Available online at www.medicinescience.org

ORIGINAL ARTICLE

Medicine Science International Medical Journal

Medicine Science 2023;12(1):231-7

Machine learning-based ovarian cancer prediction with XGboost and stochastic gradient boosting models

Onural Ozhan[1], Zeynep Kucukakcali[2], Ipek Balikci Cicek[2]

[1]Inonu University Faculty of Medicine, Department of Pharmacology, Malatya, Türkiye
[2]Inonu University Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Türkiye

Received 12 September 2022; Accepted 08 December 2022
Available online 27.02.2023 with doi: 10.5455/medscience.2022.09.207

Abstract

Ovarian cancer is one of the most common types of gynecological malignancies with its high mortality rate, silent and occult tumor growth, late onset of symptoms and diagnosis in advanced stages. Therefore, the need to develop new diagnostic techniques to predict the course of the disease and the prognosis of this malignancy has increased. In this study, ovarian cancer and benign ovarian tumor samples will be classified to create an accurate diagnostic predictive model using the machine learning method XGBoost and Stochastic Gradient Boosting and disease-related risk factors will be determined. This current study considered the open-access ovarian cancer and benign ovarian tumor samples data set. For this purpose, data from 349 patients were included. The data set was divided as 80:20 as a training and test dataset. XGBoost and Stochastic Gradient Boosting were constructed for the classification via five-fold cross-validation. Accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, and negative predictive value performance metrics were evaluated for model performance. Among the performance criteria in the test stage obtained from the XGBoost model that has the best classification result; accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1 score were obtained as 89.5%, 88.7%, 85.7%, 91.7%, 85.7%, 91.7%, and 85.7%, respectively. According to the variable importance obtained as a result of the model, the variables most associated with the diagnosis were CA72-4, HE4, LYM%, ALB, EO%, BUN, RBC, NEU, and MCV, respectively. The applied machine learning model successfully classified ovarian cancer and created a highly accurate diagnostic prediction model. The results from the study revealed effective parameters that can diagnose ovarian cancer with high accuracy. With the parameters determined as a result of the modeling, the clinician will be able to simplify and facilitate the decision-making process for the diagnosis of ovarian cancer.

Keywords: Ovarian cancer, classification, machine learning, XGBoost, stochastic gradient boosting

Introduction

Cancer is the major cause of death in most parts of the world, and it is currently the most significant impediment to most countries achieving optimal life expectancy [1,2]. On the heels of cervical and uterine cancer, ovarian cancer is the third most common kind of gynecologic malignancy. In addition to this, it has the highest number of fatalities and the poorest prognosis [3]. Although ovarian cancer is not as prevalent as breast cancer, it is three times more likely to result in death, and it is projected that the fatality rate will significantly increase by the year 2040 [2,4]. Ovarian cancer was reported to have affected 295,414 women globally in 2018 and been responsible for 184,799 fatalities [2]. Ovarian cancer has a high mortality rate due to silent and

CITATION
Ozhan O, Kucukakcali Z, Cicek IB. Machine learning-based ovarian cancer prediction with XGboost and stochastic gradient boosting models. Med Science. 2023;12(1):231-7.

Corresponding Author: Zeynep Kucukakcali, Inonu University Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Türkiye
Email: zeynep.tunc@inonu.edu.tr

DOI: 10.5455/medscience.2022.09.207

secret tumor growth, delayed symptom presentation, and a lack of adequate screening, which leads to diagnosis in the advanced stages. Because of its stealthy nature, this disease has earned the moniker "the silent killer" [4,5].

Ovarian cancer incidence varies across the world, like that of many other malignancies [6]. Ovarian cancer mortality follows a varied pattern due to inequalities in access to diagnostic and treatment facilities, with African people having the greatest fatality rate [7]. Among the risk factors that cause the disease, it is thought that there are factors such demographic factors, reproductive factors, gynecologic factors, hormonal factors, genetic factors, and lifestyle Factors [8]. Despite the fact that ovarian tumors are chemo-sensitive and typically exhibit initial efficacy against platinum/taxane treatment, patients with the advanced illness have 5-year recurrence rates of 60% to 80% . Consequently, much effort has been devoted to developing new diagnostic techniques to predict the course of the disease and the prognosis of this malignancy [9].

Machine learning (ML), is a subfield of artificial intelligence that performs data-driven learning in order to achieve its goal of making predictions about new data when such data is introduced to it. In recent years, ML approaches have become one of the technologies that are widely utilized in the diagnosis of diseases and clinical decision support systems. These methods also have a wide range of potential applications in a variety of fields. The determination of genetic diseases, the early diagnosis of cancer diseases, and the identification of patterns in medical imaging are all examples of areas in which ML, which has a wide range of applications in the medical field, constitutes the basic infrastructure of application areas. In the last decade, both unsupervised and supervised machine learning methods have achieved high performance in a wide variety of situations [10,11].

In the current study, ovarian cancer and benign ovarian tumor samples will be classified using XGBoost and Stochastic Gradient Boosting, supervised machine learning method trained on 49 predictive variables, to create an accurate diagnostic prediction model. The performance metrics obtained from the 2 models used will be examined and the model with the best classification will be decided. The model with the best classification performance will be used as the diagnostic prediction model, and the variables that can be used in diagnosing the disease will be obtained with the help of the variable significance values obtained as a result of the model.

Material and Methods

Data collection and variables

The patients used in the current study consisted of patients with ovarian cancer and benign ovarian tumor. Patients who underwent surgical resection and were diagnosed by pathology after surgery were included in the research. None of the ovarian cancer patients included in the study received chemotherapy or radiotherapy before surgery. The dataset includes 349 Chinese patients, 49 variables such as demographics, blood routine tests, general chemistry, and tumor markers, and the target variable (ovarian cancer and benign ovarian tumor) [12].

XGboost algorithm

Gradient Boost is a powerful ML technique for regression and classification problems in which weak predictive models frequently produce ensemble forms of decision trees. Gradient Boost, which is based on the boosting method, aims to construct many weak learners in sequence and incorporate them into a complex model [13].

Extreme Gradient Boosting (XGBoost) is one of the applications of gradient boosting machines (GBM), one of the most effective supervised learning algorithms. Its fundamental structure is founded on gradient boosting and decision tree algorithms. When compared to other algorithms, it has a significant advantage in terms of speed and performance. XGBoost is also highly predictive, ten times faster than other algorithms, and includes several regularizations that improve overall performance while reducing overfitting or over-learning. Gradient boosting is an ensemble method for creating a robust classifier by combining a set of weak classifiers with boosting. Starting with a basic learner, the strong learner is trained iteratively. The principles of gradient boosting and XGBoost are the same. The key differences are in the implementation details. By controlling the complexity of the trees, XGBoost can achieve better performance by employing various regularization techniques [14,15].

Stochastic gradient boosting algorithm

Stochastic gradient boosting is a method developed by Friedman by incorporating randomness into the gradient boosting method. In stochastic gradient boosting, a random subsample is selected with permutation sampling strategy at each refresh. This selected subsample is used to calculate the model update instead of all learners and to reduce the correlation between trees [16]. As with other ensemble learning methods, large trees are not created in this method, but instead, each tree (usually 100–200 trees) developed during the process is summarized and each observation is grouped according to the most common classification among trees. These differences cause the stochastic gradient boosting method to differ from other augmentation methods and reduce its sensitivity to outliers and unbalanced data sets [16,17].

Biostatistical analysis

The median (minimum-maximum) and mean±standart deviation were used to summarize variables. Shapiro Wilk test of normality was used to determine normal distribution. Whether there is a statistically significant difference between the output variable and input variables was evaluated by using where appropriate Mann-Whitney U test, independent sample t-test, and Chi-square test. It was determined that a value was statistically significant if it had a p-value of less than 0.05 ($p<0.05$). IBM SPSS Statistics 26.0 was utilized in all analyses.

DOI: 10.5455/medscience.2022.09.207

Machine learning modeling and performance evaluation

In the current study, XGBoost and Stochastic Gradient Boosting were used in the modeling stage for the dataset in question. The data set was divided as 80:20 as a training and test dataset. The n-fold cross-validation approach was used for the analyses. The data is separated into n parts in the n-fold cross-validation procedure, and the model is applied to n parts. One of the n parts is used for testing, while the remaining n-1 parts are used to train the model. In this study, 5-fold cross-validation was employed for the modeling process. Accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1-score were used as performance evaluation criteria. In addition, variable importances were calculated, which gives information about how much the input variables contribute to the output variable. Modeling was done using R studio 4.2.1.

Results

The data set used in the study included a total of 349 patients with 171 ovarian cancer and 178 benign ovarian tumors. The mean age of the patients was 45.05±15.13. While the mean age of patients with ovarian cancer was 52.98±13.48 years, the mean age of patients with benign ovarian tumors was 37.44±12.5.

The results of the statistical analyses between the target variable and independent variables are given in Table 1.

According to the results obtained from the analyzes made between the independent variables and the target variable, there is a statistical difference between the target variable ovarian cancer ve benign ovarian tumors categories in terms of AFP, Age, ALB, ALP, AST, BASO%, Ca, CA125, CA72-4, DBIL, GGT, GLO, GLU, HCT, HE4, HGB, IBIL, LYM#, LYM%, MCH, MONO, Na, NEU, PCT, PDW, PLT, RBC, TBIL, TP variables ($p<0.05$). In addition, there is a statistically significant difference between the categories of the target variable for the menopause variable, which is a qualitative variable.

Table 2 shows the values of the performance metrics calculated for the training and test data sets as a result of the Stochastic Gradient Boosting model.

In the training stage accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1 score obtained from the Stochastic gradient boosting model were 97.5%, 97.3%, 96.7%, 98%, 96.7%, 98%, and 96.7%, respectively. Also in the testing stage accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1-score obtained from the Stochastic gradient boosting model were 84.2%, 81.5%, 71.4%, 91.7%, 83.3%, 94.6%, and 76.9%, respectively.

In the training stage accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value,

Table 1. The results of the statistical analyses between the target variable and independent variables

Variables	Abreviation of Variables	TYPE				p
		Cancer		No Cancer		
		Mean±SD	Median(Min-Max)	Mean±SD	Median(Min-Max)	
Alpha-Fetoprotein	AFP	22.71±145.02	2.44(0.61-1210)	2.71±2.18	2.1(0.61-20.53)	0.013**
Anion Gap	AG	19.4±4.6	19.97(6.6-28.36)	19.24±4.1	19.74(6.2-33.33)	0.728*
Age	Age	52.98±13.49	53(18-83)	37.44±12.51	36(15-69)	<0.001*
Albumin	ALB	38.89±6.33	39.2(22-51.5)	43.2±3.97	43.4(28-50.9)	<0.001*
Alkaline Phosphatase	ALP	86.78±58.45	77(26-763)	67.67±19.88	65(29-157)	<0.001**
Alanine Aminotransferase	ALT	17.96±11.93	15(5-86)	18.06±10.57	15(4-71)	0.353**
Aspartate Aminotransferase	AST	20.98±9.47	19(9-64)	17.28±6.99	16(7-78)	<0.001**
Basophil Cell Count	BASO#	0.03±0.02	0.02(0-0.12)	0.03±0.02	0.03(0-0.12)	0.312**
Basophil Cell Ratio	BASO%	0.44±0.33	0.4(0-1.58)	0.52±0.36	0.44(0-1.94)	0.046*
Blood Urea Nitrogen	BUN	4.01±1.42	3.75(1.12-10.19)	4.02±1.15	3.88(1.52-8.31)	0.449**
Calcium	Ca	2.32±0.42	2.44(0.92-2.82)	2.46±0.28	2.52(0.95-2.83)	<0.001**
Carbohydrate Antigen 125	CA125	695.89±1068.16	241.5(7.26-5000)	51.45±79.79	22.66(3.75-515.4)	<0.001**
Carbohydrate Antigen 19-9	CA19-9	71.95±178.98	14.99(0.6-1000)	25.91±40.36	14(0.6-279.5)	0.2
Carbohydrate Antigen 72-4	CA72-4	19.84±36.61	4.83(0.63-158.5)	3.12±4.06	1.74(0.2-19.37)	<0.001**
Carcinoembryonic Antigen	CEA	5.51±16.08	1.41(0.2-138.8)	1.47±0.88	1.27(0.2-6.95)	0.103**
Chlorine	CL	100.9±3.52	101(84.6-109.4)	100.73±2.36	100.8(93.1-108)	0.340**
Carban Dioxide-Combining Power	CO2CP	24.54±2.97	24.1(16.2-34.3)	24.03±2.37	24(16.2-30)	0.077*
Creatinine	CREA	63.35±12.59	62.7(38.2-114)	65.11±10.77	64(41-94.1)	0.060**

DOI: 10.5455/medscience.2022.09.207

Direct Bilirubin	DBIL	2.9±1.35	2.5(0.9-9.2)	3.35±1.48	3.15(1-12.1)	<0.001**
Eosinophil Count	EO#	0.06±0.06	0.05(0-0.34)	0.07±0.08	0.05(0-0.4)	0.421**
Eosinophil Ratio	EO%	1±0.94	0.8(0-5.1)	1.24±1.29	0.9(0-7.6)	0.134**
Gama Glutamyltransferasey	GGT	22.83±17.43	17(4-114)	19.84±18.8	15(4-176)	0.007**
Globulin	GLO	31.09±5.09	31(14.1-47.6)	29.29±3.75	28.85(20.4-40.5)	<0.001*
Glucose	GLU	5.56±1.48	5.23(3.71-12.44)	5.12±0.85	5(3.57-9.3)	0.014**
Hematocrit	HCT	0.38±0.05	0.39(0.22-0.57)	0.39±0.04	0.39(0.29-0.46)	0.033**
Human Epididymis Protein 4	HE4	342.82±517.31	140.9(29.49-3537.6)	49.17±39.07	43.77(16.71-531.8)	<0.001**
Hemoglobin	HGB	122.2±16.71	124(61.8-189)	128.34±13.7	129.5(80-158)	<0.001**
Indirect Bilirubin	IBIL	5.35±2.54	5(1.5-16)	6.56±3.2	6(1-28.4)	<0.001**
Kalium	K	4.38±0.41	4.36(3.08-5.4)	4.39±0.39	4.38(3.33-5.39)	0.918*
Lymphocyte Count	LYM#	1.41±0.54	1.34(0.35-3.2)	1.7±0.55	1.67(0.38-3.49)	<0.001*
Lymphocyte Ratio	LYM%	22.74±10.05	23.4(3.9-51.6)	29.27±9.7	29.8(5.1-49.6)	<0.001**
Mean Corpuscular Hemoglubin	MCH	28.34±2.71	28.9(17.7-33.7)	29.2±2.36	29.85(18.9-36.8)	<0.001**
Mean Corpuscular Volume	MCV	87.79±6.71	88.4(61-103.4)	88.34±5.32	89.6(65.2-107.9)	0.414**
Magnesium	Mg	0.98±0.13	0.96(0.65-1.3)	0.98±0.12	0.97(0.66-1.37)	0.571**
Mononuclear Cell Count	MONO#	0.39±0.16	0.35(0.07-0.94)	0.33±0.13	0.31(0.07-0.97)	<0.001**
Monocyte Ratio	MONO%	5.77±2.01	5.55(1.51-21.3)	5.4±1.82	5.28(0.3-12.14)	0.069*
Mean Platelet Volume	MPV	9.98±1.78	10.1(5.06-14.5)	10.1±1.7	10.4(5.98-13.8)	0.407**
Natrium	Na	140.91±3.26	141.3(125.1-147.9)	140.09±2.35	140.1(133-150.7)	<0.001**
Neutrophil Ratio	NEU	70.25±10.98	70.6(37.2-92)	59.59±9.4	59.5(42.44-81.7)	<0.001*
Thrombocytocrit	PCT	0.27±0.1	0.26(0.09-0.69)	0.23±0.07	0.23(0.07-0.42)	<0.001**
Platelet Distribution Width	PDW	13.91±3.19	13.4(8.8-22.8)	14.74±2.75	14.1(9.4-20.9)	0.002**
Phosphorus	PHOS	1.12±0.18	1.12(0.65-1.67)	1.12±0.19	1.13(0.57-1.75)	0.641**
Platelet Count	PLT	281.65±117.44	265(74-868)	230.25±57.33	223.5(88-398)	<0.001**
Red Blood Cell Count	RBC	4.31±0.52	4.32(2.62-6.74)	4.4±0.41	4.42(2.9-5.39)	0.031**
Red Blood Cell Distribution Width	RDW	13.68±2	13.1(10.92-22.2)	13.43±1.59	12.9(11.1-20)	0.376**
Total Bilirubin	TBIL	8.25±3.65	7.6(2.5-22.7)	9.91±4.47	9.2(2.7-38.3)	<0.001**
Total Protein	TP	69.62±8.88	71.9(32.9-83.8)	72.49±5.15	73(49.5-86.8)	0.015**
Urie Acid	UA	246.25±78.04	236.9(96-632)	241.26±58.18	233.25(102.3-460.6)	0.982**

*: Independent samples t-test, **: Mann Whitney U test, SD: Standart Deviation, Min: Minimum, Max: Maximum

Table 2. Performance metric values calculated from Stochastic Gradient Boosting model

Performance Metrics	Training Stage	Testing Stage
	Value (%)	Value (%)
Accuracy	97.5	84.2
Balanced Accuracy	97.3	81.5
Sensitivity	96.7	71.4
Specificity	98.0	91.7
Positive predictive value	96.7	83.3
Negative predictive value	98.0	84.6
F1-score	96.7	76.9

Table 3. Performance metric values calculated from XGBoost model

Performance Metrics	Training Stage	Testing Stage
	Value (%)	Value (%)
Accuracy	91.2	89.5
Balanced Accuracy	89.7	88.7
Sensitivity	83.3	85.7
Specificity	96.0	91.7
Positive predictive value	92.6	85.7
Negative predictive value	90.6	91.7
F1-score	87.7	85.7

DOI: 10.5455/medscience.2022.09.207

and F1-score obtained from the XGBoost model were 91.2%, 89.7%, 83.3%, 96%, 92.6%, 90.6%, and 87.7%, respectively. Also in the testing stage accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1-score obtained from the XGBoost model were 89.5%, 88.7%, 85.7%, 91.7%, 85.7%, 91.7%, and 85.7%, respectively. In Figure 1, values of performance metrics obtained from Stochastic Gradient Boosting and XGBoost models in the testing stage are shown.

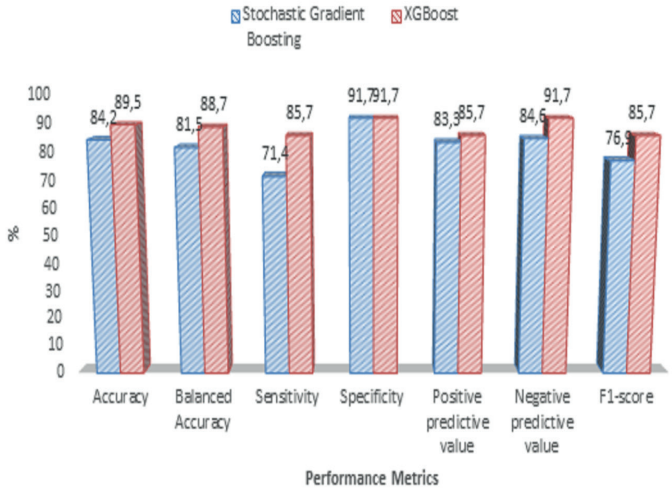

Figure 1. In the testing phase, performance metric values acquired from Stochastic Gradient Boosting and XGBoost models

When the performance metrics obtained from the XGBoost and Stochastic gradient boosting models, which are used in the modeling phase in the current study, are examined, the XGBoost model has classified the ovarian cancer data in the best way.

The variable importance values for the variables associated with the output variable obtained as a result of the XGBoost model are given in Table 4. The graph of the variable importance values calculated as a result of the XGBoost model is given in figure 2.

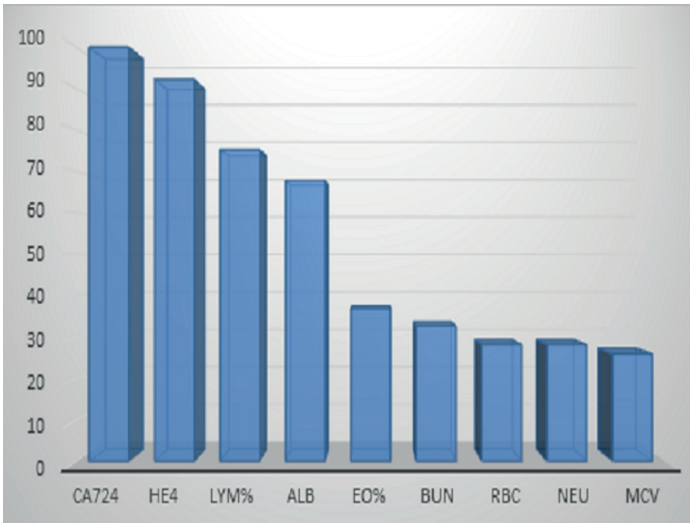

Figure 2. Variable importance values related with ovarian cancer

Table 4. The variable importance values obtained as a result of the XGBoost model

The Variables Related With Ovarian Cancer	XGBoost
CA72-4	100
HE4	92.27
LYM%	75.246
ALB	67.789
EO%	36.836
BUN	32.721
RBC	28.23
NEU	28.181
MCV	25.983

Discussion

The inability to detect the disease in its early stages and the frequency with which it recurrent make ovarian cancer the most lethal form of gynecological cancer [18,19]. Due to the lack of effective tools to detect the disease at an early stage, the disease is often diagnosed at a late stage, which can affect the overall prognosis and result in death [20]. Early-stage (I, II) ovarian cancer patients have a 5-year survival rate of over 90%, whereas late-stage (III, IV) ovarian cancer patients have a 5-year survival rate of about 20-40%. Despite breakthroughs in treatment, individuals with ovarian cancer have an overall 5-year survival rate of 45%, making it the second most fatal gynecological tumor [21-23].

When the survival rates are examined, there is a need to determine effective procedures for the diagnosis of ovarian cancer, where the survival rate is low and the diagnosis can usually be made at advanced stages. The importance of studies to be carried out in order to reveal the markers associated with the disease that can support the diagnosis in the early stages of the disease and reduce the effects of the disease with effective treatment before the prognosis progresses is increasing. Machine learning applications have become increasingly popular and applied in the diagnosis of diseases and clinical decision support systems in recent years. With machine learning techniques, which are frequently applied in the field of health, diseases detection at an early stage and revealing the factors affecting the disease are carried out [24,25]. For this reason, in the current study, the most appropriate method for classification of ovarian cancer using XGBoost and Stochastic Gradient Boosting models, which are ML methods, and the predictive factors associated with the diagnosis of the disease as a result of the classification were revealed with variable importance values.

Among the performance criteria in the test stage obtained from the XGBoost model that has the best classification result, accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1-score were obtained as 89.5%, 88.7%, 85.7%, 91.7%, 85.7%, 91.7%, and

DOI: 10.5455/medscience.2022.09.207

85.7%, respectively. Successful results for the diagnosis of ovarian cancer, and according to the variable importance obtained as a result of the model, the variables most associated with the diagnosis were CA72-4, HE4, LYM%, ALB, EO%, BUN, RBC, NEU, and MCV, respectively.

When the variable importance results are examined, it is the variable CA72-4 that explains the target variable the most. In addition, according to the results of the statistical analysis, the CA72-4 variable differed in terms of the categories of the target variable, and it was found to be at a higher level in patients with ovarian cancer than in patients with benign ovarian tumors. In a study, it was shown that the CA72-4 value is at high levels in ovarian cancer patients [26]. In another study, it was reported that CA72-4 was positive in ovarian cancer patients [27]. The second variable that most explains the target variable according to the variable significance values is the HE4 parameter. Human Epididymis Protein 4 (HE4) is a new biomarker being studied for the diagnosis of ovarian malignant tumors. According to the results obtained in the current study, the HE4 value increased in patients with ovarian cancer. In one study, it was shown that the HE4 parameter is found to be overexpressed in ovarian tumors, particularly endometrioid ovarian cancer [28]. In another study, it was reported that HE4 increased in ovarian cancer patients [29]. When the studies for the LYM % parameter, which is another important variable, were examined, it was found that it was associated with ovarian cancer [30]. In studies, the ALB parameter has been associated with the survival of ovarian cancer patients and has been said to be a determinant of survival [31,32]. In one study, it was determined that a high NEU parameter was associated with poor survival for ovarian cancer [33]. In a meta-analysis study, 12 studies were examined and it was shown that the NEU parameter is reliable and satisfactory for predicting the prognosis of patients with ovarian cancer. And it was concluded that it is associated with survival [34].

With the machine learning models applied in this study, a prediction model has been developed to diagnose ovarian cancer. The results obtained revealed effective parameters that can diagnose ovarian cancer with high accuracy. With the parameters determined as a result of the modeling, the clinician will be able to simplify and facilitate the decision-making process for the diagnosis of ovarian cancer.

Conflict of interests
The authors declare that there is no conflict of interest in the study.

Financial Disclosure
The authors declare that they have received no financial support for the study.

Ethical approval
Ethics committee approval was not obtained because the dataset used in the present study was open access.

References

1. Momenimovahed Z, Ghoncheh M, Pakzad R, et al. Incidence and mortality of uterine cancer and relationship with Human Development Index in the world. Cukurova Med J. 2017;42:233-40.
2. Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2018;68:394-424.
3. Coburn S, Bray F, Sherman M, Trabert B. International patterns and trends in ovarian cancer incidence, overall and by histologic subtype. International journal of cancer. 2017;140:2451-60.
4. Yoneda A, Lendorf ME, Couchman JR, Multhaupt HA. Breast and ovarian cancers: a survey and possible roles for the cell surface heparan sulfate proteoglycans. Journal of Histochemistry & Cytochemistry. 2012;60:9-21.
5. Badgwell D, Bast Jr RC. Early detection of ovarian cancer. Disease markers. 2007;23:397-410.
6. Lowe KA, Chia VM, Taylor A, et al. An international assessment of ovarian cancer incidence and mortality. Gynecologic oncology. 2013;130:107-14.
7. Chornokur G, Amankwah EK, Schildkraut JM, Phelan CM. Global ovarian cancer health disparities. Gynecologic oncology. 2013;129:258-64.
8. Momenimovahed Z, Tiznobaik A, Taheri S, Salehiniya H. Ovarian cancer in the world: epidemiology and risk factors. International journal of women's health. 2019;11:287-99.
9. Lu M, Fan Z, Xu B, et al. Using machine learning to predict ovarian cancer. International Journal of Medical Informatics. 2020;141:104195.
10. Polikar R. Ensemble learning. Ensemble machine learning: Springer; 2012. p. 1-34.
11. Akman M, Genç Y, Ankarali H. [Random Forests Methods and an Application in Health Science]. Turkiye Klinikleri J Biostat. 2011;3:36.
12. Lu M, Fan Z, Xu B, et al. Using machine learning to predict ovarian cancer. International journal of medical informatics. 2020;141:104195.
13. Wang J, Li P, Ran R, et al. A short-term photovoltaic power prediction model based on the gradient boost decision tree. Appl Sci. 2018;8:689.
14. Dikker J. Boosted tree learning for balanced item recommendation in online retail 2017.
15. Salam Patrous Z. Evaluating XGBoost for user classification by using behavioral features extracted from smartphone sensors [Master Thesis]: KTH Royal Institute of Technology; 2018.
16. Friedman JH. Stochastic gradient boosting. Computational statistics & data analysis. 2002;38:367-78.
17. Lawrence R, Bunn A, Powell S, Zambon M. Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. Remote sensing of environment. 2004;90:331-6.
18. Rojas V, Hirshfield KM, Ganesan S, Rodriguez-Rodriguez L. Molecular characterization of epithelial ovarian cancer: implications for diagnosis and treatment. International journal of molecular sciences. 2016;17:2113.
19. Bhatt P, Vhora I, Patil S, et al. Role of antibodies in diagnosis and treatment of ovarian cancer: Basic approach and clinical status. Journal of controlled release. 2016;226:148-67.
20. Chandra A, Pius C, Nabeel M, et al. Ovarian cancer: Current status

and strategies for improving therapeutic outcomes. Cancer medicine. 2019;8:7018-31.

21. Zeppernick F, Meinhold-Heerlein I. The new FIGO staging system for ovarian, fallopian tube, and primary peritoneal cancer. Archives of gynecology and obstetrics. 2014;290:839-42.

22. Schiavone MB, Herzog TJ, Lewin SN, et al. Natural history and outcome of mucinous carcinoma of the ovary. American journal of obstetrics and gynecology. 2011;205:480. e1-. e8.

23. Howlader N, Krapcho M, Miller D, et al. SEER Cancer Statistics Review, 1975-2014, based on November 2016 SEER data submission, posted to the SEER web site. Bethesda, MD, National Cancer Institute. 2017.

24. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future healthcare journal. 2019;6:94.

25. Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nature biomedical engineering. 2018;2:719-31.

26. Anastasi E, Manganaro L, Granato T, et al. Is CA72-4 a Useful Biomarker in Differential Diagnosis between Ovarian Endometrioma and Epithelial Ovarian Cancer? Disease Markers. 2013;35:984641.

27. Kobayashi H. [The clinical usefulness of serum CA72-4 analysis in patients with ovarian cancer]. Nihon Sanka Fujinka Gakkai Zasshi. 1989;41:585-9.

28. Drapkin R, Von Horsten HH, Lin Y, et al. Human epididymis protein 4 (HE4) is a secreted glycoprotein that is overexpressed by serous and endometrioid ovarian carcinomas. Cancer research. 2005;65:2162-9.

29. Anastasi E, Giovanna Marchei G, Viggiani V, et al. HE4: a new potential early biomarker for the recurrence of ovarian cancer. Tumor Biology. 2010;31:113-9.

30. Prodromidou A, Andreakos P, Kazakos C, Vlachos DE, Perrea D, Pergialiotis V. The diagnostic efficacy of platelet-to-lymphocyte ratio and neutrophil-to-lymphocyte ratio in ovarian cancer. Inflammation Research. 2017;66:467-75.

31. Asher V, Lee J, Bali A. Preoperative serum albumin is an independent prognostic predictor of survival in ovarian cancer. Medical oncology. 2012;29:2005-9.

32. Parker D, Bradley C, Bogle S, et al. Serum albumin and CA125 are powerful predictors of survival in epithelial ovarian cancer. BJOG: An International Journal of Obstetrics & Gynaecology. 1994;101:888-93.

33. Williams KA, Labidi-Galy SI, Terry KL, et al. Prognostic significance and predictors of the neutrophil-to-lymphocyte ratio in ovarian cancer. Gynecologic oncology. 2014;132:542-50.

34. Chen G, Zhu L, Yang Y, et al. Prognostic role of neutrophil to lymphocyte ratio in ovarian cancer: a meta-analysis. Technology in cancer research & treatment. 2018;17:1533033818791500.