Journal of Advanced Research 30 (2021) 113–122

Contents lists available at ScienceDirect

Journal of Advanced Research

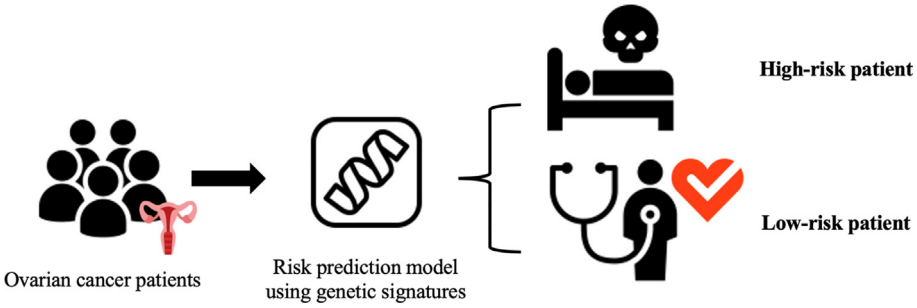
journal homepage: www.elsevier.com/locate/jare

A risk prediction model of gene signatures in ovarian cancer through bagging of GA-XGBoost models

Yi-Wen Hsiao [a,1], Chun-Liang Tao [a,1], Eric Y. Chuang [b,c], Tzu-Pin Lu [a,b,*]

[a] Department of Public Health, Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan
[b] Bioinformatics and Biostatistics Core, Center of Genomic and Precision Medicine, National Taiwan University, Taipei, Taiwan
[c] Graduate Institute of Biomedical Electronics and Bioinformatics, Department of Electrical Engineering, National Taiwan University, Taiwan

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 8 May 2020
Revised 10 October 2020
Accepted 5 November 2020
Available online 11 November 2020

Keywords:
Ovarian cancer
Risk prediction
Gene expression
Machine learning
GA-XGBoost
Bagging algorithm

ABSTRACT

Introduction: Ovarian cancer (OC) is one of the most frequent gynecologic cancers among women, and high-accuracy risk prediction techniques are essential to effectively select the best intervention strategies and clinical management for OC patients at different risk levels. Current risk prediction models used in OC have low sensitivity, and few of them are able to identify OC patients at high risk of mortality, which would both optimize the treatment of high-risk patients and prevent unnecessary medical intervention in those at low risk.
Objectives: To this end, we have developed a bagging-based algorithm with GA-XGBoost models that predicts the risk of death from OC using gene expression profiles.
Methods: Four gene expression datasets from public sources were used as training (n = 1) or validation (n = 3) sets. The performance of our proposed algorithm was compared with fine-tuning and other existing methods. Moreover, the biological function of selected genetic features was further interpreted, and the response to a panel of approved drugs was predicted for different risk levels.
Results: The proposed algorithm showed good sensitivity (74–100%) in the validation sets, compared with two simple models whose sensitivity only reached 47% and 60%. The prognostic gene signature used in this study was highly connected to *AKT*, a key component of the PI3K/AKT/mTOR signaling pathway, which influences the tumorigenesis, proliferation, and progression of OC.

Peer review under responsibility of Cairo University.
* Corresponding author at: Department of Public Health, Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan.
E-mail address: tplu@ntu.edu.tw (T.-P. Lu).
[1] These authors contributed equally to the present study.

https://doi.org/10.1016/j.jare.2020.11.006

Conclusion: These findings demonstrated an improvement in the sensitivity of risk classification of OC patients with our risk prediction models compared with other methods. Ongoing effort is needed to validate the outcomes of this approach for precise clinical treatment.

Introduction

Ovarian cancer (OC) is the seventh most common malignancy, and it causes the eighth highest mortality rate of all cancer types worldwide; 295,414 new cases were diagnosed, and 184,799 patients died of this disease in 2018 [1]. OC was initially divided into epithelial and non-epithelial types, but some recent literature has indicated that epithelial OC also has histological subtypes including high-grade serous (>70%), endometrioid (10%), clear cell (10%), mucinous (<5%), and low-grade serous (<5%) [2]. These histologically distinct tumor types have shown a wide range of different prognoses. For example, epithelial tumors classified as low-grade serous, endometrioid, mucinous, or clear cell usually present themselves at an early stage and have a good prognosis, while the high-grade serous type mostly presents itself at an advanced stage with a poor prognosis [3]. It has been revealed that the five-year survival of OC patients diagnosed at an early stage is about 90%, whereas that of patients at a late stage is less than 30% after surgery [4,5]. However, most OC patients are diagnosed at the advanced stage due to the asymptomatic features of the early stage. As a result, more sophisticated research into both the diagnostic and predictive aspects of OC is urgently needed.

According to a clinical guideline from the National Comprehensive Cancer Network (NCCN), whether OC patients should receive post-surgery chemotherapy mainly depends on their clinical features, such as tumor stage and tumor grade [6]. In general, it is recommended for OC patients at stages II to IV to receive chemotherapy after surgery. OC patients at stages IA or IB with grade 1 tumors are recommended to have follow-up tests after surgery, while those with grade 2 tumors are suggested for either follow-up with the regular investigative tests or post-surgery chemotherapy. However, there is still some controversy regarding which OC patients, especially advanced-stage patients, will obtain the most clinical benefit from post-surgery adjuvant chemotherapy. National cancer statistics from the Taiwan Cancer Registry reported that 72.56% of OC patients had received post-surgery chemotherapy in 2016, and 60.65% of these patients were diagnosed at stage I [7], revealing that decisions regarding medical intervention do not always follow the NCCN guideline. To date, there are no entirely acceptable criteria to guide treatment decisions, especially in terms of post-surgery treatment in patients with low risk.

Due to the complexity and heterogeneity of cancer, gene expression profiling can provide biological insights into cancer prognosis, over and above the use of clinical features [8]. Hence, more and more cancer-related studies are taking these molecular indicators into account [9–12]. For example, a commercially available 70-gene signature test (MammaPrint) has been able to distinguish breast cancer patients at high versus low risk of recurrence, based on their 5- or 10-year recurrence rate [13], which can assist with clinical decision-making for early-stage patients [14]. Oncotype DX is another example of a genomic test that uses a clinically validated set of 21 genes to assess the risk of breast cancer recurrence [15]. Nevertheless, this kind of test may only be applicable to a particular set of patients (e.g., those with a particular hormone expression pattern) and may not fully explain the eventual clinical outcome, suggesting that unbiased approaches with a full prognostic gene signature are needed for accurate cancer risk assessment [16].

Prior studies on OC [17,18] have proposed models for predicting survival and have discussed hazard ratios (HRs) based on gene expression data. However, very few classifiers have been built to predict high risk of mortality in OC patients with high sensitivity. Several recent studies have extensively investigated robust machine learning-based methods for the identification of prognostic molecules in breast cancer, which shares many standard pathological features with OC [19–22]. However, few of these novel approaches have been applied to OC [23]. Therefore, the purpose of our study was to incorporate a bagging-based algorithm with GA-XGboost models into a comprehensive risk prediction model, using prognosis-related genes to arrive at a clinically meaningful classification of OC patients. The accuracy of our prediction model was evaluated in comparison with that of other conventional methods. The primary objective of this study was to effectively identify high-risk OC patients, with the long-term goal of reducing unnecessary preventive treatments in low-risk patients.

Materials and methods

An overview of the workflow is illustrated in Fig. 1. With the aim of identifying high- risk patients with OC, we constructed a complex set of procedures, including data preprocessing, dimensional reduction, a bagging-based algorithm with GA-XGBoost models, and external validation, to construct a comprehensive prediction model.

Datasets and data preprocessing

For the evaluation of the predictive model, four gene expression datasets (GSE26193 [24,25], GSE30161 [26], GSE19829 [27], GSE63885 [28]) that had OC outcomes were collected in this work (Table 1). All datasets used were from the publicly available Gene Expression omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/), and the platform used for these datasets was the Affymetrix Human Genome U133 Plus 2.0 Array (GPL570). In each dataset, patients who lacked 3-year follow-up information were excluded. Then, the GSE26193 dataset (n = 106) was divided into a training set and a validation set for building the prediction model. The remaining datasets, including GSE63885 (n = 73), GSE30161 (n = 50) and GSE19829 (n = 23), were used for external validation. Based on the clinical data, we further stratified patients in each dataset into two groups. Two previous studies have suggested that around 50% of OC patients suffer from recurrence within 1.5–2 years [29,30]; hence, we set three years as a cut-off to ensure most patients with recurrence were included in the following analyses. The low-risk group was defined as patients with overall survival of three years or more, whereas the high-risk group was defined as patients with overall survival less than three years.

For the minimization of batch effects among different datasets, raw intensity-level data merged from all datasets were first normalized using robust multichip averaging (RMA) and then by quantile normalization with default parameters using the affy (version 1.62.0) [31] and preprocessCore (version 1.46.0) [32] R packages.

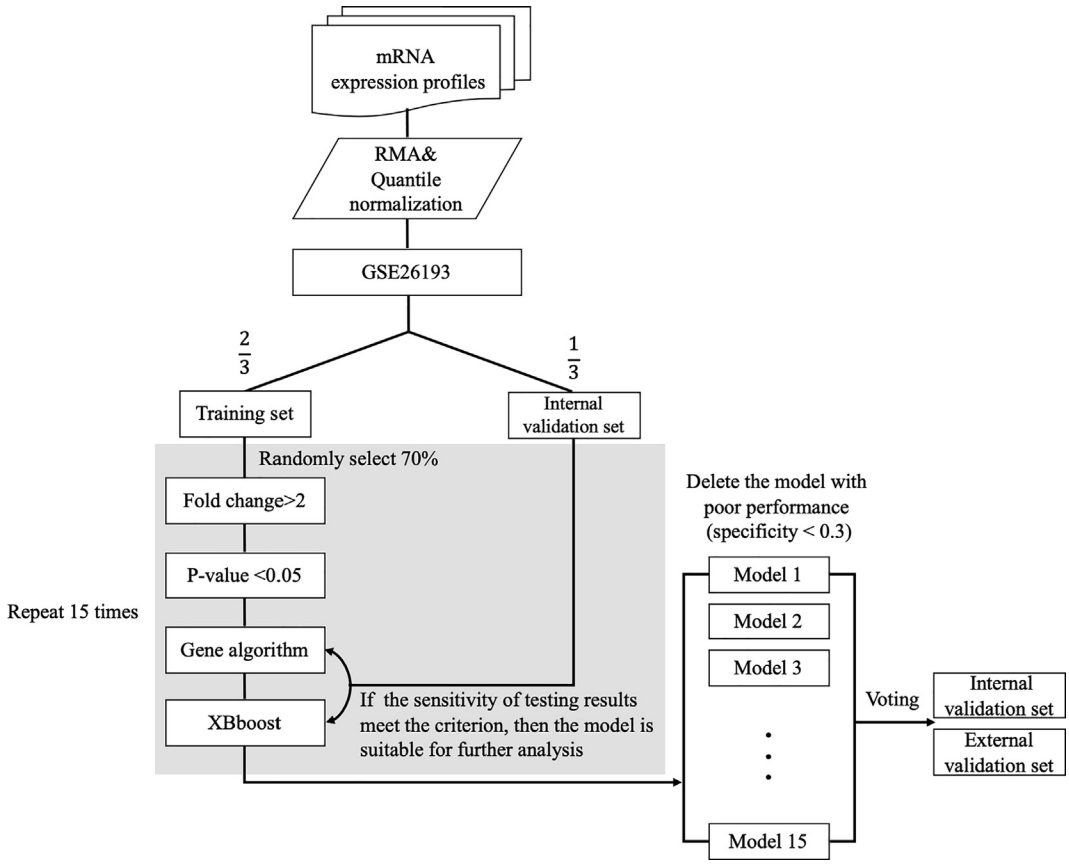


Fig. 1. The pipeline of our bagging-based algorithm with GA-XGBoost models.

Table 1
Summary of GEO datasets used in this study.

Datasets (GEO Accession)	Year	Country	Number of total samples	Number of used samples	Chemotherapy
GSE26193	2011	France	107	106	Yes
GSE30161	2012	United States	58	50	Yes
GSE19829	2010	United States	28	23	Yes
GSE63885	2014	Poland	75	73	Yes

Variable selection of gene expression patterns for dimension reduction

For feature reduction of the training dataset from GSE26193, the *t*-test and fold change method was used as a criterion to identify differentially expressed genes between low- and high-risk OC patients. An absolute $\log_2$ fold change ≥ 2 and a P-value < 0.05 were set as the cut-off values to screen for these probes.

XGBoost

The XGBoost (extreme gradient boosting) algorithm is a learning framework based on gradient boosted decision trees [33]. Compared with traditional boosting tree models implemented with only first order derivative information [34], this boosting model uses a second-order Taylor expansion for calculating the loss function and its scalability to enhance not only computational speed but also the model performance. Therefore, XGBoost was used for risk prediction classifiers for OC patients in this paper.

Genetic algorithm for the most suitable combination of selected gene expression patterns

Genetic algorithms (GAs) have been designed to replicate the concept of natural selection by searching for an available combination of gene expression profiling probes which will produce a predictive model with superior performance [35,36]. Therefore, in terms of feature selection, the XGBoost algorithm could be further improved by using a GA, a process that we call GA-XGBoost. As shown in Fig. 2, a GA involves five main phases: initial population, fitness function, selection, crossover, and mutation. In the GA, genetic coding segments of a chromosome are represented using a string of zeros and ones. Therefore, in this study, in order to correspond to expression being either on or off, significantly expressed probes were denoted as 1, whereas the rest were assigned as 0 (Figure S1).

First, we randomly sampled a combination of probes from those significant probes that were determined in the previous step to be a chromosome (i.e., a string of zeros and ones corresponding to the

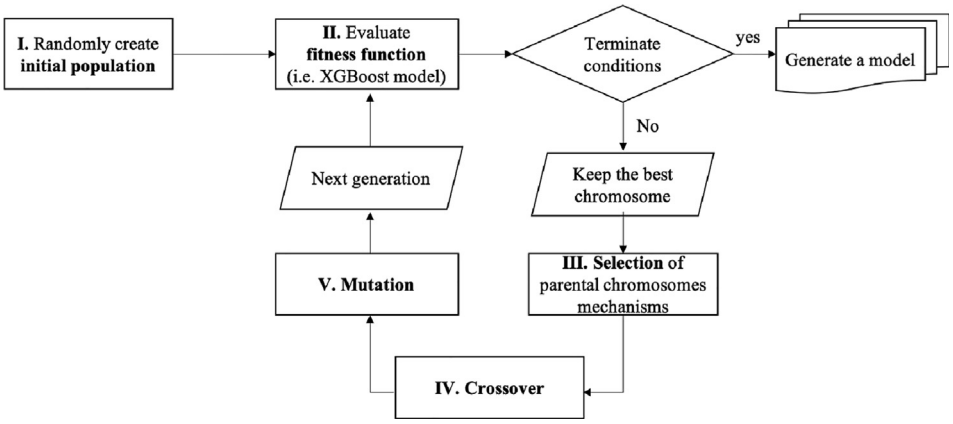


Fig. 2. Genetic Algorithm flowchart. GA algorithm includes five main steps: initial population, fitness function, selection, crossover, and mutation.

expression status of each gene expression profiling probe), and then repeated this procedure to generate a population of chromosomes defined as the first generation. Second, the fitness values (i.e., sensitivity and specificity in this study) of each chromosome was calculated by the fitness function (i.e., the XGBoost model), and only the ones with the highest fitness were retained in the next generation. In the roulette wheel selection, the wheel is divided according to the fitness values; that is, the fittest chromosome has the largest share, whereas the weakest chromosome has the smallest percentage (Figure S2A). The underlying assumption of this step is that the fitter chromosomes will tend to have a better chance of survival among the whole population, and then will mate to create the next generation. As a result, the fittest individuals will be stochastically selected from a particular population to form the next generation.

The chromosome showing the best fitness value (i.e., highest sensitivity) among the models (i.e., XGBoost model) from the first generation is either passed directly to the next generation, or crossover and mutation operators are performed to generate the next generation. In the crossover step, two parental chromosomes with the best fitness are selected from the original population, and a random threshold (for example, 20% of genetic information from parent chromosomes) is defined to determine the proportion of values within the chromosome that should be swapped to form two offspring chromosomes (Figure S2B). For emulating the dispersion of a mutation in a population, a proportion of the values (such as 10%) in a chromosome should be flipped, which means if it is a zero it now becomes a one and vice versa (Figure S2C and S3).

Finally, the conditions for evaluating when the GA should be stopped are defined as follows:

$$\begin{aligned} &training\ \ sensitivity - validation\ \ sensitivity \le 0.05 \\ &training\ \ specificity - validation\ \ specificity \le 0.05 \end{aligned} \quad (1)$$

The process of crossover and mutation will be repeated until these criteria are met or the final generation is reached (a predetermined number), unless the chromosomes with the best fitness of all generations meet the criteria and are outputted as a prediction model. By this process, the outputted model with the highest sensitivity for the classification of OC patients into risk response groups (high risk/low risk) is developed. The GA-XGBoost model was performed using R (version 3.5.2) and the xgboost R package (version 0.82.1) [37].

Bagging-based algorithm and external validation

To construct a robust bagging algorithm [38], GSE26193 was first divided into training and internal validation datasets, with a 2:1 split. Then, 70% of the training data was randomly selected to perform variable selection as described above, which generated 15 GA-XGBoost models for bagging. Those with a specificity < 0.3 were dropped, and based on each model's performance in both internal and external validation sets, the voting system was used to further identify OC patients with high risk.

Other existing methods

Other proposed methods can also be used in risk prediction based on gene expression values. Two traditional methods used in this study were least absolute shrinkage and selection operator (LASSO) regression [39] and forward stepwise logistic regression [40]. The performance evaluation was conducted by comparison of the predictive results, including accuracy, specificity, sensitivity, and F1-score, between GA-XGBoost and these two methods.

Survival analysis

Through this bagging algorithm of GA-XGBoost models, the common differentially expressed genes from all models were identified for the classification of two risk groups. Survival analyses were performed by the survival R package [41], and Kaplan-Meier survival curves were plotted to compare whether those expression profiles could distinguish between high- and low-risk groups of OC patients in internal and external validation sets. A Cox proportional hazards model was also used to compare the difference between survival curves for different risk groups.

Drug prediction for the identification of effective drugs

To further identify potential drugs effectively targeting each risk group, the dataset GSE36133, including gene expression profiles and a drug sensitivity indicator represented by activity area in the Cancer Cell Line Encyclopedia (CCLE) project, was used [42]. This project collected the drug response of 44 OC cell lines exposed to 24 commercially available drugs. The values of the activity area quantify the drug responses of each cell line. For this analysis, the expression profiles of the 44 OC cell lines were used as the inputs of our model to identify their potential risk level (high or low). Then, the Wilcoxon rank-sum test was used to evaluate

which drugs have a significant difference in the activity area between high- and low-risk groups.

Functional analysis

To understand the relationship between the respective differentially expressed genes obtained from this bagging algorithm of GA-XGBoost models and OC, we also used the Ingenuity® Pathway Analysis (IPA®) software program (QIAGEN Inc., https://www.qiagenbio-informatics.com/products/ingenuity-pathway-analysis) to identify their potential functional role in biological processes.

Statistical analysis

Categorical variables, such as stages, grades, clinical signatures, and subtypes, were reported as counts and percentages. Between-group comparisons (i.e., high- and low-risk groups) were performed by a Fisher's exact test. A P-value below 0.05 was defined as statistically significant. The analyses were conducted using R (version 3.5.2).

Results

Clinical characteristics for the training set

Table 2 presents the clinical characteristics of the training set (GSE26193; n = 106). The majority of samples in this dataset were from patients with stage III and grade III, constituting 55.7% and 63.2% of the samples, respectively. However, there were no significant differences, in terms of stage (P = 0.2828) and grade (P = 0.2665), between the two risk groups. Similarly, clinical signatures (P = 0.2515) and subtypes (P = 0.8113) also showed no difference between the two groups. Therefore, these clinical variables do not account for the risk of death from OC.

Parameter optimization

After dimension reduction of gene expression features, 507 differentially expressed probes (i.e., 406 genes) were extracted and used to inform the bagging-based algorithm that uses GA-XGBoost models. The bagging results using an internal validation set, three individual external validation sets, and a combined external validation set for different combinations of parameters are displayed in Table 3. To determine the best combination of parameters for our model, it is possible to fix all the settings except one and then decide which one has the strongest effect on model performance. The optimum combination of parameters has moderate specificity when the maximum sensitivity is reached, and these outcomes need to be supported by at least two external validation sets. First, we adjusted the number of GA-XGBoost models used in the bagging algorithm, and it can be seen that using 15 models showed the best performance, in terms of both specificity and sensitivity. Using more than 15 models may cause overfitting, while it may not be stable due to the small sample size when the number of models is less than 15. Second, the proportion of the GSE26193 dataset used for training (50%, 70%, or 90%) was adjusted, and 70% was optimal. Fewer training samples (50%) may generate an unstable bagging algorithm; on the contrary, a larger sample size may have an overfitting issue due to less variation among the models. Then, four tunable parameters used in GA-XGBoost were adjusted: the number of chromosomes in a generation, the number of generations, the mutation rate, and the number of tree layers. It can be observed that the combination of 300 chromosomes, 500 generations, a 50% mutation rate, and three tree layers are the best conditions. Lastly, imposing a requirement for high specificity (>70%), led to a less robust model with extremely low sensitivity and accuracy in many validation sets; for example, in GSE63885, the specificity, sensitivity, and accuracy were 0.784, 0.417, and 0.603, respectively.

Validation of the bagging-based algorithm that uses GA-XGBoost models

Table 4 presents the prediction ability of the 15 GA-XGBoost models used in the bagging algorithm using the internal validation set (GSE26193; n = 35). The range of the number of selected gene expression patterns among these models was 24–150 based on 15 cycles of variable selection using fold-change and P-value cut-offs after randomly selecting 70% of the training dataset, and the sensitivity of each model was over 0.8. A patient was considered as "high risk" when there were over seven models supporting this. As shown in Table 5, the bagging algorithm also maintains high sensitivity (100%) and specificity (52.4%) in the internal validation set. Among 35 patients in this validation set, 24 of them were predicted as high risk, and 11 were low risk. Kaplan-Meier survival analysis was also performed to determine the prognostic outcome, and the result indicated a significant difference (P = 0.0024) between high-risk and low-risk groups (Fig. 3A). Notably, the survival time of the high-risk group decreased, while that of the low-risk group was maintained as time passed.

In order to confirm that the high sensitivity of our bagging algorithm in predicting the risk level was not caused by model overfit-

Table 2
Statistical analysis of clinical variables in the GSE26193 dataset.

		Overall survival < 3 years (N = 52) No. (%)	Overall survival ≥ 3 years (N = 54) No. (%)	P-values
Stage	I	15(14.15)	22(20.76)	0.2828
	II	4(3.77)	6(5.66)	
	III	33(31.13)	26(24.53)	
Grade	I	2(1.89)	5(4.72)	0.2665
	II	19(17.92)	13(12.26)	
	III	31(29.25)	36(33.96)	
Signature	Oxidative stress	22(20.75)	29(27.36)	0.2515
	Fibrosis	30(28.30)	25(23.59)	
Subtype	Adenocarcinoma	1(0.94)	2(1.88)	0.8113
	Brenner Tumor	1(0.94)	0(0)	
	Carcinosarcoma	2(1.88)	0(0)	
	Clear Cell	3(2.83)	3(2.83)	
	Endometrioid	3(2.83)	5(4.72)	
	Mucinous	5(4.72)	3(2.83)	
	Serous	37(34.91)	41(38.69)	

Table 3
Parameter tuning of bagging-based algorithm with GA-XGBoost models.

	GSE26193 (internal validation set)			Combined external validation set*		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
1. Number of GA-XGBoost models for bagging						
10 models/1 model deleted	0.829	1	0.667	0.5	0.853	0.192
15 models/1 model deleted	0.714	1	0.524	0.589	0.824	0.385
20 models/1 model deleted	0.829	1	0.684	0.541	0.588	0.5
2. The size of training data (Proportion of samples in GSE26193)						
50%/3 models deleted	0.857	0.95	0.733	0.623	0.544	0.692
70%/1 model deleted	0.714	1	0.524	0.589	0.824	0.385
90%/1 model deleted	0.857	1	0.667	0.603	0.544	0.654
3. Number of chromosomes in each generation						
100 chromosomes/0 model deleted	0.857	1	0.737	0.582	0.72	0.462
300 chromosomes/1 model deleted	0.714	1	0.524	0.589	0.824	0.385
500 chromosomes/1 model deleted	0.943	1	0.894	0.555	0.632	0.487
4. Number of generations in GA-XGBoost						
300 generations/0 model deleted	0.886	1	0.789	0.596	0.544	0.641
500 generations/1 model deleted	0.714	1	0.524	0.589	0.824	0.385
1000 generations/0 model deleted	0.857	1	0.737	0.562	0.456	0.654
5. Mutation rates in GA-XGBoost						
0.3/1 model deleted	0.8	1	0.632	0.527	0.765	0.321
0.5/1 model deleted	0.714	1	0.524	0.589	0.824	0.385
0.7/2 models deleted	0.771	1	0.579	0.514	0.735	0.321
6. Number of tree layers used in GA-XGBoost						
two layers/1 model deleted	0.714	1	0.524	0.589	0.824	0.385
three layers/0 model deleted	0.914	1	0.85	0.589	0.706	0.487
five layers/0 model deleted	0.914	1	0.842	0.562	0.574	0.551
7. Model with high specificity						
High specificity	0.943	1	0.895	0.582	0.485	0.667

*This combined dataset included three external validation sets (GSE30161, GSE19829, and GSE63885; n = 146).

Table 4
Results of individual models for an internal validation set based on bagging of GA-XGBoost models.

Model No.	Number of variables	GSE26193 (internal validation set)			
		Sensitivity	Specificity	Accuracy	F1-score
1	81	0.929	0.429	0.629	0.667
2	117	0.929	0.81	0.857	0.839
3	39	1	0.143	0.486	0.609
4	140	0.929	0.619	0.743	0.743
5	52	0.929	0.524	0.686	0.703
6	25	1	0.81	0.886	0.875
7	53	1	0.524	0.714	0.737
8	70	1	0.476	0.686	0.718
9	129	1	0.333	0.600	0.667
10	73	1	0.333	0.600	0.667
11	59	1	0.524	0.714	0.737
12	87	1	0.476	0.686	0.718
13	30	1	0.476	0.686	0.718
14	24	0.929	0.524	0.686	0.703
15	150	0.857	0.619	0.714	0.706

Table 5
Performance comparison between the bagging of GA-XGBoost models and two existing models.

	Accuracy	Sensitivity	Specificity	F1-score
GA-XGBoost				
GSE26193 (internal validation set)	**0.714**	**1**	0.524	0.737
Combined external validation set*	**0.589**	**0.824**	0.385	0.651
GSE30161	**0.580**	**0.739**	0.444	0.618
GSE19829	**0.478**	**1**	0.143	0.600
GSE63885	**0.630**	**0.833**	0.432	0.690
Forward logistic regression				
GSE26193 (internal validation set)	0.600	0.533	0.650	0.533
Combined external validation set*	0.514	0.456	0.564	0.466
GSE30161	0.620	0.652	0.593	0.612
GSE19829	0.478	0.556	0.429	0.455
GSE63885	0.452	0.306	0.595	0.355
LASSO regression				
GSE26193 (internal validation set)	0.543	0.643	0.476	0.529
Combined external validation set*	0.555	0.544	0.564	0.532
GSE30161	0.520	0.478	0.556	0.478
GSE19829	0.565	0.889	0.357	0.615
GSE63885	0.575	0.500	0.649	0.537

*This combined dataset included three external validation sets (GSE30161, GSE19829, and GSE63885; n = 146).

ting to the training set, we also tested the same bagging algorithm using a combined external validation set (GSE30161, GSE19829, and GSE63885; n = 146) (Table 5). The sensitivity and specificity values of the bagging algorithm in this combined set were 82.4% and 38.5%. The Kaplan-Meier survival analysis was performed after combining all external validation sets, displaying that there is a significant difference between the two risk groups (P = 0.014; Fig. 3B). The individual external validation sets (GSE30161, n = 50; GSE19829, n = 23; GSE63885, n = 73) were also tested (Table S1). The sensitivity values of the bagging algorithm in these sets were 73.9%, 100%, and 83.3%, respectively, while the specificity values were 44.4%, 14.3% and 43.2%. The Kaplan-Meier survival analysis also showed a distinct difference in the survival time between the two groups in GSE63885 (P = 0.035), while the other two datasets (GSE30161, P = 0.2; GSE19829, P = 0.29) did not have a significant difference, likely due to the small sample size. The values of the HR and corresponding 95% confidence interval for each validation set were further visualized using forest plots, except

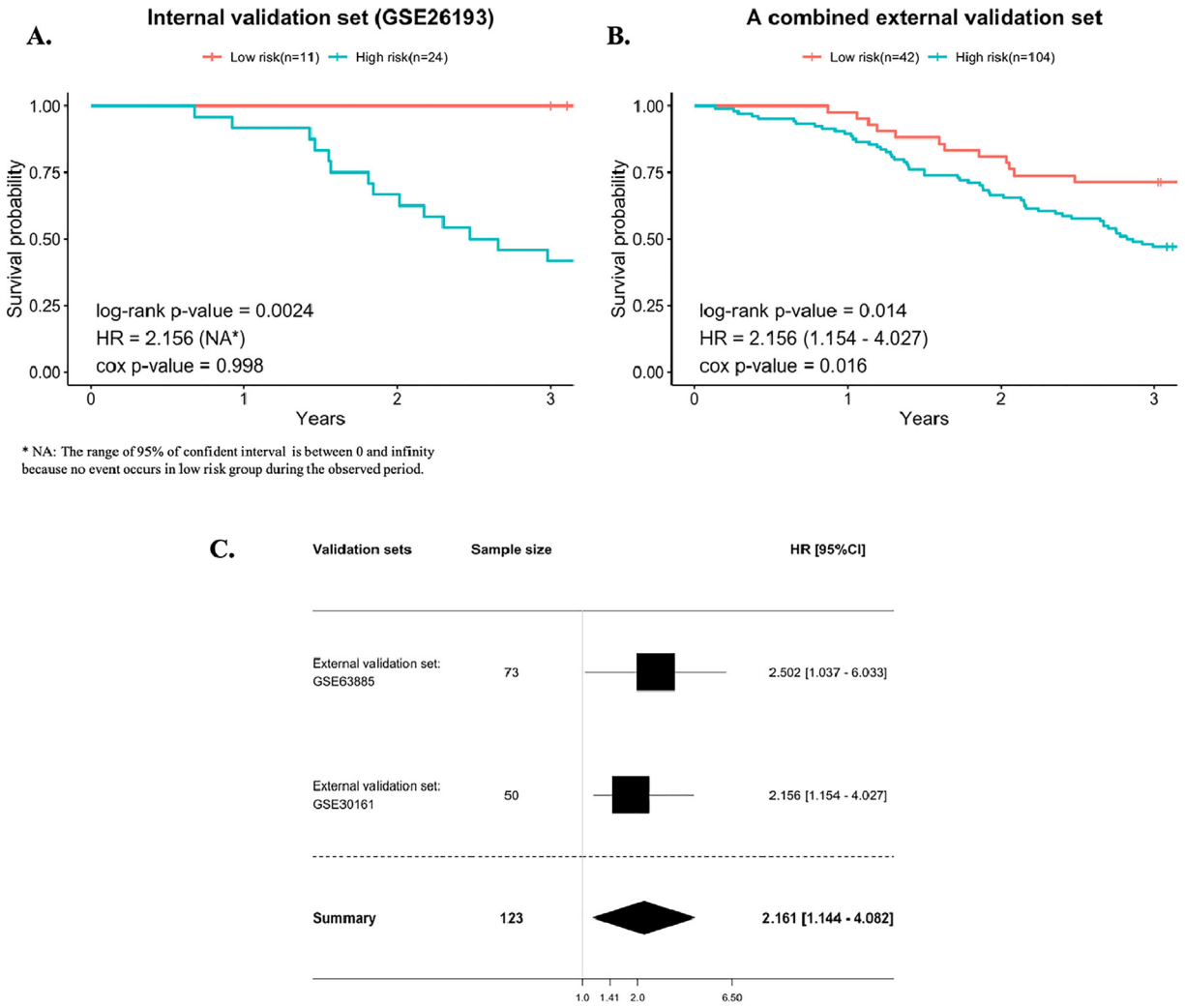


Fig. 3. Survival analysis of the internal validation set and external validation sets. (A) Kaplan-Meier analysis was conducted for the internal validation set (GSE26193; n = 25), and the patients were divided into two groups based on their risk scores. Significant differences (P < 0.05) were identified between the two groups over time. (B) Similar results are shown in the combined external validation set (GSE30161, GSE19829 and GSE63885; n = 146). (C) Forest plot of the hazard ratio (HR) and corresponding 95% confidence interval (CI) for individual external validation sets, excepting GSE19829. The vertical line indicates the null value (HR = 1). Each box indicates an individual study point estimate of the HR, and horizontal lines crossing these boxes indicate the 95% confident intervals. The diamond denotes the overall summary estimate of pooled studies.

for one individual external validation set (GSE19829) with an extremely large HR because no events happened in the low-risk group during the observed period (Fig. 3C). The HR point estimates of GSE63885 and GSE30161 were 2.502 (1.037, 6.033) and 2.156 (1.154, 4.027). Overall, the pooled HR of these external validation sets was 2.161 (1.144, 4.082).

Performance comparison

To verify the necessity and effectiveness of constructing a highly sensitive prediction model using such complicated GA-XGBoost models in the bagging algorithm, we replaced the GA-XGBoost model with two simple models: forward stepwise logistic regression and LASSO regression. The performance of these two models in both internal and external validation sets is displayed in Table 5. It can be seen that the results of the GA-XGBoost model in both internal and external validation sets achieved higher sensitivity and accuracy than the other two models, showing that the GA-XGBoost model is superior for risk prediction using gene expression values.

Functional analysis

To identify the biological function associated with the differentially expressed probes, we uploaded our probe list to the Ingenuity® Pathway Analysis (IPA®) server. The top disease/function annotations were significantly enriched in female genital tract serous carcinoma (P = 2.33E-34), and 64 differentially expressed genes (DEGs) were involved (Table S2 and S3). Additionally, based on a network analysis, it is noteworthy that the top regulatory network constructed by the DEGs was mainly regulated by the hub gene, *AKT*, which is implicated in many cancers (Fig. 4).

Effective drug prediction

For the identification of drugs effective in either high- or low-risk OC patients, the expression profiles and drug responses of 44 OC cell lines in the CCLE project were used. Through our bagging-based algorithm with GA-XGBoost models, 35 cell lines were defined as high risk, whereas the others were classified as low risk. Regarding the drug response, however, only 17-AAG

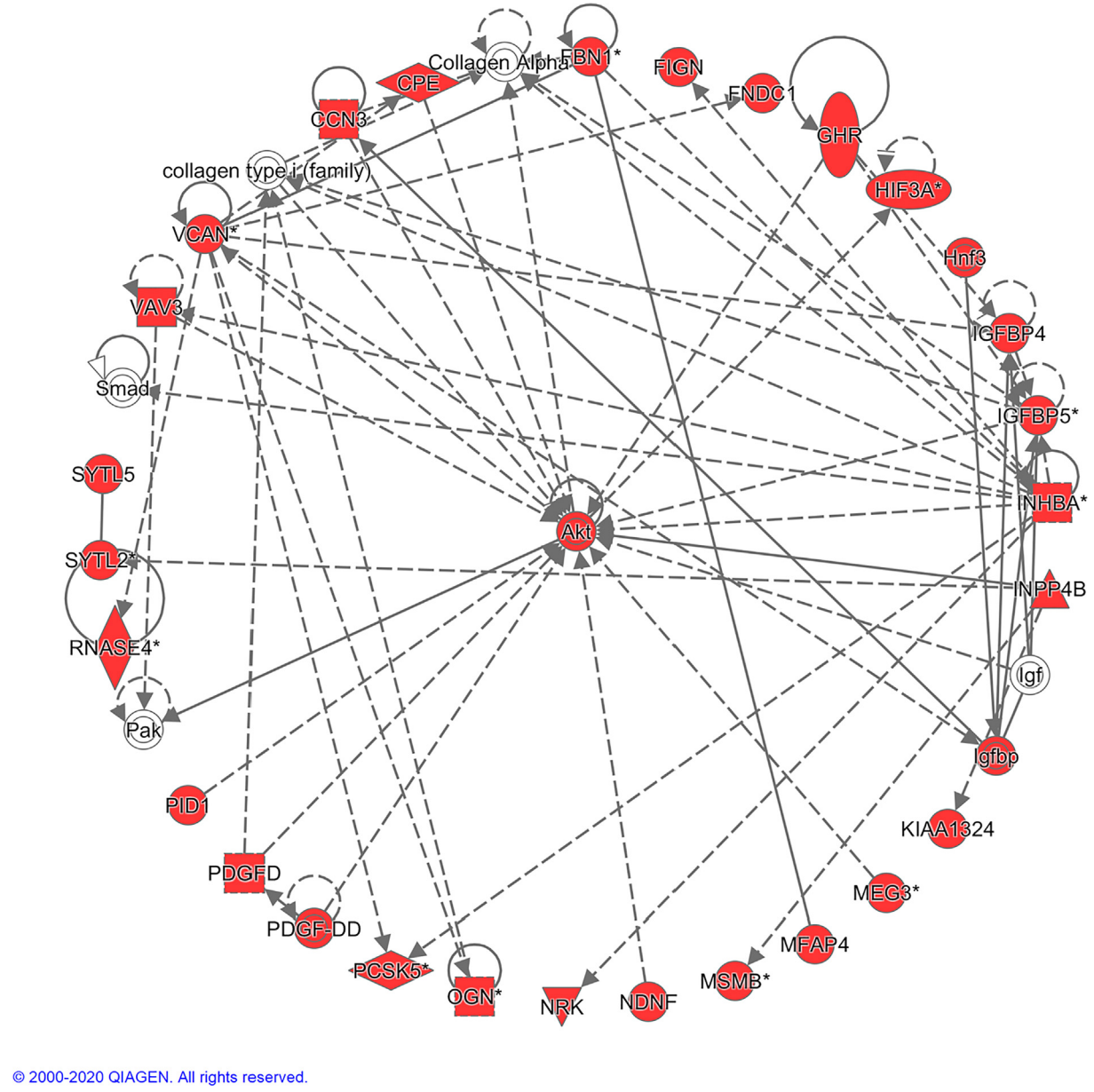


Fig. 4. The top network result from the Ingenuity® Pathway Analysis (IPA®) program. Red molecules represent the respective differentially expressed genes in our dataset, while white molecules indicate the putative genes that may be possibly involved in this network based on the IPA® database. Solid lines infer a direct interaction while dashed lines infer an indirect interaction.

(17-N-allylamino-17-demethoxygeldanamycin/ Tanespimycin), an antitumor antibiotic, and RAF265, a novel RAF/VEGFR2 inhibitor, had a slight difference in treatment efficacy at killing tumor cells between high-risk and low-risk cell lines, with P-values of 0.08 and 0.055, respectively (Table S4).

Discussion

As previously mentioned, few risk prediction models can predict high-risk OC patients with excellent sensitivity, and most of these models are not machine learning-based approaches. Therefore, we coupled t-tests and 2-fold changes to select features that were fitted by a GA-XGBoost model within a bagging algorithm. This method exhibited high sensitivity and moderate specificity in identifying high-risk patients who qualify for chemotherapy. Also, the combined HR point estimate of external validation sets indicated that the selected predictors are effective to distinguish the two risk groups.

Although our bagging algorithm successfully showed a high sensitivity for detecting high-risk OC patients, the low specificity of 38.5% in the external validation sets inferred a low accuracy for identifying the low-risk groups. Yet, few studies have focused on risk prediction models using gene expression for OC, so it is not feasible to compare the performance of our method with other models. However, two prediction models for breast cancer were amenable to comparison. Naderi et al. [43] used a Cox-ranked classifier with a prognostic signature of 70 genes and found it to have sensitivities of 77% and 63% in two external datasets, suggesting it may tend to ignore some high-risk patients who need to take chemotherapy. Similarly, another breast cancer study [44] developed three predictive models with good sensitivities (0.97–1) but

low specificity (30%), suggesting that the issue of low specificity in current risk prediction models remains a challenge in these female-specific cancers.

The specificity of the bagging algorithm was lower in the combined external validation sets than in the internal validation set, showing that some overfitting issues may exist in this approach. GAs themselves tend to overfit the training set, and unfortunately, there is no solution to this problem in GAs [45]. Overfitting may also arise from the complexity of the GA-XGBoost model [46] or from model diversity (Table S5), which limits the prediction performance [47]. We tried various strategies to avoid overfitting, including random sampling of the training set to increase the variety of each GA-XGBoost model, combining the GA with XGBoost via the shrinkage method [48], and comparison with forward stepwise logistic regression and LASSO regression. These methods produced an improved but still suboptimal prediction model, showing that the process of risk prediction is imperfect and iterative. Future research should balance the complexity and diversity of the prediction model with the performance of the bagging algorithm.

Regarding the biological evidence of significantly differentially expressed genes involved in our bagging algorithm, the network analysis from Ingenuity® Pathway Analysis (IPA®) revealed that the *AKT* (AKT serine/threonine kinase) gene is a hub for many of these genes. This gene is a crucial molecule in the PI3K/AKT/mTOR signaling pathway, which is vital in regulating cell proliferation, survival, and migration [49]. It has been reported that this pathway is frequently deregulated and associated with poor prognosis at advanced tumor stage in OC [50]; as a result, this pathway has become one of the famous anticancer targets in OC [51,52]. Both *PAK* (P21 activated kinase) and *INHBA* (inhibin subunit beta A) showed direct interactions with AKT, but only the latter was present in our dataset. A recent study revealed that higher expression of *INHBA* was connected to higher risk of death in patients with late-stage OC; hence, it is a potential target for blocking tumor progression [53]. The IPA results also revealed that *AKT* has many indirect interactions with insulin-like growth factor binding protein family members (e.g., IGFBP-4 and IGFBP-5), which can modulate insulin-like growth factors that have endocrine, autocrine, or paracrine functions [54]. *IGFBP-4* expression is elevated in the early tumor stage [55], and *IGFBP-5* is known to be a tumor suppressor by inhibiting expression of *AKT* [56]. In addition, several other genes connected to *AKT* also play important roles in the prognosis of OC. For example, *GHR* (growth hormone receptor), including estrogen or progesterone receptors, has been associated with better survival outcomes [57]. *VAV3* (vav guanine nucleotide exchange factor 3) overexpression in cancer stem cells is a biomarker for poor survival outcomes in OC [58]. Moreover, *PCSK5* encodes a proprotein convertase, and the increased expression of this protein family was related to poor survival outcomes in OC [59]. These findings suggest that the function of selected genes in this study is highly associated with the survival of OC patients.

Several factors may explain the slight difference in the response of high-risk and low-risk cell lines to 17-AAG and RAF265. First, only 55 of the 1457 cell lines (3.77%) in the CCLE are ovary cell lines, illustrating that a small number of samples may produce biased performance estimates when performing cross-validation of such high-dimensional data [60]. Second, although it has been reported that various types of OC, such as clear cell carcinoma, serous carcinoma, and endometrioid and mucinous carcinoma, showed different drug responses [61] the cell lines in CCLE were not classified into these subtypes, suggesting that the heterogeneity of the ovary cell lines may also affect the drug response results. Lastly, the expression patterns in tumor tissues and normal cell lines are not the same [62], so a model trained by tumor tissue samples may not be generalized to cell line samples.

Some drawbacks exist in this study. First, the insufficient number of samples may influence the accuracy of this method, and expanding the sample size of OC patients is still necessary. Second, unmeasured and residual confounders may exist that affect the results. Finally, quantile normalization did not remove the batch effect across these datasets. Moreover, the feature combinations from the GA were different even when the same parameters were set. Because the risk prediction approach we present here is not comprehensive enough to extend into other cancers, further research is required to fully develop a risk prediction model that considers cancer heterogeneity, cancer subtypes, and functional pathways. In the future, we will work to apply our method to other data sources, such as gene expression profiles from next-generation sequencing data in OC.

Conclusion

Predictive modeling using gene signatures for the early identification of high-risk individuals has shed light on personalized medicine, especially in stratified prevention strategies and clinical management. Considering that there are few risk-prediction models of OC using gene expression, we developed a bagging-based algorithm with GA-XGBoost models to predict the mortality risk of OC patients based on their gene expression patterns. Our method accurately predicted high-risk OC patients and has the potential to reduce unnecessary healthcare for those with low risk. However, several limitations still need to be addressed. Therefore, in the future, further investigations are necessary and warranted to validate the outcomes before clinical application.

Compliance with Ethics Requirements

This article contains gene expression profiles from human subjects and all analyzed datasets in this study were retrieved from a public database, Gene Expression Omnibus (GEO). All studied subjects were deidentified and thus it is not applicable to trace personal information from the IDs.

Declaration of Competing Interests

The authors declared no conflicts of interest.

Acknowledgments

We thank Dr. Melissa Stauffer for editing this manuscript. This work has been supported in part by the Center of Genomic and Precision Medicine, National Taiwan University, Taiwan [106R8400]; the Center for Biotechnology, National Taiwan University, Taiwan [GTZ300], and the Ministry of Science and Technology, Taiwan, (MOST-106-2314-B-002-134-MY2 and MOST-104-2314-B-002-107-MY2).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jare.2020.11.006.

References

[1] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: Cancer J Clin 2018;68:394–424.
[2] Gilks CB, Prat J. Ovarian carcinoma pathology and genetics: Recent advances. Hum Pathol 2009;40:1213–23.
[3] Prat J. Pathology of cancers of the female genital tract. Int J Gynecol Obstetrics 2015;131:S132–45.

[4] Rauh-Hain JA, Krivak TC, del Carmen MG, Olawaiye AB. Ovarian cancer screening and early detection in the general population. Rev Obstetr Gynecol 2011;4:15.
[5] Holland JF, Pollock RE. Holland-frei cancer medicine 8. PMPH-USA 2010.
[6] Morgan RJ, Armstrong DK, Alvarez RD, Bakkum-Gamez JN, Behbakht K, Chen L-m, et al. Ovarian cancer, version 1.2016, nccn clinical practice guidelines in oncology. J Natl Compr Canc Netw 2016;14:1134–63.
[7] Health promotion administration ministry of health and welfare Taiwan. Cancer registry annual report, 2016 Taiwan; 2018.
[8] Szalat R, Avet-Loiseau H, Munshi NC. Gene expression profile in clinical practice. Clin Cancer Res: Off J Am Assoc Cancer Res 2016;22:5434.
[9] Marisa L, de Reyniès A, Duval A, Selves J, Gaub MP, Vescovo L, et al. Gene expression classification of colon cancer into molecular subtypes: Characterization, validation, and prognostic value. PLoS Med 2013;10.
[10] Yasui W, Oue N, Aung PP, Matsumura S, Shutoh M, Nakayama H. Molecular-pathological prognostic factors of gastric cancer: A review. Gastric Cancer 2005;8:86–94.
[11] Markert EK, Mizuno H, Vazquez A, Levine AJ. Molecular classification of prostate cancer using curated expression signatures. Proc Natl Acad Sci 2011;108:21276–81.
[12] Stratford JK, Bentrem DJ, Anderson JM, Fan C, Volmar KA, Marron J, et al. A six-gene signature predicts survival of patients with localized pancreatic ductal adenocarcinoma. PLoS Med 2010;7.
[13] Van't Veer LJ, Dai H, Van De Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415:530–6.
[14] Pohl H, Kotze MJ, Grant KA, van der Merwe L, Pienaar FM, Apffelstaedt JP, et al. Impact of mammaprint on clinical decision-making in south african patients with early-stage breast cancer. Breast J 2016;22:442–6.
[15] Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 2004;351:2817–26.
[16] Ming C, Viassolo V, Probst-Hensch N, Chappuis PO, Dinov ID, Katapodi MC. Machine learning techniques for personalized breast cancer risk prediction: Comparison with the bcrat and boadicea models. Breast Cancer Res 2019;21:75.
[17] Crijns AP, Fehrmann RS, de Jong S, Gerbens F, Meersma GJ, Klip HG, et al. Survival-related profile, pathways, and transcription factors in ovarian cancer. PLoS Med 2009;6.
[18] Bonome T, Levine DA, Shih J, Randonovich M, Pise-Masison CA, Bogomolniy F, et al. A gene signature predicting for survival in suboptimally debulked patients with ovarian cancer. Cancer Res 2008;68:5478–86.
[19] Tabl AA, Alkhateeb A, ElMaraghy W, Rueda L, Ngom A. A machine learning approach for identifying gene biomarkers guiding the treatment of breast cancer. Front Genet 2019;10.
[20] Xu X, Zhang Y, Zou L, Wang M, Li A. A gene signature for breast cancer prognosis using support vector machine. In: 2012 5th International conference on biomedical engineering and informatics. p. 928–31.
[21] Ferroni P, Zanzotto FM, Riondino S, Scarpato N, Guadagni F, Roselli M. Breast cancer prognosis using a machine learning approach. Cancers. 2019;11:328.
[22] Abdou Y, Baird A, Dolan J, Lee S, Park S, Lee S. 176o machine learning-assisted prognostication based on genomic expression in the tumour microenvironment of estrogen receptor positive and her2 negative breast cancer. Ann Oncol 2019;30(mdz240):002.
[23] Gao Y-C, Zhou X-H, Zhang W. An ensemble strategy to predict prognosis in ovarian cancer based on gene modules. Front Genet 2019;10:366.
[24] Gentric G, Kieffer Y, Mieulet V, Goundiam O, Bonneau C, Nemati F, et al. Pml-regulated mitochondrial metabolism enhances chemosensitivity in human ovarian cancers. Cell Metab 2019;29(156–173):e110.
[25] Mateescu B, Batista L, Cardon M, Gruosso T, De Feraudy Y, Mariani O, et al. Mir-141 and mir-200a act on ovarian tumorigenesis by controlling oxidative stress response. Nat Med 2011;17:1627.
[26] Ferriss JS, Kim Y, Duska L, Birrer M, Levine DA, Moskaluk C, et al. Multi-gene expression predictors of single drug responses to adjuvant chemotherapy in ovarian carcinoma: Predicting platinum resistance. PLoS ONE 2012;7.
[27] Konstantinopoulos PA, Spentzos D, Karlan BY, Taniguchi T, Fountzilas E, Francoeur N, et al. Gene expression profile of brcaness that correlates with responsiveness to chemotherapy and with outcome in patients with epithelial ovarian cancer. J Clin Oncol 2010;28:3555.
[28] Lisowska KM, Olbryt M, Dudaladava V, Pamuła-Piłat J, Kujawa K, Grzybowska E, et al. Gene expression analysis in ovarian cancer–faults and hints from DNA microarray study. Front Oncol 2014;4:6.
[29] Colombo N, Lorusso D, Scollo P. Impact of recurrence of ovarian cancer on quality of life and outlook for the future. *International Journal of Gynecologic.* Cancer 2017;27.
[30] Ushijima K. Treatment for recurrent ovarian cancer—at first relapse. J Oncol 2010;2010.
[31] Irizarry RA, Gautier L, Bolstad BM, Miller C. Methods for affymetrix oligonucleotide arrays. R package version 2006;1(12):1.
[32] Bolstad BM. Preprocesscore: a collection of pre-processing functions. R Package Version 2013;1.
[33] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. p. 785–94.
[34] Friedman JH. Stochastic gradient boosting. Comput Stat Data Anal 2002;38:367–78.
[35] Holland JH. Genetic algorithms. Sci Am 1992;267:66–73.
[36] Goldberg DE. Genetic algorithms. Pearson Education India; 2006.
[37] Chen T, He T, Benesty M, Khotilovich V, Tang Y. Xgboost: Extreme gradient boosting. R package version 2015(4-2):1–4.
[38] Breiman L. Bagging predictors. Mach Learn 1996;24:123–40.
[39] Tibshirani R. Regression shrinkage and selection via the lasso. J Roy Stat Soc: Ser B (Methodol) 1996;58:267–88.
[40] In Lee K, Koval JJ. Determination of the best significance level in forward stepwise logistic regression. Commun Statist-Simulat Comput 1997;26:559–75.
[41] Lin H, Zelterman D. Modeling survival data: Extending the cox model; 2002.
[42] Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 2012;483:603–7.
[43] Naderi A, Teschendorff A, Barbosa-Morais N, Pinder S, Green A, Powe D, et al. A gene-expression signature to predict survival in breast cancer across independent data sets. Oncogene 2007;26:1507–16.
[44] Haibe-Kains B, Desmedt C, Piette F, Buyse M, Cardoso F, Van't Veer L, et al. Comparison of prognostic gene expression signatures for breast cancer. BMC Genom 2008;9:394.
[45] Allen F, Karjalainen R. Using genetic algorithms to find technical trading rules. J Financ Econ 1999;51:245–71.
[46] Myung IJ. The importance of complexity in model selection. J Math Psychol 2000;44:190–204.
[47] Kuncheva LI, Whitaker CJ. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Mach Learn 2003;51:181–207.
[48] Liao X, Cao N, Li M, Kang X. Research on short-term load forecasting using xgboost based on similar days. 2019 International conference on intelligent transportation, big data & smart city (ICITBS) 2019:675–8.
[49] Owusu-Brackett N, Shariati M, Meric-Bernstam F. Role of pi3k/akt/mtor in cancer signaling. In: Badve S, Kumar GL, editors. Predictive biomarkers in oncology: Applications in precision medicine. Cham: Springer International Publishing; 2019. p. 263–70.
[50] Musa F, Schneider R. Targeting the pi3k/akt/mtor pathway in ovarian cancer. Trans Cancer Res 2015;4:97–106.
[51] Cheaib B, Auguste A, Leary A. The pi3k/akt/mtor pathway in ovarian cancer: Therapeutic opportunities and challenges. Chinese J Cancer 2015;34:4–16.
[52] Ghoneum A, Said N. Pi3k-akt-mtor and nfκb pathways in ovarian cancer: Implications for targeted therapeutics. Cancers 2019;11:949.
[53] Li X, Yang Z, Xu S, Wang Z, Jin P, Yang X, et al. Targeting inhba in ovarian cancer cells suppresses cancer xenograft growth by attenuating stromal fibroblast activation. Dis Markers 2019;2019.
[54] Grimberg A, Cohen P. Role of insulin-like growth factors and their binding proteins in growth control and carcinogenesis. J Cell Physiol 2000;183:1–9.
[55] Mosig R. Igfbp-4 is a candidate serum biomarker for detection and surveillance of early stage epithelial ovarian cancer. Research 2015.
[56] Hwang JR, Cho Y-J, Lee Y, Park Y, Han HD, Ahn HJ, et al. The c-terminus of igfbp-5 suppresses tumor growth by inhibiting angiogenesis. Sci Rep 2016;6:39334.
[57] Chen S, Dai X, Gao Y, Shen F, Ding J, Chen Q. The positivity of estrogen receptor and progesterone receptor may not be associated with metastasis and recurrence in epithelial ovarian cancer. Sci Rep 2017;7:1–7.
[58] Kwon A-Y, Kim G-I, Jeong J-Y, Song J-Y, Kwack K-B, Lee C, et al. Vav3 overexpressed in cancer stem cells is a poor prognostic indicator in ovarian cancer patients. Stem Cells Dev 2015;24:1521–35.
[59] Page RE, Klein-Szanto AJ, Litwin S, Nicolas E, Al-Jumaily R, Alexander P, et al. Increased expression of the pro-protein convertase furin predicts decreased survival in ovarian cancer. Anal Cell Pathol 2007;29:289–99.
[60] Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. PLoS ONE 2019;14:e0224365.
[61] Kim S, Han Y, Kim SI, Kim H-S, Kim SJ, Song YS. Tumor evolution and chemoresistance in ovarian cancer. NPJ Precis Oncol 2018;2:1–9.
[62] Ochs MF, Ertel A, Verghese A, Byers SW, Tozeren A. Pathway-specific differences between tumor cell lines and normal and tumor tissue cells; 2006.