Preventive Medicine Reports 35 (2023) 102296

Contents lists available at ScienceDirect

Preventive Medicine Reports

journal homepage: www.elsevier.com/locate/pmedr

Development of prediction model to estimate future risk of ovarian lesions: A multi-center retrospective study

Bilin Jing [a,1], Gaowen Chen [a,1], Miner Yang [b], Zhi Zhang [c], Yue Zhang [d], Jingyao Zhang [d], Juncheng Xie [d], Wenjie Hou [e], Yong Xie [f], Yi Huang [g], Lijie Zhao [h], Hua Yuan [i], Weilin Liao [c,*], Yifeng Wang [a,*]

[a] Zhujiang Hospital of Southern Medical University, Guangzhou 510280, China
[b] Guangzhou Women and Children's Medical Center, Guangzhou 510620, China
[c] Geography and Planning of Sun Yat-sen University, Guangzhou 510275, China
[d] Second Clinical Medical College, Guangzhou 510599, China
[e] Soochow University Medical Center, Suzhou 215125, China
[f] Foshan First People's Hospital, Foshan 528010, China
[g] Nanhai District People's Hospital, Foshan 528099, China
[h] Foshan Maternal and Child Health Hospital, Foshan 528099, China
[i] Wuxi Maternal and Child Health Hospital, Wuxi 214002, China

ARTICLE INFO

Keywords:
Ovarian Disease
Lasso Regression
Machine Learning
Disease Prediction
AUC

ABSTRACT

Background: To develop the preoperative prediction of ovarian lesions using regression-based statistics analyses and machine learning methods based on multiple serological biomarkers in China.
Methods: 1137 patients with ovarian lesions in Zhujiang Hospital and 518 patients in others hospital in China were randomly assigned to training, test and external validation cohorts. Five machine learning classifiers, including Random Forest (RF), Extreme Gradient Boosting (XGB), Support Vector Classifier (SVC), K-nearest Neighbor (KN), Multi-Layer Perceptron (MLP) and the Lasso-Logistics prediction model (LLRM) were used to derive diagnostic information from 23 predictors.
Results: The RF model had a high diagnostic value (AUC = 0.968) in predicting benign and malignant ovarian disease. Age and MLR were also potential diagnostic indicators for predicting ovarian disease except tumor indicators. The RF model well distinguished borderline ovarian tumors (AUC = 0.742). The RFM had a high predictive power to identify ovarian serous adenocarcinoma (AUC = 0.943) and ovarian endometriosis cysts (AUC = 0.914).
Conclusions: The RF models can effectively predict adnexal lesions, promising to be adjuncts to the preoperative prediction of ovarian cancer.

1. Introduction

Ovarian tumors are the most common disease for adnexal diseases, divided into the benign, the malignant and the borderline. The ovarian cancer (OC) is one of the three major malignant tumors of the female reproductive system and mortality is the top of gynecological malignant tumor (Yuyu et al., 2011). According to statistics, 310,000 new OC cases were detected worldwide in 2020, among which, about 210,000 patients died from OC, accounting for 4.7% of cancer deaths in all female systems (Sung et al., 2021). And there were 60,000 new OC patients and 66.7% of patients died in China in 2020 (Sung et al., 2021). The OC with high morbidity and mortality has become the biggest serious threat to women's health. OC has an insidious onset and lacks specific clinical symptoms in the early stage, which is not easy to detect and cause

* Corresponding authors.
E-mail addresses: jingbilin@163.com (B. Jing), cgw2012@163.com (G. Chen), yangme95@126.com (M. Yang), zhangzh49@mail2.sysu.edu.cn (Z. Zhang), zhangy1228@hotmail.com (Y. Zhang), 1960275656@qq.com (J. Zhang), gallenxie@qq.com (J. Xie), 55189246@qq.com (W. Hou), whitefox2013@163.com (Y. Xie), 123hy@126.com (Y. Huang), 13202902121@163.com (L. Zhao), yuanhua62099@163.com (H. Yuan), liaoweilin@mail.sysu.edn.cn (W. Liao), wangyifeng@smu.edu.cn (Y. Wang).
[1] Bilin Jing and Gaowen Chen contributed equally to this manuscript.

https://doi.org/10.1016/j.pmedr.2023.102296
Received 6 February 2023; Received in revised form 12 June 2023; Accepted 21 June 2023
Available online 23 June 2023

attention. With the progression of tumor and dissemination of intra-abdominal, abdominal distension, abdominal pain and abdominal mass may appear. About 60% of patients diagnosed with OC due to abdominal distension or ascites were in the advanced stage (stages III ~ IV) (Zhang et al., 2021). Patients with advanced OC lose the optimal treatment time, are unable to perform comprehensive staging surgery or even lose the opportunity for surgery, leading to poor prognosis, and their 5-year survival rate is less than 30%. However, a small number of patients with ovarian mass removal found by physical examination and post-operative pathological indication of OC are in the early stage (stage I ~ II), only accounted for 19% of the number of OC (Yaqin and Tan., 2022). But they had won the opportunity for the early intervention and treatment, which improved the 5-year survival rate of early OC patients and reached as high as 90% (Qin and Pca-, 2021). Early detection and early diagnosis are crucial to improve the survival rate of patients.

In 1981, the scholars first detected cancer antigen 125 (CA125) with murine anti-human monoclonal antibodies, which was the earliest marker for the clinical application of OC. Studies showed that CA125 had no high sensitivity or specificity in the diagnosis of OC (Chunfang, 2013). With the development of molecular biology and imaging technology, the methods for OC preoperative diagnosis are improved continuously. The combined application of multiple tumor markers can effectively improve the detection rate of OC. Ying Tang confirmed that CA125 combined lymphocyte/monocyte ratio as a predictor in OC diagnosis has a high diagnostic specificity (AUC = 0.782) (Tang et al., 2021). The application of imaging examination and serological indicators can effectively improve the prognosis rate of OC in early stage. Wu Meng evaluated a combined method of *trans*-vaginal contrast-enhanced ultrasound, HE4 and RI to diagnose OC (AUC = 0.888), which significantly improved the diagnostic sensitivity (87.5%) of stage I to II in OC (Meng et al., 2017). In recent years, a large number of research has focused on the field of machine learning, Zhang Tongshuo established a multi-index joint diagnosis model of ovarian cancer based on artificial neural network (AUC = 0.948), which confirmed that the diagnostic value was significantly better than CA125 alone (Tongshuo, 2018). The emerging development of machine learning also provides us with a powerful tool to establish machine learning model for clarifying the disease histological type and selecting the appropriate treatment modality. Therefore, we establish non-invasive predictive models for the early screening of OC through the multi-index combined screening and multiple machine learning methods.

Ovarian borderline tumor is an epithelial ovarian tumor with malignant potential between benign adenoma and carcinoma. Most studies generalize it to malignancy for calculation. For the current clinical development for the early diagnosis of ovarian borderline tumors, all the methods are still in the exploratory stage. At present, the effective prediction methods and specific diagnosis methods have not been found in healthy people or even in the high-risk groups (Pharoah, 2012). Even studies that distinguish between borderline tumors and malignancies are rare, not to mention the differentiation of preoperative histological type of ovarian mass.

Therefore, we strive to seek new strategies for the preoperative diagnosis of OC and borderline tumor, even to predict the type of ovarian mass. This study aims at evaluating and predicting ovarian lesions through the construction of clinical prediction model and machine learning prediction model, using clinical statistics and multiple machine learning methods based on the clinically accessible hematology indicators. The best model was selected to further distinguish the borderline ovarian tumors and predict the histology. It is expected to provide new methods for early diagnosis, to provide a new basis for personalized treatment of patients and to provide new ideas for clinical diagnosis and treatment.

2. 2.Materials and methods

2.1. Information of the patients

Our study retrospectively collected the clinical information and serological indicators about patients, who were confirmed by surgical pathology to have adnexal disease (e.g., ovarian serous adenocarcinoma and mature teratoma), came from Zhujiang Hospital of Southern Medical University, Nanhai District People's Hospital, Foshan Maternal and Child Health Hospital, Foshan First People's Hospital, Soochow University Medical Center and Wuxi Maternal and Child Health Hospital from January 2015 to December 2020 in China. Data were collected from the electronic medical record system.

Inclusion criteria included: Patients who underwent surgery were pathologically diagnosed with adnexal lesion. Patients with complete preoperative serological index data and all blood samples were collected within one week before surgery. Exclusion criteria included: Pathology confirmed benign and malignant mixed tumors, such as left ovarian serous carcinoma with right ovarian endometriosis cyst. At the same time, patient will suffer from other systemic tumor diseases and primary pathogenesis diseases, such as breast cancer and renal failure, etc. Patients would undergo the emergency surgery for ovarian tumors and recurrent.

The stage of malignant tumors in the adnexal area was based on the Surgery-Pathological Stage of Ovarian Cancer, Tubal Cancer and Primary peritoneal Cancer (FIGO, 2014) (Mutch and Prat, 2014), and the histological type of diseases was based on the classification of female genital organ tumors issued by WHO tumor classification in 2020 (McCluggage et al., 2022).

The study was conducted according to the ethical principles of the Declaration of Helsinki, reviewed and approved by the Medical Ethics Committee of Zhujiang Hospital, Southern Medical University (Approval Number: 2022-KY-141–01). Informed consent was waived due to the noninvasive and retrospective nature of our study. Identity information (i.e., patient name) were replaced as numeric codes to ensure the data confidentiality.

2.2. Selection of the predictors

Serological tumor indicators, such as CA125 and human epididymis secretory protein4 (HE4), are the most valuable tumor markers in ovarian epithelial carcinoma, which can be used for auxiliary diagnosis, efficacy monitoring and recurrence monitoring (Mutch and Prat, 2014). Blood routine indicators, such as platelets and derivatives, have been proven valuable in tumors (Qundi, 2021; Jia Jiyun, 2022; Shen et al., 2014). In 1863, The German physicist Vichow proved that Cancer susceptibility and severity may be contacted with functional polymorphisms of inflammatory cytokine genes, and deletion or inhibition of inflammatory cytokines inhibits development of experimental cancer (Balkwill and Mantovani, 2001). Seungjoo Chon indicates that platelet/lymphocyte ratio (PLR) can be used as an independent significant prognostic factor in advanced epithelial ovarian cancer (Chon et al., 2021). At present, there are few studies that can predict the risk of ovarian tumors directly and simply combined with the above indicators. So we took all factors into account in our study.

Twenty-three clinical predictors of patients including general demographic indicators, tumor markers indicators and routine blood markers were retrospectively collected in our study. The general demographic variables included age, blood type (i.e., A RH+, B RH+, O RH+, AB RH+, none), pregnancy, gravidity, menopausal state. The blood routine variables included Leukocyte count (WBC), neutrophil count (Neut), lymphocyte count (Lymph), red blood cell count (RBC), hemoglobin (Hb), monocyte count (Mono), monocyte/lymphocyte ratio (MLR), neutrophil/lymphocyte ratio (NLR),PLR. The tumor markers included CA125, HE4, carbohydrate antigen 199 (CA199), carcinoembryonic antigen (CEA), alpha-fetoprotein (AFP), Roman index before

menopause (ROMA_pro) and Roman index after menopause (ROMA_post).

2.3. Data splitting

All missing values were less than 20% and filled with the mean of the feature. The shuffled patients data of Zhujiang Hospital was then stratified and sampled into training and testing sets, until there was no significant difference (P value > 0.200) between the two sets with respect to all outcome variables.The P value was calculated using Pearson's chi-squared test for categorical variables. This resulted in allocation of 795 patients to the training cohort and 342 patients to the test cohort. 518 patients from the other hospitals were included in the external validation set (Table 1).

2.4. Establish of machine learning and statistical analysis models

2.4.1. The Lasso-logistics regression predict model (LLRM)

Statistical analysis was performed using R4.1.2. All the R package used in this article are available at https://cran.r-project.org/web/packages/available_packages_by_name.html. Table 1 shows the distribution of the three major datasets in bi-classification models. We used the training set to filter the most appropriate variables by Lasso regression (Fig. 1a) and achieved optimal λ value by cross validation (Fig. 1b) in "glmnet" R package, then we chose log λ = 0.035 corresponding to a standard square error, when selecting four significant correlation indicators (Menopausal state, APF, ROMA_pro and ROMA_post) with non-zero coefficients. Next, we joined the internationally recognized major relevant antigen for CA125 and combined with the above as independent variables, taking benign and malignant ovarian disease as the outcome variable, then constructed stepwise Logistic regression model, and the non-zero coefficient of each independent variable was 0.001, 1.603, 0.022, 0.022, 0.062. Using "survival" and "rms" R package to build the LLRM equation, was calculated as: $In[P/(1-P)] = -4.699 + 0.001*CA125 + 1.603*Menopausal\,state + 0.022*APF + 0.022*ROMA\_prp + 0.062*ROMA\_post$. The 'P' represents the probability of OC, and the '1-P' represents the probability of benign ovarian disease.

To use the "DynNom" R package to draw a dynamic nomogram (Fig. 1c). On the test set, the LLRM was differentiated, calibration and clinical effectiveness evaluated by calculating AUC, Mean Absolute Error (MAE) and Decision Curve Analysis (DCA) in "pROC", "rms" and "rmda" R package. On the external validation set, the LLRM was externally verified by drawing ROC and calculating AUC.

2.4.2. The supervised machine learning models

Python 3.8.5 was used for machine learning to build five machine learning prediction models: Random Forest model (RFM), Extreme Gradient Boosting model (XGBM), Support Vector Classifier model (SVCM), K-nearest Neighbor model (KNM) and Multi-Layer Perceptron model (MLPM). On the training datasets, the hyper-parameters of each model were parameterized by grid search. Only the best simulated hyper-parameters of each model were selected to make predictions on the testing set and the external validation set.

First, construction of the binary classification model, we constructed the above five machine learning models, drew the ROC curve, calculated the AUC, and selected the optimal classifier for subsequent predictor importance analysis. Secondly, Table 1 shows the sets data of the triple classification predict model to distinguish between borderline tumors and predict the preoperative diagnosis probability of borderline and malignant ovarian tumors for further model deepening and optimization. Finally, using the histological type as the outcome variable, the RF classifier was used to predict the preoperative histological type of ovarian mass. The AUC and F1 values were calculated for models' evaluation and validation by scikit-learn 1.1.2. The Z-test was used to compare between the models, and P < 0.050 was considered statistically significant. The specific procedures can be seen in Fig. 2.

Table 1
Basic information statistics of 1655 patients with adnexal lesions.

Categorical	The training set (n = 795)	The test set (n = 342)	P value	The external validation set (n = 518)
Bi-classification model				
Benign	612(82.7%)	263 (82.4%)	0.990	201
Malignant	128(17.3%)	56(17.6%)		209
Tri-classification model				
Benign	612(80.0%)	263 (76.9%)	0.988	201
Malignant	128(16.1%)	56(16.3%)		209
Borderline	55(6.9%)	23(6.7%)		108
Multi-classification model				
Serous cystadenoma (B1)	56(7.2%)	24(6.9%)	1	62
Mucinous cystadenoma (B2)	51(6.5%)	23(6.6%)		17
Chocolate cyst (B3)	207(26.5%)	90(25.9%)		68
Mature teratoma (B4)	128(16.4%)	56(16.1%)		47
Fibroma,follicular membranatoma (B5)	28(3.6%)	12(3.5%)		0
Physiological cyst (B6)	67(8.6%)	30(8.6%)		7
Inflammatory (B7)	72(9.2%)	31(8.9%)		0
Pulsar carcinoma (M1)	64(8.2%)	28(8.1%)		133
Mucous carcinoma (M2)	16(2%)	7(2%)		16
Endometrioid carcinoma (M3)	9(1.2%)	4(1.2%)		31
Malignant germ cell tumors (M4)	9(1.2%)	4(1.2%)		6
Sex cord stromal tumor (M5)	7(0.9%)	4(1.2%)		9
Metastatic carcinoma (M6)	11(1.4%)	6(1.7%)		0
Others (M7)	4(0.5%)	3(0.9%)		0
Serous borderline tumor (L1)	20(2.6%)	9(2.6%)		49
Mucinous borderline tumor (L2)	30(3.8%)	14(4%)		30
Clear-cellular borderline tumors (L3)	1(0.1%)	1(0.3%)		1
Mixed borderline tumors (L4)	2(0.3%)	1(0.3%)		12

3. Results

3.1. Identification of OC from ovarian benign lesions

The LLRM quality was evaluated as follows. Distinction was verified according to the ROC, and AUC was 0.946 (95 %CI:0.922–0.971). The calibration curve revealed good predictive accuracy between the actual probability and predicted probability and MAE was 0.013 (Fig. 3a). Fig. 3b was DCA cure. The validated of the LLRM was confirmed by ROC as follow. It had an internal validation AUC of 0.946 (95 % CI:0.862–0.975) and an external validation AUC of 0.896 (95 % CI:0.863–0.929). The internal validation AUC and F1 values for RFM, XGBM, SVCM, KNM and MLPM were 0.968 and 0.909, 0.951 and 0.893, 0.821 and 0.855, 0.909 and 0.893, 0.935 and 0.922, respectively (Fig. 3c). The external validation AUC and F1 values of machine learning were 0.944 and 0.825, 0.936 and 0.822, 0.875 and 0.689, 0.808 and 0.717, 0.890 and 0.811, respectively (Fig. 3d).

By comparing the five machine learning prediction models with

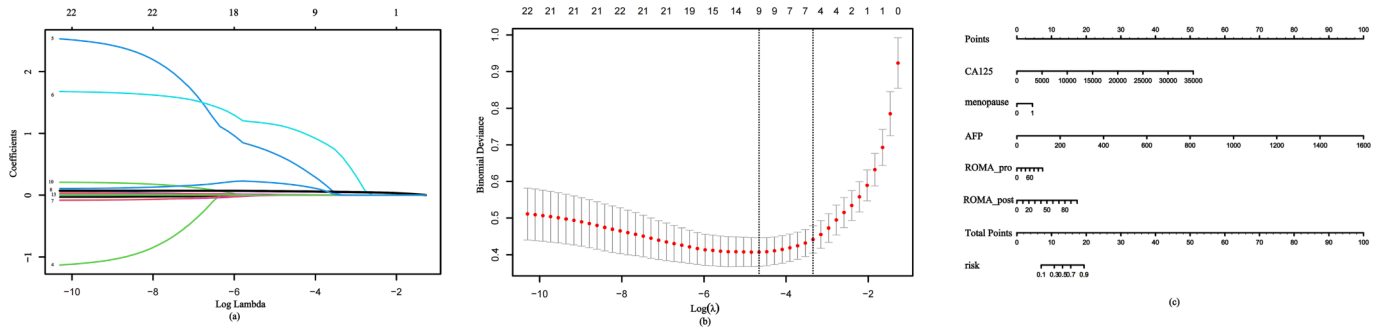

Fig. 1. The structure of LLRM. (a) Lasso regression variable screening. Each curve represents the changing trajectory of each independent variable coefficient, the ordinate is coefficient and the upper abscissa is the number of nonzero coefficient of the model. Variables are constantly compressed as λ increase, and the final coefficient is compressed to 0, indicating that the independent variable with the earliest one compressed to 0 has a low degree on the dependent variable. **(b) The process of Lasso regression screening λ for the most appropriate values by a cross-validation method.** The log (λ) is the abscissa, the deviation value is the ordinate, the number of variables is the upper coordinate, and the red dot in the figure represents each corresponding target parameter. We choose the λ corresponding to a standard square error, that is log (λ) = 0.035. At this time, the selected parameters are 4, which constructed an excellent performance and the minimum number of independent variables model. **(c) The nomogram model.** It was a visual display of the LLRM. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

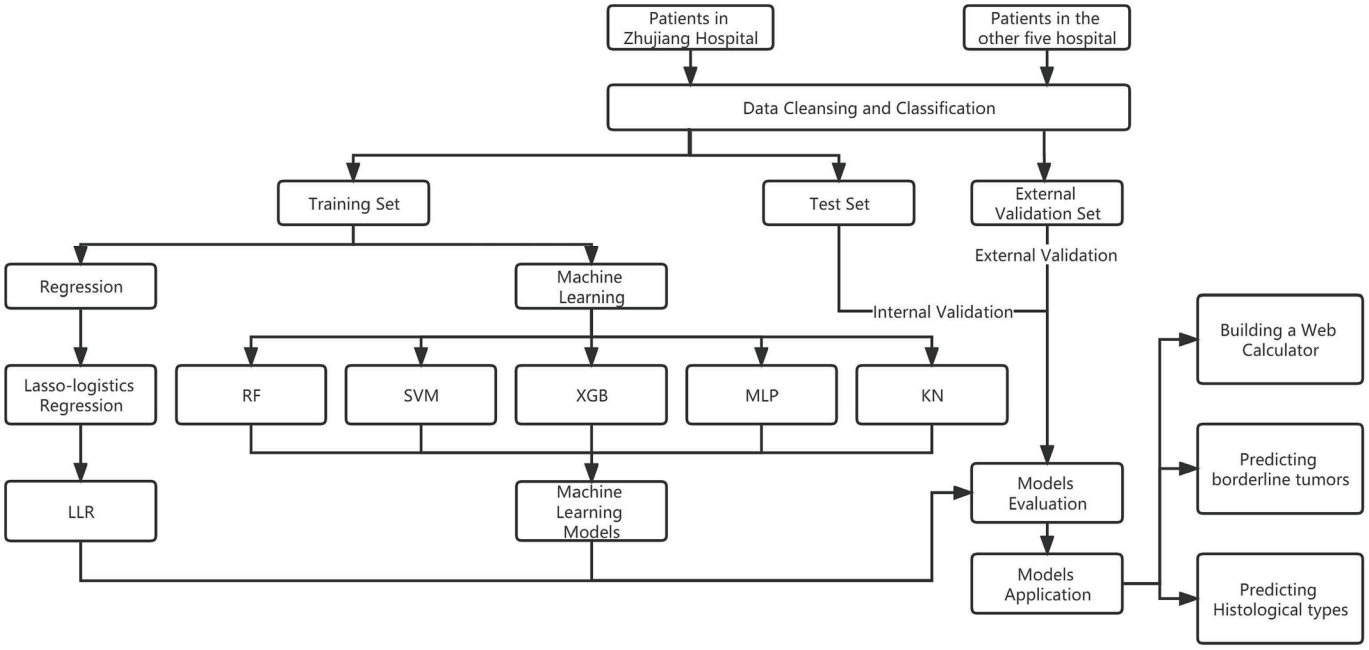

Fig. 2. Flow chart of model construction and verification. Firstly, the collected clinical data of patients with ovarian disease from six hospitals in China were collated and divided into training set, test set and validation set. Next, Lasso-Logistic regression, RF algorithm, SVM algorithm, XGB algorithm, MLP algorithm and KN algorithm were selected to construct the bi-classification prediction model for the predicition of benign and malignant ovarian diseases. Finally, the optimal algorithm was selected to construct three-classification and multi-classification prediction models to make preoperative prediction of ovarian borderline diseases and histological types of diseases.

LLRM, the AUC of RFM were significantly higher than LLRM (Z = -6.098, P < 0.050), indicating that the predictive power of machine learning models is significantly higher than that of statistical models, which provided a guiding significance for clinical application. The internal validation value of the RFM is as high as 0.968, indicating that the predictive power in the same distribution set is very high. Its external validation value also shows a good generalization ability and a high differentiate evaluation of external data, which is conducive to the promotion of the model in different hospitals. In the same way, the model examines more diverse populations will help improve the generalizability of the model (Moore et al., 2019). The results suggested that RFM have a high diagnostic value for the distinction between ovarian benign lesions and ovarian malignangt lesions. So we further choose the RF machine algorithm to optimize the prediction ability.

3.2. *Predictor importance ordering*

We computed the important predictors in the RFM (Fig. 3e). The top ten were: ROMA_post, HE4, ROMA_pro, CA125, AFP, CA199, CEA, age, MLR and RBC. Their importance coefficients were 0.284, 0.188, 0.148, 0.067, 0.049, 0.044, 0.037, 0.024, 0.021 and 0.021, respectively. First, RFM identified all tumor indicators, while LLRM did not screen out HE4, CA199 and CEA. Next, age, MLR and RBC are more important predictors besides the serum tumor index in the RFM but not statistically significant in the LLRM, indicating that machine learning is more powerful than clinical statistical prediction in the background of big data, which can identify the main indicators that can not be identified by statistically logistic. Katharina Anic proposed perioperative RBC transfusions have been associated with increased morbidity in some solid neoplasms (Anic

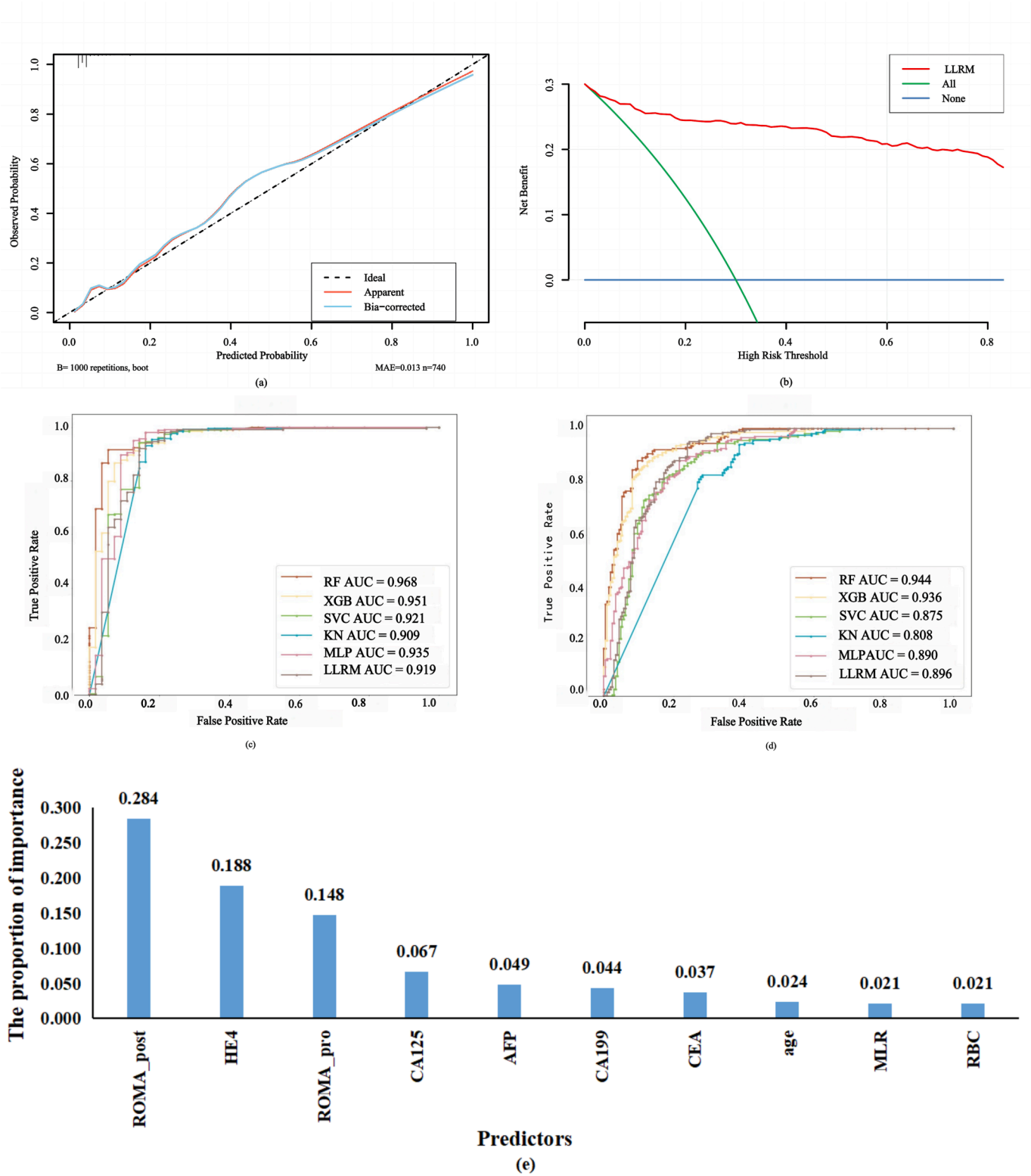

Fig. 3. The result of bi-classification models. (a) The calibration curve of LLRM. Red line was the performance of the LLRM, while the blue line correceted for any bias in LLRM. Dashed line was the reference line where the prediction of LLRM would like. Clinical effectiveness (Fig. 3b) indicated that when the predicted probability is 60% of non-benign is diagnosed and treated, then 20 per 100 people can benefit without harming any others. **(b) The DCA curve of LLRM.** The abscissa for prediction probability diagnosis threshold, ordinate for the clinical intervention benefit minus disadvantages net benefit rate, horizontal line for no treatment, the arc curve for all treatment, the benefit rate down to 0 finally, red solid prediction model DCA, the curve deviation from the two extreme cases, proves that the prediction model has certain clinical effectiveness. Hence, we can conclude that, treat patients on the basis of the prediction model leads to higher benefit than treating all patients, no patients, or only those patients who are positive on the diagnostic test. **(c) ROC and AUC of the machine learning models and LLRM in internal validation. (d) ROC and AUC of the machine learning models and LLRM in external validation. (e) Ranks the importance of the top 10 important predictors in the RF model.** (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

et al., 2022). Eiryo Kawakami proved that age was a critical variable in discriminating ovarian benign and malignant tumors (Kawakami et al., 2019). Cai Zhenzhen confirmed getting older and increase of PLR were independent risk factors for epithelial ovarian cancer (Cai Zhenzhen, 2021). Li Feixia confirmed that NLR can assist CA125 and HE4 in identifying early EOC (Xiafeng, 2020). Menopause state appeared less important in the RFM but important in LLRM, indicating that maybe because the RF builds a weak decision tree by selecting a subset of the variables, and obtains accuracy without over-fitting and multiple collinearity (Kawakami et al., 2019). Although the LLRM also selected a subset of variables that was statistically different, it relied entirely on the selected variables chosen to construct the equations for prediction. It shows that the mutual influence between independent variables is reduced in the establishment process of machine learning model, and the accuracy of the RFM is high.

3.3. Predicting the ovarian borderline diseases

Next, we selected the RF algorithm with the best predictive power to further predict borderline tumors, the sets of the model-building can be found in Table 1. Fig. 4a shows the ROC of the test set and the external validation set with predicted benign, malignant, and borderline diseases, their AUC are 0.927, 0.961, 0.742 and 0.838, 0.921, 0.527, and F1 values of 0.936, 0.782, 0.148 and 0.697, 0.782, 0.018, respectively. The triple-classification RF model also has some diagnostic value for the differentiation of borderline diseases, with an AUC of 0.742, which provides a new method for the preoperative diagnosis of borderline ovarian disease. The AUC for distinguishing the borderline diseases in the external validation was 0.527, which suggests that our predictive ability in the different population distribution needs to be further optimized and improved. Predictor importance analysis of the triple-classification RF model showed that ROMA_post, HE4, ROMA_pro, CA125, AFP, CA199, CEA, PLR, age and MLR were the top ten. Their importance coefficients were 0.127, 0.121, 0.120, 0.085, 0.065, 0.050, 0.045, 0.041, 0.041, 0.036, respectively. PLR has increased its importance, which has some research significance for predicting ovarian disease.Fig. 4b-4i showed the top eight important blood markers.

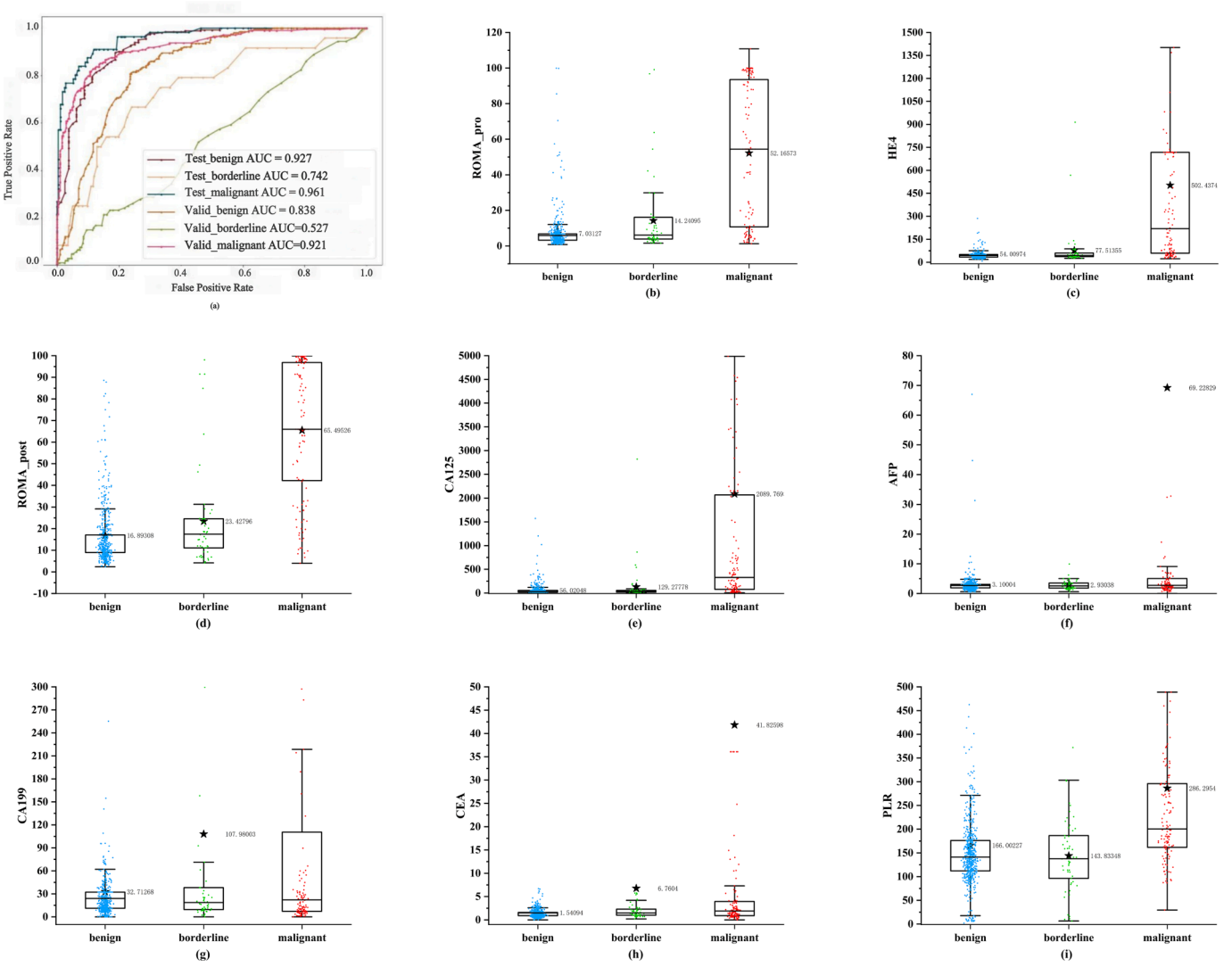

Fig. 4. The result of three classification model. (a) ROC and AUC of the RF three classification model in internal and external validation. The AUC for predicting benign, borderline, and malignant in the internal validation were 0.927, 0.742 and 0.961. The AUC for predicting benign, borderline, and malignant in the external validation were 0.838, 0.527 and 0.921. **(b) - (i) The Box-plot represents the distribution of the top eight important blood markers in benign, borderline and malignant diseases.** Blue dots, green dots and red dots represent the benign data, borderline data and malignant data. The asterisk refers to the mean value. The (b) - (i) represents ROMA-pro, HE4, ROMA-post, CA125, AFP, CA199, CEA and PLR, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.4. Predicting the ovarian histological types

We used the same approach to assess the predictive ability of the disease histology types. With the histology types as the outcome variables, the RF algorithm was used to establish the RF multi-classification prediction model in the training set. On the test set and the external validation set, the model evaluation (Fig. 5a) and the external validation (Fig. 5b) were conducted through the ROC curve. The results showed that the RF multi-classification model could significantly distinguish between ovarian endometriosis cysts, ovarian mature teratoma, and ovarian serous adenocarcinoma. Its AUC values in internal and external validation were 0.914 and 0.865, 0.903 and 0.897, 0.943 and 0.910, respectively. Fig. 5 showed RF can also better distinguish between other types of diseases, with their AUC values of more than 0.800. We can conclude that the RF multi-classification prediction model had the highest predictive value for ovarian serous adenocarcinoma, and had a certain predictive power for different histology types of diseases. Ovarian sex cord interstitial tumor is a group of tumors with sex hormone secretion function, as sex hormone indicators were not included in our study, the predictive ability of ovarian sex cord interstitial tumor needs to be further improved.

4. Discussion

Early detection and early diagnosis of OC have become a research hotspot. The diagnosis of OC mainly relies on histopathological specimens, systematic screening of a large population using invasive diagnostic techniques such as tissue biopsy is impractical, and non-invasive techniques such as ultrasound imaging require sonographer with high expertise to distinguish them manually. These diagnosis method all exist judgment bias (Yue et al., 2021). Secondly, the discovery of the earliest CA125 has laid the foundation for the early diagnosis of OC, but neither diagnostic specificity nor sensitivity was high (Chunfang, 2013). The emergence of ROMA is combining HE4 to define the risk stratification of benign and malignant in order to improve the specificity and sensitivity of CA125 diagnosis (Dochez et al., 2019; Yan et al., 2019). The RMI was established to budget OC risk by combining CA125 with ultrasound (Karlsen et al., 2015). The UKCTOCS has developed and tested a new ovarian cancer risk algorithm, named ROCA, to assess CA125 changes over time to predict the risk of OC (Naumann and Brown, 2018). At present, it has entered the era of big data, and massive data and information is constantly generated. Machine learning develops a new model after learning potential patterns from large samples of data information, and multiple exercises and learns to achieve the mode of AI judgment (Kawakami et al., 2019). It also has a good application in other systems of cancer diagnosis, which contributes to the preoperative understanding of cancer diagnosis and classification, and provides great opportunities for the precise treatment of OC (Xiangyuan et al., 2021). Therefore, building machine learning diagnosis and prediction models with simple and accessible serological indicators is worth exploring.

This study through clinical statistics and various machine learning

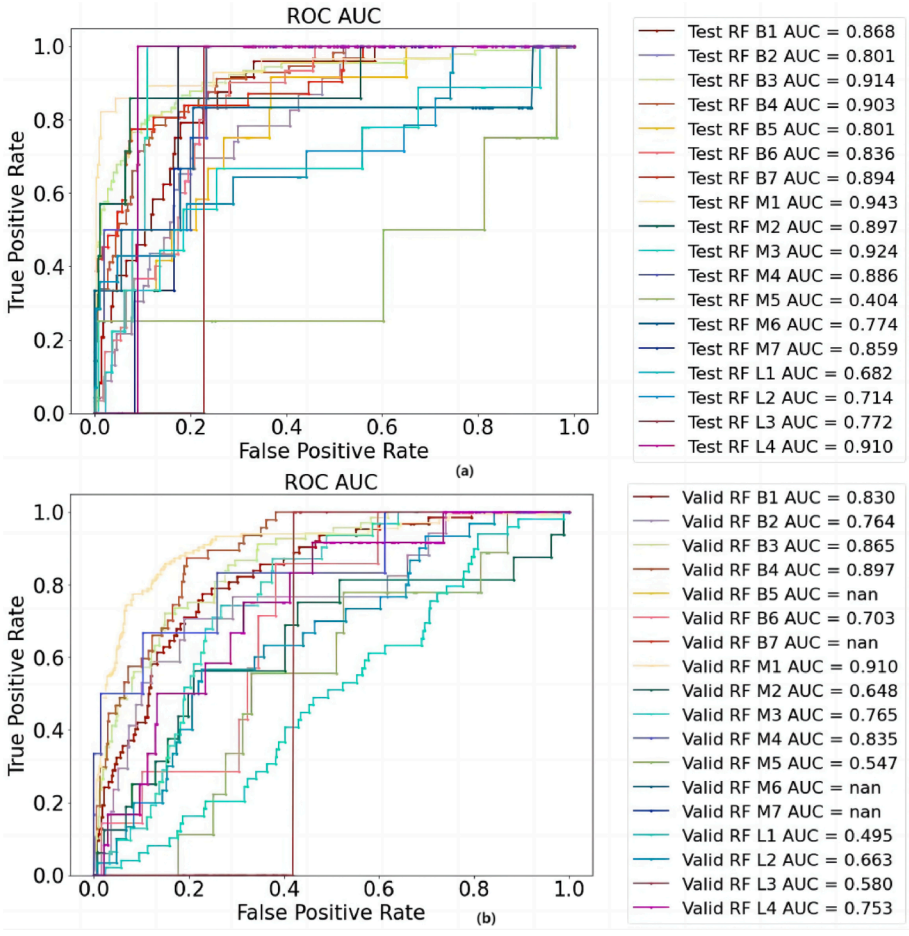

Fig. 5. The result of multi-classification model. (a) ROC and AUC of the RF multi-classification model in internal validation. (b) ROC and AUC of the RF multi-classification model in external validation. B1 represents ovary serous cystadenoma (AUC in internal validation = 0.868, AUC in external validation = 0.830). B2 represents ovary mucinous cystadenoma (AUC in internal validation = 0.801, AUC in external validation = 0.764). B3 represents ovary chocolate cyst (AUC in internal validation = 0.914, AUC in external validation = 0.865). B4 represents ovary mature teratoma (AUC in internal validation = 0.903, AUC in external validation = 0.897). B5 represents ovarian fibroma and follomoma (AUC in internal validation = 0.836). B6 represents ovary mature teratoma (AUC in internal validation = 0.903, AUC in external validation = 0.703). B7 represents ovary inflammatory (AUC in internal validation = 0.894). M1 represents ovary pulsar carcinoma (AUC in internal validation = 0.943, AUC in external validation = 0.910). M2 represents ovary mucous carcinoma (AUC in internal validation = 0.897, AUC in external validation = 0.648). M3 represents ovary endometrioid carcinoma (AUC in internal validation = 0.924, AUC in external validation = 0.765). M4 represents ovary malignant germ cell tumors (AUC in internal validation = 0.886, AUC in external validation = 0.835). M5 represents ovary sex cord stromal tumor (AUC in internal validation = 0.404, AUC in external validation = 0.547). M6 represents ovary metastatic carcinoma (AUC in internal validation = 0.774). M7 represents the remaining species of OC (AUC in internal validation = 0.859). L1 represents ovary serous borderline tumor (AUC in internal validation = 0.682, AUC in external validation = 0.495). L2 represents ovary mucinous borderline tumor (AUC in internal validation = 0.714, AUC in external validation = 0.663). L3 represents clear-cellular borderline ovary tumor (AUC in internal validation = 0.772, AUC in external validation = 0.580). L1 represents mixed ovary borderline tumor (AUC in internal validation = 0.910, AUC in external validation = 0.752).

methods, based on 1655 patients of clinical population informatics and serological index, clinical prediction model and machine learning prediction model were constructed, evaluating the benign, malignant and borderline ovarian lesion, and further distinguishing the histological type of malignant tumor, and providing the basis for early diagnosis of ovarian tumor and further personalized treatment. We distinguished a benign and malignant ovarian disease by establishing the LLRM, RFM, XGBM, SVMM, KNM and MLPM. The results show that the RFM have higher predictive power(AUC = 0.968) and provide more valuable diagnostic information, which may facilitate the development of personalized treatment strategies before the primary treatment method for OC. In addition, we know from the important predictor analysis that age and MLR are not statistically different in the LLRM, but they are important predictors of machine learning, indicating that machine learning can help identify new biomarker that cannot be identified by the regression models and further improve the efficacy of the prediction (Kawakami et al., 2019). Finally, we use the RF to conduct triple classification and even multiple classification models. Based on the difficulty in clinical preoperative diagnosis of borderline tumors, and rare studies have predicted borderline ovarian tumors separately from OC, our study constructed an RF triple-classification model to distinguish the benign, malignant and borderline characteristics of ovarian diseases, with AUC: 0.927, 0.961 and 0.742. But this model needs to be further improved, probably because of the lack of strong distinguishing features of borderline diseases at the level of serological indicators. The RF multiple classification model has the highest predictive value for ovarian epithelial tumors (such as ovarian serous adenocarcinoma), and has a certain predictive power for different histological types of diseases. These results suggest that both LLRM and machine learning predictive models can provide valuable diagnostic information for ovarian lesions based on preoperative serological markers, which may facilitate the corresponding personalized treatment and determine the surgical scope for OC. And also provide valuable information to clinicians on the examination of patients' disease stratification.

Limitations of this study: First, this study belongs to a retrospective study design, and there may be selection bias and differences in disease types and clinical practice between different levels of hospitals (Hwangbo et al., 2021). Second, we used only patient clinical pathological information in this study, and the lack of gene data may affect the predictive efficacy (Paik et al., 2019). We will deepen the current work and further improve the above deficiencies in the follow-up researches.

5. Conclusions

In conclusion, based on the traditional statistical method and machine learning algorithm, we developed stable and powerful prediction models for evaluating ovarian lesions. RFM is undeniably powerful forecasting tool that can distinguish benign and malignant ovarian ovarian disease. In particular, RFM has predictive value for ovarian borderline disease and clinical guidance for predicting the histological types of ovarian lesions.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Medical Ethics Committee of Zhujiang Hospital, Southern Medical University (protocol code 2022-KY-141–01 and Sept. 21, 2022).

Informed Consent Statement: Informed consent was waived due to the noninvasive and retrospective of our study.

Disclosure of funding and conflicts of Interest: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. And all authors disclosed no relevant relationships.

CRediT authorship contribution statement

Bilin Jing: Conceptualization, Methodology, Formal analysis, Writing – original draft, Visualization. **Gaowen Chen:** Writing – review & editing, Supervision. **Miner Yang:** Writing – review & editing. **Zhi Zhang:** Software, Formal analysis. **Yue Zhang:** Formal analysis, Data curation. **Jingyao Zhang:** Formal analysis, Data curation. **Juncheng Xie:** Formal analysis, Data curation. **Wenjie Hou:** Data curation. **Yong Xie:** Data curation. **Yi Huang:** Data curation. **Lijie Zhao:** Data curation. **Hua Yuan:** Data curation. **Weilin Liao:** Validation, Writing – review & editing. **Yifeng Wang:** Conceptualization, Validation, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data presented in this study are available on request from the corresponding author. The data are not publicly available due to restrictions ethical.

References

Anic, K., Schmidt, M.W., Schmidt, M., Krajnak, S., Löwe, A., Linz, V.C., Schwab, R., Weikel, W., Brenner, W., Westphalen, C., Rissel, R., Hartmann, E.K., Conradi, R., Hasenburg, A., Battista, M.J., 2022. Impact of perioperative red blood cell transfusion, anemia of cancer and global health status on the prognosis of elderly patients with endometrial and ovarian cancer[J]. Front Oncol 12, 967421. https://doi.org/10.3389/fonc.2022.967421.

Balkwill, F., Mantovani, A., 2001. Inflammation and cancer: back to Virchow?[J]. Lancet 357 (9255), 539–545. https://doi.org/10.1016/S0140-6736(00)04046-0.

Cai Zhenzhen, Z.J., 2021. Value of PLR, D-dimer, and CA125 in the diagnosis and prognostic evaluation of epithelial ovarian cancer[J]. International Journal of Laboratory Medicine 42 (24), 2999–3003.

Chon, S., Lee, S., Jeong, D., Lim, S., Lee, K., Shin, J., 2021. Elevated platelet lymphocyte ratio is a poor prognostic factor in advanced epithelial ovarian cancer[J]. Journal of Gynecology Obstetrics and Human Reproduction 50 (6), 101849.

Chunfang, D.u., 2013. A new progress in the early diagnosis of serological markers of ovarian cancer[J]. Modern Oncology Medicine 21 (12), 2866–2869.

Dochez, V., Caillon, H., Vaucel, E., Dimet, J., Winer, N., Ducarme, G., 2019. Biomarkers and algorithms for diagnosis of ovarian cancer: CA125, HE4, RMI and ROMA, a review[J]. J Ovarian Res 12 (1). https://doi.org/10.1186/s13048-019-0503-7.

Hwangbo, S., Kim, S.I., Kim, J.-H., Eoh, K.J., Lee, C., Kim, Y.T., Suh, D.-S., Park, T., Song, Y.S., 2021. Development of Machine Learning Models to Predict Platinum Sensitivity of High-Grade Serous Ovarian Carcinoma[J]. Cancers 13 (8), 1875.

Jia Jiyun HXXX. The relationship between preoperative NLR, PLR, and serum CEA and the pathological features of epithelial ovarian cancer and their clinical predictive value for prognosis[J]. The Practical Journal of Cancer, 2022, 37(8): 1359-63. DOI: 10.3969/j.issn.1001-5930.2022.08.038.

Karlsen, M.A., Høgdall, E.V.S., Christensen, I.J., Borgfeldt, C., Kalapotharakos, G., Zdrazilova-Dubska, L., Chovanec, J., Lok, C.A.R., Stiekema, A., Mutz-Dehbalaie, I., Rosenthal, A.N., Moore, E.K., Schodin, B.A., Sumpaico, W.W., Sundfeldt, K., Kristjansdottir, B., Zapardiel, I., Høgdall, C.K., 2015. A novel diagnostic index combining HE4, CA125 and age may improve triage of women with suspected ovarian cancer — An international multicenter study in women with an ovarian mass[J]. Gynecol Oncol 138 (3), 640–646.

Kawakami E, Tabata J, Yanaihara N, et al. Application of Artificial Intelligence for Preoperative Diagnostic and Prognostic Prediction in Epithelial Ovarian Cancer Based on Blood Biomarkers[J]. Clin Cancer Res, 2019, 25(10): 3006-15. DOI: 10.1158/1078-0432.CCR-18-3378.

McCluggage, W.G., Singh, N., Gilks, C.B., 2022. Key changes to the World Health Organization (WHO) classification of female genital tumours introduced in the 80, 762–778.

Meng, W.u., Ying, W., Qichao, Z., Ping, L.i., Jie, T., 2017. Clinical value of combining transvaginal contrast-enhanced ultrasonography with serum human epididymisprotein-4 and the resistance index for early-stage epithelial ovarian cancer[J]. Saudi Med J 38 (6), 592–597.

Moore, R.G., Blackman, A., Miller, M.C., Robison, K., DiSilvestro, P.A., Eklund, E.E., Strongin, R., Messerlian, G., 2019. Multiple biomarker algorithms to predict epithelial ovarian cancer in women with a pelvic mass: Can additional makers improve performance?[J]. Gynecol Oncol 154 (1), 150–155.

Mutch, D.G., Prat, J., 2014. 2014 FIGO staging for ovarian, fallopian tube and peritoneal cancer[J]. Gynecol Oncol 133 (3), 401–404. https://doi.org/10.1016/j.ygyno.2014.04.013.

Naumann, R.W., Brown, J., 2018. Ovarian cancer screening with the Risk of Ovarian Cancer Algorithm (ROCA): Good, bad, or just expensive?[J]. Gynecol Oncol 149 (1), 117–120. https://doi.org/10.1016/j.ygyno.2018.01.029.

Paik, E.S., Lee, J.-W., Park, J.-Y., Kim, J.-H., Kim, M., Kim, T.-J., Choi, C.H., Kim, B.-G., Bae, D.-S., Seo, S.W., 2019. Prediction of survival outcomes in patients with

epithelial ovarian cancer using machine learning methods[J]. J Gynecol Oncol 30 (4). https://doi.org/10.3802/jgo.2019.30.e65.

Pharoah, P.D.P., 2012. The Potential for Risk Stratification in the Management of Ovarian Cancer Risk[J]. Int J Gynecol Cancer 22, S16–S17. https://doi.org/10.1097/IGC.0b013e318251caaf.

Qin, M.O., Pca-,, 2021. MPL- -ANN model in the identification of benign and malignant ovarian tumors[J]. Medical. Information Volume 34(Issue 7).

Qundi YXAS. Construction of prediction model for malignancy risk in ovarian tumor patients based on blood routine and CA125 index[J]. Chinese Joumal of Hospital Statistics, 2021, 28(4). DOI:10. 3969/j. issn. 1006-5253, 2021. 04. 005.

Shen, L., Zhang, H., Liang, L., Li, G., Fan, M., Wu, Y., Zhu, J.i., Zhang, Z., 2014. Baseline neutrophil-lymphocyte ratio (≥2.8) as a prognostic factor for patients with locally advanced rectal cancer undergoing neoadjuvant chemoradiation[J]. Radiat Oncol 9 (1). https://doi.org/10.1186/s13014-014-0295-2.

Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F., 2021. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries[J]. CA: a cancer journal for clinicians 71 (3), 209–249.

Tang, Y., Hu, H.-Q., Tang, Y.-L., Tang, F.-X., Zheng, X.-M., Deng, L.-H., Yang, M.-T., Yin, S.u., Li, J., Xu, F., 2021. Preoperative LMR and Serum CA125 Level as Risk Factors for Advanced Stage of Ovarian Cancer[J]. J Cancer 12 (19), 5923–5928.

Tongshuo, Z., 2018. a multi-test index combined diagnosis model for ovarian cancer based on integrated machine learning[J]. Journal of Clinical Laboratory 36 (12), 908–913.

Xiafeng, L., 2020. Clinical value of serum CA125 and HE4 combined with PLR, NLR and MLR in the diagnosis of epithelial ovarian cancer[J]. Shandong Medicine 60 (10), 70–72.

Xiangyuan, L.i., Jinbao, Z., Penghua, L., 2021. Application status and prospect of deep learning in medical imaging field[J]. Journal of. Clinical Radiology 40 (12), 2423–2429.

Yan, L., Yifeng, W., Gaowen, C., 2019. Diagnostic accuracy of ROMA index in diagnosing ovarian malignancy in pelvic mass patients[J]. Natl Med J China 27, 2141–2144.

Yaqin, L.L., Tan., 2022. The diagnostic value and clinical significance of mesothelin and carbohydrate antigen 125 alone and in combination for ovarian cancer[J]. Cancer progression 20 (11), 1125–1128.

Yue Z, Sun C, Chen F, et al. Machine learning-based LIBS spectrum analysis of human blood plasma allows ovarian cancer diagnosis[J]. Biomed Opt Express, 2021, 12(5): 2559. DOI:10.1364/BOE.421961.

Yuyu, Z., Yanqiu, W., Xiaoping, W., 2011. The selection of serum markers for the early diagnosis of ovarian cancer in different tissue types[J]. Progress in Modern Obstetrics and Gynecology 20 (08), 663–665.

Zhang M, Cheng S, Jin Y, et al. Roles of CA125 in diagnosis, prediction, and oncogenesis of ovarian cancer[J]. Biochimica et Biophysica Acta (BBA) - Reviews on Cancer, 2021, 1875(2): 188503. DOI:10.1016/j.bbcan.2021.188503.