BMC Medical Informatics
and Decision Making

RESEARCH Open Access

Decision tree model for predicting ovarian tumor malignancy based on clinical markers and preoperative circulating blood cells

Yingjia Li[1], Xingping Zhao[1], Yanhua Zhou[1], Lina Gong[1] and Enuo Peng[1*]

Abstract

Objective Ovarian cancer is a serious malignant tumor threatening women's health. The early diagnosis and effective treatments of ovarian cancer remain inadequate, and about 70% of ovarian cancers are in advanced stages when discovered. This study aimed to use the decision tree method of artificial intelligence machine learning to build a model for predicting the benign and malignant degree of ovarian cancer patients.

Study design A total of 758 patients were included in the study. These patients were diagnosed by B-ultrasound, CT or MR. The clinicopathological features and circulating blood cell indexes were recorded and analyzed. The prediction model of benign and malignant ovarian tumors was constructed by CART decision tree, and the receiver operating characteristic (ROC) curve was drawn to evaluate the predictive value of the decision tree model.

Results It was found that significant predictor variables included age, disease duration, patient general condition and menopausal status, ascites, tumor size, HE4, CA125, ROMA index, and blood routine related indicators (except for basophil count percentage and absolute value). In the constructed decision tree model, ROMA_after was the root node with the maximum information gain. ROMA_after, Mass size (MR/CT), HE4, CA125, platelet number, lymphocyte ratio, white blood cell count, post-menopause, hematocrit and mean platelet volume were important indicators in the decision tree model. The area under the receiver operating characteristic curve of this model for predicting benign and malignant ovarian cancer was 0.86.

Conclusions The decision tree model was successfully constructed based on clinical indicators and preoperative circulating blood cells, and showed better results in predicting benign and malignant ovarian cancer than alone imaging indicators or biomarkers among our data, which means that our model can more accurately predict benign and malignant ovarian cancer.

Keywords Decision tree model, Ovarian cancer, Preoperative circulating blood cells, Machine learning, Prediction

*Correspondence:
Enuo Peng
pengena@163.com
[1]The Third Xiangya Hospital of Central South University,
Changsha 410013, China

Introduction

Ovarian cancer, the seventh most common cancer worldwide, is the second leading cause of gynecological cancer death after cervical cancer with the mortality-to-incidence ratio (MIR) of over 0.6 [1, 2]. A study estimates that about 1/6 women have died within three months of being diagnosed with ovarian cancer [2]. Therefore, it is very essential to diagnose and treat ovarian cancer early.

At present, clinical diagnosis of ovarian cancer mainly relies on imaging examinations and tumor marker detection [3, 4]. However, imaging examinations (CT, MRI, etc.) have the disadvantage of not being able to identify benign and malignant tumors and transvaginal ultrasound (TVUS) combined with two-dimensional vaginal color Doppler ultrasound can only detect about 17% of patients with ovarian tumors [4–6]. Ovarian tumor markers include CA125, HE4, and the Risk of Ovarian Malignancy Algorithm (ROMA) combining HE4 and CA125, but the sensitivity and specificity of CA125 and HE4 are not high, with less than 50% in the diagnostic sensitivity of stage I ovarian cancer [7–9]. There is an urgent need to develop new methods for the early clinical diagnosis of ovarian cancer. Routine blood examination is one of the simplest, most economical, and most convenient examinations in clinical practice. It is mainly used to measure white blood cells, platelets, and red blood cells related indicators [10]. This examination can quickly reflect the patient's blood status and inflammation at the time of blood collection [11]. Some studies show that routine blood examination can reflect the status of the tumor micro-environment and has certain value in the diagnosis and prognosis of tumors [12–14]. However, The evidence of circulating blood cell in predicting early diagnosis of ovarian cancer is unclear.

This study aimed to analyze the differences in preoperative circulating blood cell indicators between benign and malignant ovarian cancer patients and to explore the role of routine blood indicators as potential biomarkers for the diagnosis of benign and malignant tumors. In addition, this study used artificial intelligence machine learning algorithms to build a decision tree model for predicting benign and malignant ovarian cancer patients.

Materials and methods

General information collection and ethics

The experimental framework design of this experiment was shown in Fig. 1. This study retrospectively analyzed the information of ovarian cancer patients who visited the department of gynecology of the Third Xiangya Hospital of Central South University from January 2018 to December 2020. These patients were diagnosed with ovarian cancer through B-ultrasound, CT or MR examinations with the age of 14 years or older and a history of menstruation. Patients were excluded with severe heart

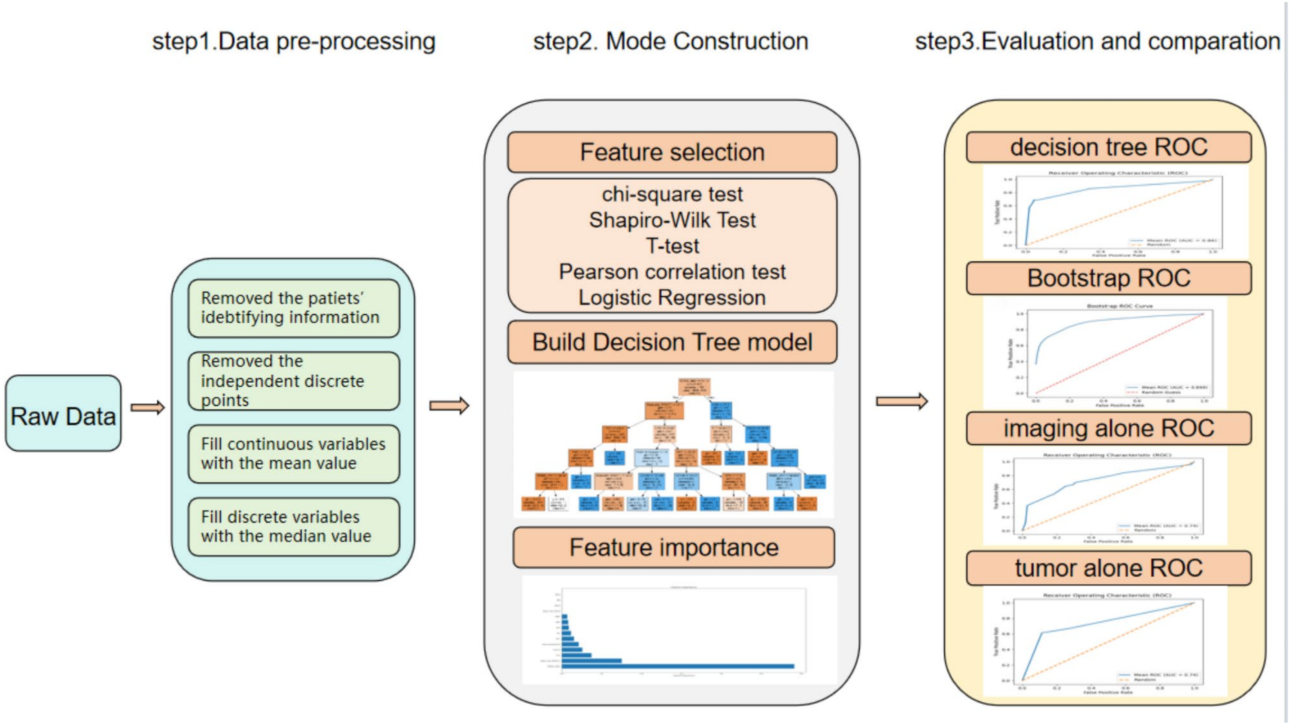
step1.Data pre-processing
step2. Mode Construction
step3.Evaluation and comparation
Raw Data
Removed the patiets' idebtifying information
Removed the independent discrete points
Fill continuous variables with the mean value
Fill discrete variables with the median value
Feature selection
chi-square test
Shapiro-Wilk Test
T-test
Pearson correlation test
Logistic Regression
Build Decision Tree model
Feature importance
decision tree ROC
Bootstrap ROC
imaging alone ROC
tumor alone ROC

Fig. 1 Experimental framework design. This flowchart detailed a three-step process: data preprocessing, model construction, and evaluation. Step 1 involved removing features, handling outliers, and filling missing values. Step 2 focused on feature selection using various tests and building a decision tree model with Bootstrap ROC. Step 3 evaluated the model by comparing ROC curves for imaging and tumor data

disease, liver and kidney disease, diabetes, and non-neoplastic ovarian cysts with a history of radiotherapy and chemotherapy [15]. And all ovarian tumors were initially diagnosed by cryopathology before surgery, and postoperative specimens were evaluated by at least two gynecological pathologists. Preoperative blood samples (5 ml/person) from each participant were collected, processed, and stored at − 80 °C for subsequent analysis. HE4 and CA125 were detected by using the electrochemiluminescence technology of the COBAS E411 analyzer (Elecsys; Roche Diagnostics, Mannhein, Germany), with detection ranges of 15-1500 pmol/L and 0.600–5000 U/mL. This project was a retrospective study conducted in the department of gynecology of the Third Xiangya Hospital of Central South University, which was qualified for tumor treatment. Informed consent: all participants gave informed written informed consent. This study had been approved by the Ethics Committee of the Third Xiangya Hospital of Central South University (IRB No. 2018-S355). All methods were carried out in accordance with relevant guidelines and regulations.

ROMA calculation

The calculation is based on the patient's menopausal status and the concentrations of HE4 and CA125 [16]. The calculation formula is as follows:

Premenopausal: Predictive Index (PI) = $-12.0 + 2.38 \times \ln[\mathrm{HE4}] + 0.0626 \times \ln[\mathrm{CA125}]$

Postmenopausal: Predictive Index (PI) = $-8.09 + 1.04 \times \ln[\mathrm{HE4}] + 0.732 \times \ln[\mathrm{CA125}]$

ROMA (%) = Prediction Index (PI) × 100%

For premenopausal women, ROMA value 11.4% is used as the cut-off point, $\geq 11.4\%$ indicates a high risk of ovarian cancer, and $< 11.4\%$ indicates a low risk of ovarian cancer. For postmenopausal women, the ROMA value is 29.9% as the cut-off point, $\geq 29.9\%$ indicates a high risk of ovarian cancer, and $< 29.9\%$ indicates a low risk of ovarian cancer [16].

Decision tree model

The decision tree model in this study was constructed based on the classification and regression tree (CART) analysis method developed in 1984 [17]. CART analysis is a non-parametric, nonlinear method that gradually divides samples into subsets according to certain criteria [18]. The decision tree generated by CART consists of nodes and leaf nodes. The nodes are divided according to attributes to form sub-nodes or leaf nodes. Each sub-node corresponds to a subset under the attribute, with the purpose of improving the purity of the subset as much as possible [19].

Statistical analysis

Mean ± standard deviation was applied to describe continuous variables. For discrete variables, the chi-square test or Fisher's exact test was used to detect differences between benign and malignant groups. A p value less than 0.05 was considered to have a significant difference. For continuous variables, a normality test was first performed by using the Shapiro-Wilk test. If the data conformed to the normal distribution, the independent samples t test was applied for verification; otherwise, the Mann-Whitney U test was applied.

For statistical analysis, the scikit-learn machine learning library in Python 3.8 was used to implement machine learning algorithms. The univariate logistic regression was to determine the dominant variables of the benign and malignant tumor prediction model. Then a decision tree model was built for predicting ovarian cancer malignancy and the Graphviz library was used to visualize the decision tree. The model evaluation was performed through an internal validation technique using 1000 bootstrapped resamples to obtain bias-corrected estimation for predictive performance [20]. Finally, the receiver operating characteristic curve (ROC) was applied to further evaluate the predictive performance of the decision tree model.

Results

Clinical characteristics and circulating blood cell indicators of patients with benign and malignant ovarian tumors

A total of 758 patients were included in the study (Table 1). 534 patients were diagnosed with benign tumors and 224 patients were diagnosed with malignant tumors. The age of patients with benign tumors was significantly lower than that of patients with malignant tumors ($p < 0.001$). In addition, the course time of patients with benign tumors was significantly longer than that of patients with malignant tumors ($p < 0.001$), which may be related to the lower survival rate of patients with ovarian cancer diagnosed as malignant. Under different health conditions, 6.74% of patients with benign tumors were in good condition, 93.26% were in moderate condition. Among patients with malignant tumors, 2.23% were in good condition, 95.98% were in moderate condition, and 1.79% were in cachexia. The overall condition of benign patients was better than that of malignant patients ($p < 0.001$). There was no statistical difference in body mass index (BMI) ($p = 0.257$).

Table 1 showed that the size of malignant tumors detected by ultrasound was significantly larger than that of benign tumors ($p < 0.001$), which was consistent with the detection results of CT or MR. Both benign and malignant tumors tended to occur on one side. Among patients with malignant tumors, 64.73% patients were post-menopausal, and 14.73% patients developed signs

Table 1 Statistical analysis of baseline indicators and clinical characteristics of patients with benign and malignant ovarian tumors

Variable	Total	Malignant tumor (*n* = 224)	Benign tumor (*n* = 534)	*P*-value
Normal information				
Age	41.97 ± 16.16	50.50 ± 14.09	38.39 ± 15.64	< 0.001
BMI	22.46±2.45	22.60±2.47	22.40±2.44	0.257
Course of disease (days)	307.72±740.25	149.49±464.49	374.09±820.40	< 0.001
General condition				< 0.001
good(1)	41(5.41%)	5(2.23%)	36(6.74%)	
medium(2)	713(94.06%)	215(95.98%)	498(93.26%)	
cachexia (3)	4(0.53%)	4(1.79%)	-	
Imaging indicators				
Mass size (BUS)	85.58±51.97	103.10±46.04	78.24±52.59	< 0.001
Mass size (MR/CT)	89.25±65.57	105.67±62.45	82.36±65.68	< 0.001
Mass locations (BUS)				0.284
Unilateral (1)	682(89.97%)	197(87.95%)	485(90.82%)	
Bilateral (2)	76(10.03%)	27(12.05%)	49(9.18%)	
Mass locations (MR/CT)				1
Unilateral (1)	647(85.36%)	191(85.27%)	456(85.39%)	
Bilateral (2)	111(14.64%)	33(14.73%)	78(14.61%)	
Post-menopause				< 0.001
No (1)	461(60.82%)	79(35.27%)	382(71.54%)	
Yes (2)	297(39.18%)	145(64.73%)	152(28.46%)	
Ascites (palpation)				< 0.001
Yes (1)	39(5.15%)	33(14.73%)	6(1.12%)	
No (2)	719(94.85%)	191(85.27%)	528(98.88%)	
Tumor markers				
HE4 (pmol/L)	212.10±537.10	515.23±893.93	84.95±143.39	< 0.001
CA125 (U/mL)	332.66±1120.92	971.25±1911.41	64.79±112.53	< 0.001
ROMA_before	25.52±28.60	53.86±36.55	13.64±11.13	< 0.001
ROMA_after	30.65±28.30	60.85±33.65	17.98±10.92	< 0.001
Red blood cell related indicators				
RBC	4.18 ± 0.43	4.06 ± 0.47	4.24 ± 0.40	< 0.001
HGB	121.82 ± 14.68	116.47 ± 15.76	124.07 ± 13.60	< 0.001
HCT	37.51 ± 3.98	36.19 ± 4.42	38.06 ± 3.65	< 0.001
RDW	13.12 ± 1.62	13.37 ± 1.97	13.01 ± 1.44	< 0.001
MCV	89.88 ± 6.85	89.39 ± 6.68	90.08 ± 6.91	0.016
MCH	29.18 ± 2.76	28.75 ± 2.60	29.36 ± 2.80	< 0.001
MCHC	324.28 ± 12.12	321.37 ± 11.23	325.50 ± 12.28	< 0.001
White blood cell related indicators				
WBC	6.45 ± 2.17	7.03 ± 2.67	6.20 ± 1.87	< 0.001
NE%	60.69 ± 11.26	67.76 ± 11.00	57.73 ± 9.98	< 0.001
NE	4.04 ± 2.06	4.95 ± 2.55	3.66 ± 1.68	< 0.001
MO%	6.50 ± 1.94	6.82 ± 2.07	6.37 ± 1.86	0.006
MO	0.41 ± 0.16	0.46 ± 0.19	0.38 ± 0.13	< 0.001
BA%	0.43 ± 0.24	0.40 ± 0.23	0.44 ± 0.24	0.051
BA	0.03 ± 0.01	0.03 ± 0.02	0.03 ± 0.01	0.994
EO%	2.20 ± 1.89	1.91 ± 2.00	2.32 ± 1.84	< 0.001
EO	0.13 ± 0.12	0.12 ± 0.12	0.14 ± 0.12	0.005
LY	1.83 ± 0.65	1.46 ± 0.55	1.99 ± 0.63	< 0.001
LY%	30.10 ± 10.22	23.01 ± 9.94	33.08 ± 8.78	< 0.001
Platelet related indicators				
PLT	247.02 ± 83.61	288.03 ± 112.71	229.82 ± 60.08	< 0.001
MPV	10.51 ± 1.35	10.22 ± 1.35	10.64 ± 1.34	< 0.001

of ascites. Both proportions were higher than those in patients with benign tumors ($p < 0.001$).

Common biomarkers for ovarian cancer including CA125 and HE4 had significant differences in identifying benign and malignant tumors. The CA125 concentration in patients with malignant tumors was 15 times higher than that of benign tumors and the HE4 concentration in the serum of patients with malignant tumors was 6 times higher. ROMA index was used to evaluate the difference between patients with benign and malignant tumors ($p < 0.001$). Before and after menopause, the average ROMA index of patients with benign tumors was significantly lower than that of patients with malignant tumors ($p < 0.001$).

In addition to commonly used clinical diagnostic methods for ovarian cancer, we also studied the differences in blood routine indicators between patients with benign and malignant ovarian cancer. Among these indicators, the values of patients with malignant tumors were significantly higher than those of patients with benign tumors, including red blood cell distribution width (RDW) among red blood cell-related indicators, white blood cell count (WBC) and neutrophil count (NE) and percentage (NE%) among white blood cell-related indicators, monocyte number (MO) and percentage (MO%); and platelet number (PLT) among platelet-related indicators ($p < 0.001$). On the contrary, these indicators were significantly lower in patients with malignant tumors, including red blood cell-related indicators such as red blood cell count (RBC), hemoglobin concentration (HGB), hematocrit (HCT), mean corpuscular volume (MCV), and mean hemoglobin content (MCH) and mean hemoglobin concentration (MCHC); eosinophil number (EO) and percentage (EO%), lymphocyte number (LY) and percentage (LY%) in leukocyte-related indicators; and platelet-related indicators in mean platelet volume (MPV) ($p < 0.001$). There was no statistical difference in basophil count (BA) and percentage (BA%) between patients with benign and malignant tumors ($p > 0.05$).

In summary, important clinical variables to distinguish benign and malignant ovarian tumors included age, disease course, patient's general condition and menopausal status, ascites, tumor size, HE4, CA125, ROMA index, and blood routine related indicators (except basophils percentage and absolute value).

Univariate logistic regression analysis on the impact of outcomes

Next, univariate logistic regression was used to make preliminary judgments to determine the indicators that affected the outcomes. The results were shown in Table 2. According to the p value ($p > 0.05$), the values of BMI, tumor location, BA, BA%, EO and MCV did not affect the benign and malignant outcomes of ovarian cancer. Therefore, these indicators were excluded in subsequent model construction.

Decision tree model for predicting benign and malignant ovarian cancer

After processing the data using Scikit-Learn, ROMA_after was the root node with the largest information gain (Fig. 2). The Gini coefficient of the node indicates the impurity of the data, and the smaller the better [21]. In the root node, sample represented the total number of samples before division, and value represented the number of malignant and benign ovarian cancer samples in the node. According to whether ROMA_after ≤ 54.13, the 454 samples were divided into two groups, one group contained 380 samples and the other group contained 74 samples. Mass size (MR/CT) and HE4 were used as classification criteria for first-level nodes. Then, entered the second, third, fourth and fifth layer nodes respectively until there were no leaf nodes. And when the decision reached each leaf node, the probabilities of benign and malignant tumors were determined by the number of class samples divided by the total number of samples in the current subset. The detailed explanation of the decision tree model can be found in the supplementary materials.

The indicators involved in this decision tree model included ROMA_after, mass size (MR/CT), LY%, CA125, post-menopause, HCT, PLT, HE4, MPV, WBC, mass size (BUS), MO%, MO and NE%. ROMA_after was of highest importance to the decision tree model, while mass size (MR/CT), LY%, CA125, post-menopause, HCT, PLT, HE4, MPV, WBC had smaller contribution to the model (Fig. 3). Furthermore, the predictive ability of this decision tree model for benign and malignant ovarian cancer could be evaluated by the Area Under Curve (AUC). The AUC of the decision tree model was 0.86 (Fig. 4A) and the average AUC of internal validation technique using 1000 bootstrapped resamples was 0.899 (Fig. 4B), while the AUCs of both benign and malignant ovarian cancers predicted by imaging indicators or ovarian tumor biomarkers (such as CA125, HE4, and ROMA) were 0.74 (Fig. 4C and D), which showed that the decision tree model constructed by introducing preoperative circulating blood cell indicators had a higher predictive value.

Discussion

Approximately 70% of ovarian malignant tumors are diagnosed at an advanced stage, resulting in a poor prognosis, due to the lack of specificity of its main symptoms [22]. Therefore, early detection of benign and malignant tumors is of great significance to improve the prognosis of ovarian cancer patients. Increasing studies have confirmed that blood routine indexes have important predictive roles in the early diagnosis of cancer [12–14]. This

Table 2 Single-factor logistic regression analysis of differential variables

Variable	Estimated value	Standard deviation	*P*-value	OR	Lower limit	Upper limit
Normal information						
Age	0.049	0.006	< 0.001	1.051	1.039	1.062
General condition	1.502	0.489	0.002	4.491	1.723	11.708
Course of disease (days)	-0.001	0.000	< 0.001	0.999	0.999	1.000
Imaging indicators						
Mass size (BUS)	0.011	0.002	< 0.001	1.011	1.007	1.015
Mass size (MR/CT)	0.005	0.001	< 0.001	1.005	1.003	1.008
Mass locations (BUS)	0.305	0.254	0.230	1.357	0.824	2.232
Mass locations (MR/CT)	0.010	0.225	0.964	1.010	0.650	1.569
Post-menopause	1.529	0.170	< 0.001	4.613	3.308	6.431
Ascites (palpation)	-2.722	0.452	< 0.001	0.066	0.027	0.159
Tumor markers						
HE4	0.007	0.001	< 0.001	1.007	1.006	1.009
CA125	0.005	0.001	< 0.001	1.005	1.004	1.007
ROMA_before	0.070	0.006	< 0.001	1.073	1.059	1.086
ROMA_after	0.082	0.007	< 0.001	1.085	1.070	1.101
Red blood cell related indicators						
RBC	-1.009	0.197	< 0.001	0.364	0.248	0.537
HGB	-0.035	0.006	< 0.001	0.966	0.955	0.976
HCT	-0.118	0.021	< 0.001	0.889	0.853	0.926
RDW	0.126	0.048	0.009	1.135	1.032	1.248
MCV	-0.014	0.011	0.206	0.986	0.964	1.008
MCH	-0.076	0.028	0.006	0.927	0.878	0.979
MCHC	-0.029	0.007	< 0.001	0.972	0.959	0.985
White blood cell related indicators						
WBC	0.167	0.036	< 0.001	1.182	1.101	1.270
NE%	0.091	0.009	< 0.001	1.095	1.076	1.115
NE	0.303	0.043	< 0.001	1.354	1.243	1.474
MO%	0.118	0.04	0.004	1.126	1.040	1.219
MO	3.253	0.544	< 0.001	25.870	8.903	75.178
EO%	-0.136	0.05	0.007	0.873	0.791	0.964
EO	-1.327	0.733	0.070	0.265	0.063	1.117
LY%	-0.115	0.010	< 0.001	0.891	0.874	0.910
LY	-1.618	0.171	< 0.001	0.198	0.142	0.277
Platelet related indicators						
PLT	0.009	0.001	< 0.001	1.009	1.007	1.011
MPV	-0.248	0.064	< 0.001	0.780	0.688	0.885

study also found that most routine blood indicators had significant differences in counts and percentages between benign and malignant tumors, and changes in these indicators are also closely related to the clinicopathological characteristics of ovarian tumors [23].

At present, machine learning has been widely used in the medical field, and the predictive performance of decision tree model for cancer diagnosis has been increasingly recognized [24, 25]. In this study, we successfully established a decision tree model based on preoperative routine blood indicators and serum tumor markers. Some studies report the ROMA index constructed by combining CA125 and HE4 can not only improve the diagnostic specificity of ovarian cancer [26], but can also be used to identify different types of ovarian tumors [7]. Similarly, ROMA_after was also the root node with the maximum information gain in our decision tree model. Morely, LY%, HCT, PLT, MPV and WBC in routine blood indicators also played important roles in the model. Research reports tumor-infiltrating lymphocytes with high expression of CD3 + and CD8 + can significantly improve the prognosis of endometrioid ovarian cancer [27] and lymphocytes can promote synergistic anti-tumor responses of humoral immunity and cellular immunity in mouse ovarian cancer models [28]. In recent years, the recognition of platelets as markers of inflammation has increased, which can induce the transformation of tumor epithelial cells into mesenchymal cells and promote the extravasation and metastasis of tumor cells [29]. Ovarian cancer patients with thrombocytosis have a

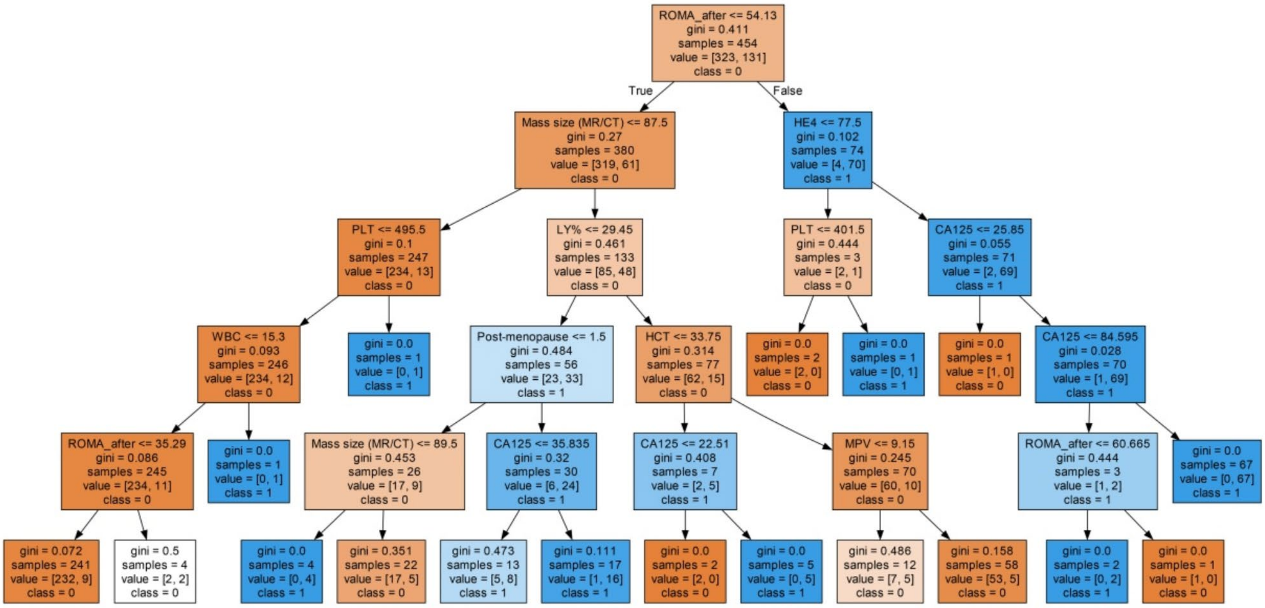

Fig. 2 Visualization of the decision tree for predicting benign and malignant ovarian cancer. ROMA_after was the root node. Mass size (MR/CT) and HE4 were the classification standards for the first-layer nodes; PLT, LY%, and CA125 were the second-layer nodes; WBC, Post-menopause, HCT and CA125 were the third-layer nodes; ROMA_after, Mass size (MR/CT), CA125 and MPV were the fourth-layer nodes; the fifth layer was the leaves. The probability of benign and malignant was the number of class samples divided by the total number of samples in the current subset. 1 represented a benign tumor, and 0 represented a malignant tumor. Figure 2 represented a single decision tree generated from the complete dataset, not a collection of trees from the 1000 bootstrap samples

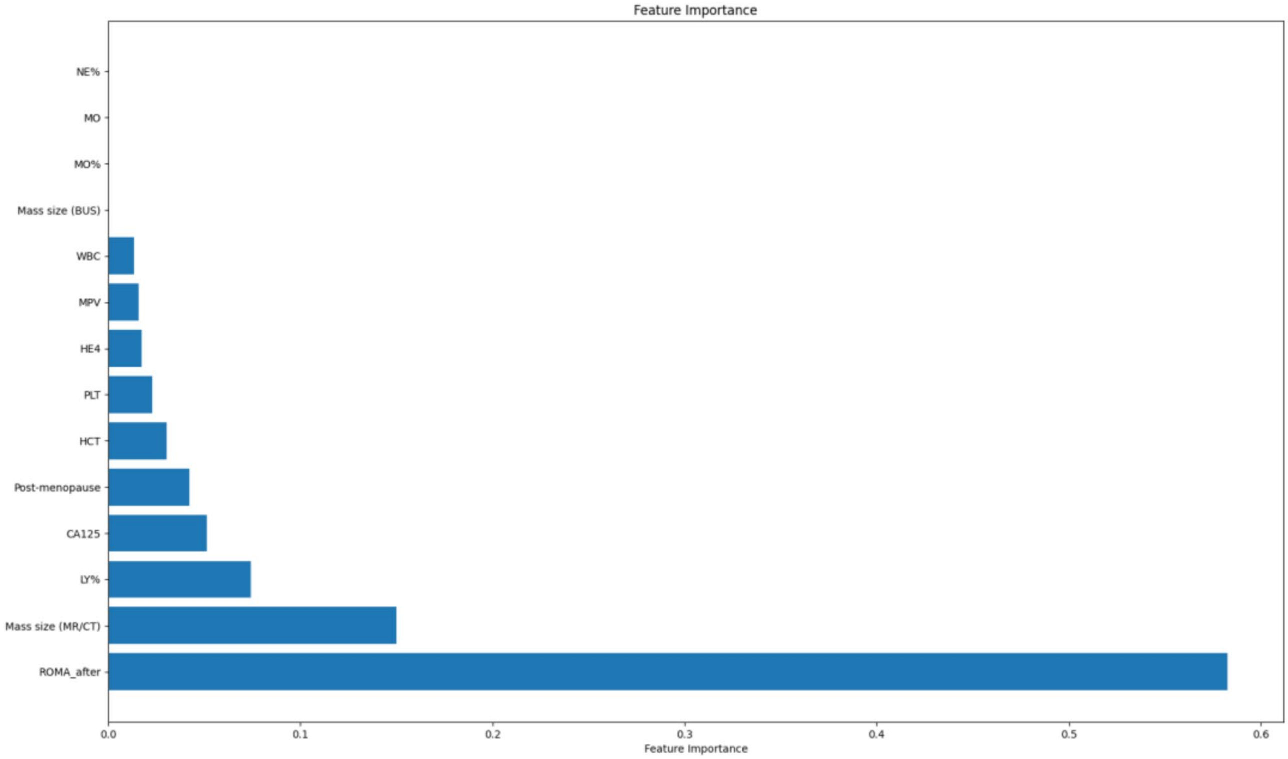

Fig. 3 The importance of features was expressed by the weight distribution rate of decision tree

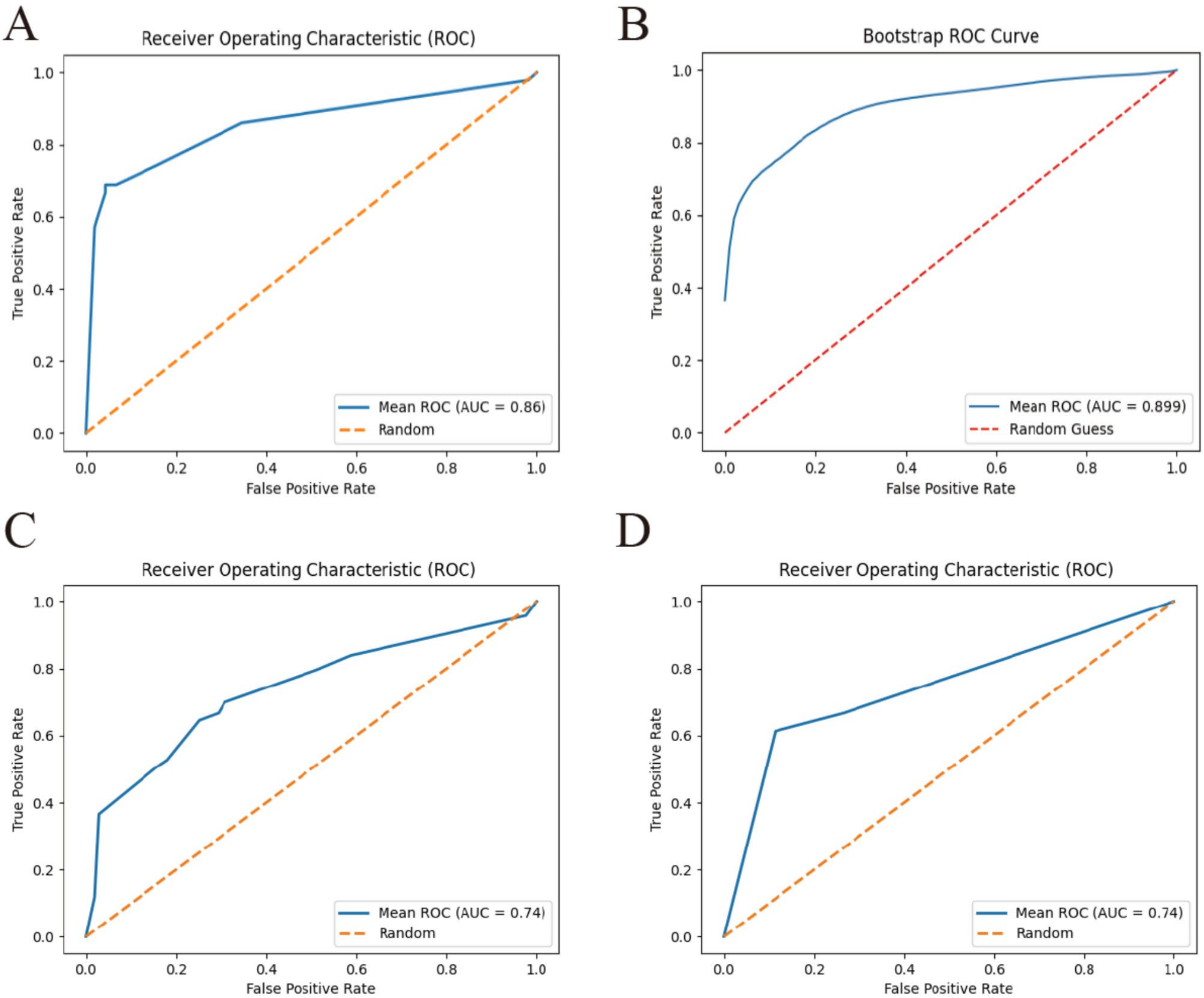

Fig. 4 ROC for predicting benign and malignant ovarian cancer models. (**A**) a decision tree model; (**B**) the average value of internal validation technique using 1000 bootstrapped resamples; (**C**) a prediction model using imaging indicators alone; (**D**) a prediction model using tumor markers alone

poor prognosis whether at first diagnosis or at recurrence [30]. And WBC is considered to be a prognostic indicator of ovarian cancer [31]. This decision tree model was built by combining imaging, tumor serum markers, and blood routine related indicators. Therefore, in view of the importance of these indicators in differentiating benign and malignant ovarian cancer, this decision tree model will further improve the diagnostic ability of ovarian cancer in primary hospitals.

For the purpose of relevance and effectiveness maintained over time, it is essential to implement a systematic approach for updating and improving the model with new data and variables. As additional patient data is collected, the model can be retrained using larger, more diverse datasets, which would enhance its accuracy and generalizability across different populations. Furthermore, incorporating emerging clinical markers, genetic information, or additional blood indices could significantly improve predictive capabilities. Establishing a continuous learning framework would allow the model to adapt to evolving clinical practices and patient demographics by periodically integrating new data and outcomes. Furthermore, creating feedback mechanisms for clinicians utilizing the model in practice can provide invaluable insights into its performance and areas needing refinement. Regular evaluation against real-world outcomes will help identify any decline in predictive accuracy, facilitating timely adjustments. By adopting these strategies, the decision tree model can evolve and maintain its utility in clinical settings, ultimately leading to improved patient management and outcomes. With the aim of dataset expansion, we plan to establish multi-center collaborations with various hospitals and research institutions to collect data from diverse regions and populations. Additionally, we will leverage existing public datasets and utilize data-sharing platforms to encourage contributions from other researchers. Patient recruitment efforts will focus on engaging underrepresented groups, while data augmentation techniques, such as Generative Adversarial Networks (GANs), will be employed to create synthetic samples. We will also conduct longitudinal studies to gather more comprehensive data over time and leverage electronic health records (EHR) to quickly increase sample size, all while ensuring adherence to ethical guidelines and data privacy regulations.

The decision tree model incorporates various strategies to effectively handle missing data, which is crucial for maintaining its robustness and applicability. One common approach is imputation, where missing values are filled using techniques such as mean, median, or mode imputation, as well as more advanced methods like K-Nearest neighbors (KNN) or regression imputation [32, 33]. This allows the model to utilize all available data without losing valuable cases. Moreover, the model can employ surrogate splits, identifying alternate splitting criteria for instances with missing values, thus ensuring that decisions can still be made even when some data points are incomplete. During the tree-building process, missing values can be accommodated by creating separate branches for instances with missing data, allowing the model to address the inherent uncertainty of missing information. While a complete case analysis may exclude cases with missing data, this approach risks losing important information and introducing bias. By implementing these strategies, the decision tree model enhances its predictive accuracy and remains effective in real-world applications.

The performance of model was compared favorably with existing diagnostic tools, specifically imaging indicators and conventional tumor biomarkers. The decision tree model achieved an AUC of 0.86 for predicting benign and malignant ovarian cancer, which indicated a strong ability to distinguish between the two classes, comparing with the AUCs of 0.74 for predictions based solely on imaging indicators or tumor biomarkers. The model combined clinical markers and preoperative circulating blood cell indicators with traditional imaging and biomarker data, enhancing predictive accuracy and making it more effective than models relying solely on a single type of diagnostic tool. Importantly, the internal validation does not guarantee the generalizability of the model to external populations, meaning our findings are interpreted with caution. The further validation is essential to confirm the model's applicability in diverse clinical settings. Therefore, we will conduct prospective studies that utilize independent datasets to validate our model's performance. Such studies will provide critical insights into the model's robustness and its potential role in clinical decision-making. Overall, while our decision tree model demonstrates potential, ongoing research and validation efforts are necessary to establish its clinical utility.

The decision tree model can adapt to new data changes and multi-center applications through strategic approaches such as incremental learning, which allows updates with new patient data without complete retraining. Regularization and pruning help prevent overfitting, ensuring good generalization to unseen data. Incorporating ensemble methods like Random Forests enhances robustness by aggregating predictions from multiple trees, beneficial for diverse patient populations. Cross-center validation ensures effective performance across different healthcare settings, while continuous monitoring and feedback loops enable timely adjustments based on user insights.

For the sake of effectiveness, concrete plans for periodically updating data and retraining can be established. A regular schedule for collecting new patient data—quarterly or biannually—keeps the model current with demographic changes. Automated data pipelines facilitate seamless integration, minimizing errors. Continuous monitoring tracks performance metrics, triggering automatic retraining when performance declines. Feedback from clinical users informs retraining decisions, while version control systems ensure stability during updates. Conducting cross-validation after retraining assesses performance on updated data, ensuring generalization across clinical settings. Together, these strategies ensure the decision tree model remains relevant and effective in a dynamic healthcare environment.

In order to investigate the potential impact of confounding variables such as comorbidities and medication use on the model, we plan to conduct further analyses that will incorporate these variables in future iterations of the model. Specifically, we will utilize electronic health records and structured patient interviews to systematically collect comprehensive histories regarding patients' comorbidities and medications, including dosage and duration of use. We will also employ multivariable regression analysis and machine learning techniques to assess how these confounding variables influence the model's predictive capabilities. This process will help us identify and control for factors that may significantly impact the outcomes, thereby enhancing the model's accuracy and applicability. Besides, we plan to conduct external validation across diverse populations to ensure the model's broad applicability and reliability in varied clinical settings. Through these measures, we aim to build a more comprehensive and robust model that better serves clinical practice.

This study has the following limitations: (1) This model was not suitable for the benign and malignant judgment of all ovarian tumor patients. This study excluded patients with severe heart disease, liver and kidney disease, diabetes, non-neoplastic ovarian cysts and a history of radiotherapy and chemotherapy; (2) Known risk factors for ovarian cancer (such as family history, hormone replacement therapy) and ovulation factors (lifetime Ovulatory cycle, breastfeeding, irregular menstruation and fallopian tube ligation) [15] were not included in this study; (3) The pathological subtypes of tumors were not further graded; (4) The study does not explicitly account for confounding variables such as patient comorbidities or medication use in the final decision tree mode; (5) We recognize that feature selection is a critical aspect of model development that can introduce significant limitations. While we used statistical methods like univariate logistic regression to identify significant predictors, this reliance may inadvertently exclude relevant indicators that could enhance model performance, resulting in a model that, despite its robustness with selected features, may not fully capture the complexity of ovarian tumor malignancy. Likewise, the feature selection process can introduce bias if the chosen indicators do not adequately represent the broader patient population, potentially compromising the model's validity and applicability in diverse clinical settings. In future research, we aim to enhance our understanding of minimum sample size requirements for reliable model performance across various clinical contexts. Establishing these requirements is essential for ensuring robust and generalizable outcomes. We plan to incorporate statistical power analysis to determine the necessary sample sizes based on effect sizes and variability within patient populations. By providing context-specific recommendations and conducting pilot studies, we hope to refine these estimates and adapt our model to evolving clinical practices. This ongoing evaluation will ultimately strengthen the applicability of our decision tree model in diverse settings.

Conclusion

A decision tree model based on clinical markers and preoperative circulating blood cells was constructed. Compared with using imaging indicators and biomarkers (such as CA125, HE4, ROMA) alone to predict benign and malignant ovarian cancer, this decision tree model showed higher predictive value. In future research, it's planning to expand the scope of research objects, add external validation, and add more indicators to further improve the results of decision tree analysis.

Abbreviations

AUC	The area under the receiver operating characteristic curve
BA	Basophil count
BC	Breast cancer
CART	The classification and regression tree
CRC	Colorectal cancer
HCT	Hematocrit
HGB	Hemoglobin concentration
LY	Lymphocyte number
MCH	Mean hemoglobin. content
MCHC	Mean hemoglobin concentration
MCV	Mean corpuscular volume
MIR	Mortality-to-incidence ratio
MO	Monocyte number
MPV	Mean platelet volume
NE	Neutrophil count
NLR	Neutrophil/lymphocyte ratio
PLR	Platelet count/lymphocyte ratio
PLT	Platelet number
RBC	Red blood cell count
RDW	Red blood cell distribution width
ROC	The receiver operating characteristic curve
ROMA	Risk of Ovarian Malignancy Algorithm
TVUS	Transvaginal ultrasound
WBC	White blood cell count

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-025-02934-8.

Supplementary Material 1

Acknowledgements

Not applicable.

Author contributions

YL, XZ, YZ, LG and EP: data collection, formal analysis, investigation and methodology. YL, XZ, LG and EP: project development, manuscript writing. EP: Manuscript revision. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by grants from Provincial Natural Science Foundation of Hunan (2025JJ80128).

Data availability

The data that support the findings of this study are available on request from the corresponding author.

Declarations

Ethics approval and consent to participate

This project was a retrospective study conducted in the department of gynecology of the Third Xiangya Hospital of Central South University. All subjects provided written informed consent for inclusion before they participated in the study to ensure confidentiality of patient personal data. This study had been approved by the Ethics Committee of the Third Xiangya Hospital of Central South University (IRB No. 2018-S355). All methods were carried out in accordance with relevant guidelines and regulations.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Clinical trial number

Not applicable.

Received: 7 September 2024 / Accepted: 12 February 2025
Published online: 20 February 2025

References

1. Kuroki L, Guntupalli SR. Treatment of epithelial ovarian cancer [J]. BMJ (Clinical Res ed). 2020;371:m3773.
2. Lheureux S, Braunstein M, Oza AM. Epithelial ovarian cancer: evolution of management in the era of precision medicine [J]. Cancer J Clin. 2019;69(4):280–304.
3. Gilbert L, Basso O, Sampalis J, et al. Assessment of symptomatic women for early diagnosis of ovarian cancer: results from the prospective DOvE pilot project [J]. Lancet Oncol. 2012;13(3):285–91.
4. Irajizad E, Han CY, Celestino J, et al. A blood-based Metabolite Panel for distinguishing ovarian Cancer from Benign Pelvic masses [J]. Clin cancer Research: Official J Am Association Cancer Res. 2022;28(21):4669–76.
5. Shetty M. Imaging and Differential diagnosis of ovarian Cancer [J]. Semin Ultrasound CT MR. 2019;40(4):302–18.
6. Castellucci P, Perrone AM, Picchio M, et al. Diagnostic accuracy of 18F-FDG PET/CT in characterizing ovarian lesions and staging ovarian cancer: correlation with transvaginal ultrasonography, computed tomography, and histology [J]. Nucl Med Commun. 2007;28(8):589–95.
7. Felder M, Kapur A, Gonzalez-Bosquet J, et al. MUC16 (CA125): tumor biomarker to cancer therapy, a work in progress [J]. Mol Cancer. 2014;13:129.
8. Lycke M, Ulfenborg B, Malchau Lauesgaard J, et al. Consideration should be given to smoking, endometriosis, renal function (eGFR) and age when interpreting CA125 and HE4 in ovarian tumor diagnostics [J]. Clin Chem Lab Med. 2021;59(12):1954–62.
9. Swiatly A, Plewa S, Matysiak J, et al. Mass spectrometry-based proteomics techniques and their application in ovarian cancer research [J]. J Ovarian Res. 2018;11(1):88.
10. Ziv-Baran T, Wasserman A, Goldiner I, et al. The association between C-reactive protein and common blood tests in apparently healthy individuals undergoing a routine health examination [J]. Clin Chim Acta. 2020;501:33–41.
11. Zhang J, Zhou X, Ding H, et al. The prognostic value of routine preoperative blood parameters in muscle-invasive bladder cancer [J]. BMC Urol. 2020;20(1):31.
12. Tanriver G, Kocagoncu E. Additive pre-diagnostic and diagnostic value of routine blood-based biomarkers in the detection of colorectal cancer in the UK Biobank cohort [J]. Sci Rep. 2023;13(1):1367.
13. Liu Q, Luo D, Cai S, et al. Circulating basophil count as a prognostic marker of tumor aggressiveness and survival outcomes in colorectal cancer [J]. Clin Translational Med. 2020;9(1):6.
14. Sun H, Yin CQ, Liu Q, et al. Clinical significance of Routine Blood Test-Associated Inflammatory Index in breast Cancer patients [J]. Med Sci Monitor: Int Med J Experimental Clin Res. 2017;23:5090–5.
15. Gao B, Zhao X, Gu P, et al. A nomogram model based on clinical markers for predicting malignancy of ovarian tumors [J]. Front Endocrinol (Lausanne). 2022;13:963559.
16. Hada A, Han LP, Chen Y, et al. Comparison of the predictive performance of risk of malignancy indexes 1–4, HE4 and risk of malignancy algorithm in the triage of adnexal masses [J]. J Ovarian Res. 2020;13(1):46.
17. Magnini M, Ciatto G, Cantürk F, et al. Symbolic knowledge extraction for explainable nutritional recommenders [J]. Comput Methods Programs Biomed. 2023;235:107536.
18. Henrard S, Speybroeck N, Hermans C. Classification and regression tree analysis vs. multivariable linear and logistic regression methods as statistical tools for studying haemophilia [J]. Haemophilia: Official J World Federation Hemophilia. 2015;21(6):715–22.
19. Rutkowski L, Jaworski M, Pietruczuk L, et al. A new method for data stream mining based on the misclassification error [J]. IEEE Trans Neural Networks Learn Syst. 2015;26(5):1048–59.
20. Henderson AR. The bootstrap: a technique for data-driven statistics. Int J Clin Chem. 2005;359(1–2):1–26. Using computer-intensive analyses to explore experimental data [J]Clinica chimica acta.
21. Quinlan JR. Induction of decision trees [J]. Mach Learn. 1986;1(1):81–106.
22. Chen YN, Ma F, Zhang YD, et al. Ultrasound features improve diagnostic performance of Ovarian Cancer predictors in distinguishing Benign and malignant ovarian tumors [J]. Curr Med Sci. 2020;40(1):184–91.
23. Asante DB, Calapre L, Ziman M, et al. Liquid biopsy in ovarian cancer using circulating tumor DNA and cells: ready for prime time? [J]. Cancer Lett. 2020;468:59–71.
24. Saberi-Karimian M, Khorasanchi Z, Ghazizadeh H, et al. Potential value and impact of data mining and machine learning in clinical diagnostics [J]. Crit Rev Clin Lab Sci. 2021;58(4):275–96.
25. Reix N, Lodi M, Jankowski S, et al. A novel machine learning-derived decision tree including uPA/PAI-1 for breast cancer care [J]. Clin Chem Lab Med. 2019;57(6):901–10.
26. Moore RG, McMeekin DS, Brown AK, et al. A novel multiple marker bioassay utilizing HE4 and CA125 for the prediction of ovarian cancer in patients with a pelvic mass [J]. Gynecol Oncol. 2009;112(1):40–6.
27. Gallego A, Mendiola M, Hernando B, et al. Prognostic markers of inflammation in endometrioid and clear cell ovarian cancer [J]. Int J Gynecol Cancer. 2022;32(8):1009–16.
28. Ukita M, Hamanishi J, Yoshitomi H et al. CXCL13-producing CD4 + T cells accumulate in the early phase of tertiary lymphoid structures in ovarian cancer [J]. JCI Insight, 2022, 7(12).
29. Cedervall J, Hamidi A, Olsson AK, Platelets. NETs and cancer [J]. Thromb Res. 2018;164(Suppl 1):S148–52.
30. Gao Y, Liu CJ, Li HY, et al. Platelet RNA enables accurate detection of ovarian cancer: an intercontinental, biomarker identification study [J]. Protein Cell. 2023;14(6):579–90.
31. Bishara S, Griffin M, Cargill A, et al. Pre-treatment white blood cell subtypes as prognostic indicators in ovarian cancer [J]. Eur J Obstet Gynecol Reprod Biol. 2008;138(1):71–5.
32. Srisuradetchai P, Suksrikran K. Random kernel k-nearest neighbors regression [J]. Front big data. 2024;7:1402384.
33. Zahid FM, Heumann C. Multiple imputation with sequential penalized regression [J]. Stat Methods Med Res. 2019;28(5):1311–27.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.