benjamin, E., Singh, N., et al. (2017). United Kingdom Familial Ovarian Cancer Screening Study collaborators. Evidence of stage shift in women diagnosed with ovarian cancer during phase II of the United Kingdom familial ovarian cancer screening study. J. Clin. Oncol. *35*, 1411–1420.
5. Skates, S.J., Greene, M.H., Buys, S.S., Mai, P.L., Brown, P., Piedmonte, M., Rodriguez, G., Schorge, J.O., Sherman, M., Daly, M.B., et al. (2017). Early detection of ovarian cancer using the risk of ovarian cancer algorithm with frequent CA125 testing in women at increased familial risk–combined results from two screening trials. Clin. Cancer Res. *23*, 3628–3637.
6. Greene, M.H., Piedmonte, M., Alberts, D., Gail, M., Hensley, M., Miner, Z., Mai, P.L., Loud, J., Rodriguez, G., Basil, J., et al. (2008). A prospective study of risk-reducing salpingo-oophorectomy and longitudinal CA-125 screening among women at increased genetic risk of ovarian cancer: Design and baseline characteristics: A gynecologic oncology group study. Cancer Epidemiol. Biomarkers Prev. *17*, 594–604. https://doi.org/10.1158/1055-9965.EPI-07-2703.
7. Erekson, E.A., Martin, D.K., and Ratner, E.S. (2013). Oophorectomy: The debate between ovarian conservation and elective oophorectomy. Menopause *20*, 110–114. https://doi.org/10.1097/gme.0b013e31825a27ab.
8. Sánchez-Prieto, M., Sánchez-Borrego, R., Lubián-López, D.M., and Pérez-López, F.R. (2022). Etiopathogenesis of ovarian cancer. An inflamm-aging entity? Gynecol Oncol *42*, 101018. https://doi.org/10.1016/j.gore.2022.101018.
9. Multhoff, G., Molls, M., and Radons, J. (2012). Chronic inflammation in cancer development. Front. Immunol. *2*, 98. https://doi.org/10.3389/fimmu.2011.00098.
10. Zhao, H., Wu, L., Yan, G., Chen, Y., Zhou, M., Wu, Y., and Li, Y. (2021). Inflammation and tumor progression: signaling pathways and targeted intervention. Signal Transduct. Targeted Ther. *6*, 263. https://doi.org/10.1038/s41392-021-00658-5.
11. Fathalla, M.F. (1971). Incessant ovulation—a factor in ovarian neoplasia? Lancet *298*, 163. https://doi.org/10.1016/S0140-6736(71)92335-X.
12. Morrison, J.H., Brinton, R.D., Schmidt, P.J., and Gore, A.C. (2006). Estrogen, menopause, and the aging brain: How basic neuroscience can inform hormone therapy in women. J. Neurosci. *26*, 10332–10348. https://doi.org/10.1523/JNEUROSCI.3369-06.2006.
13. Kozieł, M.J., and Piastowska-Ciesielska, A.W. (2023). Estrogens, Estrogen Receptors and Tumor Microenvironment in Ovarian Cancer. Int J Mol Sci *24*, 14673. https://doi.org/10.3390/ijms241914673.
14. Ponnappan, S., and Ponnappan, U. (2011). Aging and Immune Function: Molecular Mechanisms to Interventions. Antioxid. Redox Signal. *14*, 1551–1585. https://doi.org/10.1089/ars.2010.3228.
15. Zhang, B., Chen, F., Xu, Q., Han, L., Xu, J., Gao, L., Sun, X., Li, Y., Li, Y., Qian, M., et al. (2018). Revisiting ovarian cancer microenvironment: A friend or a foe? Protein Cell *9*, 674. https://doi.org/10.1007/s13238-017-0466-7.
16. Price, J., Lord, J.M., and Harrison, P. (2020). Inflammaging and platelet hyperreactivity: A new therapeutic target? J. Thromb. Haemost. *18*, 3–5. https://doi.org/10.1111/jth.14670.
17. Faria, A.V.S., Andrade, S.S., Peppelenbosch, M.P., Ferreira-Halder, C.V., and Fuhler, G.M. (2020). Platelets in aging and cancer—"double-edged sword". Cancer Metastasis Rev. *39*, 1205–1221. https://doi.org/10.1007/s10555-020-09926-2.
18. Bartlett, D.B., Firth, C.M., Phillips, A.C., Moss, P., Baylis, D., Syddall, H., Sayer, A.A., Cooper, C., and Lord, J.M. (2012). The age-related increase in low-grade systemic inflammation (Inflammaging) is not driven by cytomegalovirus infection. Aging Cell *11*, 912–915. https://doi.org/10.1111/j.1474-9726.2012.00849.x.
19. Hufnagel, D.H., Cozzi, G.D., Crispens, M.A., and Beeghly-Fadiel, A. (2020). Platelets, thrombocytosis, and ovarian cancer prognosis: Surveying the landscape of the literature. Int. J. Mol. Sci. *21*, 8169. https://doi.org/10.3390/ijms21218169.
20. Stone, R.L., Nick, A.M., McNeish, I.A., Balkwill, F., Han, H.D., Bottsford-Miller, J., Rupairmoole, R., Armaiz-Pena, G.N., Pecot, C.V., Coward, J., et al. (2012). Paraneoplastic Thrombocytosis in Ovarian Cancer. N. Engl. J. Med. *366*, 610–618. https://doi.org/10.1056/nejmoa1110352.
21. Davis, A.N., Afshar-Kharghan, V., and Sood, A.K. (2014). Platelet effects on ovarian cancer. Semin. Oncol. *41*, 378–384. https://doi.org/10.1053/j.seminoncol.2014.04.004.
22. In 't Veld, S.G.J.G., Arkani, M., Post, E., Antunes-Ferreira, M., D'Ambrosi, S., Vessies, D.C.L., Vermunt, L., Vancura, A., Muller, M., Niemeijer, A.-L.N., et al. (2022). Detection and localization of early- and late-stage cancers using platelet RNA. Cancer Cell *40*, 999–1009.e6. https://doi.org/10.1016/j.ccell.2022.08.006.
23. Gao, Y., Liu, C.J., Li, H.Y., Xiong, X.M., Li, G.L., In 't Veld, S.G.J.G., Cai, G.Y., Xie, G.Y., Zeng, S.Q., Wu, Y., et al. (2023). Platelet RNA enables

accurate detection of ovarian cancer: an intercontinental, biomarker identification study. Protein Cell *14*, 579–590. https://doi.org/10.1093/procel/pwac056.

24. Fields, A.T., Lee, M.C., Mayer, F., Santos, Y.A., Bainton, C.M.V., Matthay, Z.A., Callcut, R.A., Mayer, N., Cuschieri, J., Kober, K.M., et al. (2022). A new trauma frontier: Exploratory pilot study of platelet transcriptomics in trauma patients. J. Trauma Acute Care Surg. *92*, 313–322. https://doi.org/10.1097/TA.0000000000003450.
25. GJG, S., and Wurdinger, T. (2019). Tumor-educated platelets. Blood *22*, 2359–2364. https://doi.org/10.1182/blood-2018-12-852830.
26. Nassa, G., Giurato, G., Cimmino, G., Rizzo, F., Ravo, M., Salvati, A., Nyman, T.A., Zhu, Y., Vesterlund, M., Lehtiö, J., et al. (2018). Splicing of platelet resident pre-mRNAs upon activation by physiological stimuli results in functionally relevant proteome modifications. Sci. Rep. *8*, 498. https://doi.org/10.1038/s41598-017-18985-5.
27. Best, M.G., In 't Veld, S.G.J.G., Sol, N., and Wurdinger, T. (2019). RNA sequencing and swarm intelligence–enhanced classification algorithm development for blood-based disease diagnostics using spliced blood platelet RNA. Nat. Protoc. *14*, 1206–1234. https://doi.org/10.1038/s41596-019-0139-5.
28. Li, L., Mao, B., Yan, M., Wu, S., Ge, R., Lian, Q., and Cheng, C.Y. (2019). Planar cell polarity protein Dishevelled 3 (Dvl3) regulates ectoplasmic specialization (ES) dynamics in the testis through changes in cytoskeletal organization. Cell Death Dis. *10*, 194. https://doi.org/10.1038/s41419-019-1394-7.
29. Zhou, Z., Yang, Z., Zhou, L., Yang, M., and He, S. (2023). The Versatile Roles of Testrapanins in Cancer from Intracellular Signaling to Cell–Cell Communication: Cell Membrane Proteins without Ligands. Cell Biosci *13*, 59. https://doi.org/10.1186/s13578-023-00995-8.
30. Wang, J., Ma, S., Ma, R., Qu, X., Liu, W., Lv, C., Zhao, S., and Gong, Y. (2014). KIF2A silencing inhibits the proliferation and migration of breast cancer cells and correlates with unfavorable prognosis in breast cancer. BMC Cancer *14*, 461. https://doi.org/10.1186/1471-2407-14-461.
31. Pastuszak, K., Supernat, A., Best, M.G., In 't Veld, S.G.J.G., Łapińska-Szumczyk, S., Łojkowska, A., Różański, R., Żaczek, A.J., Jassem, J., Würdinger, T., and Stokowy, T. (2021). imPlatelet classifier: image-converted RNA biomarker profiles enable blood-based cancer diagnostics. Mol. Oncol. *15*, 2688–2701. https://doi.org/10.1002/1878-0261.13014.
32. Best, M.G., Sol, N., In 't Veld, S.G.J.G., Vancura, A., Muller, M., Niemeijer, A.L.N., Fejes, A.V., Tjon Kon Fat, L.A., Huis In 't Veld, A.E., Leurs, C., et al. (2017). Swarm Intelligence-Enhanced Detection of Non-Small-Cell Lung Cancer Using Tumor-Educated Platelets. Cancer Cell *32*, 238–252.e9. https://doi.org/10.1016/j.ccell.2017.07.004.
33. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.
34. Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. *37*, 907–915. https://doi.org/10.1038/s41587-019-0201-4.
35. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.
36. Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. *33*, 290–295. https://doi.org/10.1038/nbt.3122.
37. Tu, J.J., Ou-Yang, L., Zhu, Y., Yan, H., Qin, H., and Zhang, X.F. (2021). Differential network analysis by simultaneously considering changes in gene interactions and gene expression. Bioinformatics *37*, 4414–4423. https://doi.org/10.1093/bioinformatics/btab502.
38. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.
39. Höhn, A.K., Brambs, C.E., Hiller, G.G.R., May, D., Schmoeckel, E., and Horn, L.C. (2021). 2020 WHO Classification of Female Genital Tumors. Geburtshilfe Frauenheilkd. *81*, 1145–1153. https://doi.org/10.1055/a-1545-4279.
40. Glass, K., Huttenhower, C., Quackenbush, J., and Yuan, G.C. (2013). Passing Messages between Biological Networks to Refine Predicted Interactions. PLoS One *8*, e64832. https://doi.org/10.1371/journal.pone.0064832.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
blood platelet samples	This study	see Tables S2 and S3
Chemicals, peptides, and recombinant proteins		
RNAlater™ stabilization solution	Thermo Fisher	cat. no. AM7020
Critical commercial assays		
*mir*Vana miRNA Isolation Kit	Thermo Fisher	cat. no. AM1560, AM1561
SMART-Seq® v4 Ultra® Low Input RNA Kit for Sequencing	Takara Bio	cat. no. 634897
High Sensitivity DNA kit and reagents, Bioanalyzer 2100	Agilent Technologies	cat. no. 5067-4626
TruSeq Nano DNA Library Prep Kit	Illumina	cat. no. 20015964
Deposited data		
Raw RNA-seq data	(Pastuszak et al., 2021)[31]	GEO: GSE158508
Raw RNA-seq data	(Best et al., 2017)[32]	GEO: GSE89843
Raw RNA-seq data	(GJG et al., 2022)[22]	GEO: GSE183635
Raw RNA-seq data	This study	GEO: GSE278917
Software and algorithms		
Trimmomatic (version 0.39)	(Bolger et al., 2014)[33]	http://www.usadellab.org/cms/?page=trimmomatic
HISAT2 (version 2.1.0)	(Kim et al., 2019)[34]	https://daehwankimlab.github.io/hisat2/
Samtools (version 1.9)	(Li et al., 2009)[35]	http://samtools.sourceforge.net
stringtie (version 2.1.7)	(Pertea et al., 2015)[36]	https://github.com/gpertea/stringtie
chNet (version 4.3.2)	(Tu et al., 2021)[37]	https://github.com/Zhangxf-ccnu/chNet
Python (version 3.8.13)	Python3	https://www.python.org/doc/
scikit-learn (version 1.1.1)	(Pedregosa et al., 2011)[38]	https://github.com/scikit-learn/scikit-learn
Other		
BD Vacutainer® Blood Collection Tubes (10 mL)	BD	cat. no. 367863

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Clinical specimens

Blood samples were procured from participants who were prospectively enrolled in the study, including those diagnosed with gynecological cancer (n=85), benign tumors (n=66), and healthy women (n=44), at Seoul National University Hospital (SNUH) and Boaz Medical Center at Handong Global University between August 2022 and February 2023. Samples were excluded from patients under medication or failing to meet quality control criteria (Table S1). As a result, 46 samples from gynecological cancer patients, 22 samples from benign tumor patients, and 18 samples from healthy women were included in the final analysis.

Participants were categorized into three groups based on selection criteria. The gynecological cancer group consisted of individuals diagnosed with ovarian, or uterine cancer, confirmed through clinical assessment and histopathological examination. The benign tumor group included individuals diagnosed with non-malignant gynecological tumors, who were also evaluated through clinical assessment and histopathological confirmation to ensure accurate diagnosis. In contrast, the healthy control group comprised women who had no history of gynecological diseases or ovarian-related symptoms. This group was primarily recruited from Boaz Medical Center at Handong Global University, and their eligibility was determined based on self-reported medical history and routine health check-up records. We retrieved 1 cm3-sized cubes of fresh-frozen cancer tissues, cut from viable portions of the primary ovarian cancer tissues at the time of surgery under gross examination and frozen section procedures by pathologists for each patient, from Seoul National University Hospital Human Biobank. One expert gynecologic pathologist (Cheol Lee) reviewed our study population according to the World Health Organization Classification of Tumors, 5th edition.[39] The study was approved by the Ethics Committee of SNUH, Seoul, Korea (No. 2206-148-1335) and Handong Global University, Pohang, Korea (No. 2022-HGUA025). Written informed consent was obtained from the guardians of all the patients. The metadata of the dataset is shown in Table S2 (Model1 – tumor prediction) & S3 (Model2 – malignancy prediction).

Ethics approval and consent to participate

We strictly followed all guidelines from the Ethics Committee of Seoul National University Hospital, Republic of Korea, and compliance with the Declaration of Helsinki has been paramount. The study received approval from the Institutional Review Boards (No. H-1807-037-956), ensuring patient anonymity and excluding personally identifiable information.

METHOD DETAILS

Hematology analysis and platelet RNA extraction

Blood samples were collected using 10-mL EDTA-coated purple-capped BD Vacutainers (BD). Following collection, samples were processed in accordance with standard protocols for platelet isolation, which involved a two-step centrifugation procedure at 4°C within 48 hours of collection, as per established procedures.[27] Simultaneously, hematology parameters of whole blood cells and platelet-rich plasma were measured using XP-300-Hematology-Analyzer (Sysmex Corporation). To prevent RNA degradation, the extracted platelets were subsequently stored in RNAlater (Thermo Scientific, Waltham, MA, USA) at 4°C overnight, and then frozen at -80°C. RNA extraction from stored platelet samples was performed approximately every two months using the miRVana RNA isolation kit (Thermo Scientific, Waltham, MA, USA).

RNA-sequencing

For the assessment of total RNA quality, Bioanalyzer analysis utilizing BioAnalyzer 2100 (Agilent, Santa Clara, CA, USA) was performed. High-quality platelet RNA was identified based on a RNA Integrity Number (RIN) exceeding 6 and/or the presence of distinct ribosomal peaks. A total of 500 picograms of platelet RNA were used for cDNA synthesis and amplification. The SMART-Seq v4 Ultra Low Input RNA Kit (Takara Bio, Mountain View, CA, USA) was employed for the cDNA synthesis and amplification process. The quality assessment of the cDNA was conducted using the DNA High Sensitivity chip on the Agilent Bioanalyzer 2100. The amplified cDNA samples underwent shearing via sonication by Covaris Inc. and were subsequently labeled with index barcodes suitable for Illumina sequencing using the Truseq Nano DNA Sample Prep Kit platform (Illumina, San Diego, CA, USA). The libraries were amplified with 8 cycles and purified with AMPure XP beads. The concentration and size of final library were determined with the TapeStation 4200 (Agilent, Santa Clara, CA, USA) instrument and High Sensitivity D1000 screen tape. High-quality samples exhibiting distinct product sizes within the range of 500-600 base pairs were equimolarly pooled. The pooled samples were then sequenced on the Illumina NovaSeq6000 platform (Illumina, San Diego, CA, USA) with 150 bp paired-end.

Opensource dataset

To obtain gene expression values from the opensource dataset,[22,27,31,32] we downloaded RNA-seq data provided in fastq format and corresponding patient information from Gene Expression Omnibus (GEO). We downloaded public platelet transcriptome data of gynecological cancer (n=186) including ovarian cancer, benign tumor (n=30), and healthy counterparts (n=391) from GEO: GSE158508, GSE89843, GSE183635. Additionally, after removing male samples and patients at the fourth stage cancer to avoid capturing noisy signal derived from advanced metastatic cancer, samples from gynecological cancer (n=151), benign tumors (n=30), and healthy women (n=218) were included in the feature selection and model development for our research. The metadata of the dataset is described in Table S4.

Preprocessing of RNA-seq data

The same NGS workflow was utilized to process the Opensource and our own clinical RNA-seq datasets. In this study, we processed the Fastq files using Trimmomatic (version 0.39) to clip sequence adapters and trim reads based on quality metrics. Subsequently, we aligned the sequencing reads to the human GRCh38 genome reference using HISAT2 (version 2.1.0). The resulting SAM files were converted to BAM format and filtered to retain only primary alignments using Samtools (version 1.9). The BAM files containing only primary alignments were then used to count reads meeting the following criteria at each junction: 1. Reads spanning from 150 bases upstream to 150 bases downstream of the junction, 2. Reads with N-containing cigar strings indicative of splicing events, and 3. Reads whose spliced regions match the position of the junction. We obtained count values for 2,855,955 junctions in Opensource and our own clinical RNA-seq datasets. Expression levels (Fragments Per Kilobase per Million mapped fragments: FPKM and Transcripts Per kilobase Million: TPM) of four genes (*PTPRC*, *PF4*, *SELP*, and *CD63*) were automatically calculated by using stringtie (2.1.7) in our preprocessing pipeline. *PTPRC* is a leukocyte marker to measure the possible contamination in the experiment, and *PF4*, *SELP*, and *CD63* are markers for platelet activation.

Developing machine learning-based gynecological tumor diagnosis model

In this study, a machine learning-based model was developed for diagnosing gynecological tumors. The training and validation datasets consisted of opensource data (n=365 and n=242, respectively), while our own clinical data was used for training (n=54) and testing (n=53). Two normalization methods were employed, including the in-house developed batch-invariant normalization known as Binning FMH. This method selected relevant variables by excluding junctions with a count of 0 and those associated with the Y chromosome. Unique values from the selected junctions in both the opensource and clinical RNA-seq datasets were used to establish the 100th percentile interval. Each section was assigned a rank value based on the corresponding junction value.

To develop the model, Counts Per Million (CPM) values, obtained by dividing the value of each junction by the sum of all junctions in the sample and multiplying by 10^6 were utilized on a log2 scale. The resulting CPM values were then transformed into the log2 scale using the formula log2 (CPM + 1). Variable selection and model development were conducted using the public training set (n=243), public validation set (n=156), and in-house clinical training set (n=47). Through the Binning FMH normalization on the training set, junction variables showing significant differences between the tumor group and the non-tumor group were selected using the Mann-Whitney U test ($|\log2FC| > 1$, FDR < 0.05). Ultimately, 198 junctions with the concordant direction of changes from both public and clinical data were chosen.

The SVM model with the bootstrap method was employed for model development, with 100 rounds of bootstrapping performed. Each bootstrapped set randomly selected 94 public training samples and 47 internal clinical training samples for model learning. A total of 39 combinations of 'kernel', 'C', and 'gamma' hyperparameters were used in each set. The optimal hyperparameter combination was determined by performing verification bootstrapping 100 times, with 94 public validation samples randomly selected and 47 internal clinical training samples in each bootstrapped verification set. The combination with the highest average AUC from 100 bootstrap validation sets was selected as the optimal hyperparameter combination. Forward selection was then applied to choose the best model combination with the highest validation AUC from the 100 models learned using the optimal hyperparameter combination. The final value for clinical test data was obtained by averaging the probability values from the 13 selected models. Performance evaluation of the clinical test data was conducted using a predefined cutoff that ensured a specificity of 0.99 in the clinical training data.

Developing machine learning-based malignancy prediction model

The methodology employed to discriminate between malignant and non-malignant tumors involves an ensemble of two models. The first model utilizes platelet RNA sequencing data, while the second model relies on hematology measurements.

In the first model using platelet RNA sequencing data, analysis was performed on ovarian cancer (OC, n=18), endometrial cancer (EC, n=26), benign patients (n=21), and healthy control samples (HC, n=18). Other cancer types, symptomatic control samples, and male samples were excluded from the study. The entire dataset was divided into training (n=29), validation (n=17), and test (n=37) sets. In the first step, highly expressed junctions were selected. The expression training data from our own clinical dataset was transformed into log2CPM values, and highly expressed junctions were identified by selecting those with expression levels higher than a reference gene (PTP4A2). Only OC, EC, and benign patient data were used in this step, resulting in a total of 604 junctions being selected. In the second step, the selected 604 junctions were paired in combinations of two to generate marker combinations, resulting in 182,106 marker pairs. For each pair, the two marker values, x and y, were compared, with a value of 1 assigned if x was greater than y, and 0 otherwise. Markers with a variance of 0 were excluded, leaving 114,425 markers after this step. In the third step, univariate logistic regression analysis was performed to assess the performance of individual markers and filter for stable performance. Markers with an AUC value $\geq$ 0.7 in both the training and validation datasets were selected, yielding 1,038 markers. To ensure marker stability, random noise was added to the data in the form of a normal distribution, N(0, std), and the analysis was repeated 10 times. Markers with an AUC $\geq$ 0.6 in both the training and validation datasets across all repetitions were retained, resulting in 276 markers. In the fourth step, Fisher's Exact Test was applied to the 276 selected markers, extracting those with a p-value < 0.01. Markers related to platelet activation genes were excluded, leaving 36 markers. From these, only junctions involving genes that appeared at least twice were retained, resulting in 27 markers in total (29 junctions in 10 genes). In the fifth step, the model was trained using a Random Forest algorithm. Hyperparameter tuning and training were performed using 5-fold cross-validation (CV). The predicted probabilities from all folds were averaged to calculate the final predicted probabilities.

In the second model using hematology measurements, model training for distinguishing between malignant and non-malignant tumors was conducted using our clinical training set (n=68; OC 9, EC 17, Benign 24, HC 18) and evaluated on our clinical validation set (n=45) and test set (n=48). Among the 31 hematology variables, 26 are univariable, and 5 are combination variables. Model development employed GridSearch with 5-fold Cross Validation using RandomForestClassifier. The predicted probabilities from all folds were averaged to obtain the final predicted values.

The final prediction was made by combining the prediction probabilities of both models through soft voting. Weights were assigned based on AUC maximization from the clinical validation set. The final prediction probability was calculated by multiplying the prediction probabilities of the two models by their respective weights and then adding them together. A cutoff value of 0.5 was applied to evaluate the clinical test data.

Functional analysis

Differential network analysis of the 198 junction markers (95 genes) was conducted using chNet (v. 4.3.2).[37] For gene ontology analysis, either significantly over- or under-expressed were analyzed using the Ingenuity Pathway Analysis (IPA) core pathway analysis (Qiagen Ingenuity Systems). For Gene Regulatory Network (GRN) analysis, we utilized PANDA to build a gene regulatory network comprising 198 junction markers representing 94 genes.[40] From this network, we extracted the top 100 edge weights to highlight key connections within each of the 47 cancer groups and 40 noncancer groups analyzed in our study.

QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical analyses were conducted in Python (v. 3.8.13). Continuous data was compared using a Mann-Whitney U test, applied to analyze read composition, hematological parameters, and blood markers utilized for cancer diagnosis. A p-value of <0.05 was considered statistically significant. Junction markers were selected using the same statistical test, with statistical significance determined by a false discovery rate value of <0.05. The performance of model was evaluated using metrics of accuracy, sensitivity, specificity, balanced accuracy, and area under the curve (AUC) of receiver operating characteristics curves, calculated with Python library scikit-learn (v. 1.1.1). The Wilson method was applied to calculate 95% confidence intervals of all AUC values.