Contents lists available at ScienceDirect

International Journal of Medical Informatics

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

Using machine learning to predict ovarian cancer

Mingyang Lu[a,b,c,1], Zhenjiang Fan[d,1], Bin Xu[a,b,c], Lujun Chen[a,b,c], Xiao Zheng[a,b,c], Jundong Li[e], Taieb Znati[d], Qi Mi[f,*], Jingting Jiang[a,b,c,*]

[a] *Department of Tumor Biological Treatment, the Third Affiliated Hospital of Soochow University, Changzhou, Jiangsu, People's Republic of China*
[b] *Jiangsu Engineering Research Center for Tumor Immunotherapy, Changzhou, Jiangsu, People's Republic of China*
[c] *Institute of Cell Therapy, Soochow University, Changzhou, Jiangsu, People's Republic of China*
[d] *Department of Computer Science, University of Pittsburgh, Pittsburgh, PA, USA*
[e] *Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA, USA*
[f] *Department of Sports Medicine and Nutrition, University of Pittsburgh, Pittsburgh, PA, USA*

ARTICLE INFO

Keywords:
Ovarian Cancer
Tumor Marker
Machine Learning

ABSTRACT

Objective: Ovarian cancer (OC) is one of the most common types of cancer in women. Accurately prediction of benign ovarian tumors (BOT) and OC has important practical value.
Methods: Our dataset consists of 349 Chinese patients with 49 variables including demographics, blood routine test, general chemistry, and tumor markers. Machine learning Minimum Redundancy – Maximum Relevance (MRMR) feature selection method was applied on the 235 patients' data (89 BOT and 146 OC) to select the most relevant features, with which a simple decision tree model was constructed. The model was tested on the rest of 114 patients (89 BOT and 25 OC). The results were compared with the predictions produced by using the risk of ovarian malignancy algorithm (ROMA) and logistic regression model.
Results: Ten notable features were selected by MRMR, among which two were identified as the top features by the decision tree model: human epididymis protein 4 (HE4) and carcinoembryonic antigen (CEA). Particularly, CEA is a valuable marker for OC prediction in patients with low HE4. The model also yields better prediction result than ROMA.
Conclusion: Machine learning approaches were able to accurately classify BOT and OC. Our goal is to derive a simple predictive model which also carries a good performance. Using our approach, we obtained a model that consists of just two biomarkers, HE4 and CEA. The model is simple to interpret and outperforms the existing OC prediction methods. It demonstrates that the machine learning approach has good potential in predictive modeling for the complex diseases.

1. Introduction

Ovarian cancer (OC) is one of the most common types of cancer in women [1]. It was found that 295,414 women were diagnosed with ovarian cancer and 184,799 deaths were caused by it worldwide in the year of 2018 [2]. Most patients with ovarian cancer have advanced disease at the time of diagnosis because early-stage tumors are typically asymptomatic, resulting in a poorer long-term survival [3]. Although ovarian cancers are chemo-sensitive and generally show initial efficacy against platinum/taxane treatment, the 5-year recurrence rates are 60% to 80% in those with advanced disease [4]. Therefore, considerable effort has been focused on the establishment of novel methods for the prediction of disease progression and prognosis for this malignancy.

To diagnose whether a patient is carrying ovarian cancer, besides a physical examination (including a pelvic examination) or transvaginal ultrasound [5,6], the tumor biomarkers such as carbohydrate antigen 125 (CA125), carbohydrate antigen 72-4 (CA72-4) [7], and human epididymis protein 4 (HE4) [8] are often assessed clinically. Many studies have been conducted to evaluate the performance of these biomarkers and the indices that utilizing them (e.g. Risk Ovarian Malignancy Algorithm (ROMA) [9], a dual marker algorithm utilizing HE4 and CA125; Risk Malignancy Index (RMI) based on the image features, menopausal status, and CA125 [10]) to differentiate benign status from OC. Anton et al. compared the sensitivities of CA125, HE4, ROMA and RMI on 128 patients [11], and showed that HE4 demonstrated the best overall sensitivity for the evaluation of the malignant ovarian tumor.

* Corresponding author.
E-mail addresses: qi.mi@pitt.edu (Q. Mi), jiangjingting@suda.edu.cn (J. Jiang).

[1] Mingyang Lu and Zhenjiang Fan contributed equally to this work.

https://doi.org/10.1016/j.ijmedinf.2020.104195
Received 1 March 2020; Received in revised form 24 April 2020; Accepted 21 May 2020

Moore et al. compared RMI and ROMA to predict epithelial ovarian cancer (EOC) in 457 women and concluded that ROMA achieves a significantly higher sensitivity for identifying women with EOC than does RMI [6]. Wang *et al.* performed a meta-analysis to evaluate the diagnostic value of HE4, CA125, and ROMA based on 32 studies, and they suggested HE4 is useful for diagnosing OC, especially in the premenopausal population. CA125 and ROMA are more suitable for diagnosing OC in the postmenopausal population [12]. Zhang et al. developed a linear multi-marker model by combining HE4, CA125, progesterone (Prog), and estradiol (E2) and the last two markers were suggested by the numbers of epidemiological evidence in the development and progression of OC [13,14]. Their multi-marker model showed significant improvement when compared to CA125 or HE4 to differentiate benign pelvic masses (BPM) and EOC patients [15]. In summary, the current clinical tests for diagnosis of OC focus on a small number of biomarkers that were either selected by meeting over-expression of genes/proteins criteria [16] or the epidemiological evidence. Machine learning (ML) is an emerging research area which offers a variety of useful methodologies that can handle large dimensional dataset and it excels in providing methods which can efficiently and effectively evaluate a large number of variables to construct an accurate model for prediction [17,18,19,20]. In the medical research, the ML methods such as decision trees have been applied successfully to predict mortality of trauma [21,23], and breast cancer [22]. In this study, we will employ the methodologies in machine learning to develop a predictive model by investigating a panel of 49 measures from 349 patients which include demographics, blood routine test, general chemistry, and tumor marker for classifying BOT and OC.

2. Materials and methods

2.1. Patients and tumor characteristics

The samples were collected from 171 ovarian cancer patients and 178 patients of benign ovarian tumors who underwent surgical resection between July 2011 and July 2018 in the Third Affiliated Hospital of Soochow University (supplementary data 1). 235 patients' data collected from July 2011 to July 2017 were used as the training data for model development (see supplementary data 3 for imputed data, supplementary data 4 for raw data, and supplementary Table 1 for description). The remaining 114 patients' data collected from Aug 2017 to July 2018 were used as the testing data (the description of the testing data is shown in supplementary Table 2, see supplementary data 5 for raw data). All the patients were diagnosed by pathology after surgery. None of the ovarian cancer patients received pre-operative chemotherapy or radiotherapy. The histological type was classified according to the World Health Organization (WHO) criteria. The patient's demographics and histological type information are listed in Table 1. This study was approved by the ethics committee of the Third Affiliated Hospital of Soochow University.

2.2. Blood routine test, the general chemistry tests, and tumor markers detection method

The blood routine tests were analyzed by Sysmex XE-2100 automated hematology analyzer (Sysmex, Japan) with full blood. The general chemistry tests were analyzed by Beckman Coulter AU5800 series of clinical chemistry analyzers (Beckman Coulter, USA) with serum. The tumor markers were analyzed by Roche Cobas 8000 modular analyzer series (Roche, Switzerland) with serum. The detailed experiment method for each test item is listed in the supplementary data 2.

2.3. Statistical analysis

All statistical analyses were performed using SPSS 22.0 software (SPSS, Inc., Chicago, IL) and GraphPad Prism 8.0 software package (GraphPad Software, Inc., San Diego, USA). Non-normal distribution data are represented by median and IQR (Interquartile Range), and Mann-Whitney U Test was used to compare the difference of two groups. The *P* value of less than 0.05 based on the two-sided test was considered to be statistically significant.

3. Machine learning analysis

3.1. Missing value

For the variables in the dataset, most have the low missing value rate (less than 7% missing), except CA74-4 which has 69% missing and Neutrophil Ratio(%) (100% complete in training and 80% missing in testing data). For the variables that have the low missing value rate, they were imputed by using their mean.

3.2. Feature selection

Given in our dataset, it consists of 49 variables which include histology, demographic, blood routine test, general chemistry, tumor markers as shown in Table 1. To effectively prepare this high-dimensional data for ML, we employed a data dimension reduction strategy, namely feature selection, that can help derive a clean and the most relevant subset of data to predict the outcome (benign/OC).

3.3. Minimum Redundancy - Maximum Relevance (MRMR)

Generally, filter type feature selection methods (based on mutual information [24] or statistical tests (t-test, F-test) [25–27]) use a simple ranking approach to identify the most relevant features to the target class, where only the top-ranked features are selected. They ignore the fact that the selected futures could be highly correlated among themselves, creating the issue of "redundancy" among the selected features. Given the selected features may be highly correlated, the underlying 'relevance' space to the target classes is not maximally represented by them. MRMR is a filter type feature selection method proposed by Ding and Peng et al [28]. MRMR aims to find the maximal relevant features to the target classes while it is also trying to make sure that these selected features are mutually maximally dissimilar to each other. Given the two conditions above, MRMR optimizes them simultaneously by combining them into a single criterion function. The first condition (minimum redundancy condition), which aims to find a set of features that are mutually maximally dissimilar from each other, can be denoted as:

$$Min\ W_I,\ W_I = \frac{1}{|S|^2} \sum_{i,j \in S} I\,(i, j) \tag{1}$$

where *S* is the subset of selected features, *i* and *j* represent variable *i* and variable *j* in the subset respectively, and *I (i, j)* is the mutual information function between variable *i* and variable *j*.

The second condition (the maximum relevance condition) is given by:

$$Max\ V_I,\ V_I = \frac{1}{|S|} \sum_{i \in S} I\,(i, h) \tag{2}$$

where *h* is the target variable.

The mutual information I of two variables x and y are defined based on their joint probabilistic distribution p(x, y) and the respective marginal probabilities p(x) and p(y):

$$I\ (x,\ y) = \sum_{i,j} p(x_i, y_j) log \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \tag{3}$$

MRMR algorithm selects the significant feature variables that are mutually far away from the target variable while still having "high"

Table 1
Statistics summary of the full data set for 349 patients.

	Ovarian Cancer			Benign Ovarian Tumors			
	N = 171			N = 178			
Histology							
OC, *n*(%)		171 (100%)			NA		
Epithelial Ovarian Cancer		157(91.81%)					
Ovarian Germ Cell Cancer		5 (2.92%)					
Others		9 (5.26%)					
BOT, *n*(%)		NA			178 (100%)		
Ovarian Teratoma					36 (20.22%)		
Ovarian Cyst					142 (79.78%)		
Demographics	**N**	**Median**	**IQR**	**N**	**Median**	**IQR**	***P***
Age	171	53.00	45-63	178	36.00	27.00-46.00	< 0.001
Premenopausal	75			155			
Postmenopausal	96			23			
Blood Routine Test							
Mean Platelet Volume(fL)	171	10.10	9.10-11.20	176	10.40	8.98-11.20	0.407
Basophil Cell Count(10ˆ9/L)	171	0.02	0.01-0.04	178	0.03	0.01-0.04	0.312
Basophil Cell Ratio(%)	171	0.40	0.20-0.60	178	0.44	0.20-0.72	0.054
Eosinophil Count(10ˆ9/L)	171	0.05	0.02-0.08	178	0.05	0.02-0.10	0.421
Mean Corpuscular Hemoglubin(Pg)	171	28.90	27.40-30.00	178	29.85	28.30-30.50	< 0.001
Red Blood Cell Distribution Width(%)	171	13.10	12.30-14.40	178	12.90	12.40-14.40	0.376
Platelet Distribution Width(%)	171	13.40	11.20-16.52	176	14.10	12.50-17.00	0.002
Hemoglobin(g/L)	171	124.00	116.00-132.00	178	129.50	121.00-138.00	< 0.001
Lymphocyte Count(10ˆ9/L)	171	1.34	1.07-1.67	178	1.67	1.31-2.02	< 0.001
Lymphocyte Ratio(%)	171	23.40	15.42-28.40	178	29.80	23.40-36.43	< 0.001
Mononuclear Cell Count(10ˆ9/L)	171	0.35	0.28-0.47	178	0.31	0.24-0.39	< 0.001
Platelet Count(10ˆ9/L)	171	265.00	205.00-326.00	178	223.50	192.00-267.00	< 0.001
Neutrophil Ratio(%)	169	70.60	63.50-78.20	89	59.50	52.80-65.70	< 0.001
Eosinophil Ratio(%)	171	0.80	0.30-1.40	178	0.90	0.40-1.60	0.134
Red Blood Cell Count(10ˆ12/L)	171	4.32	4.08-4.61	178	4.42	4.14-4.71	0.031
Thrombocytocrit(L/L)	171	0.26	0.21-0.32	176	0.23	0.19-0.27	< 0.001
Hematocrit(L/L)	171	0.39	0.35-0.40	178	0.39	0.37-0.41	0.033
Monocyte Ratio(%)	171	5.55	4.70-6.69	178	5.28	4.02-6.41	0.068
Mean Corpuscular Volume(fL)	171	88.40	85.60-91.30	178	89.60	85.80-91.50	0.414
General Chemistry							
Phosphorus(mmol/L)	171	1.12	1.00-1.21	178	1.13	0.99-1.26	0.641
Glucose(mmol/L)	171	5.23	4.64-5.85	178	5.00	4.66-5.42	0.014
Kalium(mmol/L)	171	4.36	4.17-4.65	178	4.38	4.15-4.67	0.908
Aspartate Aminotransferase(U/L)	167	19.00	14.00-24.00	172	16.00	13.00-20.00	< 0.001
Magnesium(mmol/L)	171	0.96	0.89-1.05	178	0.97	0.91-1.04	0.571
Chlorine(mmol/L)	171	101.00	99.10-102.80	178	100.80	99.20-102.40	0.340
Albumin(g/L)	167	39.20	34.70-44.20	172	43.40	41.40-46.00	< 0.001
Indirect Bilirubin(μmol/L)	167	5.00	3.50-6.40	172	6.00	4.80-7.60	< 0.001
Gama Glutamyltransferasey(U/L)	167	17.00	13.00-24.00	172	15.00	12.00-22.00	0.007
Globulin(g/L)	167	31.00	27.50-34.10	172	28.85	26.50-32.00	0.001
Alanine Aminotransferase(U/L)	167	15.00	11.00-21.00	172	15.00	12.00-20.00	0.353
Direct Bilirubin(μmol/L)	167	2.50	2.00-3.40	172	3.15	2.40-3.80	< 0.001
Creatinine(μmol/L)	171	62.70	54.30-70.00	178	64.00	57.00-72.00	0.060
Natrium(mmol/L)	171	141.30	139.00-143.20	178	140.10	138.60-141.70	< 0.001
Blood Urea Nitrogen(mmol/L)	171	3.75	2.99-4.84	178	3.88	3.36-4.62	0.449
Calcium(mmol/L)	171	2.44	2.29-2.53	178	2.52	2.35-2.60	< 0.001
Anion Gap(mmol/L)	171	19.97	17.25-22.67	177	19.74	17.30-21.95	0.339
Total Protein(g/L)	167	71.90	65.30-75.90	172	73.00	69.20-76.10	0.015
Urie Acid(μmol/L)	171	236.90	198.80-279.90	178	233.25	200.30-276.10	0.982
Carban Dioxide-combining Power(mmol/L)	171	24.10	22.20-27.00	177	24.00	22.70-25.50	0.329
Total Bilirubin(μmol/L)	167	7.60	5.70-9.60	172	9.20	7.30-11.50	< 0.001
Alkaline Phosphatase(U/L)	167	77.00	64.00-97.00	172	65.00	55.00-77.00	< 0.001
Tumor Marker							
Carbohydrate Antigen 72-4(U/mL)	46	4.83	1.98-12.18	63	1.74	0.79-3.40	< 0.001
Carcinoembryonic Antigen(ng/mL)	149	1.41	0.85-2.63	178	1.27	0.83-1.90	0.103
Carbohydrate Antigen 19-9(U/mL)	147	14.99	7.14-43.48	178	14.00	8.00-25.00	0.200
Alpha-fetoprotein(ng/mL)	150	2.47	1.80-3.58	177	2.10	1.00-3.00	0.012
Human Epididymis Protein 4(pmol/L)	151	140.90	59.36-393.10	178	43.77	38.80-51.57	< 0.001
Carbohydrate Antigen 125(U/mL)	154	241.5	48.14-758.10	178	22.66	14.00-49.00	< 0.001

correlation to the target variable. Mutual information can be used to quantify the dependency between each of the feature variables and the target variable. MRMR is a supervised feature selection model where it maximizes the dependency between the joint distribution of the selected features and the target variable. The numerical features in our dataset were converted to 10-binned categorical variables when MRMR was performed in this study.

To determine the optimal number of features that will be used for building the decision tree model, forty-eight MRMR feature selection experiments were performed which is based on the number of variables in our dataset (CA72-4 was not included due to its high percentage of missing value, see supplementary data 3). In each *i*th ($i = 1$ to 48) experiment, the algorithm divides the samples into 10 groups with equal size. The top *i* number of features were selected by MRMR based on the random 9 groups of samples, followed by a 5-level CART decision tree model built using only selected features on the same 9 groups

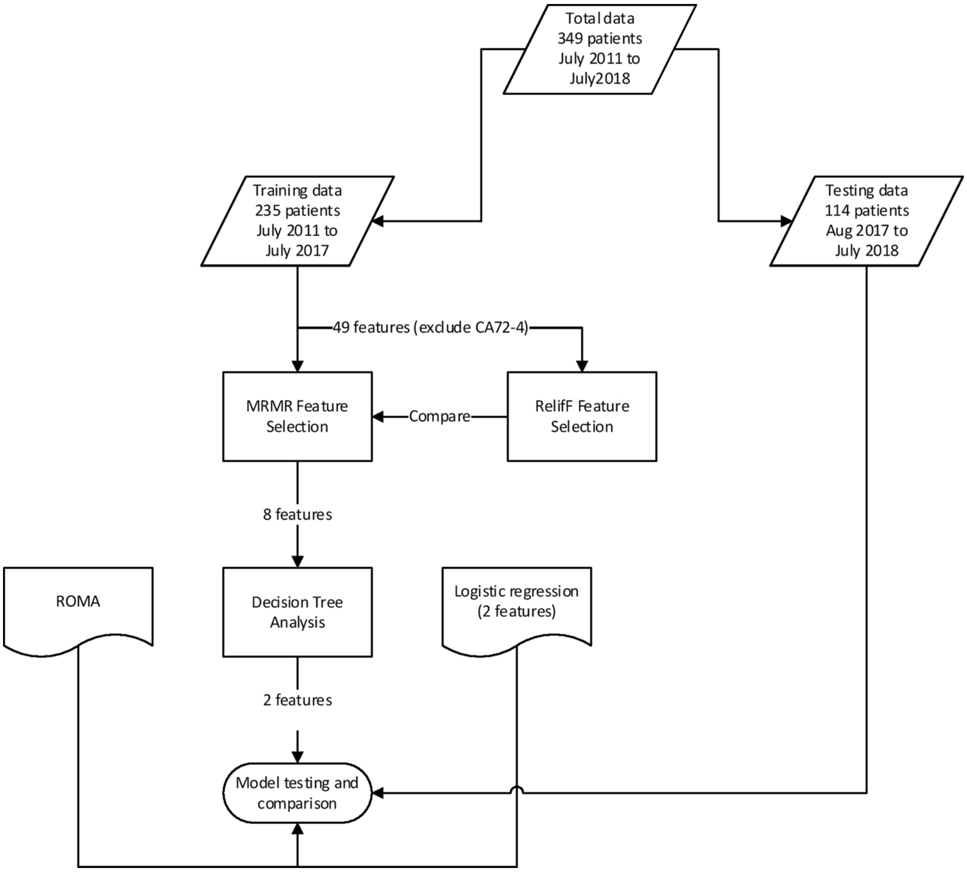

Fig. 1. A flowchart that represents the modeling procedure in this study including the description of the patient assignment to training and testing groups, the feature selection method, decision tree model development and comparison the results with ROMA, and the logistic regression model.

of samples. Then the CART model was tested on the remaining 1 group of samples and the accuracy and F1 score were evaluated for the model performance. This process was repeated for each experiment 10 times. The average and standard deviation values for accuracy and F1 score, as well as the selected features for each experiment, were recorded.

3.4. ReliefF feature selection

ReliefF is to select features to separate instances from different classes. It is also regarded as a similarity based method as it is equivalent to selecting features that preserve a special form of data similarity matrix. The method can be used to handle both continuous and discrete feature variables. We performed ReliefF feature selection using the same procedure as described in MRMR to compare their feature selection results.

3.5. Risk of Ovarian Malignancy Algorithm (ROMA) calculation

The risk of ovarian malignancy algorithm (ROMA) was proposed by Moore et al [9], which is calculated from HE4 and CA125 concentrations with the combination of the patient's menopausal status. According to [9], the predictive index of ROMA is calculated as follows:

$$\text{Pre-menopausal Predictive Index (PI)} = 12.0 + 2.38*\text{LN(HE4)} + 0.0626*\text{LN(CA 125)} \quad (4)$$

$$\text{Post-menopausal Predictive Index (PI)} = -8.09 + 1.04*\text{LN(HE4)} + 0.732*\text{LN(CA 125)} \quad (5)$$

And ROMA%, Predicted Probability (PP) is calculated as: $\text{PP} = \exp(\text{PI})/[1+\exp(\text{PI})] \text{ x } 100\%$ (6)

For ROMA for high-risk premenopausal and postmenopausal women was 13.1% and 27.7%, respectively [9].

3.6. Decision Tree Model

Decision Trees are a non-parametric supervised learning method used for classification and regression. The reasons why we choose decision trees are (1) it is simple to understand and to interpret. It can formulate if-then type of rules to explain the underlying relationships between biomarkers and ovarian cancer based on the built tree, (2) it requires minimal data preprocessing, (3) it can handle both numerical and categorical data since our dataset contains both numerical and categorical biomarkers.

In the constructed decision tree, each leaf node corresponds to a class label of the target variable and each internal node represents a feature (a biomarker in this study). The constructed decision tree is a binary tree which is built using a classification algorithm called Classification and Regression Trees (CART) [29]. The implementation package of the decision tree model used in this study is provided by Python Scikit-learn [30], where it optimized the CART algorithm by pre-pruning and pre-sorting.

3.7. Logistic Regression

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable. In our present study, two top features selected by decision tree model were used to fit a logistic regression model, the prediction probability of this model and the true classification were used to construct the confusion matrix in Table 2. Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value and Accuracy Rate were calculated based on the confusion matrix.

3.8. Modelling Procedure Illustration

Fig. 1 is a flowchart that represents the above modeling steps, including the description of the patient assignment to training and testing groups, the feature selection methods, decision tree model development and comparison the results with ROMA, and logistic regression model.

4. Result

4.1. Statistics

The t-test or the rank-sum test of two independent samples were used to achieve the comparison between BOT and OC groups. There were 22 variables had statistically significant differences ($P < 0.05$). They were 8 blood routine test (Platelet Distribution Width, Hemoglobin, Lymphocyte Count, Lymphocyte Ratio, Mononuclear Cell Count, Platelet Count, Red Blood Cell Count, Thrombocytocrit), 9 general Chemistry (Aspartate Aminotransferase, Chlorine, Albumin, Gama Glutamyltransferasey, Globulin, Calcium, Anion Gap, Carban Dioxide-combining Power, Alkaline Phosphatase) and 5 tumor marker (Carbohydrate Antigen 72-4, Carbohydrate Antigen 19-9, Alpha-fetoprotein, Human Epididymis Protein 4, Carbohydrate Antigen 125).

4.2. Feature selection

The MRMR feature section was performed on thetraining dataset with 235 patients data. Fig. 2 shows the feature selection result from our 48 MRMR experiments. In both A) and B), the x-axis represents the number of features we used in each experiment while the y-axis in A) is the average accuracy from the 10-fold cross-validation and the y-axis in B) is the average F1 score from the 10-fold cross-validation.

As shown in Fig. 2 A&B, both the average F1 score and the average accuracy increase as the number of features included in the model increases, and they become stable when the number of features reaches around 10. Thus, we examined the 10^{th} experiment which selected 10 features. There are eight features occurred most frequent in all 10-fold simulations which include Menopause, Age, AFP, CEA, HE4, CA19-9, LYM%, and CO2CP (supplementary data 6).

For the ReliefF feature selection, as shown in the Fig. 2 C&D, the accuracy and F1 score of the classification model in the ReliefF reach the maximum until it includes low ranked features such as HE4 and CEA (supplementary data 7). This result can highlight HE4 and CEA are the important predictors. It also shows the top features of ReliefF is not as good for prediction as the ones selected by MRMR. Therefore, we used these eight features derived from MRMR to build our final decision tree model.

4.3. Decision Tree

A simple two-level CART decision tree model was developed using the training dataset and the result is shown in Fig. 3. The model has four terminal nodes and only two biomarkers (HE4 and CEA) from the eight features derived from MRMR were appeared in the top two-level of trees. The remaining six features were not selected by the decision tree model in the top two levels. Their rules and results are depicted as the followings:

Node C: If HE4 ≤ 70.73 pmol/L and CEA ≤ 3.12 ng/mL, predict Begin (85/111 or 76.6%)

Node D: If HE4 ≤ 70.73 pmol/L and CEA > 3.12 ng/mL, predict OC (16/18 or 88.9%)

Node E: If HE4 > 70.73 pmol/L and HE4 < = 91.24 pmol/L, predict OC (4/6 or 66.7%)

Node F: If HE4 > 91.24 pmol/L, predict OC (100/100 or 100%)

4.4. Model Evaluation and Comparison

As shown in our decision tree model (Fig. 2) and some other work in the literature, CEA [31] and HE4 [8,7,32,9,33,34,15,5,20] show strong predictive power for distinguishing OC from BOC. A logistic regression model was developed using HE4 and CEA on the training dataset. The equation for logistic regression is:

$$P(y = ovarian\ cancer) = \frac{\exp(-4.111 + 0.369 \times CEA + 0.055 \times HE4)}{1 + \exp(-4.111 + 0.369 \times CEA + 0.055 \times HE4)} \quad (10)$$

$$\text{Or}\ \text{logit}(P) = \exp(-4.111 + 0.369 \times CEA + 0.055 \times HE4) \quad (11)$$

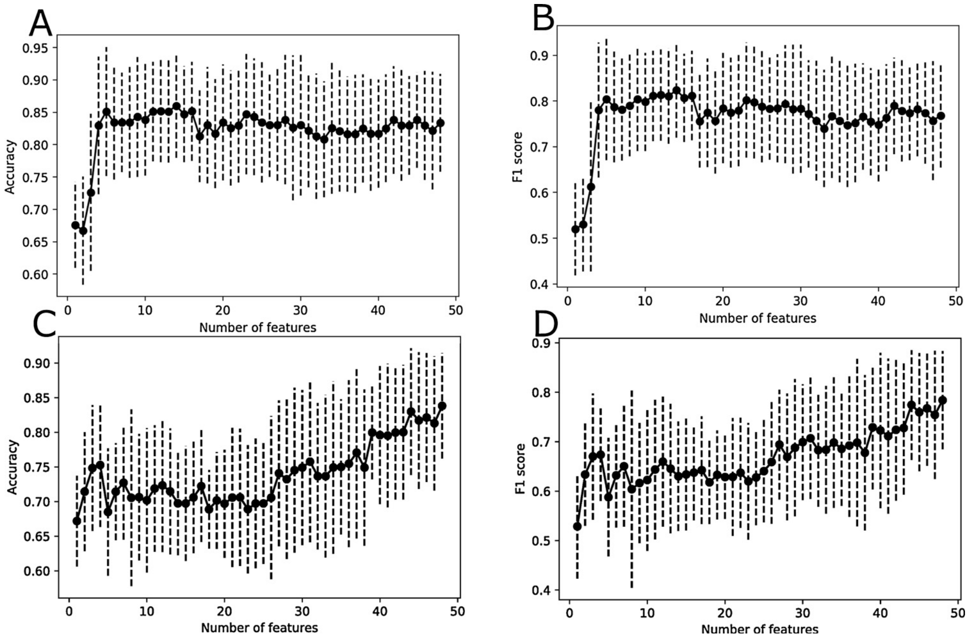

Fig. 2. Feature selection results from our 48 experiments using MRMR and ReliefF. In all the figures (A, B, C, and D), the x-axis represents the number of features we used in each experiment while the y-axis in A and C is the average accuracy from the 10-fold cross-validation and the y-axis in B and D is the average F1 score from the 10-fold cross-validation. The results in A and B are from MRMR and the results in C and D are from ReliefF. From ReliefF's results (C and D), both the average accuracy and F1 score reach to the maximum value until it includes the low ranked features. Whereas the results from MRMR (A and B), both the average accuracy and F1 score increase following the same trend and become stable when selecting the top 10 ranked features.

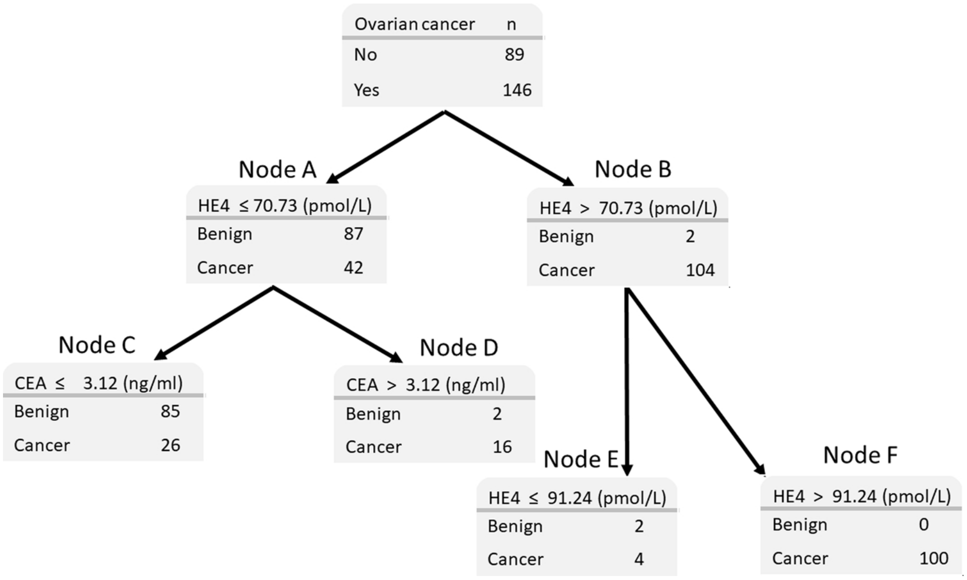

Fig. 3. The decision tree model consisting of two features: HE4 and CEA for classification of BOT and OC.

Table 2
The model performance metrics for ROMA, Decision Tree and Logistic Regression separated by training and test data.

Training data		Ture Benign	True Cancer	Total	Sensitivity	Specificity	Positive predictive Value	Negative Predictive Value	Accuracy Rate
		Count	Count	Count					
ROMA	Predicted Benign	79	38	117					
	Predicted Cancer	10	108	118	0.740	0.888	0.915	0.675	0.796
	Total	89	146	235					
Decision Tree	Predicted Benign	85	26	111					
	Predicted Cancer	4	120	124	0.822	0.955	0.968	0.766	0.872
	Total	89	146	235					
Logistic Regression	Predicted Benign	89	36	125					
	Predicted Cancer	0	110	110	0.753	1	1	0.712	0.847
	Total	89	146	235					
Test data		Ture Benign Count	True Cancer Count	Total Count	Sensitivity	Specificity	Positive predictive Value	Negative Predictive Value	Accuracy Rate
ROMA	Predicted Benign	80	0	80					
	Predicted Cancer	9	25	34	1	0.899	0.735	1	0.921
	Total	89	25	114					
Decision Tree	Predicted Benign	86	2	88					
	Predicted Cancer	3	23	26	0.92	0.966	0.885	0.977	0.956
	Total	89	25	114					
Logistic Regression	Predicted Benign	87	1	88					
	Predicted Cancer	2	24	26	0.96	0.978	0.923	0.989	0.974
	Total	89	25	114					

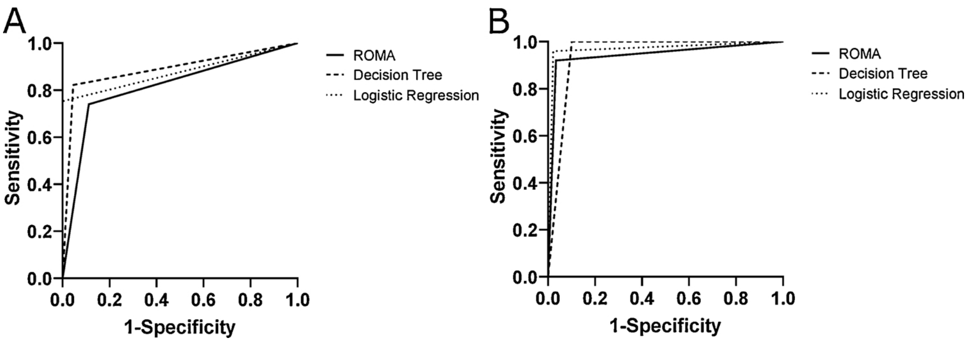

Fig. 4. ROC analysis of ROMA, logistic regression model, and decision tree model on both training data (A) and testing data (B).

The performance of the prediction of the decision tree model is compared with the ROMA and a logistic regression model on both training and testing datasets. The confusion matrix, sensitivity, specificity, positive predictive value, negative predictive value, and accuracy rate result for each model on both training and testing data are shown in Table 2. The sensitivity, specificity, positive predictive value, negative predictive value and accuracy rate value on the training dataset for the decision tree model are 0.822, 0.955, 0.968, 0.766 and 0.872 respectively, for ROMA are 0.74, 0.888, 0.915, 0.675 and 0.796 respectively, and for the logistic regression model are 0.753, 1, 1, 0.712 and 0.847. On the testing dataset with 114 patients data, the sensitivity, specificity, positive predictive value, negative predictive value and accuracy rate value for decision tree model are 0.92, 0.966, 0.885, 0.977 and 0.956 respectively, for ROMA are 1, 0.899, 0.735, 1, and 0.921 respectively, and for the logistic regression model are 0.96, 0.978, 0.923, 0.989 and 0.974 respectively. The ROC result for each model is shown in Fig. 4 where our decision tree model has the best AUC (0.888, SE: 0.023) on the training data comparing to ROMA (0.814, SE: 0.029, $p = 0.045$) and the logistic regression model (0.877, SE: 0.023, $p = 0.735$). On the testing data, the decision tree model has AUC (0.949, SE: 0.020, $p = 0.876$) comparing to ROMA (0.943, SE 0.033) and the logistic regression model (0.969, SE: 0.024, $p = 0.522$).

5. Conclusions and discussion

As one of the most common types of cancer in women, accurately classifying benign and malignant ovarian neoplasm using biomarkers has a profound effect on many people's lives. Therefore, it’s of great importance for us to improve the accuracy of early diagnosis and detection of ovarian cancer. We applied a machine learning feature selection method MRMR to systematically investigate the list of variables containing patients’ demographics, blood routine test, general chemistry, and the tumor markers to select the most relevant features for predicting ovarian cancer. Then a straightforward two-level decision tree was derived by using the selected features on the training dataset (235 patients) and tested on an independent dataset (114 patients).

Our decision tree model suggests the combination of the two biomarkers, HE4 and CEA, have the most significant prediction power when it comes to the classification of ovarian cancer vs the benign ovarian tumors. Interestingly, the cut-off value for HE4 derived from our decision tree model is 70.73 pmol/L and it is consistent with the HE4 cut-off value used in other studies [11,6]. Also, for the patients with HE4 value less than 70.73 pmol/L, our model suggests that if their CEA level is great than 3.12 ng/mL, they may also highly likely to have OC. CEA is one of the commonly used serum biomarkers for solid tumor diagnosis such as colorectal cancer, breast cancer, lung cancer etc. Based on our study, it reveals that CEA is a valuable marker for prediction of OC for the patients in HE4 low group. This founding may provide new perspective for studying CEA for its role in ovarian cancer.

Using only these two biomarkers, as shown in the result section, a decision tree model produces a better prediction result in both training and testing data comparing to the ROMA (AUC of ROC for decision tree and ROMA on training data are 0.888 and 0.814 respectively and on testing data are 0.949 and 0.943 respectively) which uses value of HE4, CA125 and menopause status and is a commonly used index in the diagnosis of ovarian cancer [35]. The results are also comparable to the logistic regression model (AUC of ROC for decision tree and logistic regression on training data are 0.888 and 0.877 respectively and on testing data are 0.949 and 0.969 respectively). Though logistic regression shows a slightly higher AUC value in the testing data, the advantage of the decision tree model is that it has if-then rules which makes it easier to understand and interpret for the medical professionals. Also the logistic regression utilized the two features HE4 and CEA that derived from feature selection and decision tree analysis which in another aspect indicates the high predictive value for the HE4 and CEA. In conclusion, our contribution to the ovarian cancer research area is that we discoverd that using HE4 and CEA can produce a simple and high performance model for the classification of BOT and OC. Our study demonstrates that the ML models can be used to uncover pattern in the complex diseases such as ovarian cancer. Future work of the study can be to develop a tool that incorporates the decision tree model developed in this study and apply it to a large cohort of patients for the further validation.

Declarations of interest

None.

Competing interests

The authors declare no conflict of interest.

All authors have read and approved the above statement.

Acknowledgements

This work was supported in part by grants from the National Key R&D Plan (2018YFC1313400) to J. J., the National Natural Science Foundation of China (31570877, 31729001, 81972869) to J. J., the National Key Technology R&D Program (2015BAI12B12) to J. J., Jiangsu Engineering Research Center for Tumor Immunotherapy (BM2014404) to J. J., the Key R&D Project of Science and Technology Department of Jiangsu Province (BE2018645) to J. J., the National Natural Science Foundation of China (81902386) to X. Z., Young Medical Talents Program of Jiangsu Province (QNRC2016286) to X. Z., Changzhou Science and Technology Project (Applied Based Research, CJ20190094) to X.Z., and Changzhou High-Level Medical Talents Training Project (2016CZBJ001) to L.C.

References

[1] R.L. Siegel, K.D. Miller, A. Jemal, Cancer statistics, 2018, CA: a cancer journal for clinicians 68 (1) (2018) 7–30.

[2] F. Bray, J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre, A. Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA: A Cancer Journal for Clinicians 68 (6) (2018) 394–424.

[3] D.A. Fishman, K. Bozorgi, The scientific basis of early detection of epithelial ovarian cancer: the National Ovarian Cancer Early Detection Program (NOCEDP), Cancer treatment and research 107 (2002) 3–28.

[4] C. Marchetti, C. Pisano, G. Facchini, G.S. Bruni, F.P. Magazzino, S. Losito, S. Pignata, First-line treatment of advanced ovarian cancer: current research and perspectives, Expert review of anticancer therapy 10 (1) (2010) 47–60.

[5] K. Aslan, M.A. Onan, C. Yilmaz, N. Bukan, Erdem M, Comparison of HE 4, CA 125, ROMA score and ultrasound score in the differential diagnosis of ovarian masses, Journal of Gynecology Obstetrics and Human Reproduction (2020) 101713.

[6] R.G. Moore, M. Jabre-Raughley, A.K. Brown, K.M. Robison, M.C. Miller, W.J. Allard, R.J. Kurman, R.C. Bast, S.J. Skates, Comparison of a novel multiple marker assay vs the Risk of Malignancy Index for the prediction of epithelial ovarian cancer in patients with a pelvic mass, American Journal of Obstetrics and Gynecology 203 (3) (2010) 228.e221-228.e226.

[7] T. Granato, C. Midulla, F. Longo, B. Colaprisca, L. Frati, E. Anastasi, Role of HE4, CA72.4, and CA125 in monitoring ovarian cancer, Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine 33 (5) (2012) 1335–1339.

[8] I. Hellstrom, J. Raycraft, M. Hayden-Ledbetter, J.A. Ledbetter, M. Schummer, M. McIntosh, C. Drescher, N. Urban, K.E. Hellstrom, The HE4 (WFDC2) protein is a biomarker for ovarian carcinoma, Cancer research 63 (13) (2003) 3695–3700.

[9] R.G. Moore, D.S. McMeekin, A.K. Brown, P. DiSilvestro, M.C. Miller, W.J. Allard, W. Gajewski, R. Kurman, R.C. Bast, S.J. Skates, A novel multiple marker bioassay utilizing HE4 and CA125 for the prediction of ovarian cancer in patients with a pelvic mass, Gynecologic Oncology 112 (1) (2009) 40–46.

[10] I. Jacobs, D. Oram, J. Fairbanks, J. Turner, C. Frost, J.G. Grudzinskas, A risk of malignancy index incorporating CA 125, ultrasound and menopausal status for the accurate preoperative diagnosis of ovarian cancer, British journal of obstetrics and gynaecology 97 (10) (1990) 922–929.

[11] C. Anton, F.M. Carvalho, E.I. Oliveira, G.A.R. Maciel, E.C. Baracat, J.P. Carvalho, A comparison of CA125, HE4, risk ovarian malignancy algorithm (ROMA), and risk malignancy index (RMI) for the classification of ovarian masses, Clinics (Sao Paulo) 67 (5) (2012) 437–441.

[12] J. Wang, J. Gao, H. Yao, Z. Wu, M. Wang, J. Qi, Diagnostic accuracy of serum HE4, CA125 and ROMA in patients with ovarian cancer: a meta-analysis, Tumor Biology

35 (6) (2014) 6127–6138.

[13] A. Lukanova, R. Kaaks, Endogenous hormones and ovarian cancer: epidemiology and current hypotheses, Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology 14 (1) (2005) 98–107.

[14] S.M. Ho, Estrogen, progesterone and epithelial ovarian cancer, Reproductive biology and endocrinology : RB&E 1 (2003) 73.

[15] P. Zhang, C. Wang, L. Cheng, P. Zhang, L. Guo, W. Liu, Z. Zhang, Y. Huang, Q. Ou, X. Wen, et al., Development of a multi-marker model combining HE4, CA125, progesterone, and estradiol for distinguishing benign from malignant pelvic masses in postmenopausal women, Tumor Biology 37 (2) (2016) 2183–2191.

[16] L.J. Havrilesky, C.M. Whitehead, J.M. Rubatt, R.L. Cheek, J. Groelke, Q. He, D.P. Malinowski, T.J. Fischer, A. Berchuck, Evaluation of biomarker panels for early stage ovarian cancer detection and monitoring for disease recurrence, Gynecol Oncol 110 (3) (2008) 374–382.

[17] I.H. Witten, E. Frank, M.A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann Publishers Inc., 2011.

[18] P.-N. Tan, M. Steinbach, Kumar V, Introduction to Data Mining, (First Edition), Addison-Wesley Longman Publishing Co. Inc., 2005.

[19] M.D. Ganggayah, N.A. Taib, Y.C. Har, P. Lio, S.K. Dhillon, Predicting factors for survival of breast cancer patients using machine learning techniques, BMC Medical Informatics and Decision Making 19 (1) (2019) 48.

[20] R. Miao, T.C. Badger, K. Groesch, P.L. Diaz-Sylvester, T. Wilson, A. Ghareeb, J.A. Martin, M. Cregger, M. Welge, C. Bushell, et al., Assessment of peritoneal microbial features and tumor marker levels as potential diagnostic tools for ovarian cancer, PLOS ONE 15 (1) (2020) e0227707.

[21] L.H. Kim, J.L. Quon, T.A. Cage, M.B. Lee, L. Pham, H. Singh, Mortality prediction and long-term outcomes for civilian cerebral gunshot wounds: A decision-tree algorithm based on a single trauma center, Journal of Clinical Neuroscience 75 (2020) 71–79.

[22] R. Sumbaly, N. Vishnusri, S. Jeyalatha, Diagnosis of Breast Cancer using Decision Tree Data Mining Technique, International Journal of Computer Applications 98 (2014) 16–24.

[23] C.-S. Rau, S.-C. Wu, P.-C. Chien, P.-J. Kuo, Y.-C. Chen, H.-Y. Hsieh, H.-Y. Hsieh, Prediction of Mortality in Patients with Isolated Traumatic Subarachnoid Hemorrhage Using a Decision Tree Classifier: A Retrospective Analysis Based on a Trauma Registry System, International Journal of Environmental Research and Public Health 14 (11) (2017) 1420.

[24] J. Cheng, R. Greiner, Comparing Bayesian network classifiers, In: Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc., Stockholm, Sweden, 1999, pp. 101–108.

[25] C.H.Q. Ding, Analysis of gene expression profiles: class discovery and leaf ordering, Proceedings of the sixth annual international conference on Computational biology, ACM, Washington, DC, USA, 2002, pp. 127–136.

[26] S. Dudoit, J. Fridlyand, T.P. Speed, Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data, Journal of the American Statistical Association 97 (457) (2002) 77–87.

[27] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, et al., Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, Science 286 (5439) (1999) 531–537.

[28] C. Ding, H. Peng, Minimum redundancy feature selection from microarray gene expression data. In: Computational Systems Bioinformatics CSB2003 Proceedings of the 2003 IEEE Bioinformatics Conference CSB2003: 11-14 Aug. 2003, (2003) 2003: 523-528.

[29] L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen, Classification and Regression Trees, Taylor & Francis, 1984.

[30] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, et al., API design for machine learning software: experiences from the scikit-learn project. arXiv:13090238, (2013).

[31] F. Zhang, Zhang Z-l, The Diagnostic Value of Transvaginal Sonograph (TVS), Color Doppler, and Serum Tumor Marker CA125, CEA, and AFP in Ovarian Cancer, Cell Biochemistry and Biophysics 72 (2) (2015) 353–357.

[32] X.-Y. Li, J.-Y.-Y, J.-B.-B. Qin, X.-Y.-Y. Li, P. Dong, B.-D. Yin, Diagnostic Value of Human Epididymis Protein 4 Compared with Mesothelin for Ovarian Cancer: a Systematic Review and Meta-analysis, Asian Pacific Journal of Cancer Prevention 13 (11) (2012) 5427–5432.

[33] L. Wu, Z.-Y.-Y. Dai, Y.-H.-H. Qian, Y. Shi, F.-J.-J. Liu, Yang C, Diagnostic Value of Serum Human Epididymis Protein 4 (HE4) in Ovarian Carcinoma: A Systematic Review and Meta-Analysis, International Journal of Gynecologic Cancer 22 (7) (2012) 1106–1112.

[34] S. Yu, Yang H-j, Xie S-q, Bao Y-X, Diagnostic value of HE4 for ovarian cancer, a meta-analysis 50 (8) (2012) 1439.

[35] R. Cui, Y. Wang, Y. Li, Y. Li, Clinical value of ROMA index in diagnosis of ovarian cancer: meta-analysis, Cancer management and research 11 (2019) 2545–2551.