World Journal of Clinical Cases

Submit a Manuscript: https://www.f6publishing.com

World J Clin Cases 2025 July 16; 13(20): 104556

DOI: 10.12998/wjcc.v13.i20.104556

ISSN 2307-8960 (online)

ORIGINAL ARTICLE

Case Control Study

Prediction of genomic biomarkers for endometriosis using the transcriptomic dataset

Zeynep Kucukakcali, Sami Akbulut, Cemil Colak

Specialty type: Medicine, research and experimental

Provenance and peer review: Invited article; Externally peer reviewed.

Peer-review model: Single blind

Peer-review report's classification

Scientific Quality: Grade C, Grade C

Novelty: Grade B, Grade B

Creativity or Innovation: Grade B, Grade C

Scientific Significance: Grade C, Grade C

P-Reviewer: Huang Y; Li X

Received: December 24, 2024

Revised: March 3, 2025

Accepted: March 13, 2025

Published online: July 16, 2025

Processing time: 106 Days and 7.6 Hours

Zeynep Kucukakcali, Sami Akbulut, Cemil Colak, Department of Biostatistics and Medical Informatics, Inonu University Faculty of Medicine, Malatya 44280, Türkiye

Sami Akbulut, Surgery and Liver Transplant Institute, Inonu University Faculty of Medicine, Malatya 44280, Türkiye

Corresponding author: Sami Akbulut, MD, PhD, Professor, Surgery and Liver Transplant Institute, Inonu University Faculty of Medicine, Elazig Yolu 10. Km, Malatya 44280, Türkiye. akbulutsami@gmail.com

Abstract

BACKGROUND

Endometriosis is a clinical condition characterized by the presence of endometrial glands outside the uterine cavity. While its incidence remains mostly uncertain, endometriosis impacts around 180 million women worldwide. Despite the presentation of several epidemiological and clinical explanations, the precise mechanism underlying the disease remains ambiguous. In recent years, researchers have examined the hereditary dimension of the disease. Genetic research has aimed to discover the gene or genes responsible for the disease through association or linkage studies involving candidate genes or DNA mapping techniques.

AIM

To identify genetic biomarkers linked to endometriosis by the application of machine learning (ML) approaches.

METHODS

This case-control study accounted for the open-access transcriptomic data set of endometriosis and the control group. We included data from 22 controls and 16 endometriosis patients for this purpose. We used AdaBoost, XGBoost, Stochasting Gradient Boosting, Bagged Classification and Regression Trees (CART) for classification using five-fold cross validation. We evaluated the performance of the models using the performance measures of accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value and F1 score.

RESULTS

Bagged CART gave the best classification metrics. The metrics obtained from this model are 85.7%, 85.7%, 100%, 75%, 75%, 100% and 85.7% for accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive va-

lue and F1 score, respectively. Based on the variable importance of modeling, we can use the genes *CUX2*, *CLMP*, *CEP131*, *EHD4*, *CDH24*, *ILRUN*, *LINC01709*, *HOTAIR*, *SLC30A2* and *NKG7* and other transcripts with inaccessible gene names as potential biomarkers for endometriosis.

CONCLUSION

This study determined possible genomic biomarkers for endometriosis using transcriptomic data from patients with/without endometriosis. The applied ML model successfully classified endometriosis and created a highly accurate diagnostic prediction model. Future genomic studies could explain the underlying pathology of endometriosis, and a non-invasive diagnostic method could replace the invasive ones.

Key Words: Endometriosis; RNA-seq; Transcriptomics; Machine learning; Classification

©The Author(s) 2025. Published by Baishideng Publishing Group Inc. All rights reserved.

Core Tip: Genetic research has aimed to discover the gene or genes responsible for the disease through association or linkage studies involving candidate genes or DNA mapping techniques. This study aimed to determine genomic biomarkers associated with endometriosis by using machine learning models (AdaBoost, XGBoost, Stochasting Gradient Boosting, Bagged Classification and Regression Trees). According to the variables' importance in the modeling, *CUX2*, *CLMP*, *CEP131*, *EHD4*, *CDH24*, *ILRUN*, *LINC01709*, *HOTAIR*, *SLC30A2*, and *NKG7* genes and transcripts whose other gene names are inaccessible can be used as candidate biomarkers for endometriosis.

Citation: Kucukakcali Z, Akbulut S, Colak C. Prediction of genomic biomarkers for endometriosis using the transcriptomic dataset. *World J Clin Cases* 2025; 13(20): 104556
URL: https://www.wjgnet.com/2307-8960/full/v13/i20/104556.htm
DOI: https://dx.doi.org/10.12998/wjcc.v13.i20.104556

INTRODUCTION

Endometriosis is a disease defined by developing endometrial glands outside the uterine cavity[1]. The symptoms include severe dysmenorrhea, pelvic discomfort, and decreased fertility[2]. Though its prevalence is still primarily unclear, endometriosis affects more than 180 million women globally[3]. Diagnosing endometriosis is difficult and frequently takes years. The gold standard for the diagnosis of endometriosis continues to be a visual examination of the pelvis by laparoscopy and biopsy[4]. A recent meta-analysis calculated the general prevalence of endometriosis to be 18%, while infertile patients, patients with chronic pelvic pain, and asymptomatic patients accounted for 31%, 42%, and 23%, respectively[5].

Symptoms related to the disease seriously affect women's daily life activities and quality of life. Additionally, the disease significantly impairs the mental quality of life in patients. Endometriosis imposes a significant financial burden due to the high healthcare costs associated with hospitalization, outpatient visits, and medications. We should design endometriosis management very well to enhance the patient's quality of life, mitigate the disease's negative effects, and lower treatment costs. Therefore, early intervention is essential for lowering disease-related suffering and costs[6,7].

Despite numerous epidemiological and clinical hypotheses, the exact mechanism underlying endometriosis, a complex and common gynecological disorder, remains unclear[6,8]. Therefore, recent studies have focused on the genetic aspect of the disease[9]. Previous studies for endometriosis have shown that there is a familial predisposition to the development of the disease and may have a genetic basis[10,11]. The majority of researchers believe that inheritance occurs in a polygenic/multifactorial manner. Polygenic/multifactorial inheritance occurs when a mix of numerous genes and environmental factors determine the phenotype[11]. The risk for an individual with a sibling with endometriosis was 15 times higher than for the general population[12].

Through association or linkage research with candidate genes or DNA mapping techniques, endometriosis-related genetic research has sought to identify the gene or genes responsible for the disease. In addition, several genomic studies have shown significant changes in gene expression in endometriosis[9,11].

In recent years, machine learning (ML) algorithms, widely used in diagnosis and clinical decision support systems, have pioneered the development of meaningful biological models from microarray expression data and next-generation sequencing (NGS) data[13-16]. Researchers have successfully employed ML technologies to classify data from demographic, clinical, and omics technologies for use as biomarkers in endometriosis disease[17,18]. Researchers have investigated various biomarkers for the early detection or prediction of endometriosis, a disease where environmental and genetic factors are believed to play a role in its pathophysiology. To this end, numerous studies have focused on genomics, epigenomics, transcriptomics, proteomics, metabolomics, lipidomics, secretomics, and microbiomic[19-21]. The use of transcriptomic data to correlate disease has dramatically increased in recent years, creating an opportunity to use this data in clinical diagnosis. The application of ML classifiers on transcriptomic data has achieved various successes[6,18,22,23].

The challenge of diagnosing endometriosis lies in its heterogeneous presentation and the absence of specific biomarkers, which has spurred extensive research into potential diagnostic tools, including the exploration of biomarkers and ML applications.

Recent studies have highlighted the potential of circulating microRNAs (miRNAs) as non-invasive biomarkers for endometriosis. Moga *et al*[24] emphasize that miRNAs play a crucial role in the pathogenesis of endometriosis by regulating various biological processes such as cell survival, matrix remodeling, and angiogenesis, making them promising candidates for early diagnosis. Furthermore, Chen *et al*[25] demonstrated that combining circulating serum miRNAs with traditional biomarkers like CA125 can enhance diagnostic sensitivity and specificity, indicating that a multi-biomarker approach may be more effective than single biomarkers alone. This aligns with findings from Sarria-Santamera *et al*[26], who noted the variability in endometriosis presentation and the limitations of existing diagnostic methods, underscoring the need for reliable biomarkers. In parallel, the integration of ML into endometriosis research has shown promise in improving diagnostic accuracy and efficiency. Sivajohan *et al*[27]conducted a scoping review that outlined how AI and ML techniques could enhance research efficacy by identifying potential biomarkers and improving the understanding of endometriosis pathophysiology. Additionally, studies by Blass *et al*[28] and Mbuguiro *et al*[29] have proposed predictive models that leverage clinical, self-reported, and genetic data to better understand the etiology of endometriosis and its associated symptoms. Moreover, ML algorithms have been employed to create screening tools that utilize patient-reported symptoms and clinical parameters. For instance, a study by Konrad *et al*[30] developed a predictive model based on clinical parameters that demonstrated high sensitivity and specificity for diagnosing endometriosis.

Differences in gene expressions in transcriptomic data obtained from studies may provide new avenues for developing endometriosis diagnostic methods. This could contribute to the development of more efficient and early diagnostic methods, as well as targeted therapies. To make a good diagnostic prediction model, the control and endometriosis samples in this study will be put into groups using supervised ML methods [AdaBoost, XGBoost, Stochasting Gradient Boosting, Bagged Classification and Regression Trees (CART)] trained on transcriptomic data. Thus, we will use bioinformatic analyses to obtain and associate biomarker candidates with the disease.

MATERIALS AND METHODS

Data collection and variables

The current study's subjects, ranging in age from 18 to 49 years, underwent diagnostic laparoscopy due to pain or infertility. Exclusion criteria in the study were patients who had visual observation of lesions and diagnostic laparoscopy without visual observation of endometriotic lesions. Under general anesthesia, we used suction tubing to collect endometrial biopsies yielding 250 mg of tissue before laparoscopy. Endometrial biopsy is a 5-minute, minimally invasive procedure with minimal risk of infection, uterine perforation, or bleeding. The professionals thoroughly investigated the peritoneal cavity during the laparoscopic surgery and visually validated the existence or absence of endometriosis. Pathology evaluated at least one endometriotic lesion for histological confirmation of endometriosis. We processed tissue samples using Illumina next Seq NGS technology to generate high-throughput mRNA (RNA-Seq) data. The current study preprocessed data using several widely accepted bioinformatics tools. The transcriptomics dataset underwent five phases of processing. We first verified all raw data for quality control using FastQC. The Illumina next Seq NGS technology generated high-throughput mRNA (RNA-Seq) data and enrichment-based DNA methylation (MBD-seq) data from the tissue samples. The transcriptomics dataset consisted of 38 single-end RNA-seq samples (22 controls and 16 endometriosis)[6]. In the second stage, Cutadapt eliminated reads containing poor-quality bases, adapter sequences, and other contaminant sequences. In step three, we used Bowtie2 to match the sequencing reads to the reference genome hg38. In the fourth step of RNA-seq, we used TopHat to find the positions of short sequence reads relative to the reference. We used HTSeq in step five of RNA-seq to generate the read count data, which we then filtered to exclude very low-count genes. The filtering criterion was to keep the genes that have at least 1 count per million reads mapped in at least n samples, where n is the smallest group size[6].

Feature selection for genomics data

Feature selection is a crucial phase in predictive modeling systems, and one of the key matters in constructing a statistical model is determining which data to incorporate in the modeling. Determining the most valuable features of the dataset to use in the study, before working with vast datasets and models with high computing costs, will lead to high efficiency in terms of results. Feature selection determines the most significant factors influencing the dependent variable in data[31]. Feature selection becomes crucial when analyzing high-dimensional datasets, including metabolomic, genomic, epigenomic, or proteomic datasets. The curse of dimensionality, which arises from a large number of input dimensions, complicates many analysis methods with such high-dimensional datasets. For instance, as the number of data characteristics available to an ML classifier increases, the likelihood of a feature arbitrarily splitting training samples into positive and negative classes increases. This leads to robust performance on training data, but results in poor generalization to non-trained data due to overfitting. Therefore, feature selection methods aim to overcome such problems by shifting data from higher dimensions to lower dimensions. Several regularization methods, such as the least absolute shrinkage and selection operator (LASSO), Ridge, and Elastic Net, are available for variable selection. LASSO regularization leads many regression coefficients to precisely zero, enabling automatic variable selection that selects only one predictor from the associated predictors. The current study employs the Elastic Net regularization method, which simultaneously employs ridge and LASSO penalties, to leverage the benefits of both regulation techniques. The Elastic Net method is a powerful

variable selection method often used in genetic data, as it is a combination of lasso and ridge regression methods. It has the advantage of being able to handle extreme multicollinearity in high-dimensional data sets such as genome-wide association studies. LASSO is aggressive in variable selection by reducing only certain variables to zero, while ridge preserves the influence of genes by shrinking the coefficients. Elastic Net combines the advantages of both methods to create a more stable model[32-34].

CART method

CART is a decision tree method used for classification, forming the basis for systems like Random Forest. It splits data into binary parts and uses the Gini index to identify important variables. We prioritize variables with lower Gini indices [35]. However, small changes in the data can result in different outcomes, leading to instability. To overcome this, ensemble methods like bagging (bootstrap aggregation) combine multiple classifiers, using a majority vote to improve accuracy[36]. Bagging is a popular ensemble method that boosts decision tree accuracy. The models determine the final class by majority voting, with each classifier training on a subset of the data[37,38]. The hypermarameters used in Torbalı CART are as follows: Nbagg = 100 (Number of decision trees created for bagging); coob = TRUE (Calculates the Out-of-Bag error); control = rpart.control (...) (Control settings for decision trees); minsplit = 5 (Minimum number of observations required to split a node); cp = 0.01 (Complexity parameter for pruning); maxdepth = 10 (Maximum depth of the decision tree); xval = 10 (Number of cross-validation folds). These settings help prevent overfitting and optimize model performance.

Adaptive Boosting

Adaptive Boosting (AdaBoost) is an ensemble technique that integrates several weak classifiers to form a robust classifier. It operates by sequentially implementing weak classifiers on the training data, concentrating on cases that were misclassified in prior iterations. This iterative procedure modifies the weights of the training samples, assigning greater significance to those that are challenging to categorize. AdaBoost has demonstrated considerable efficacy across several applications, attaining high accuracy in tasks including face detection and the classification of imbalanced datasets[39,40].

Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is a sophisticated implementation of gradient boosting that enhances both speed and performance. It utilizes a gradient descent approach to reduce the loss function, rendering it very efficient for extensive datasets. XGBoost employs regularization approaches to mitigate overfitting, hence improving its generalization abilities relative to conventional boosting algorithms[41].

Stochastic Gradient Boosting

Stochastic Gradient Boosting is a form of gradient boosting that incorporates randomness into the model training procedure. Randomly selecting a data subset for each iteration mitigates the danger of overfitting and enhances the model's robustness. This stochastic method facilitates expedited training durations and may enhance generalization on unfamiliar material[41].

Bioinformatics analysis to identify differentially expressed genes and enrichment analysis

To find differentially expressed genes (DEGs) in the transcriptome dataset collected from control and endometriosis cases, the Linear Models for Microarray Analysis (limma) package in the programming language R was used for differential expression analyses. We can apply differential expression analysis to normalized read count data by statistically detecting quantitative variations in expression profile levels among experimental groups. For instance, we utilize statistical testing to assess the statistical significance of an observed variation in reading counts for a particular gene, determining if it surpasses the random prediction. Therefore, differential expression analysis, which derives from the term differential expression, seeks to confirm the distinct levels of gene expression under various conditions. These genes can provide biological information about the processes impacted by the condition(s) of interest.

We built a pipeline for the relevant analysis using the R software environment. We used the limma package to write the necessary codes for the analysis, aiming to identify DEGs in the 2-group data. We also used the ggplot2, ggrepel, and DT packages in the code writing process. Limma is a bioconductor package for the analysis of microarray gene expression data, particularly the use of linear models to analyze designed experiments and evaluate differential expression[42]. The system displays the obtained results as a table of genes ranked in order of relevance and a graph that illustrates genes with differential expression. A volcano plot plots significance *vs* fold change on a log2 basis to quickly observe DEGs on the y- and x-axes. The results include corrected P and log2-fold change (log2FC) values, with genes with the lowest P values being the most reliable. Upregulated genes were identified using log2FC > 1, and downregulated genes were determined using log2FC < -1[43]. We generated volcano plots for relevant genes with differential expression (Figure 1).

For enrichment analysis, the clusterProfiler package was used for enrichment analysis on gene ontology terms, and the org.Hs.eg.db database was used to access human gene annotations. In the visualization phase, the enrichplot and DOSE packages were used to display the enrichment results in various graphs (dotplot, emapplot, cnetplot, ridgeplot and gseaplot) and the ggplot2 package was used for additional visualization functions. Gene cluster enrichment analysis was performed by ranking the genes according to their log-fold change values and the *P*-value cut-off point was set at 0.05. Figure 2 and Figure 3 show the results of the enrichment analysis.

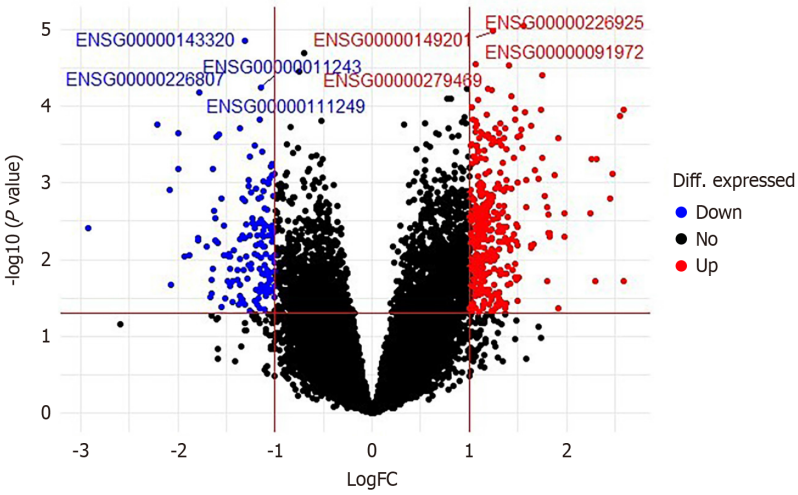


Figure 1 A volcano plot graph.

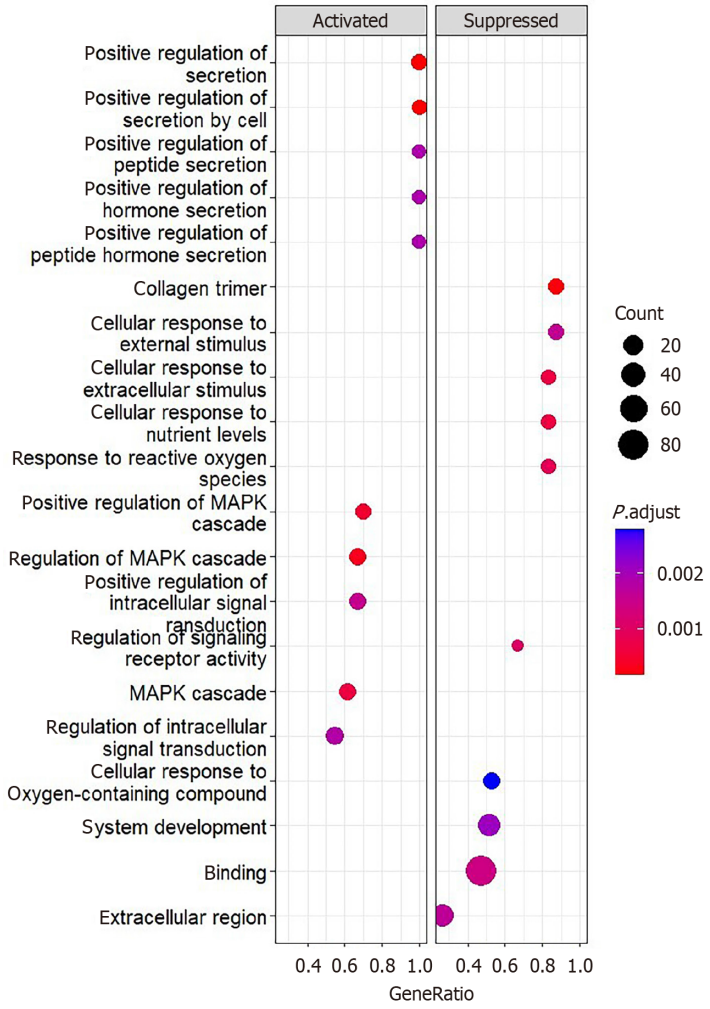


Figure 2 Gene ontology enrichment analysis for control and endometriosis groups.

ML modeling and performance evaluation

The current study used Bagged CART to model the dataset in question. We divided the data set 80:20 into training and test datasets. We conducted analyses using the n-fold cross-validation technique. The n-fold cross-validation approach divides the data set into n parts and then implements the model in each part. We use one of the n components for testing and the remaining n-minus-one components for training the model. We performed 5-fold cross-validation for the modeling procedure in this work. As performance assessment criteria, we employed accuracy, balanced accuracy,

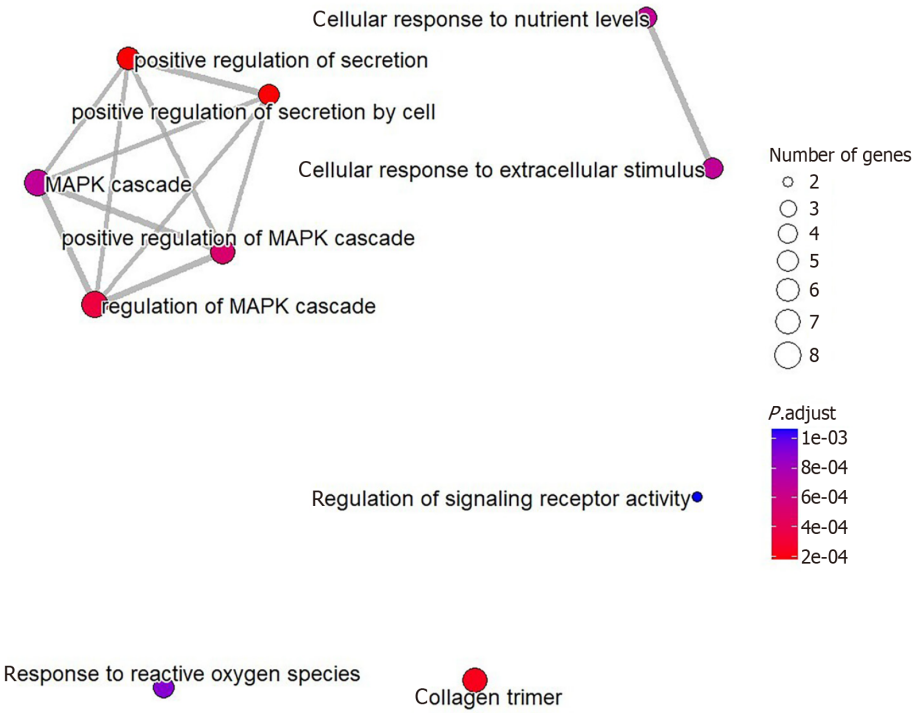


Figure 3 Network map of semantic relationships between gene ontology terms.

sensitivity, specificity, positive predictive value, negative predictive value, and F1 score. We also determined the variable importance, providing insight into the degree to which the factors prioritize the outcome variable. We used R Studio 4.2.1 for modeling.

Study protocol and ethics committee approval

This study, which included human participants and utilized the National Center for Biotechnology Information Gene Expression Omnibus open-access dataset, complied with the institutional and national research committee's ethical standards, as well as the 1964 Helsinki Declaration and its associated regulations or other comparable ethical standards. The Inonu University Institutional Review Board granted ethical permission for non-interventional clinical research (2022/3842). To assess the risk of bias and the general quality of this study, the STROBE standard was used[44].

Biostatistical analysis

We determined whether or not the variables had a normal distribution using the Shapiro-Wilk test of normality. We used the median (minimum-maximum) to summarize the quantitative data. We employed the Mann-Whitney *U* test on the data. We used logistic regression analysis (LRA) to calculate the odds ratio (OR) for each gene, which serves as a measure of effect size. We calculated the Hosmer and Lemeshow's goodness of fit test for logistic regression modeling and the omnibus test of model coefficients. A *P*-value < 0.05 was considered significant. IBM SPSS Statistics 25.0 program was employed in the analysis. A post-hoc power analysis revealed 0.927 power (1-β) considering type I error (α) of 0.05, parent distribution of Logistic, estimated effect size of 1.1, the sample sizes of 22 and 16 based on the Wilcoxon-Mann-Whitney test (two groups).

RESULTS

The transcriptomics dataset includes 38 (22 controls and 16 endometriosis) single-end RNA-seq samples. The dataset used contains 58050 expressions. We analyzed bioinformatics to detect DEGs in the transcriptomic data and summarized the top ten results for minimum corrected p-values in Table 1. The numbers in Table 1 show that three genes (*CRABP2*, *AKAP8 L*, and *MROH5*) went down, four genes (*IL1R1-AS1*, *CCDC81*, *CD200*, and *ESYT3*) went up, and three genes (*ABCG5*, *KBTBD2*, and *NIPBL*) did not change. They were 1.5643, 1.2448, 1.4053, -1.3148, -1.1421, 0.9750, -1.7847, 1.4350, 0.8138, and 0.7743 for the *IL1R1-AS*, *CCDC81*, *CD200*, *CRABP2*, *AKAP8 L*, *ABCG5*, *MROH5*, *ESYT3*, *KBTBD2*, and *NIPBL* genes, in that order. The results of the enrichment analysis are summarized in Figures 2 and 3.

Using the Elastic Net technique, we selected 21 genes from the dataset of 58050 expressions and created a new one. Table 2 shows descriptive statistics for the selected genes in endometriosis and normal groups. Table 2 also presents the OR obtained using the selected expressions and the examined target variable. One hundred eighteen gene expressions were statistically different between the groups ($P < 0.005$). The only genes that were not statistically different were *ENSG00000085741* ($P = 0.093$), *ENSG00000158014* ($P = 0.637$), and *ENSG000000183729* ($P = 0.166$). In addition, the

Table 1 The results of the bioinformatics analysis

ID	Gene symbol	Adjust *P* value	*P* value	*t*	Log2FC	Diff. expressed
ENSG00000226925	IL1R1-AS1	0.08254818	0.0000092762	5.1238	1.5643	Up
ENSG00000149201	CCDC81	0.08254818	0.0000105433	5.2962	1.2448	Up
ENSG00000091972	CD200	0.08254818	0.0000300938	4.7445	1.4053	Up
ENSG00000143320	CRABP2	0.08254818	0.0000141606	-5.8310	-1.3148	Down
ENSG00000011243	AKAP8 L	0.08254818	0.0000572588	-4.5349	-1.1421	Down
ENSG00000138075	ABCG5	0.08254818	0.0000608491	4.5150	0.9750	No
ENSG00000226807	MROH5	0.08254818	0.0000663279	-4.5179	-1.7847	Down
ENSG00000158220	ESYT3	0.08254818	0.0000758312	4.4428	1.4350	Up
ENSG00000170852	KBTBD2	0.08254818	0.0000796663	4.4265	0.8138	No
ENSG00000164190	NIPBL	0.08254818	0.0000805697	4.4228	0.7743	No

FC: Fold change.

Table 2 Descriptive statistics for input variables[1]

Gene ID	Gene symbol	Median		OR	*P*
		Normal (*n* = 22)	Endometriosis (*n* = 16)		
ENSG00000085741	WNT11	1 (0-6)	0 (0-2)	-	0.093
ENSG00000103966	EHD4	10 (2-39)	4.5 (0-28)	0.828	< 0.0001
ENSG00000105374	NKG7	4 (0-15)	1 (0-7)	0.619	0.003
ENSG00000111249	CUX2	2 (0-6)	0 (0-1)	0.151	< 0.0001
ENSG00000139880	CDH24	1 (0-5)	0 (0-2)	0.145	< 0.0001
ENSG00000141577	CEP131	5 (1-14)	2 (0-9)	0.545	< 0.0001
ENSG00000158014	SLC30A2	1 (0-60)	1 (0-8)	-	0.637
ENSG00000166250	CLMP	7 (1-11)	12 (8-38)	2.617	< 0.0001
ENSG00000180385	EMC3-AS1	20 (10-40)	14 (2-20)	0.751	< 0.0001
ENSG00000183729	No found	1 (0-39)	0 (0-39)	-	0.166
ENSG00000196821	ILRUN	4 (1-8)	2 (0-7)	0.682	0.021
ENSG00000201207	Y_RNA	11.5 (3-25)	6 (1-12)	0.802	0.012
ENSG00000211935	IGHV1-3	1 (0-12)	0 (0-3)	-	0.022
ENSG00000226715	LINC01709	19 (0-53)	36 (3-84)	1.050	0.014
ENSG00000228630	HOTAIR	3 (0-12)	1 (0-5)	0.603	0.009
ENSG00000229245	No found	1 (0-35)	23.5 (0-59)	1.100	< 0.0001
ENSG00000232153	Novel transcript	31 (2-631)	323 (36-926)	1.007	0.000
ENSG00000256249	Novel transcript	8.5 (1-1653)	3 (0-22)	-	0.011
ENSG00000264630	PRKCA-AS1	0 (0-2)	1.5 (0-19)	2.812	0.003
ENSG00000270591	PPP2R5C	1 (0-3)	3 (0-8)	2.736	< 0.0001
ENSG00000279271	TEC	1 (0-9)	10 (0-45)	1.486	< 0.0001

[1]Continuous variables are given as median (min-max).

univariate LRA results were not statistically significant for five genes. The OR ratios obtained for other genes were significant.

Table 3 presents the results of the performance metrics from the ML models (Bagged CART, AdaBoost, XGBoost, Stochasting Gradient Boosting). In the testing phase, the Bagged CART model gave the best results for Accuracy, Balanced accuracy, Sensitivity, Specificity, Positive predictive value, Negative predictive value and F1 score with values of 85.7%, 85.7%, 100%, 75%, 75%, 100% and 85.7% respectively. Figure 4 depicts the values of performance criteria calculated from the Bagged CART. Figure 5 illustrates the variable importance of selected genes, which serve as input variables to explain the output variable. The *ENSG00000279271* (*TEC*) gene had the highest predictor importance of 100%, followed by *ENSG00000111249* (*CUX2*; 78.491%), *ENSG00000166250* (*CLMP*; 59.954%), *ENSG00000232153* (novel transcript; 51.852%), and *ENSG00000229245* (32.655%).

DISCUSSION

Endometriosis is an important gynecological disease that affects the quality of life of women, such as chronic pelvic pain and infertility. Despite numerous clinical and genetic studies suggesting the involvement of hormonal, neurological, and immunological factors in the underlying causes, the etiology of endometriosis remains unclear[11,45]. It is known that uncertainties in pathophysiology lead to delays of 4 to 11 years in the diagnosis of endometriosis[6,46]. Early intervention in endometriosis is critical to reducing the symptoms and costs of the disease. A novel, less invasive diagnostic technique, such as endometrial biopsy, might be advantageous in reducing diagnostic delay[6]. Numerous familial and epidemiological research studies support the notion that this disease is a polygenic/multifactorial genetic condition[11]. Determining the number and location of disease-causing genes is a challenging aspect of disease genomic analysis. However, recent advances in molecular technology now make it possible to identify and explain these genes[47]. With the advent of genomic technologies, the acquisition of transcriptomic data has dramatically increased in recent years, opening the way to correlated diseases and creating the opportunity to use this data in clinical diagnosis[48-51]. Endometriosis patients have different transcriptome (RNA-seq) levels. These changes in gene expression may help find biomarkers for the creation of a minimally invasive method for diagnosing endometriosis[6,52,53].

In the dataset of the present study, Log2FC values used to assess the expression fold differences between the both groups, the *CRABP2* gene was expressed 2.48-fold less in endometriosis group than in the control ones. Similarly, the *AKAP8 L* gene had 2.20-fold lower gene expression; the *MROH5* gene had 3.43-fold lower gene expression. The *IL1R1-AS1* gene had 2.95-fold, the *CCDC81* gene 2.36-fold, the *CD200* gene 2.63-fold, and the *ESYT3* gene 2.69-fold upper gene expression in endometriosis group than the control ones. Finally, the *ABCG5* gene, the *KBTBD2* gene, and the *NIPBL* gene had the same expression between the both groups because of the enormous amount of gene expression data, modeling with big datasets might result in lengthy analytical durations and computational inefficiencies. Before modeling using the current data set, therefore, the most significant expression connected with the output variable was chosen using the Elastic Net variable selection approach.

Twenty-one expression selected by the Elastic Net method was used in building Xgboost modeling. Accuracy, Balanced accuracy, Sensitivity, Specificity, Positive predictive value, Negative predictive value and F1 score obtained from the Bagged CART model were 85.7 %, 85.7%, 100%, 75%, 75%, 100%, and 85.7%, respectively. In brief, this study indicated that Bagged CART can accurately classify endometriosis and is a reliable approach to classification. Among the genes whose OR values were calculated, it was determined that nine genes were down-regulated, seven were upregulated, and five did not vary between the both groups. Based on the results of variable importance values from the Bagged CART method, *CUX2*, *CLMP*, *CEP131*, *EHD4*, *CDH24*, *ILRUN*, *LINC01709*, *HOTAIR*, *SLC30A2*, *NKG7* genes, and transcripts whose other gene names are inaccessible and included in the variable importance graph can be used as candidate predictive biomarkers for endometriosis. According to the results of the enrichment analysis, the activated and suppressed processes are as follows:

Activated processes: Activation of secretion and hormone regulation is consistent with the inflammatory environment seen in endometriosis. Endometriosis lesions secrete a variety of cytokines, growth factors and hormones different from normal endometrium. These active secretory mechanisms may contribute to disease progression and pain generation. Activation of the MAPK cascade: The MAPK signaling pathway is critical for cell proliferation, survival and invasion in endometriosis. Activation of this pathway can promote the abnormal growth of ectopic endometrial cells and their invasion into the pelvic regions.

Suppressed processes

Suppression of collagen trimer and extracellular matrix structures may be associated with tissue remodeling and abnormal fibrosis seen in endometriosis lesions. This may facilitate implantation and growth of ectopic tissue. Suppression of cellular response mechanisms: Suppression of the response to reactive oxygen species and nutrient levels may explain the oxidative stress state and abnormal metabolism observed in endometriosis. Moreover, the network of relationships of the processes obtained explains the following:

Centrally located MAPK cascade: The central location of MAPK signaling in the network plot suggests that this pathway coordinates many pathological processes in endometriosis. This suggests that MAPK inhibitors may be potential therapeutic targets in the treatment of endometriosis.

Table 3 The results of performance metrics for the model

	AdaBoost	XGBoost	Stochasting Gradient Boosting	Bagged CART
Accuracy	71.4	85.7	71.4	85.7
Balanced accuracy	70.8	83.3	66.7	85.7
Sensitivity	66.7	66.7	33.3	100
Specificity	75.0	100	100	75
PPV	66.7	100	100	75
NPV	75.0	80.0	66.7	100
F1 score	66.7	80.0	50.0	85.7

CART: Classification and regression trees; PPV: Positive predictive value; NPV: Negative predictive value.

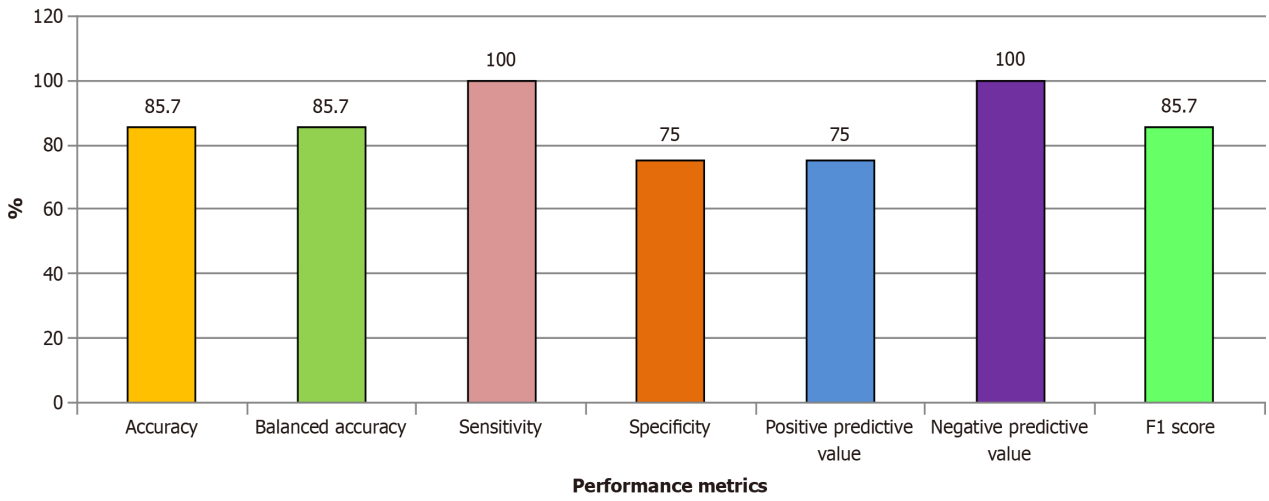


Figure 4 Graph of values for performance criteria obtained from Bagged Classification and Regression Trees models.

Response to extracellular stimuli and signal transduction: These processes may indicate an abnormal response of cells to the microenvironment in endometriosis. Ectopic endometrial cells respond differently from normal endometrial cells to steroid hormones and inflammatory stimuli.

Secretory regulation: The intensity of secretory processes shown in the graph may explain how endometriosis lesions may contribute to disease progression by altering the local microenvironment.

In conclusion, this gene expression profile supports the abnormal cell behavior and inflammatory environment seen in endometriosis. The importance of the MAPK signaling pathway highlights the value of therapeutic approaches targeting this pathway. The activation of secretory processes suggests the presence of specific proteins that could be used as biomarkers in endometriosis. Changes in extracellular matrix and collagen structures explain the fibrotic nature of endometriotic lesions and how this may contribute to disease symptoms. This analysis may contribute to understanding the molecular basis of endometriosis pathogenesis and identify potential therapeutic targets.

In addition, the calculated OR values and the variable importance values in the study support each other. The genes from which OR values were obtained were determined as the genes contributing to the development of endometriosis according to their variable importance values. The suggested pipeline resulted in a volcano graphic, which represents the up-and-downregulation of the genes in this study. The available plots are commonly employed in omics datasets such as genomics, proteomics, and metabolomics, where thousands of duplicate data points are usually available between the two conditions and provide a visualization of DEGs[54].

CUX2, *CLMP*, *CEP131*, and *EHD4* are biomarkers that play significant roles in various molecular pathways, particularly in cancer biology and cellular signaling. CUX2 is implicated in several oncogenic processes, particularly through its involvement in the PI3K-AKT-mTOR signaling pathway. Studies have shown that CUX2 functions as an oncogene in papillary thyroid cancer, where it may facilitate cell metastasis by reversing epithelial-mesenchymal transition processes [55]. This is supported by findings that *CUX2* acts as an accessory factor in the repair of oxidative DNA damage, suggesting that its expression is crucial for maintaining genomic stability in cancer cells[56]. Furthermore, *CUX2* has been associated with gastric cancer risk, where its expression is modulated by genetic variants that influence susceptibility to oxidative stress[57]. The role of *CUX2* in DNA repair mechanisms highlights its potential as a therapeutic target in

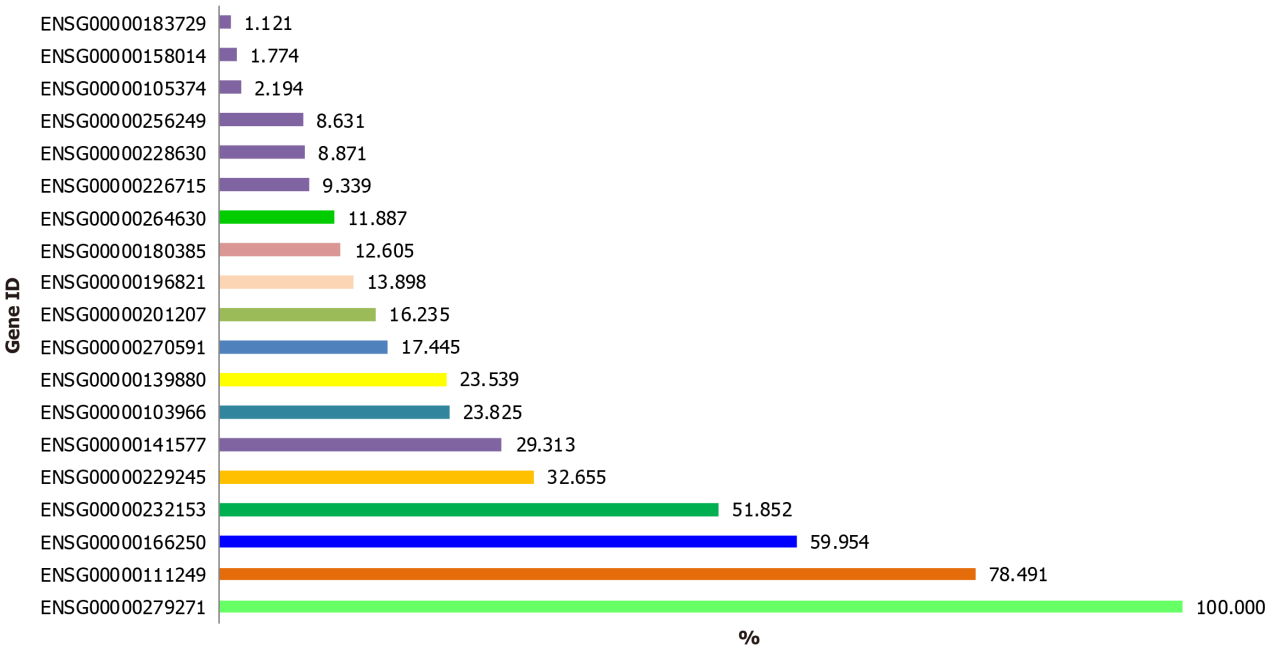


Figure 5 The graphic of gene importance values for predicting the output variable.

cancers characterized by high levels of reactive oxygen species[58].

CLMP (Coxsackie and Adenovirus Receptor-like Membrane Protein) is a member of the CTX family and is primarily involved in cell adhesion and the formation of tight junctions in epithelial cells. It has been shown to regulate the expression of connexins, which are critical for intercellular communication[59,60]. Additionally, *CLMP* is implicated in adipocyte maturation and obesity, where it modulates cell adhesion dynamics and actin polymerization, thereby influencing metabolic pathways[61].

CEP131 and *EHD4* are biomarkers that play crucial roles in various molecular pathways, particularly in cellular processes such as ciliogenesis, endocytosis, and cancer progression. *CEP131* is primarily known for its involvement in the regulation of centrosome function and ciliogenesis. It interacts with several proteins that are essential for the assembly and stability of centrioles, which are critical for proper cell division and signaling[62]. *CEP131*'s overexpression has been linked to centrosome amplification, a phenomenon often observed in cancer cells, particularly in colon cancer, where it regulates the stability of Plk4, a protein crucial for centriole duplication[63]. *EHD4*, on the other hand, is primarily associated with endosomal trafficking and has been shown to mediate the internalization of various receptors, including neurotrophin receptors, which are vital for neuronal signaling[64,65]. Additionally, *EHD4* has been shown to interact with other proteins involved in endosomal dynamics, suggesting that it plays a broader role in cellular signaling and receptor recycling[65]. The dysregulation of *EHD4* has been associated with various cancers, indicating its potential as a biomarker for tumor progression and a target for therapeutic intervention[64,65].

A study reported that *HOTAIR* showed higher expression in patients with endometriosis[66]. Another study measured the expression of *HOTAIR* in different endometrial tissues and observed its upregulation in ectopic endometrial tissues [67]. In addition, all genetic studies on *HOTAIR* have shown that this RNA is associated with endometrial carcinoma. This RNA was found to be upregulated in both endometrial cancer and carcinoma conditions. As a general approach, studies believe that *HOTAIR* reveals a new mechanism for developing endometriosis and that this will provide a new therapeutic target for the disease[67-70]. Based on variable significance values, some clinical and experimental studies have revealed the associations of *CUX2* with endometrial cancer and *EHD4* with endometrial epithelium, among other genes that may be associated with the disease[71,72]. *CEP131* has been found to be associated with tumor development and cancer development in many studies. In this study, it was determined among the most important genes as a result of modeling, and its relationship with endometriosis can be clarified with comprehensive studies[63,73,74]. Studies have shown that the *CDH24* gene is associated with ovarian carcinoma, adenomyosis, endometrial receptivity. In this study, it was determined that it may be associated with endometriosis, and its relationship with gynecological diseases can be examined more comprehensively[75-77].

With this important finding, designing studies by increasing the number of patients and conducting studies, such as testing the reliability of candidate biomarkers obtained from studies related to the disease, can expand the scope of genetic information and the power of studies. Drug treatments can be developed, and the effects of the disease that develop with severe pain can be reduced by studies on genes and their pathways that have been studied and expanded in scope. Thus, individuals whose living standards are affected may spend their days. The current study was conducted using an open-access data set, and the limited demographic and clinical information of the patients and the incomplete accessibility may be two of the limitations of this study. Also, there is also a need for larger patient datasets and external validation to ensure the validity and control of the study results.

CONCLUSION

This study suggested potential genomic biomarkers for endometriosis using transcriptomic data. The applied ML model successfully classified endometriosis and created a highly accurate diagnostic prediction model. By conducting more thorough analyses, we can test the reliability of the obtained genes, develop treatment approaches based on these genes, and detail their clinical uses. Increasing extensive genomic studies and revealing specific markers associated with the disease can explain the underlying pathology of endometriosis, paving the way for a non-invasive diagnosis to replace the invasive laparoscopy procedure.

FOOTNOTES

Author contributions: Akbulut S and Kucukakcali Z collected data; Kucukakcali Z and Colak C analyzed statistical analysis; Akbulut S and Kucukakcali Z wrote manuscript; Akbulut S and Kucukakcali Z projected development and reviewed final version.

Institutional review board statement: This study was reviewed and approved by the Inonu University institutional review board for non-interventional studies (Approval No: 2022/3842).

Informed consent statement: Not applicable, as this study was retrospective.

Conflict-of-interest statement: The authors declare that they have no conflicts of interest regarding this study.

Data sharing statement: There are no additional data available for this study.

STROBE statement: The authors have read the STROBE Statement—checklist of items, and the manuscript was prepared and revised according to the STROBE Statement—checklist of items.

Open Access: This article is an open-access article that was selected by an in-house editor and fully peer-reviewed by external reviewers. It is distributed in accordance with the Creative Commons Attribution NonCommercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: https://creativecommons.org/Licenses/by-nc/4.0/

Country of origin: Türkiye

ORCID number: Zeynep Kucukakcali 0000-0001-7956-9272; Sami Akbulut 0000-0002-6864-7711; Cemil Colak 0000-0001-5406-098X.

S-Editor: Liu H
L-Editor: A
P-Editor: Zheng XM

REFERENCES

1 **Koninckx PR**, Ussia A, Adamyan L, Tahlak M, Keckstein J, Wattiez A, Martin DC. The epidemiology of endometriosis is poorly known as the pathophysiology and diagnosis are unclear. *Best Pract Res Clin Obstet Gynaecol* 2021; **71**: 14-26 [PMID: 32978068 DOI: 10.1016/j.bpobgyn.2020.08.005]

2 **Becker CM**, Bokor A, Heikinheimo O, Horne A, Jansen F, Kiesel L, King K, Kvaskoff M, Nap A, Petersen K, Saridogan E, Tomassetti C, van Hanegem N, Vulliemoz N, Vermeulen N; ESHRE Endometriosis Guideline Group. ESHRE guideline: endometriosis. *Hum Reprod Open* 2022; **2022**: hoac009 [PMID: 35350465 DOI: 10.1093/hropen/hoac009]

3 **Greene AD**, Lang SA, Kendziorski JA, Sroga-Rios JM, Herzog TJ, Burns KA. Endometriosis: where are we and where are we going? *Reproduction* 2016; **152**: R63-R78 [PMID: 27165051 DOI: 10.1530/REP-16-0052]

4 **Montgomery GW**, Nyholt DR, Zhao ZZ, Treloar SA, Painter JN, Missmer SA, Kennedy SH, Zondervan KT. The search for genes contributing to endometriosis risk. *Hum Reprod Update* 2008; **14**: 447-457 [PMID: 18535005 DOI: 10.1093/humupd/dmn016]

5 **Moradi Y**, Shams-Beyranvand M, Khateri S, Gharahjeh S, Tehrani S, Varse F, Tiyuri A, Najmi Z. A systematic review on the prevalence of endometriosis in women. *Indian J Med Res* 2021; **154**: 446-454 [PMID: 35345070 DOI: 10.4103/ijmr.IJMR_817_18]

6 **Akter S**, Xu D, Nagel SC, Bromfield JJ, Pelch K, Wilshire GB, Joshi T. Machine Learning Classifiers for Endometriosis Using Transcriptomics and Methylomics Data. *Front Genet* 2019; **10**: 766 [PMID: 31552087 DOI: 10.3389/fgene.2019.00766]

7 **Barbara G**, Buggio L, Facchin F, Vercellini P. Medical Treatment for Endometriosis: Tolerability, Quality of Life and Adherence. *Front Glob Womens Health* 2021; **2**: 729601 [PMID: 34816243 DOI: 10.3389/fgwh.2021.729601]

8 **Barnhart K**, Giudice L, Young S, Thomas T, Diamond MP, Segars J, Youssef WA, Krawetz S, Santoro N, Eisenberg E, Zhang H; NICHD Cooperative Reproductive Medicine Network. Evaluation, validation and refinement of noninvasive diagnostic biomarkers for endometriosis (ENDOmarker): A protocol to phenotype bio-specimens for discovery and validation. *Contemp Clin Trials* 2018; **68**: 1-6 [PMID: 29524590 DOI: 10.1016/j.cct.2018.03.002]

9 **Rahmioglu N**, Mortlock S, Ghiasi M, Møller PL, Stefansdottir L, Galarneau G, Turman C, Danning R, Law MH, Sapkota Y, Christofidou P, Skarp S, Giri A, Banasik K, Krassowski M, Lepamets M, Marciniak B, Nõukas M, Perro D, Sliz E, Sobalska-Kwapis M, Thorleifsson G, Topbas-Selcuki NF, Vitonis A, Westergaard D, Arnadottir R, Burgdorf KS, Campbell A, Cheuk CSK, Clementi C, Cook J, De Vivo I, DiVasta

A, Dorien O, Donoghue JF, Edwards T, Fontanillas P, Fung JN, Geirsson RT, Girling JE, Harkki P, Harris HR, Healey M, Heikinheimo O, Holdsworth-Carson S, Hostettler IC, Houlden H, Houshdaran S, Irwin JC, Jarvelin MR, Kamatani Y, Kennedy SH, Kepka E, Kettunen J, Kubo M, Kulig B, Kurra V, Laivuori H, Laufer MR, Lindgren CM, MacGregor S, Mangino M, Martin NG, Matalliotaki C, Matalliotakis M, Murray AD, Ndungu A, Nezhat C, Olsen CM, Opoku-Anane J, Padmanabhan S, Paranjpe M, Peters M, Polak G, Porteous DJ, Rabban J, Rexrode KM, Romanowicz H, Saare M, Saavalainen L, Schork AJ, Sen S, Shafrir AL, Siewierska-Górska A, Słomka M, Smith BH, Smolarz B, Szaflik T, Szyłło K, Takahashi A, Terry KL, Tomassetti C, Treloar SA, Vanhie A, Vincent K, Vo KC, Werring DJ, Zeggini E, Zervou MI; DBDS Genomic Consortium; FinnGen Study; FinnGen Endometriosis Taskforce; Celmatix Research Team; 23andMe Research Team, Adachi S, Buring JE, Ridker PM, D'Hooghe T, Goulielmos GN, Hapangama DK, Hayward C, Horne AW, Low SK, Martikainen H, Chasman DI, Rogers PAW, Saunders PT, Sirota M, Spector T, Strapagiel D, Tung JY, Whiteman DC, Giudice LC, Velez-Edwards DR, Uimari O, Kraft P, Salumets A, Nyholt DR, Mägi R, Stefansson K, Becker CM, Yurttas-Beim P, Steinthorsdottir V, Nyegaard M, Missmer SA, Montgomery GW, Morris AP, Zondervan KT. The genetic basis of endometriosis and comorbidity with other pain and inflammatory conditions. *Nat Genet* 2023; **55**: 423-436 [PMID: 36914876 DOI: 10.1038/s41588-023-01323-z]

10 **Méar L**, Herr M, Fauconnier A, Pineau C, Vialard F. Polymorphisms and endometriosis: a systematic review and meta-analyses. *Hum Reprod Update* 2020; **26**: 73-102 [PMID: 31821471 DOI: 10.1093/humupd/dmz034]

11 **Hansen KA**, Eyster KM. Genetics and genomics of endometriosis. *Clin Obstet Gynecol* 2010; **53**: 403-412 [PMID: 20436317 DOI: 10.1097/GRF.0b013e3181db7ca1]

12 **Bouaziz J**, Mashiach R, Cohen S, Kedem A, Baron A, Zajicek M, Feldman I, Seidman D, Soriano D. How Artificial Intelligence Can Improve Our Understanding of the Genes Associated with Endometriosis: Natural Language Processing of the PubMed Database. *Biomed Res Int* 2018; **2018**: 6217812 [PMID: 29750165 DOI: 10.1155/2018/6217812]

13 **Goshisht MK**. Machine Learning and Deep Learning in Synthetic Biology: Key Architectures, Applications, and Challenges. *ACS Omega* 2024; **9**: 9921-9945 [PMID: 38463314 DOI: 10.1021/acsomega.3c05913]

14 **Mishra R**, Li B. The Application of Artificial Intelligence in the Genetic Study of Alzheimer's Disease. *Aging Dis* 2020; **11**: 1567-1584 [PMID: 33269107 DOI: 10.14336/AD.2020.0312]

15 **Mbonyinshuti F**, Nkurunziza J, Niyobuhungiro J, Kayitare E. Application of random forest model to predict the demand of essential medicines for non-communicable diseases management in public health facilities. *Pan Afr Med J* 2022; **42**: 89 [PMID: 36034003 DOI: 10.11604/pamj.2022.42.89.33833]

16 **Vincent ACSR**, Sengan S. Edge computing-based ensemble learning model for health care decision systems. *Sci Rep* 2024; **14**: 26997 [PMID: 39506092 DOI: 10.1038/s41598-024-78225-5]

17 **Zhao N**, Hao T, Zhang F, Ni Q, Zhu D, Wang Y, Shi Y, Mi X. Application of machine learning techniques in the diagnosis of endometriosis. *BMC Womens Health* 2024; **24**: 491 [PMID: 39237940 DOI: 10.1186/s12905-024-03334-2]

18 **Zhang H**, Zhang H, Yang H, Shuid AN, Sandai D, Chen X. Machine learning-based integrated identification of predictive combined diagnostic biomarkers for endometriosis. *Front Genet* 2023; **14**: 1290036 [PMID: 38098472 DOI: 10.3389/fgene.2023.1290036]

19 **Goulielmos GN**, Matalliotakis M, Matalliotaki C, Eliopoulos E, Matalliotakis I, Zervou MI. Endometriosis research in the -omics era. *Gene* 2020; **741**: 144545 [PMID: 32165309 DOI: 10.1016/j.gene.2020.144545]

20 **Cho SB**. Molecular Mechanisms of Endometriosis Revealed Using Omics Data. *Biomedicines* 2023; **11**: 2210 [PMID: 37626707 DOI: 10.3390/biomedicines11082210]

21 **Samare-Najaf M**, Razavinasab SA, Samareh A, Jamali N. Omics-based novel strategies in the diagnosis of endometriosis. *Crit Rev Clin Lab Sci* 2024; **61**: 205-225 [PMID: 37878077 DOI: 10.1080/10408363.2023.2270736]

22 **Johnson NT**, Dhroso A, Hughes KJ, Korkin D. Biological classification with RNA-seq data: Can alternatively spliced transcript expression enhance machine learning classifiers? *RNA* 2018; **24**: 1119-1132 [PMID: 29941426 DOI: 10.1261/rna.062802.117]

23 **Thompson JA**, Tan J, Greene CS. Cross-platform normalization of microarray and RNA-seq data for machine learning applications. *PeerJ* 2016; **4**: e1621 [PMID: 26844019 DOI: 10.7717/peerj.1621]

24 **Moga MA**, Bălan A, Dimienescu OG, Burtea V, Dragomir RM, Anastasiu CV. Circulating miRNAs as Biomarkers for Endometriosis and Endometriosis-Related Ovarian Cancer-An Overview. *J Clin Med* 2019; **8**: 735 [PMID: 31126056 DOI: 10.3390/jcm8050735]

25 **Chen G**, Guo J, Li W, Zheng R, Shang H, Wang Y. Diagnostic value of the combination of circulating serum miRNAs and CA125 in endometriosis. *Medicine (Baltimore)* 2023; **102**: e36339 [PMID: 38050316 DOI: 10.1097/MD.0000000000036339]

26 **Sarria-Santamera A**, Orazumbekova B, Terzic M, Issanov A, Chaowen C, Asúnsolo-Del-Barco A. Systematic Review and Meta-Analysis of Incidence and Prevalence of Endometriosis. *Healthcare (Basel)* 2020; **9**: 29 [PMID: 33396813 DOI: 10.3390/healthcare9010029]

27 **Sivajohan B**, Elgendi M, Menon C, Allaire C, Yong P, Bedaiwy MA. Clinical use of artificial intelligence in endometriosis: a scoping review. *NPJ Digit Med* 2022; **5**: 109 [PMID: 35927426 DOI: 10.1038/s41746-022-00638-1]

28 **Blass I**, Sahar T, Shraibman A, Ofer D, Rappoport N, Linial M. Revisiting the Risk Factors for Endometriosis: A Machine Learning Approach. *J Pers Med* 2022; **12**: 1114 [PMID: 35887611 DOI: 10.3390/jpm12071114]

29 **Mbuguiro W**, Gonzalez AN, Mac Gabhann F. Computational Models for Diagnosing and Treating Endometriosis. *Front Reprod Health* 2021; **3**: 699133 [PMID: 36303959 DOI: 10.3389/frph.2021.699133]

30 **Konrad L**, Fruhmann Berger LM, Maier V, Horné F, Neuheisel LM, Laucks EV, Riaz MA, Oehmke F, Meinhold-Heerlein I, Zeppernick F. Predictive Model for the Non-Invasive Diagnosis of Endometriosis Based on Clinical Parameters. *J Clin Med* 2023; **12**: 4231 [PMID: 37445265 DOI: 10.3390/jcm12134231]

31 **Saeys Y**, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics* 2007; **23**: 2507-2517 [PMID: 17720704 DOI: 10.1093/bioinformatics/btm344]

32 **Daneshvar A**, Mousa G. Regression shrinkage and selection *via* least quantile shrinkage and selection operator. *PLoS One* 2023; **18**: e0266267 [PMID: 36795659 DOI: 10.1371/journal.pone.0266267]

33 **Tay JK**, Narasimhan B, Hastie T. Elastic Net Regularization Paths for All Generalized Linear Models. *J Stat Softw* 2023; **106**: 1 [PMID: 37138589 DOI: 10.18637/jss.v106.i01]

34 **Shortreed SM**, Ertefaie A. Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics* 2017; **73**: 1111-1122 [PMID: 28273693 DOI: 10.1111/biom.12679]

35 **Zimmerman RK**, Balasubramani GK, Nowalk MP, Eng H, Urbanski L, Jackson ML, Jackson LA, McLean HQ, Belongia EA, Monto AS, Malosh RE, Gaglani M, Clipper L, Flannery B, Wisniewski SR. Classification and Regression Tree (CART) analysis to predict influenza in primary care patients. *BMC Infect Dis* 2016; **16**: 503 [PMID: 27659721 DOI: 10.1186/s12879-016-1839-x]

36 **Mahajan P**, Uddin S, Hajati F, Moni MA. Ensemble Learning for Disease Prediction: A Review. *Healthcare (Basel)* 2023; **11**: 1808 [PMID: 37372925 DOI: 10.3390/healthcare11121808]

37 **Choubin B**, Abdolshahnejad M, Moradi E, Querol X, Mosavi A, Shamshirband S, Ghamisi P. Spatial hazard assessment of the PM10 using machine learning models in Barcelona, Spain. *Sci Total Environ* 2020; **701**: 134474 [PMID: 31704408 DOI: 10.1016/j.scitotenv.2019.134474]

38 **Nguyen D**, Sadeghnejad Barkousaraie A, Bohara G, Balagopal A, McBeth R, Lin MH, Jiang S. A comparison of Monte Carlo dropout and bootstrap aggregation on the performance and uncertainty estimation in radiation therapy dose prediction with deep learning neural networks. *Phys Med Biol* 2021; **66**: 054002 [PMID: 33503599 DOI: 10.1088/1361-6560/abe04f]

39 **Ren Z**, Li Q, Yang X, Wang J. A novel method for identifying corrosion types and transitions based on Adaboost and electrochemical noise. *Anti-Corrosion Methods Mater* 2023; **70**: 78-85 [DOI: 10.1108/acmm-11-2022-2725]

40 **Li K**, Zhou G, Zhai J, Li F, Shao M. Improved PSO_AdaBoost Ensemble Algorithm for Imbalanced Data. *Sensors (Basel)* 2019; **19**: 1476 [PMID: 30917599 DOI: 10.3390/s19061476]

41 **Lai SBS**, Binti Md Shahri NHN, Mohamad MB, Rahman HABA, Rambli AB. Comparing the Performance of AdaBoost, XGBoost, and Logistic Regression for Imbalanced Data. *Math Stat* 2021; **9**: 379-385 [DOI: 10.13189/ms.2021.090320]

42 **Ritchie ME**, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015; **43**: e47 [PMID: 25605792 DOI: 10.1093/nar/gkv007]

43 **Yan H**, Zheng G, Qu J, Liu Y, Huang X, Zhang E, Cai Z. Identification of key candidate genes and pathways in multiple myeloma by integrated bioinformatics analysis. *J Cell Physiol* 2019; **234**: 23785-23797 [PMID: 31215027 DOI: 10.1002/jcp.28947]

44 **Vandenbroucke JP**, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, Poole C, Schlesselman JJ, Egger M; STROBE Initiative. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *Int J Surg* 2014; **12**: 1500-1524 [PMID: 25046751 DOI: 10.1016/j.ijsu.2014.07.014]

45 **Saunders PTK**, Horne AW. Endometriosis: Etiology, pathobiology, and therapeutic prospects. *Cell* 2021; **184**: 2807-2824 [PMID: 34048704 DOI: 10.1016/j.cell.2021.04.041]

46 **Agarwal SK**, Chapron C, Giudice LC, Laufer MR, Leyland N, Missmer SA, Singh SS, Taylor HS. Clinical diagnosis of endometriosis: a call to action. *Am J Obstet Gynecol* 2019; **220**: 354.e1-354.e12 [PMID: 30625295 DOI: 10.1016/j.ajog.2018.12.039]

47 **Bischoff F**, Simpson JL. Genetic basis of endometriosis. *Ann N Y Acad Sci* 2004; **1034**: 284-299 [PMID: 15731320 DOI: 10.1196/annals.1335.030]

48 **Tan Y**, Flynn WF, Sivajothi S, Luo D, Bozal SB, Davé M, Luciano AA, Robson P, Luciano DE, Courtois ET. Single-cell analysis of endometriosis reveals a coordinated transcriptional programme driving immunotolerance and angiogenesis across eutopic and ectopic tissues. *Nat Cell Biol* 2022; **24**: 1306-1318 [PMID: 35864314 DOI: 10.1038/s41556-022-00961-5]

49 **Bunis DG**, Wang W, Vallvé-Juanico J, Houshdaran S, Sen S, Ben Soltane I, Kosti I, Vo KC, Irwin JC, Giudice LC, Sirota M. Whole-Tissue Deconvolution and scRNAseq Analysis Identify Altered Endometrial Cellular Compositions and Functionality Associated With Endometriosis. *Front Immunol* 2021; **12**: 788315 [PMID: 35069565 DOI: 10.3389/fimmu.2021.788315]

50 **Braga Melo VB**, Rocha Oliveira EE, Verruma CG, Sousa e Silva PV, dos Reis RM, Alves Sales SL, Cavalcante M, Libardi M Furtado C. Transcriptomic study of the genetic profile of endometriosis Fertil Steril 2024; 122: e275 [DOI: 10.1016/j.fertnstert.2024.07.862]

51 **Geng R**, Huang X, Li L, Guo X, Wang Q, Zheng Y, Guo X. Gene expression analysis in endometriosis: Immunopathology insights, transcription factors and therapeutic targets. *Front Immunol* 2022; **13**: 1037504 [PMID: 36532015 DOI: 10.3389/fimmu.2022.1037504]

52 **Eyster KM**, Klinkova O, Kennedy V, Hansen KA. Whole genome deoxyribonucleic acid microarray analysis of gene expression in ectopic versus eutopic endometrium. *Fertil Steril* 2007; **88**: 1505-1533 [PMID: 17462640 DOI: 10.1016/j.fertnstert.2007.01.056]

53 **Lee B**, Du H, Taylor HS. Experimental murine endometriosis induces DNA methylation and altered gene expression in eutopic endometrium. *Biol Reprod* 2009; **80**: 79-85 [PMID: 18799756 DOI: 10.1095/biolreprod.108.070391]

54 **Blum BC**, Emili A. Omics Notebook: robust, reproducible and flexible automated multiomics exploratory analysis and reporting. *Bioinform Adv* 2021; **1**: vbab024 [PMID: 36700091 DOI: 10.1093/bioadv/vbab024]

55 **Sun Y**, Ye D, Li Y, Chen E, Hao R, Cai Y, Wang Q, Wang O, Zhang X. CUX2 functions as an oncogene in papillary thyroid cancer. *Onco Targets Ther* 2019; **12**: 217-224 [PMID: 30636884 DOI: 10.2147/OTT.S185710]

56 **Pal R**, Ramdzan ZM, Kaur S, Duquette PM, Marcotte R, Leduy L, Davoudi S, Lamarche-Vane N, Iulianella A, Nepveu A. CUX2 protein functions as an accessory factor in the repair of oxidative DNA damage. *J Biol Chem* 2015; **290**: 22520-22531 [PMID: 26221032 DOI: 10.1074/jbc.M115.651042]

57 **Tanikawa C**, Kamatani Y, Toyoshima O, Sakamoto H, Ito H, Takahashi A, Momozawa Y, Hirata M, Fuse N, Takai-Igarashi T, Shimizu A, Sasaki M, Yamaji T, Sawada N, Iwasaki M, Tsugane S, Naito M, Hishida A, Wakai K, Furusyo N, Murakami Y, Nakamura Y, Imoto I, Inazawa J, Oze I, Sato N, Tanioka F, Sugimura H, Hirose H, Yoshida T, Matsuo K, Kubo M, Matsuda K. Genome-wide association study identifies gastric cancer susceptibility loci at 12q24.11-12 and 20q11.21. *Cancer Sci* 2018; **109**: 4015-4024 [PMID: 30281874 DOI: 10.1111/cas.13815]

58 **Ramdzan ZM**, Vickridge E, Faraco CCF, Nepveu A. CUT Domain Proteins in DNA Repair and Cancer. *Cancers (Basel)* 2021; **13**: 2953 [PMID: 34204734 DOI: 10.3390/cancers13122953]

59 **Langhorst H**, Jüttner R, Groneberg D, Mohtashamdolatshahi A, Pelz L, Purfürst B, Schmidt-Ott KM, Friebe A, Rathjen FG. The IgCAM CLMP regulates expression of Connexin43 and Connexin45 in intestinal and ureteral smooth muscle contraction in mice. *Dis Model Mech* 2018; **11**: dmm032128 [PMID: 29361518 DOI: 10.1242/dmm.032128]

60 **Raschperger E**, Engstrom U, Pettersson RF, Fuxe J. CLMP, a novel member of the CTX family and a new component of epithelial tight junctions. *J Biol Chem* 2004; **279**: 796-804 [PMID: 14573622 DOI: 10.1074/jbc.M308249200]

61 **Murakami K**, Eguchi J, Hida K, Nakatsuka A, Katayama A, Sakurai M, Choshi H, Furutani M, Ogawa D, Takei K, Otsuka F, Wada J. Antiobesity Action of ACAM by Modulating the Dynamics of Cell Adhesion and Actin Polymerization in Adipocytes. *Diabetes* 2016; **65**: 1255-1267 [PMID: 26956488 DOI: 10.2337/db15-1304]

62 **Tollenaere MAX**, Villumsen BH, Blasius M, Nielsen JC, Wagner SA, Bartek J, Beli P, Mailand N, Bekker-Jensen S. p38- and MK2-dependent signalling promotes stress-induced centriolar satellite remodelling *via* 14-3-3-dependent sequestration of CEP131/AZI1. *Nat Commun* 2015; **6**: 10075 [PMID: 26616734 DOI: 10.1038/ncomms10075]

63 **Kim DH**, Ahn JS, Han HJ, Kim HM, Hwang J, Lee KH, Cha-Molstad H, Ryoo IJ, Jang JH, Ko SK, Yang JO, Lee HG, Lee S, Song EJ, Kim JY, Huh YH, Kwon YT, Soung NK, Kim BY. Cep131 overexpression promotes centrosome amplification and colon cancer progression by regulating Plk4 stability. *Cell Death Dis* 2019; **10**: 570 [PMID: 31358734 DOI: 10.1038/s41419-019-1778-8]

64 **Mohd S**, Oder A, Specker E, Neuenschwander M, Von Kries JP, Daumke O. Identification of drug-like molecules targeting the ATPase activity of dynamin-like EHD4. *PLoS One* 2024; **19**: e0302704 [PMID: 39074100 DOI: 10.1371/journal.pone.0302704]

65 **Yap CC**, Lasiecka ZM, Caplan S, Winckler B. Alterations of EHD1/EHD4 protein levels interfere with L1/NgCAM endocytosis in neurons and disrupt axonal targeting. *J Neurosci* 2010; **30**: 6646-6657 [PMID: 20463227 DOI: 10.1523/JNEUROSCI.5428-09.2010]

66 **Chang CY**, Tseng CC, Lai MT, Chiang AJ, Lo LC, Chen CM, Yen MJ, Sun L, Yang L, Hwang T, Tsai FJ, Sheu JJ. Genetic impacts on thermostability of onco-lncRNA HOTAIR during the development and progression of endometriosis. *PLoS One* 2021; **16**: e0248168 [PMID: 33667269 DOI: 10.1371/journal.pone.0248168]

67 **Zhang L**, Yu Z, Qu Q, Li X, Lu X, Zhang H. Exosomal lncRNA HOTAIR Promotes the Progression and Angiogenesis of Endometriosis *via* the miR-761/HDAC1 Axis and Activation of STAT3-Mediated Inflammation. *Int J Nanomedicine* 2022; **17**: 1155-1170 [PMID: 35321026 DOI: 10.2147/IJN.S354314]

68 **He X**, Bao W, Li X, Chen Z, Che Q, Wang H, Wan XP. The long non-coding RNA HOTAIR is upregulated in endometrial carcinoma and correlates with poor prognosis. *Int J Mol Med* 2014; **33**: 325-332 [PMID: 24285342 DOI: 10.3892/ijmm.2013.1570]

69 **Huang J**, Ke P, Guo L, Wang W, Tan H, Liang Y, Yao S. Lentivirus-mediated RNA interference targeting the long noncoding RNA HOTAIR inhibits proliferation and invasion of endometrial carcinoma cells in vitro and in vivo. *Int J Gynecol Cancer* 2014; **24**: 635-642 [PMID: 24758900 DOI: 10.1097/IGC.0000000000000121]

70 **Łuczak A**, Supernat A, Łapińska-Szumczyk S, Jachimowicz D, Majewska H, Gulczyński J, Żaczek AJ. HOTAIR in relation to epithelial-mesenchymal transition and cancer stem cells in molecular subtypes of endometrial cancer. *Int J Biol Markers* 2016; **31**: e245-e251 [PMID: 26868332 DOI: 10.5301/jbm.5000187]

71 **Tasaki H**, Zhao L, Isayama K, Chen H, Nobuhiko Yamauchi, Yasufumi Shigeyoshi, Hashimoto S, Hattori MA. Profiling of circadian genes expressed in the uterus endometrial stromal cells of pregnant rats as revealed by DNA microarray coupled with RNA interference. *Front Endocrinol (Lausanne)* 2013; **4**: 82 [PMID: 23847593 DOI: 10.3389/fendo.2013.00082]

72 **Glubb DM**, Kho PF; Consortium ECA, Thompson D, Spurdle A, O'mara T. Abstract LB-164: Global study of chromatin interactions reveals biologically relevant candidate target genes at endometrial cancer risk loci. *Cancer Res* 2018; **78**: LB-164 [DOI: 10.1158/1538-7445.am2018-lb-164]

73 **Kim DH**, Kim HM, Huong PTT, Han HJ, Hwang J, Cha-Molstad H, Lee KH, Ryoo IJ, Kim KE, Huh YH, Ahn JS, Kwon YT, Soung NK, Kim BY. Enhanced anticancer effects of a methylation inhibitor by inhibiting a novel DNMT1 target, CEP 131, in cervical cancer. *BMB Rep* 2019; **52**: 342-347 [PMID: 31068247 DOI: 10.5483/BMBRep.2019.52.5.055]

74 **Shen M**, Li S, Zhao Y, Liu Y, Liu Z, Huan L, Qiao Y, Wang L, Han L, Chen Z, He X. Hepatic ARID3A facilitates liver cancer malignancy by cooperating with CEP131 to regulate an embryonic stem cell-like gene signature. *Cell Death Dis* 2022; **13**: 732 [PMID: 36008383 DOI: 10.1038/s41419-022-05187-9]

75 **Prašnikar E**, Kunej T, Gorenjak M, Potočnik U, Kovačič B, Knez J. Transcriptomics of receptive endometrium in women with sonographic features of adenomyosis. *Reprod Biol Endocrinol* 2022; **20**: 2 [PMID: 34980152 DOI: 10.1186/s12958-021-00871-5]

76 **Song F**, Li L, Zhang B, Zhao Y, Zheng H, Yang M, Li X, Tian J, Huang C, Liu L, Wang Q, Zhang W, Chen K. Tumor specific methylome in Chinese high-grade serous ovarian cancer characterized by gene expression profile and tumor genotype. *Gynecol Oncol* 2020; **158**: 178-187 [PMID: 32362568 DOI: 10.1016/j.ygyno.2020.04.688]

77 **Zhou W**, Santos L, Dimitriadis E. Characterization of the role for cadherin 6 in the regulation of human endometrial receptivity. *Reprod Biol Endocrinol* 2020; **18**: 66 [PMID: 32600462 DOI: 10.1186/s12958-020-00624-w]

Published by **Baishideng Publishing Group Inc**
7041 Koll Center Parkway, Suite 160, Pleasanton, CA 94566, USA
Telephone: +1-925-3991568
E-mail: office@baishideng.com
Help Desk: https://www.f6publishing.com/helpdesk
https://www.wjgnet.com

© 2025 Baishideng Publishing Group Inc. All rights reserved.