Lien *et al. Human Genomics* (2025) 19:56
https://doi.org/10.1186/s40246-025-00763-4

Human Genomics

RESEARCH Open Access

Exploring potential methylation markers for ovarian cancer from cervical scraping samples

Ju-Yin Lien[1], Lu Ann Hii[1], Po-Hsuan Su[2], Lin-Yu Chen[3], Kuo-Chang Wen[3,4], Hung-Cheng Lai[3,4,5,6*] and Yu-Chao Wang[1,7*]

Abstract

Background Ovarian cancer has the highest mortality rate among gynecological cancers, making early detection crucial, as the five-year survival rate drops from 92% with early-stage diagnosis compared to 31% with late-stage diagnosis. Current diagnostic methods such as histopathological examination and detection of cancer antigen 125 and human epididymis protein 4 biomarkers are either invasive or lack specificity and sensitivity. However, the Papanicolaou (Pap) test, which is widely used for cervical cancer screening, shows the potential for detecting ovarian cancer by identifying tumor DNA in cervical scrapings. Since aberrant DNA methylation patterns are linked to cancer progression, DNA methylation offers a promising avenue for early diagnosis. Therefore, this study aimed to develop a methylation-based machine-learning model to stratify patients with ovarian cancer from the cervical scraping samples collected via Pap test.

Results Cervical scrapings were collected by gynecologists using conventional Pap smears. In total, 160 samples were collected: 95 normal, 37 benign, and 28 malignant. Methylation data were generated using the Illumina Infinium MethylationEPIC BeadChip array, which contains approximately 850,000 CpG loci. Methylation data were initially divided into training and testing sets in a 3:1 ratio comprising 120 and 40 samples, respectively. A two-step methylation-based model was trained using the training data for classification: a principal component analysis (PCA) model, consisting of 30 features, to classify samples as normal or tumor; then a gradient boosting model, containing 16 features, to further stratify tumor samples as benign or malignant. The two-step model achieved an accuracy of 0.88 and an F1-score of 0.86 on the testing data. Furthermore, an over-representation analysis was conducted to explore the functions associated with genes mapped from differentially methylated positions (DMPs) in comparisons between normal and tumor samples, as well as between benign and malignant samples. These results suggest that DMPs may be associated with olfactory transduction when comparing normal versus tumor samples, and immune regulation when comparing benign and malignant samples.

*Correspondence:
Hung-Cheng Lai
hclai30656@gmail.com
Yu-Chao Wang
yuchao@nycu.edu.tw

Full list of author information is available at the end of the article

Conclusions Our two-step model shows promise for predicting ovarian cancer and suggests that cervical scrapings may be a viable alternative for sample collection during screening.

Keywords Ovarian cancer, Methylation, Machine learning, Papanicolaou test (Pap test), Epigenetics, Cancer screening, Biomarker

Background

According to Global Cancer Observatory (GCO) statistics, ovarian cancer is the eighth most common among all cancer types in women [1]. It has the highest mortality among gynecological malignancies because of its nonspecific symptoms, such as abdominal pain and bloating, dyspepsia, vomiting, constipation, and urinary symptoms [2, 3]. Another reason for this is the lack of detection methods in the early stages [4]; patients diagnosed with ovarian cancer at a later stage have a five-year overall survival rate of only 31%, whereas those diagnosed at an earlier stage have a survival rate of 92% [4, 5]. This disparity underscores the critical importance of early detection to improve patient outcomes and survival rates. Despite the efforts of many researchers in developing effective screening methods for ovarian cancer, the wide array of pathological and genetic variations complicates both clinical detection and treatment [6, 7].

Currently, ovarian cancer diagnosis primarily relies on histopathological examination. A potential early detection approach is detecting cancer antigen 125 (CA125), which is secreted by ovarian cancer cells and expressed in the endometrium [8]. However, the increase in serum CA125 levels is often high in women who have benign pelvic disease or inflammation [9], resulting in low specificity and sensitivity [10, 11]. Human epididymis protein 4 (HE4), another biomarker used for ovarian diagnosis, is overexpressed in ovarian cancer [9, 12]. However, HE4 levels are associated with menopausal status and age and show decreased sensitivity and specificity in older people [13]. Other potential biomarkers such as Desmoglein-2 (DSG2) are differentially expressed at different stages and grades of ovarian cancer [14]. However, the UKCTOCS trial, which observed more than 200,000 women over a median follow-up of 11 years, found that while HE4 and other biomarkers improved the early detection of ovarian cancer, they did not significantly reduce mortality rates. This finding underscores the need to develop more effective screening strategies and novel biomarkers [15]. The Papanicolaou test (Pap test) is commonly used in cervical cancer screening and is a noninvasive and low-risk method to detect cervical cancer. Previous studies have indicated that tumor DNA from the ovary can be detected using the Pap test because ovarian cancer cells are shed and collected using the Pap test [16, 17]. These studies analyzed somatic mutations in cervical scraping samples and detected mutations in 41% and 29% of patients with ovarian cancer, respectively. Studies suggest that the Pap test has the potential to be used in sampling for ovarian cancer detection [16, 17].

DNA methylation is most common in epigenetic studies, as it regulates gene expression and contributes to cancer initiation and progression [18, 19]. DNA methylation is performed by DNA methyltransferases, which add a methyl group from S-adenosyl methionine (SAM) to the fifth carbon of a cytosine nucleotide [20]. Hypermethylation of the promoter region suppresses gene transcription, whereas hypomethylation of the promoter activates gene transcription. Aberrant gene methylation has been observed in various types of cancers and is recognized as a potential biomarker for diagnosis and prognosis. For instance, *CDKN2A*, *CDH13*, *RASSF1A*, and *APC* serve as biomarkers for lung cancer to detect tumor development and recurrence [21, 22]. Additionally, the FDA-approved methylation biomarkers for colorectal cancer diagnosis include *SEPT9*, *NDRG4*, and *BMP3* [23–25]. Recent research suggests that DNA methylation plays a crucial role in the early detection of ovarian cancer, because alterations in methylation levels occur during the early stages of cancer development [26]. Previous reports have documented the aberrant methylation levels of tumor suppressor genes, such as hypermethylation of promoters in *BRCA1/2*, *TP53*, and *RASSF1A* [27–30], which are associated with tumor development and progression, indicating their potential as methylation biomarkers for detecting ovarian cancer. Furthermore, hypomethylation of promoters in the oncogene *HOXA9* leads to high expression, which is implicated in tumor development [26, 28].

To explore potential methylation biomarkers for the early detection of ovarian cancer, we used machine-learning methods to identify a set of CpG sites to predict patients with ovarian cancer. A previous study combined methylation data and machine-learning to predict patients with non-small cell lung cancer [31]. Another study developed a machine-learning ridge regression risk score model consisting of 14,000 CpG sites to predict ovarian cancer [32]. In this study, 242 tumor samples and 869 normal samples were collected from the cervical cells of Western women via Pap tests to serve as training data. The model demonstrated a performance level with an area under the receiver operating characteristic curve (AUC) of 0.78 on the internal validation dataset. For the external validation set, which included 47 tumor samples and 225 normal samples, the risk score model resulted in an AUC of 0.76 [32]. Although this study investigated the

potential of using cervical cell methylation data to predict ovarian cancer, the large number of features in the ridge regression model presents challenges for clinical application because many feature weights in the model are close to zero, indicating potential noise. To address these issues, this study aimed to construct a methylation-based model to stratify patients with ovarian cancer using cervical scraping samples collected via Pap tests. The goal was to use fewer CpG sites to predict ovarian cancer, thus making the method simpler and more economical for clinical use.

Methods

Overview of the method

In this study, we developed a two-step methylation-based model for the stratification of normal, benign, and malignant samples in ovarian cancer (Fig. 1a). Methylation data were initially divided into training and testing sets in a 3:1 ratio comprising 120 and 40 samples, respectively. Quality control, preprocessing, and feature selection procedures were performed separately for the training data. The first step of the model was constructed using the entire training dataset ($n = 120$), which included samples labeled as "normal" or "tumor" (encompassing both benign and malignant samples), allowing the stratification of normal and tumor samples. In the second step, the model was specifically trained using only benign and malignant samples from the training data to predict whether the samples were benign or malignant. Upon establishing the two-step model, the testing data ($n = 40$) were processed and used to validate the model.

Data collection

Cervical scrapings were collected by gynecologists using conventional Pap smears. In total, 160 samples were collected: 95 normal, 37 benign, and 28 malignant. Demographic characteristics of all clinical samples were shown in Table 1. For participants requiring surgery, samples were collected prior to their first surgical procedure. Specimens were collected according to institutional policies. The inclusion criteria consisted of women diagnosed with primary epithelial ovarian cancer (including serous ovarian carcinoma, clear cell carcinoma, endometrioid carcinoma, and mucinous carcinoma). Exclusion criteria included: (1) a history of other gynecological or breast cancers or previous cancer treatment; (2) prior hysterectomy; (3) incomplete clinical or pathological information; and (4) current pregnancy, postpartum status, or lactation. The protocol and informed consent forms were approved by the Taipei Medical University IRB (N201810036). Written informed consent was obtained from all participants prior to their enrollment in the study. Methylation data were generated using the Illumina Infinium MethylationEPIC BeadChip array, which contains approximately 850,000 CpG loci. The raw data were stored in the IDAT format, with each sample having two files corresponding to green (methylated) and red (unmethylated) intensity values.

Data preprocessing

The IDAT files were preprocessed using the R package *minfi* [33], and the training data for both the Step One and Step Two models underwent the same preprocessing pipeline (Fig. 1b). First, the IDAT files were read with *minfi*, and probes with NBeads < 3 in at least 5% of all samples were removed from RGChannelSetExtended [34]. Subsequently, samples were normalized using the single-sample normalization function *preprocessNoob* in *minfi*, which corrects for background intensity and probe dye bias [35]. Subsequently, the samples were further processed using the beta-mixture quantile dilation (BMIQ) function from the R package *wateRmelon* [36] to address the biases between the Infinium type I and type II probes. After BMIQ correction, the output was the beta-value for each CpG site. The signal intensity was converted into a beta-value using the following formula:

$$\text{beta-value} = \frac{M}{U+M+\alpha}$$

where M represents methylated intensity, U represents unmethylated intensity, and α is a constant added to regularize the beta-values, preventing errors when both U and M are low. This constant is set to 100, as recommended by Illumina. The range of beta-values ranges from zero to one, where a value of zero indicates that the CpG site is entirely unmethylated, while a value of one indicates that the CpG site is fully methylated [37]. The M-value is calculated from the beta-value, and the formula that illustrates their relationship is given below:

$$\text{M-value} = \log_2 \frac{\text{beta-value}}{1-\text{beta-value}}$$

The M-value was obtained through a base 2 logarithmic transformation of the beta-value. This transformation results in approximately homoscedastic data, making the M-value more appropriate than the beta-value for statistical analyses [37].

The probes were annotated with the reference genome version hg19 and filtered based on the following criteria in GenomicRatioSet:

1. CpG sites with a P-value > 0.01 in the *detectionP* function were considered failed; CpG sites with > 10% failure rate or samples with > 50% failed CpG sites were removed.
2. CpG sites located on the Y chromosome were removed, as all patients were female.
3. SNP-related CpG sites were excluded.
4. Cross-reactive loci were eliminated using the *maxprobes* R package [38, 39].

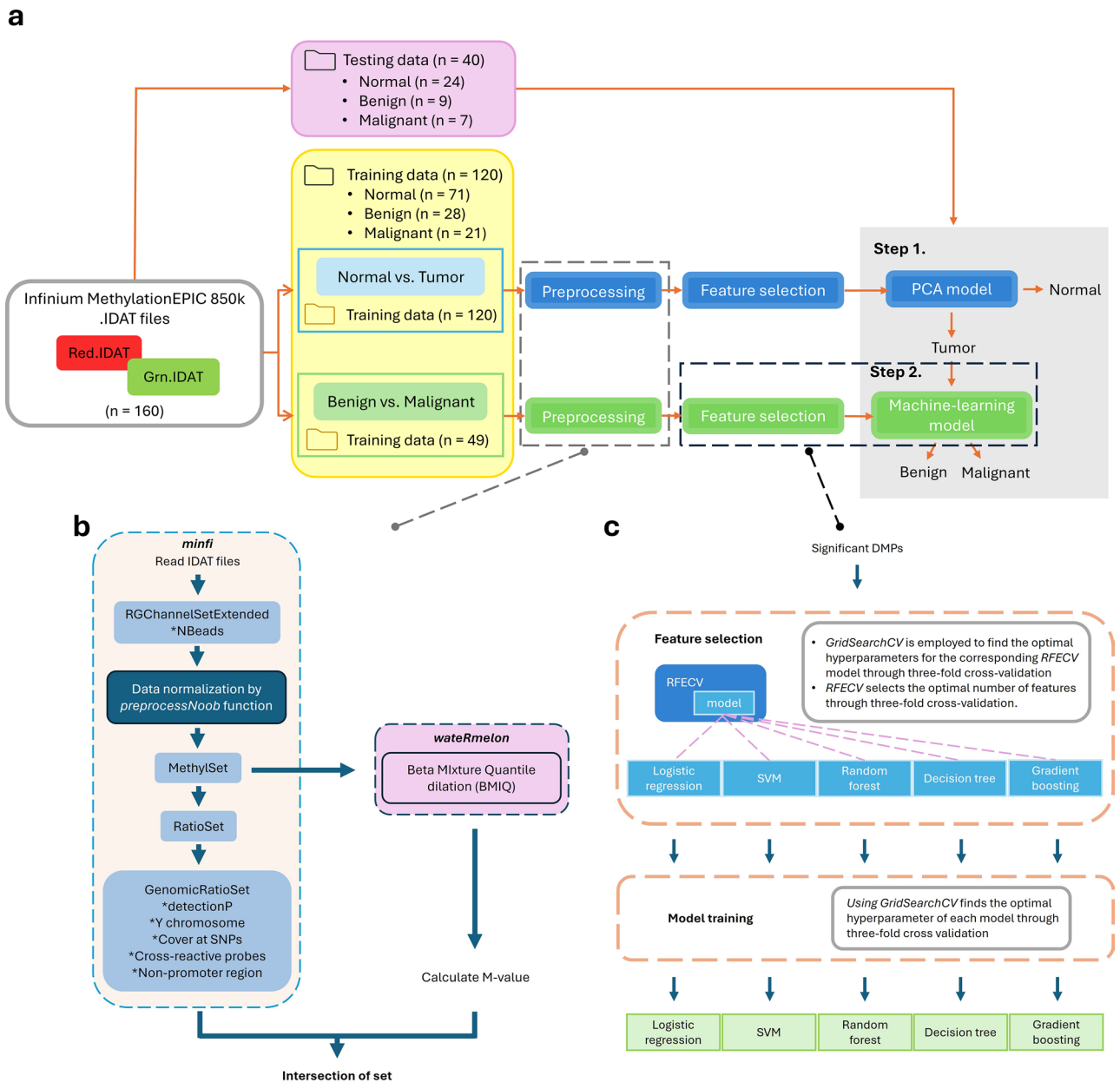

Fig. 1 Overview of the study workflow. (**a**) Study design and classification framework: Cervical scraping samples ($n = 160$) were processed using the Illumina Infinium MethylationEPIC BeadChip. The dataset was split into training data ($n = 120$) and testing data ($n = 40$). A two-step classification model was developed, where Step One classified samples as normal or tumor (benign + malignant), and Step Two further distinguished benign from malignant tumors. (**b**) Data preprocessing pipeline: Methylation data underwent quality control, normalization, and filtering to ensure high data quality. *Indicates the filtered criteria in each step. (**c**) Machine learning model development: Feature selection was performed using recursive feature elimination with cross-validation (RFECV), followed by model training. Multiple machine-learning models were evaluated for Step Two model, with gradient boosting selected as the final model for classifying benign and malignant tumors

5. CpG sites within the promoter region (TSS1500, TSS200, 5′UTR, and 1st Exon) were selected based on the "UCSC_RefGene_Group" column in the annotation data.

Subsequently, the results from BMIQ normalization and filtering intersected, and the combined data were further analyzed.

Differentially methylated positions (DMPs)

The DMP of both models, normal vs. tumor and benign vs. malignant, were identified using the same criteria: CpG sites were considered DMPs if the adjusted P-value from *dmpFinder* was less than 0.01 and the absolute value of the mean of DNA methylation differences (Δβ) was greater than 0.2. The adjusted P-values from *dmpFinder* in *minfi* were calculated using the M-value to compare

Table 1 Demographic characteristics of all clinical samples

	Normal (n = 95)	Benign (n = 37)	Malignant (n = 28)
Age	31–46	26–66	30–88
Histotype			
Endometrioma	–	17	–
Benign serous or serous cyst adenoma	–	8	–
Teratoma	–	6	–
Benign mucinous or mucinous cyst adenoma	–	5	–
Other benign tumor		1	
Serous carcinoma	–	–	10
Clear cell carcinoma	–	–	6
Endometrioid	–	–	6
Mucinous carcinoma	–	–	6
FIGO Stage			
I	–	–	15
II	–	–	3
III	–	–	6
IV	–	–	3
Unknown	–	–	1
Grade			
G1	–	–	4
G2	–	–	5
G3	–	–	12
Unknown	–	–	7

the methylation levels between the two groups using the F-test statistic, and the P-value was adjusted using Storey's method. The Δβ is the subtraction of the mean beta-value within one group from the mean beta-value within the other group.

Mann–Whitney U test

The Mann–Whitney U-test was used for feature selection in both models after DMP filtering to further select candidate CpG sites. CpG sites that were significant (adjusted P-value < 0.05, with Benjamini–Hochberg correction) in the statistical test were considered as candidate features. The Mann–Whitney U test used in this study was two-tailed and calculated using the *scipy.stats.mannwhitneyu* function.

Recursive feature elimination

Recursive feature elimination (RFE), a feature selection method that employs a machine-learning model to assess features based on their importance weights, such as *feature_importances_* and *coef_*. In each iteration, the weakest feature was removed. The *RFECV* function extends the *RFE* by evaluating the model performance across different feature subsets and stops when only one feature remains. It then identifies and returns the optimal number of features that result in the maximum F1-score. To apply the *RFECV* function to the features, we first optimized the respective machine-learning model using *GridSearchCV* in *sklearn*. Using the optimized model, we employed *RFECV* to determine the best features for a specific model.

Development of Step One model: normal versus tumor

Three types of samples—normal, benign, and malignant—comprising the training data were used to develop the Step One principal component analysis (PCA)-based model that classifies normal and tumor (benign and malignant) samples. The Step One model was trained using significant CpG sites selected from the training data using the Mann–Whitney U test, with an adjusted P-value threshold of less than 0.05. To determine the optimal number of features for the Step One PCA model, we performed PCA and identified the specific principal component (PC) that effectively divided the samples into two clusters. Since the PC indicated the linear combination of the CpG sites, we can order the CpG sites according to their corresponding absolute weights, from largest to smallest. The PCA model was then constructed, starting with the top two CpG sites and adding one CpG site in each iteration until all CpG sites were included. In each iteration, the PC that could approximately divide the samples into two clusters was identified, and the area under the precision-recall curve (AUPRC) was recorded. Consequently, the number of CpG sites corresponding to an AUPRC of 0.95 was selected as the final feature number, and the threshold for distinguishing normal and tumor samples was determined by the F1-score. The weights of this PC were then extracted to develop the Step One model which calculates the risk score for ovarian cancer.

Development of Step Two model: benign versus malignant

The Step Two model was deployed to stratify the samples identified as tumors using the Step One model, further categorizing them as benign or malignant. The training data for the Step Two model includes both benign and malignant samples. We constructed five major machine-learning models (Fig. 1c): logistic regression, support vector machine (SVM), decision tree, random forest, and gradient boosting. Applying different hyperparameters and algorithms using the *sklearn* library, such as the penalty in logistic regression and kernel in SVM, resulted in 10 models. After selecting the significant DMPs, *RFECV* was performed for each model to identify important features for model training. In the corresponding model within the *RFECV*, hyperparameter tuning was performed using *GridSearchCV* with three-fold cross-validation. The features selected by *RFECV* were then utilized to train the specific final model, with further hyperparameter tuning conducted using *GridSearchCV* in a

three-fold cross-validation. For instance, when training an SVM model, the hyperparameters were initially tuned using *GridSearchCV*, which was subsequently applied to the *RFECV* process. Subsequently, *RFECV* was employed to select the important features for the final SVM model. Subsequently, the selected features were used in conjunction with *GridSearchCV* to fine-tune the hyperparameters of the final SVM model.

Overrepresentation analysis

The overrepresentation analysis (ORA) was conducted using the R package *missMethyl*. The *gometh* function was used to perform Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses, mapping CpG sites to genes and considering the number of CpGs linked to each gene [40]. After mapping the CpG sites to genes, enrichment analysis was used to compare a gene set consisting of genes corresponding to CpG sites that were significantly different between the two groups to a background set.

In the *gometh* function, the names of DMPs from the Step One and Step Two models were used as input for the "sig.cpg" arguments, and the CpG sites after preprocessing were used as the background set input for the "all.cpg" arguments [40]. GO and KEGG terms with a false discovery rate (FDR) higher than 0.05 were filtered out, focusing specifically on the ontology of biological processes (BPs). To investigate the fundamental pathways, the GO term results were intersected with their offspring using the function *GOBPOFFSPRING* in the R package *GO.db* [41].

Results

Overview of ovarian cancer cervical scraping samples

Before constructing the ovarian cancer prediction model, we performed quality control assessment of the samples. The samples in our study utilized the Infinium MethylationEPIC methylation data derived from cervical scraping samples collected via Pap tests. To avoid the influence of poor-quality samples on the downstream analysis, we used *RGchannelSetExtended* data to check the signal intensity of each sample and employed beta-values to visualize the distribution of methylation levels in each sample. The training data for the Step One model comprised 120 samples: 71 normal, 28 benign, and 21 malignant. We detected 865,859 CpG sites using the *minfi* package. A scatter plot (Supplementary Fig. 1a) illustrates the quality of the training data for the Step One model after reading the IDAT file. The results revealed that most samples were of good quality, except for one sample identified as R03C01_205676800086, which is marked in red in Supplementary Fig. 1a. This sample was benign and was subsequently removed from downstream analysis. To further investigate methylation levels in each sample, the signal intensities of the probes were converted to beta-values using *preprocessNoob* in *minfi* and *BMIQ* in *wateRmelon*. Following this conversion, beta-values were corrected for background intensity, dye bias, and probe bias. Subsequently, a density plot revealed a bimodal distribution with two major peaks (Supplementary Fig. 1b). The training data for the Step Two model comprised 27 benign and 21 malignant samples, and 866,238 CpG sites were identified. Quality assessment via scatter plots of raw methylated and unmethylated intensity signals confirmed that all the samples were of good quality (Supplementary Fig. 1c). Beta-value distributions were analyzed for each benign and malignant sample, revealing a bimodal distribution in the density plot without any noticeable outliers after normalization (Supplementary Fig. 1d).

After excluding poor-quality samples from the training data, we conducted PCA on two training datasets, containing 865,859 and 866,238 CpG sites. The PCA results for the Step One model indicated that normal and tumor samples (benign and malignant combined) could be approximately separated along the second principal component (PC2) (Fig. 2a). In contrast, the PCA results for the Step Two model revealed that the benign and malignant samples were closely clustered together, with neither the first (PC1) nor the second principal component (PC2) effectively separating the two groups (Fig. 2b).

Samples preprocessing in training data of two-step models

The raw methylation data were not immediately suitable for model construction. Thus, we implemented a series of preprocessing steps to ensure that the dataset was clean, free from biases, and devoid of unnecessary features that were irrelevant to ovarian cancer prediction. The preprocessing steps involved filtering the probes and CpG sites to eliminate poor-quality probes and unwanted variations. Beta-values and M-values were calculated for subsequent analyses (Fig. 1b). The reduction in the number of probes and CpG sites at each step is summarized in Supplementary Table 1. For the training data of the Step One model, we began by removing 6,781 probes based on the number of "NBeads" in the *RGChannelSetExtended*, which represent noise occurring during DNA hybridization in the experiment. Next, we eliminated 1,644 CpG loci identified as low-quality CpG sites using the *detectionP* function. Following probe annotation of the reference gene, we removed 586,184 probes that could potentially affect the downstream analysis, including those located on the Y chromosome, SNP-related CpG sites, cross-reactive loci, and non-promoter CpG sites. Since DNA methylation regulates gene transcription in the promoter region, we selected 271,730 CpG loci located in the promoter regions for subsequent analysis. The preprocessing procedure for the training data in the

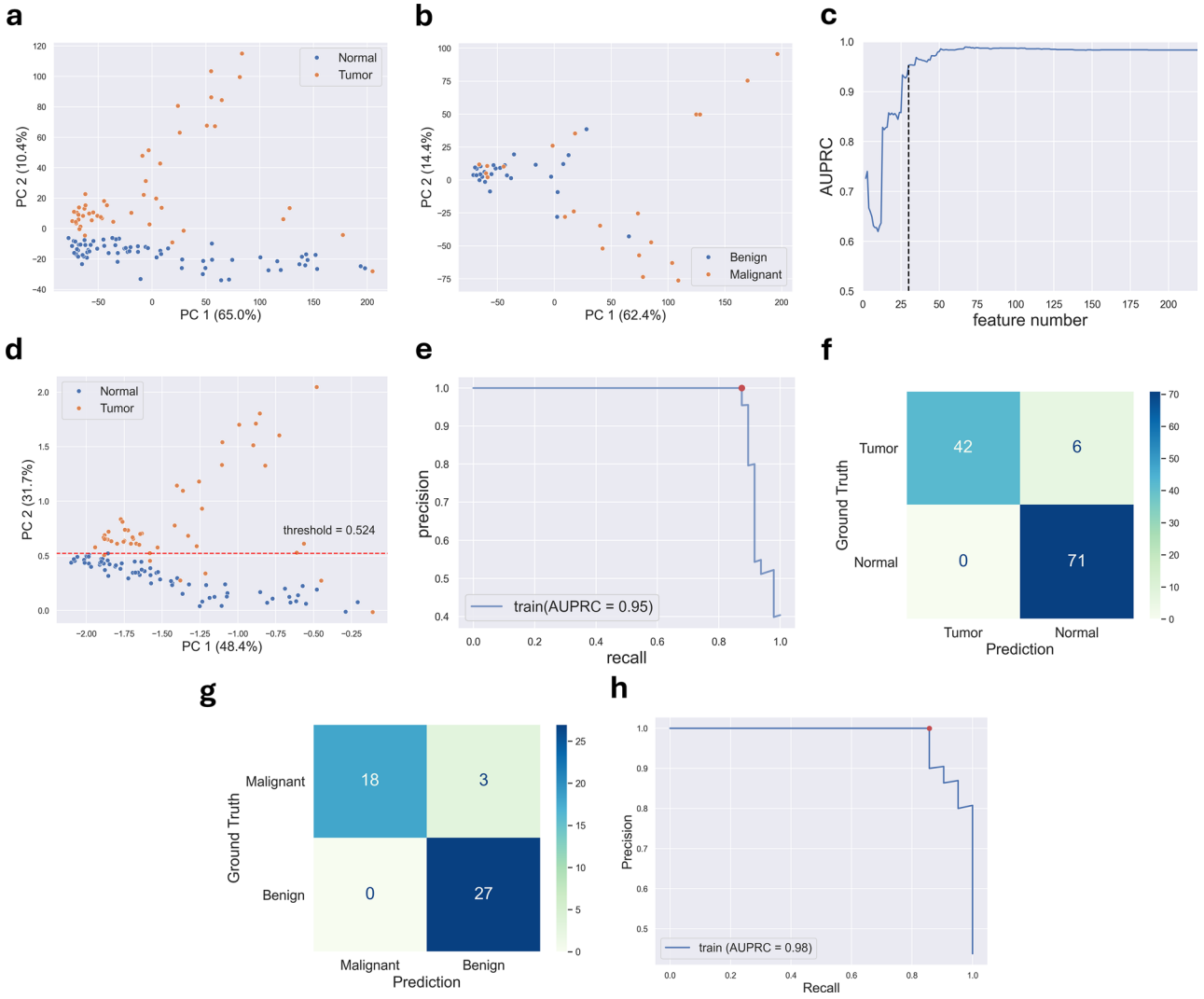

Fig. 2 Results of Step One and Step Two model development. (**a**) PCA plot of 119 samples with 865,859 CpG sites. (**b**) PCA plot of 48 samples with 866,238 CpG sites. (**c**) The value of AUPRC in 2 to 219 feature number of PCA model. The dashed line reveals the 30 features corresponding 0.95 AUPRC. (**d**) PCA of training data with 30 features (CpG sites). The dashed line indicates the classification threshold for normal and tumor. (**e**) The precision-recall curve in the PCA model. The red dot indicates the threshold corresponding to precision and recall. (**f**) Confusion matrix of prediction results in the PCA model of training data. (**g**) Confusion matrix of prediction results in the gradient boosting model by using training data. (**h**) In the precision-recall curve of the gradient boosting model, the red dot indicates the threshold corresponding to precision and recall

Step Two model follows the same steps as the Step One model (Fig. 1b). The numbers of probes and CpG sites removed at each step are shown in Supplementary Table 1. Initially, 6,643 probes were removed based on the number of "NBeads" in the training data in the Step Two model. Next, 1,383 low-quality CpG sites were eliminated using the *detectionP* function. Subsequently, unwanted variations were filtered out, resulting in the removal of 586,470 CpG sites. After this process, we retained 271,799 CpG sites located in the promoter region to develop the machine-learning model.

Construction of the two-step model for ovarian cancer stratification

After preprocessing the methylation data, an ovarian cancer prediction model was constructed in a two-step manner. The Step One model, which is a PCA-based model, stratifies normal and tumor (benign and malignant) samples. The Step Two machine-learning model further stratified benign and malignant samples from the tumor samples identified using the Step One model.

A PCA-based Step One model was constructed to stratify the tumor and normal samples. In the construction of the Step One model, we first identified DMPs between normal and tumor tissues using training data, resulting in a total of 470 DMPs, including 453 hypermethylated

DMPs and 17 hypomethylated DMPs (Supplementary Fig. 2a). We then used the Mann–Whitney U test to further select significant DMPs, leading to 219 significant DMPs. Subsequently, dimensionality reduction was performed on the DMPs using PCA. As illustrated in Supplementary Fig. 2b, PC2 effectively distinguished between normal and tumor samples, indicating its utility in the stratification of these groups. Subsequently, the weights associated with PC2 were extracted to construct the PCA-based Step One model. The absolute PC2 weights of 219 significant DMPs were ranked to determine the optimal number of features for model construction (Supplementary Table 2). A PCA model was then iteratively constructed, beginning with the top 2 features and progressively including up to 219 features. Figure 2c illustrates the AUPRC values for the models constructed with ranked feature counts ranging from 2 to 219. The number of features for the Step One model was selected by setting a cutoff at an AUPRC of 0.95, while minimizing the feature counts, leading to the selection of the top 30 features. Consequently, the Step One model was constructed using these 30 features to assess risk scores for ovarian cancer (Supplementary Table 3). The risk score for the Step One model was calculated using the following formula:

$$\text{Risk score} = \sum_{i=1}^{n} x_i \beta_i$$

where $n = 30$, β_i represents the beta-value of the ith CpG site and x_i denotes the PC2 weight at the ith CpG site. The risk score was calculated as the sum of the product of each beta-value and its corresponding PC2 weight.

The optimal threshold for the PCA risk score, which was determined by maximizing the F1-score in the training data, was 0.524 (Fig. 2d). The risk scores of the samples that exceeded this threshold were classified as tumors, whereas those below the threshold were classified as normal. The model's performance was evaluated using the training data, resulting in an AUPRC of 0.95 and F1-score of 0.93 (Fig. 2e), correctly predicting 113 samples (Fig. 2f), demonstrating its effectiveness in classification.

In the Step Two model, training data were utilized to identify DMPs between benign and malignant samples, resulting in 12,340 DMPs, including 4,346 hypermethylated and 7,994 hypomethylated DMPs (Supplementary Fig. 2c). Among these, the Mann–Whitney U test revealed 753 significant DMPs, underscoring notable differences between benign and malignant samples. However, PCA showed that the benign and malignant samples did not exhibit a clear separation in either PC1 or PC2 (Supplementary Fig. 2d). Consequently, machine-learning models were employed to stratify benign and malignant samples.

In the Step Two model, we developed five major types of machine-learning models, resulting in a total of 10 models utilizing different penalties and kernels. To construct each machine-learning model, we first performed feature selection using *RFECV* in three-fold cross-validation. After selecting the features, we used *GridSearchCV* to determine the optimal hyperparameters for the machine-learning model in a three-fold cross-validation. The performance of each model is presented in Supplementary Table 4. We chose the final Step Two model based on the mean F1-score and number of features. The gradient boosting model exhibited exceptional performance, utilizing only 16 features and achieving a mean F1-score of 0.83 (Supplementary Table 4), surpassing the performance of other machine-learning models. Consequently, we used the optimal hyperparameters obtained using *GridSearchCV* to construct the final gradient boosting model as the Step Two model. Samples with a predicted probability of >0.666 were classified as malignant, whereas those with a probability of <0.666 were classified as benign. The confusion matrix shows that 45 samples of the training data were predicted correctly (Fig. 2g), the AUPRC was 0.98, and F1-score was 0.92 (Fig. 2h). The feature importance of the 16 features utilized in the final Step Two model is shown in Supplementary Table 5.

Overall performance in training and testing data

After constructing the Step One PCA and Step Two gradient boosting models, we evaluated their overall performances using 119 training samples in the two-step approach. The multi-class confusion matrix in Fig. 3a indicates that 111 samples were correctly classified, achieving an overall precision of 0.95, recall of 0.88, and F1-score of 0.93 (Table 2).

Furthermore, the two-step model was applied to the testing data consisting of 24 normal, 9 benign, and 7 malignant samples. The testing data predictions made using Step One and Step Two models correctly classified 35 samples (Fig. 3b). Table 2 shows the performance indices for the testing data, with overall precision and overall recall of 0.81 and 0.77, respectively. The F1 scores for tumor, benign, and malignant samples were 0.98, 0.80, and 0.55, respectively, with an overall F1 score of 0.86, revealing that the classification performance for malignant samples in the two-step model was suboptimal (Table 2).

Overrepresentation analysis with DMPs in Step One and Step Two models

To investigate the function of the DMPs, we performed an ORA using GO and KEGG pathway gene sets. We separately examined the overall, hypermethylated, and hypomethylated DMPs, with a particular focus on BP

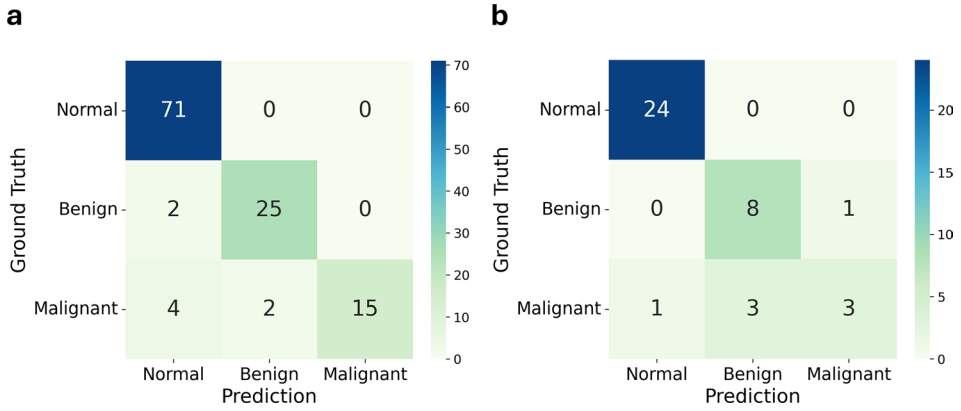

Fig. 3 The predicted results of the two-step model. (**a**) The multi-class confusion matrix of training data. (**b**) The multi-class confusion matrix of the testing data

Table 2 The performances in training data and testing data

Dataset	Normal	Benign	Malignant	Overall
Precision				
Training	0.92	0.93	1.0	[+]0.95
Testing	0.96	0.73	0.75	[+]0.81
Recall				
Training	1.0	0.93	0.71	[+]0.88
Testing	1.0	0.89	0.43	[+]0.77
F1-score				
Training	0.96	0.93	0.83	*0.93
Testing	0.98	0.80	0.55	*0.86
Accuracy				
Training	0.93			
Testing	0.88			

[+]Macro-precision and macro-recall, which was the mean of precision and recall

*The overall F1-score was calculated using *sklearn* weighted F1-score

terms in GO. The DMPs identified in the Step One model included 453 hypermethylated and 17 hypomethylated DMPs (Supplementary Fig. 2a). Enrichment analysis conducted on these DMPs revealed two significant GO terms in both the overall and hypermethylated results: "detection of chemical stimulus involved in sensory perception of smell" and "G protein-coupled receptor signaling pathway." GO analysis of hypomethylated DMPs did not yield any significant terms. In contrast, the KEGG analysis identified "olfactory transduction" as a significant pathway, which is consistent with the GO analysis results.

Subsequently, we investigated the biological functions of the DMPs in the Step Two model by performing ORA on the 12,340 DMPs identified, which included 4,346 hypermethylated and 7,994 hypomethylated DMPs when comparing malignant samples to benign samples (Supplementary Fig. 2c). For the overall and hypermethylation results, we selected the top ten significant GO terms, which are presented in Fig. 4a and b, respectively. Many of these terms are associated with immune regulation, including the "negative regulation of cytokine production," "B cell receptor signaling pathway," and "myeloid leukocyte differentiation." Conversely, hypomethylation analysis revealed two significant GO terms: "cell junction organization" and "cell adhesion." The top ten significant pathways in the KEGG results of the Step Two model for overall, hypermethylated, and hypomethylated DMPs were consistent with those from GO analysis (Fig. 4c, d, e). Many of these significant pathways were related to immune regulation, underscoring the involvement of immune-related pathways in distinguishing between benign and malignant samples in the Step Two model.

Discussion

In this study, we developed a two-step model using methylation data from cervical scrapings to stratify patients into normal, benign, or malignant categories. Cervical scraping samples were collected using a noninvasive method, the Pap test, which is commonly used in cervical cancer screening. The first step involved a PCA model consisting of 30 features trained on methylation data from normal and tumor samples. This model stratified normal and tumor samples, achieving an AUPRC of 0.95 and an F1-score of 0.93 in the training data. In the second step, a gradient boosting model with 16 features trained on benign and malignant data was employed to predict whether the samples were benign or malignant. This model achieved an AUPRC of 0.98 and an F1-score of 0.92. Consequently, we combined the two models to form our final two-step model to stratify patients into the categories. The result of the multi-class confusion matrix from the final model showed that 111 samples were correctly predicted in the training data, with an F1-score of 0.93, indicating a strong predictive ability. For the testing data, the model correctly predicted 35 samples with an F1-score of 0.86 (Fig. 3a, b, and Table 2). Overall, the two-step model utilized 46 CpG sites to effectively stratify the samples, demonstrating the potential of using methylation data from non-invasive cervical scrapings to accurately predict ovarian cancer.

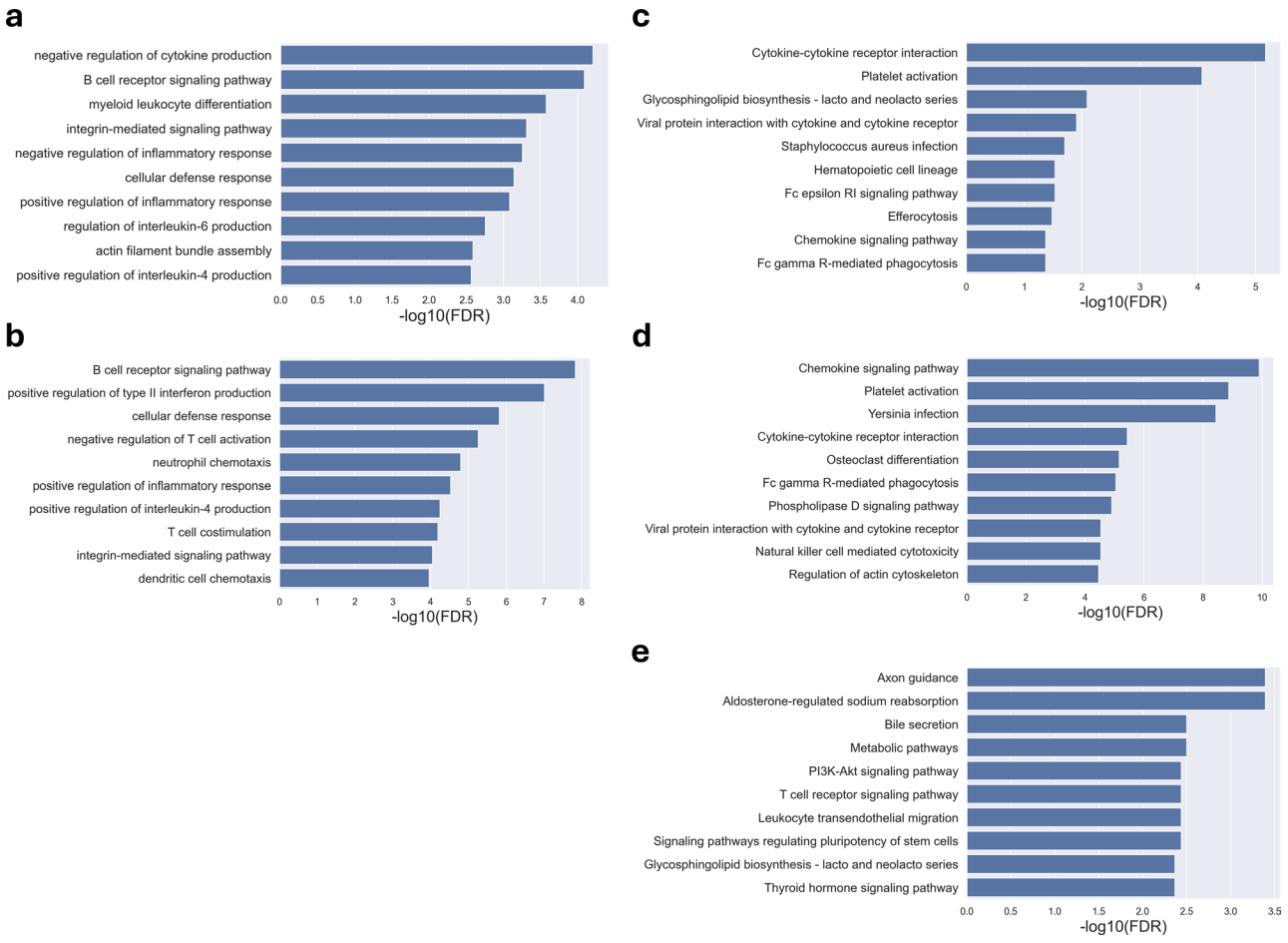

Fig. 4 The results of the overrepresentation analysis in the Step Two model. (**a**) Top ten significant GO terms in the overall DMPs. (**b**) Top ten significant GO terms in the hypermethylated DMPs. (**c**) Top ten significant KEGG pathways in the overall DMPs. (**d**) Top ten significant KEGG pathways in the hypermethylated DMPs. (**e**) Top ten significant KEGG pathways in the hypomethylated DMPs

The rationale for including benign samples, rather than only distinguishing between normal and tumor (benign + malignant) samples, stems from the clinical necessity of differentiating benign from malignant ovarian tumors for improved diagnostic accuracy and treatment planning. Ovarian tumors often present diagnostic challenges, particularly when using imaging modalities such as ultrasound, which, although highly sensitive in detecting ovarian lesions, has limited specificity. This can result in unnecessary surgical procedures for patients with benign conditions due to diagnostic uncertainty. By incorporating benign samples, our model aims to improve risk stratification, ensuring that patients receive appropriate clinical management while reducing overtreatment risks. Our two-step classification approach enhances precision in risk assessment. The Step One model (normal vs. tumor) flags tumor presence (including both benign and malignant cases), ensuring that potential tumor cases receive further evaluation. Subsequently, the Step Two model (benign vs. malignant) refines classification among tumor cases to guide clinical decision-making. Patients with benign tumors may be eligible for conservative management or minimally invasive surgery, reducing surgical morbidity and recovery time, whereas malignant cases require more aggressive interventions.

A similar work was conducted by Barrett et al. [32] They detected the methylation levels in cervical cells and developed a ridge regression model to calculate the ovarian risk score for predicting the occurrence of ovarian cancer. Their model utilized 869 normal and 242 tumor samples for training, consisting of 14,000 CpG sites, achieving an AUC of 0.78 [32]. However, the weights of many CpG sites in ridge regression were near zero, which may indicate noise in the constructed model. Comparatively, our model contains fewer features, and only four CpG sites (cg27143326, cg25736251, cg17536595, and cg15702277) overlapped with those of the Barrett et al. model. In this study, we use the F1-score as a performance index to evaluate the imbalanced data. Additionally, our strict criteria for feature selection helped identify significant CpG sites for model training and

minimize noise during model construction. Another key difference is the population focus of the studies. Barrett et al. primarily studied Europeans, whereas our samples were collected from Asian individuals. The differences in candidate CpG sites used in our model may be influenced by racial biases, underscoring the importance of further studies to identify and understand this possibility [42, 43]. Furthermore, our two-step model can be applied to different racial cohorts in future studies to assess its robustness and generalizability across diverse populations.

In the feature selection stage, we identified the DMPs for the training data in Step One and Step Two models separately. For the Step One model, which distinguished between normal and tumor samples, we identified 470 DMPs (Supplementary Fig. 2a). In the Step Two model, which differentiated between benign and malignant samples, 12,340 DMPs were found (Supplementary Fig. 2c). This significant difference in methylation levels between benign and malignant samples highlights their distinct methylation patterns.

In the overrepresentation analysis of DMPs in the Step One model, we discovered that the DMPs were enriched in the GO biological process in terms of "detection of chemical stimulus involved in sensory perception of smell" and "G protein-coupled receptor signaling pathway." These terms are involved in the conversion of olfactory chemical stimuli into molecular signals and signal transduction in cells. Genes annotated with these pathways showed significant overlap, mainly consisting of genes from the olfactory receptor family, a subset of G-protein-coupled receptors (GPCRs) [44]. Olfactory receptors are expressed in tumor cells and have been associated with tumor metastasis, differentiation, and prognosis [45–47].

The DMPs in the Step Two model were enriched in GO biological process terms related to immune regulation. Immune system regulation plays a crucial role in ovarian cancer development; previous studies have also shown that certain cytokines secreted by ovarian cancer cells are associated with prognosis, drug treatment response, and metastasis [48–50]. Interestingly, we found that the hypomethylated CpG sites in malignant samples were enriched in "cell junction organization" and "cell adhesion." Dysregulation of cell junctions and adhesion is linked to oncogenic transformation and metastasis [51, 52]. Cancer cells can modulate their adhesion strength with the extracellular matrix, promoting invasion and tumor growth [53]. ORA results for our KEGG analysis (Fig. 4c, d, e) further emphasized the pathways related to immune regulation across hypermethylated, hypomethylated, and all DMPs. For example, the results of the overall DMPs showed that alterations in glycosphingolipid biosynthesis may be involved in tumor development and metastasis [54] (Fig. 4c). Some studies have shown that abnormal platelet activation can promote cancer cell proliferation in ovarian cancer [55] (Fig. 4d). Interestingly, the KEGG analysis identified "bile secretion" as a significant pathway in the hypomethylated DMPs (Fig. 4e), which has been discovered as a potential biomarker in the early stages of ovarian cancer [56] and is associated with tumor proliferation, invasion, and epithelial-to-mesenchymal transition [57]. In contrast, when we performed an ORA using the 46 CpG sites identified in our study, neither GO nor KEGG analyses yielded significant terms or pathways.

Although our two-step model successfully stratified the three types of samples, one limitation of our study is the lack of available open-source cervical scraping methylation datasets for independent validation, given the specificity of our data. This gap highlights the need for further research and new datasets to assess the robustness and generalizability of our model. Additionally, while our model identified several CpG sites, focusing solely on individual CpG methylation levels may not fully capture transcriptional regulation. Analyzing differentially methylated regions could provide a more comprehensive understanding of gene expression changes. Furthermore, some normal samples in our study were obtained from infertile patients, which may have introduced additional confounding factors.

To strengthen the clinical applicability of our findings, experimental validation of key CpG sites is essential. Future studies should confirm the biological relevance of hypermethylated and hypomethylated CpG sites through methylation-specific PCR (MSP), bisulfite sequencing, or enzyme restriction assays in independent ovarian tumor samples. Such validation would provide crucial evidence supporting the reproducibility of these methylation markers across different patient cohorts. Additionally, functional studies examining how differential methylation affects gene expression will help determine whether these epigenetic alterations actively contribute to ovarian cancer progression or serve as passive biomarkers. Integrating experimental validation with machine learning-based prediction models will further enhance the reliability of methylation-based screening methods, ultimately advancing the development of a minimally invasive, cost-effective, and highly specific diagnostic tool for ovarian cancer.

While blood-based DNA methylation assays remain a major focus for ovarian cancer detection, researchers have explored alternative sampling methods, including cervicovaginal samples collected via routine Pap smears, uterine lavage, or endometrial brushings [58]. Cervicovaginal sampling is particularly promising, as ovarian or fallopian tube tumor cells, along with tumor-derived DNA fragments, might exfoliate into the cervical

region, enabling non-invasive detection. Previous studies have demonstrated the potential of detecting tumor-specific DNA methylation markers, such as *HOXA9* [59], *SOX17* [60], and *RASSF1A* [61] for ovarian cancer screening. Additionally, ascites fluid, which is commonly present in advanced ovarian cancer, contains abundant tumor-derived DNA, making it another valuable but stage-dependent source for methylation analysis [62]. By leveraging these diverse biological sample types alongside machine learning-based approaches, including the one presented in this study, there is significant potential to improve diagnostic accuracy, minimize unnecessary invasive procedures, and refine risk stratification for ovarian cancer detection.

Conclusions

In summary, our study indicates that cervical scrapings may offer an alternative source of samples for ovarian cancer screening. Our two-step model, comprising 46 features, demonstrated the capability to stratify normal, benign, and malignant samples, underscoring its potential as a biomarker for ovarian cancer. Furthermore, the 46 CpG sites require additional research, such as investigating the alterations in methylation levels through in vivo experiments and examining RNA expression downstream of transcription. Such studies provide a more comprehensive understanding of the mechanisms underlying these methylation biomarkers.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40246-025-00763-4.

Supplementary Material 1

Acknowledgements

We would like to thank the National Center for High-performance Computing (NCHC) of the National Applied Research Laboratories (NARLabs) of Taiwan for providing computational resources and storage resources. We also acknowledge the technical services provided by the National Genomics Center for Clinical and Biotechnological Applications of the Cancer and Immunology Research Center (National Yang Ming Chiao Tung University). The Core facility is supported by National Core Facility for Biopharmaceuticals (NCFB), National Science and Technology Council.

Author contributions

YCW and HCL conceived the study; PHS, LYC, KCW, and HCL curated the data; JYL, LAH, and YCW developed the methods; JYL, LAH, PHS, and YCW performed the analysis; JYL, LAH, PHS, LYC, KCW, HCL and YCW evaluated the results; JYL wrote the original manuscript; LAH, PHS, and YCW revised the manuscript; PHS, LYC, KCW, and HCL provided expert guidance; HCL and YCW supervised the project. All authors read and approved the final manuscript.

Funding

This study was funded by the National Science and Technology Council, Taiwan (NSTC 112-2221-E-A49-088-MY3 and NSTC 113-2221-E-A49-154-MY3 to YCW). The funder played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Data availability

The datasets used and analyzed during the current study are available from the corresponding author (Dr. Hung-Cheng Lai) on reasonable request.

Declarations

Ethics approval and consent to participate

Specimens were collected according to institutional policies, and participants with incomplete clinical/pathological results were excluded. The protocol and informed consent forms were approved by the Taipei Medical University IRB (N201810036). Written informed consent was obtained from all participants prior to their enrollment in the study.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Institute of Biomedical Informatics, National Yang Ming Chiao Tung University, Taipei, Taiwan
[2]College of Health Technology, National Taipei University of Nursing and Health Sciences, Taipei, Taiwan
[3]Department of Obstetrics and Gynecology, Shuang Ho Hospital, Taipei Medical University, New Taipei City, Taiwan
[4]Department of Obstetrics and Gynecology, School of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan
[5]Translational Epigenetics Center, Shuang Ho Hospital, Taipei Medical University, New Taipei City, Taiwan
[6]Department of Obstetrics and Gynecology, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan
[7]Digital Medicine and Smart Healthcare Research Center, National Yang Ming Chiao Tung University, Taipei, Taiwan

Received: 7 February 2025 / Accepted: 25 April 2025
Published online: 17 May 2025

References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, Jemal A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74:229–63.
2. Bankhead CR, Kehoe ST, Austoker J. Symptoms associated with diagnosis of ovarian cancer: a systematic review. BJOG. 2005;112:857–65.
3. Donovan KA, Donovan HS, Cella D, Gaines ME, Penson RT, Plaxe SC, von Gruenigen VE, Bruner DW, Reeve BB, Wenzel L. Recommended patient-reported core set of symptoms and quality-of-life domains to measure in ovarian cancer treatment trials. J Natl Cancer Inst 2014, 106.
4. Gutic B, Bozanovic T, Mandic A, Dugalic S, Todorovic J, Dugalic MG, Sengul D, Detanac DA, Sengul I, Detanac D, et al. Preliminary outcomes of five-year survival for ovarian malignancies in profiled Serbian oncology centre. Clin (Sao Paulo). 2023;78:100204.
5. National Cancer Institute. Surveillance, Epidemiology, and End Results Program. Cancer stat facts: ovarian cancer. [https://www.seer.cancer.gov/statfacts/html/ovary.html]
6. Kobel M, Kang EY. The evolution of ovarian carcinoma subclassification. Cancers (Basel) 2022, 14.
7. Stewart C, Ralyea C, Lockwood S. Ovarian cancer: an integrated review. Semin Oncol Nurs. 2019;35:151–6.
8. Zhang M, Cheng S, Jin Y, Zhao Y, Wang Y. Roles of CA125 in diagnosis, prediction, and oncogenesis of ovarian cancer. Biochim Biophys Acta Rev Cancer. 2021;1875:188503.
9. Hellstrom I, Yip YY, Darvas M, Swisher E, Hellstrom KE. Ovarian carcinomas express HE4 epitopes independently of each other. Cancer Treat Res Commun. 2019;21:100152.
10. Zhang R, Siu MKY, Ngan HYS, Chan KKL. Molecular biomarkers for the early detection of ovarian Cancer. Int J Mol Sci 2022, 23.

11. Nebgen DR, Lu KH, Bast RC Jr. Novel approaches to ovarian Cancer screening. Curr Oncol Rep. 2019;21:75.
12. Li J, Dowdy S, Tipton T, Podratz K, Lu WG, Xie X, Jiang SW. HE4 as a biomarker for ovarian and endometrial cancer management. Expert Rev Mol Diagn. 2009;9:555–66.
13. Chan KK, Chen CA, Nam JH, Ochiai K, Wilailak S, Choon AT, Sabaratnam S, Hebbar S, Sickan J, Schodin BA, Sumpaico WW. The use of HE4 in the prediction of ovarian cancer in Asian women with a pelvic mass. Gynecol Oncol. 2013;128:239–44.
14. Kim J, Beidler P, Wang H, Li C, Quassab A, Coles C, Drescher C, Carter D, Lieber A. Desmoglein-2 as a prognostic and biomarker in ovarian cancer. Cancer Biol Ther. 2020;21:1154–62.
15. Jacobs IJ, Menon U, Ryan A, Gentry-Maharaj A, Burnell M, Kalsi JK, Amso NN, Apostolidou S, Benjamin E, Cruickshank D, et al. Ovarian cancer screening and mortality in the UK collaborative trial of ovarian Cancer screening (UKCTOCS): a randomised controlled trial. Lancet. 2016;387:945–56.
16. Kinde I, Bettegowda C, Wang Y, Wu J, Agrawal N, Shih Ie M, Kurman R, Dao F, Levine DA, Giuntoli R, et al. Evaluation of DNA from the Papanicolaou test to detect ovarian and endometrial cancers. Sci Transl Med. 2013;5:167ra164.
17. Wang Y, Li L, Douville C, Cohen JD, Yen TT, Kinde I, Sundfelt K, Kjaer SK, Hruban RH, Shih IM et al. Evaluation of liquid from the Papanicolaou test and other liquid biopsies for the detection of endometrial and ovarian cancers. Sci Transl Med 2018, 10.
18. Gong G, Lin T, Yuan Y. Integrated analysis of gene expression and DNA methylation profiles in ovarian cancer. J Ovarian Res. 2020;13:30.
19. Szigeti KA, Galamb O, Kalmar A, Bartak BK, Nagy ZB, Markus E, Igaz P, Tulassay Z, Molnar B. [Role and alterations of DNA methylation during the aging and cancer]. Orv Hetil. 2018;159:3–15.
20. Moore LD, Le T, Fan G. DNA methylation and its basic function. Neuropsychopharmacology. 2013;38:23–38.
21. Belinsky SA, Nikula KJ, Palmisano WA, Michels R, Saccomanno G, Gabrielson E, Baylin SB, Herman JG. Aberrant methylation of p16(INK4a) is an early event in lung cancer and a potential biomarker for early diagnosis. Proc Natl Acad Sci U S A. 1998;95:11891–6.
22. Brock MV, Hooker CM, Ota-Machida E, Han Y, Guo M, Ames S, Glockner S, Piantadosi S, Gabrielson E, Pridham G, et al. DNA methylation markers and early recurrence in stage I lung cancer. N Engl J Med. 2008;358:1118–28.
23. Lofton-Day C, Model F, Devos T, Tetzner R, Distler J, Schuster M, Song X, Lesche R, Liebenberg V, Ebert M, et al. DNA methylation biomarkers for blood-based colorectal cancer screening. Clin Chem. 2008;54:414–23.
24. Muller D, Gyorffy B. DNA methylation-based diagnostic, prognostic, and predictive biomarkers in colorectal cancer. Biochim Biophys Acta Rev Cancer. 2022;1877:188722.
25. Imperiale TF, Ransohoff DF, Itzkowitz SH, Levin TR, Lavin P, Lidgard GP, Ahlquist DA, Berger BM. Multitarget stool DNA testing for colorectal-cancer screening. N Engl J Med. 2014;370:1287–97.
26. Wu Q, Lothe RA, Ahlquist T, Silins I, Trope CG, Micci F, Nesland JM, Suo Z, Lind GE. DNA methylation profiling of ovarian carcinomas and their in vitro models identifies HOXA9, HOXB5, SCGB3A1, and CRABP1 as novel targets. Mol Cancer. 2007;6:45.
27. Chmelarova M, Krepinska E, Spacek J, Laco J, Beranek M, Palicka V. Methylation in the p53 promoter in epithelial ovarian cancer. Clin Transl Oncol. 2013;15:160–3.
28. Fu M, Deng F, Chen J, Fu L, Lei J, Xu T, Chen Y, Zhou J, Gao Q, Ding H. Current data and future perspectives on DNA methylation in ovarian cancer (Review). Int J Oncol 2024, 64.
29. Jung Y, Hur S, Liu J, Lee S, Kang BS, Kim M, Choi YJ. Peripheral blood BRCA1 methylation profiling to predict Familial ovarian cancer. J Gynecol Oncol. 2021;32:e23.
30. Liggett TE, Melnikov A, Yi Q, Replogle C, Hu W, Rotmensch J, Kamat A, Sood AK, Levenson V. Distinctive DNA methylation patterns of cell-free plasma DNA in women with malignant ovarian tumors. Gynecol Oncol. 2011;120:113–20.
31. Zhao X, Yang M, Fan J, Wang M, Wang Y, Qin N, Zhu M, Jiang Y, Gorlova OY, Gorlov IP, et al. Identification of genetically predicted DNA methylation markers associated with non-small cell lung cancer risk among 34,964 cases and 448,579 controls. Cancer. 2024;130:913–26.
32. Barrett JE, Jones A, Evans I, Reisel D, Herzog C, Chindera K, Kristiansen M, Leavy OC, Manchanda R, Bjorge L, et al. The DNA methylome of cervical cells can predict the presence of ovarian cancer. Nat Commun. 2022;13:448.
33. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA. Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium DNA methylation microarrays. Bioinformatics. 2014;30:1363–9.
34. ChAMP. The Chip Analysis Methylation Pipeline. NBeads filter criteria. [https://bioconductor.org/packages/release/bioc/vignettes/ChAMP/inst/doc/ChAMP.html
35. Fortin JP, Triche TJ Jr., Hansen KD. Preprocessing, normalization and integration of the illumina humanmethylationepic array with Minfi. Bioinformatics. 2017;33:558–60.
36. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, Beck S. A beta-mixture quantile normalization method for correcting probe design bias in illumina infinium 450 K DNA methylation data. Bioinformatics. 2013;29:189–96.
37. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, Lin SM. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics. 2010;11:587.
38. McCartney DL, Walker RM, Morris SW, McIntosh AM, Porteous DJ, Evans KL. Identification of polymorphic and off-target probe binding sites on the illumina infinium methylationepic BeadChip. Genom Data. 2016;9:22–4.
39. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, Van Djik S, Muhlhausler B, Stirzaker C, Clark SJ. Critical evaluation of the illumina methylationepic BeadChip microarray for whole-genome DNA methylation profiling. Genome Biol. 2016;17:208.
40. Phipson B, Maksimovic J, Oshlack A. MissMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. Bioinformatics. 2016;32:286–8.
41. Carlson M. GO.db: A set of annotation maps describing the entire gene ontology. 3.8.2 edition; 2019.
42. Fraser HB, Lam LL, Neumann SM, Kobor MS. Population-specificity of human DNA methylation. Genome Biol. 2012;13:R8.
43. Heyn H, Moran S, Hernando-Herraez I, Sayols S, Gomez A, Sandoval J, Monk D, Hata K, Marques-Bonet T, Wang L, Esteller M. DNA methylation contributes to natural human variation. Genome Res. 2013;23:1363–72.
44. Crasto C, Singer MS, Shepherd GM. The olfactory receptor family album. Genome Biol. 2001;2:REVIEWS1027.
45. Kerslake R, Hall M, Vagnarelli P, Jeyaneethi J, Randeva HS, Pados G, Kyrou I, Karteris E. A Pancancer overview of FBN1, Asprosin and its cognate receptor OR4M1 with detailed expression profiling in ovarian cancer. Oncol Lett. 2021;22:650.
46. Chung C, Cho HJ, Lee C, Koo J. Odorant receptors in cancer. BMB Rep. 2022;55:72–80.
47. Weber L, Schulz WA, Philippou S, Eckardt J, Ubrig B, Hoffmann MJ, Tannapfel A, Kalbe B, Gisselmann G, Hatt H. Characterization of the olfactory receptor OR10H1 in human urinary bladder Cancer. Front Physiol. 2018;9:456.
48. Nowak M, Glowacka E, Szpakowski M, Szyllo K, Malinowski A, Kulig A, Tchorzewski H, Wilczynski J. Proinflammatory and immunosuppressive serum, Ascites and cyst fluid cytokines in patients with early and advanced ovarian cancer and benign ovarian tumors. Neuro Endocrinol Lett. 2010;31:375–83.
49. Szlosarek PW, Grimshaw MJ, Kulbe H, Wilson JL, Wilbanks GD, Burke F, Balkwill FR. Expression and regulation of tumor necrosis factor alpha in normal and malignant ovarian epithelium. Mol Cancer Ther. 2006;5:382–90.
50. Maccio A, Madeddu C. Inflammation and ovarian cancer. Cytokine. 2012;58:133–47.
51. Knights AJ, Funnell AP, Crossley M, Pearson RC. Holding tight: cell junctions and Cancer spread. Trends Cancer Res. 2012;8:61–9.
52. Yayan J, Franke KJ, Berger M, Windisch W, Rasche K. Adhesion, metastasis, and Inhibition of cancer cells: a comprehensive review. Mol Biol Rep. 2024;51:165.
53. Neville MC, Webb PG, Baumgartner HK, Bitler BG. Claudin-4 localization in epithelial ovarian cancer. Heliyon. 2022;8:e10862.
54. Cumin C, Huang YL, Everest-Dass A, Jacob F. Deciphering the importance of glycosphingolipids on cellular and molecular mechanisms associated with Epithelial-to-Mesenchymal transition in Cancer. Biomolecules 2021, 11.
55. Cho MS, Bottsford-Miller J, Vasquez HG, Stone R, Zand B, Kroll MH, Sood AK, Afshar-Kharghan V. Platelets increase the proliferation of ovarian cancer cells. Blood. 2012;120:4869–72.
56. Fan L, Yin M, Ke C, Ge T, Zhang G, Zhang W, Zhou X, Lou G, Li K. Use of plasma metabolomics to identify diagnostic biomarkers for early stage epithelial ovarian Cancer. J Cancer. 2016;7:1265–72.
57. Rezen T, Rozman D, Kovacs T, Kovacs P, Sipos A, Bai P, Miko E. The role of bile acids in carcinogenesis. Cell Mol Life Sci. 2022;79:243.
58. Hentze JL, Hogdall CK, Hogdall EV. Methylation and ovarian cancer: can DNA methylation be of diagnostic use? Mol Clin Oncol. 2019;10:323–30.

59. Taliento C, Morciano G, Nero C, Froyman W, Vizzielli G, Pavone M, Salvioli S, Tormen M, Fiorica F, Scutiero G, et al. Circulating tumor DNA as a biomarker for predicting progression-free survival and overall survival in patients with epithelial ovarian cancer: a systematic review and meta-analysis. Int J Gynecol Cancer. 2024;34:906–18.
60. Shaker N, Chen W, Sinclair W, Parwani AV, Li Z. Identifying SOX17 as a sensitive and specific marker for ovarian and endometrial carcinomas. Mod Pathol. 2023;36:100038.
61. Terp SK, Stoico MP, Dybkaer K, Pedersen IS. Early diagnosis of ovarian cancer based on methylation profiles in peripheral blood cell-free DNA: a systematic review. Clin Epigenetics. 2023;15:24.
62. Werner B, Yuwono N, Duggan J, Liu D, David C, Srirangan S, Provan P, Investigators IN, DeFazio A, Arora V, et al. Cell-free DNA is abundant in Ascites and represents a liquid biopsy of ovarian cancer. Gynecol Oncol. 2021;162:720–7.

Publisher's note