International Journal of Notural and Engineering Sciences
Uluslararası Doğa ve Mühendislik Bilimleri Dergisi
E-ISSN: 2146-0086, 11 (2): 14-22, 2017, www.nobel.gen.tr

Identification of the Tumor Markers in Ovarian Cancer using Different Data Mining Methods

Sema YILDIRIM[*1], Müşerref HOROZOĞLU[2], Hakan IŞIK[3]

[1] Computer Engineering, Graduate School of Natural Sciences, Selcuk University,
[2] Information Technology Engineering, Graduate School of Natural Sciences, Selcuk University,
[3] Department of Electrical and Electronics Engineering, Faculty of Technology, Selcuk University, Konya, Turkey

*Corresponding Author
E-mail: semayildirim@selcuk.edu.tr

Received: May 15, 2017
Accepted: July 30, 2017

Abstract

Ovarian cancer that continues the most common lethal gynecological cancer is the fourth cause of death from cancer among women in industrialized countries. Early detection of ovarian cancer is difficult because of typically diagnosed at late stage. Therefore, early detection and define the identifier has great contribute to improve clinical outcomes. In this study, we searched the best identifier(s) in early diagnosis as well as the best data mining methods by using ovarian cancer dataset that were taken from Selcuk University, Faculty of Medicine. The experimental results show that while some identifiers that include Cancer Antigen 125 (CA125), lesion 1, 2, 3 and mural lesion is the most important identifier as individual basis, combination of CA125 and lesions are very significant clinical indicators. We can say that CA125 is not considerable identifier by alone and it should be used with the other identifier although commonly used in the diagnosis of cancer. In addition to this, the classification tree method achieved the highest success in classifying ovarian cancer data, with a 92.31% success rate in both malign and benign data.

Keywords: Ovary, cancer, tumor markers, classifier, data mining.

INTRODUCTION

Ovarian cancer (OC) is the most lethal gynecological cancer [1] and is the second most common form of gynecological cancer and first cause of death from gynecological malignancy in the western hemisphere [2–4]. OC is one of most common causes of cancer-related death in women worldwide, accounting for approximately 3% of all new cancer patients in 2009 [5]. Early diagnosis is still not possible because ovaries have settled deep in the ovaries of the female pelvis and the etiology of the disease is not fully known. Therefore, OC still continues to be a major problem in gynecological cancers. Many factors may be a risk in OC thanks to studies is determined at worldwide. There are a lot of risk factors for OC such as family status, age, menstruation, menopause, pregnancy, infertility, breast-feeding, daily living habits, socio-demographic characteristics [6–10]. In addition to these, this research suggests that the risk factors identified in the results was identified the differences between countries. Therefore, the detection of OC risk factors and identifiers is required basis separately for each country. Thus, the development of community-specific strategies and taken measure will be possible by determining early diagnosis and diagnostic process. In this context, as a method for early diagnosis of OC, pelvic examination, CA125 tumor biomarker and transvaginal ultrasound are among the most frequently used methods. However, none of these tests are not sufficient for the early diagnosis of cancer when especially the CA125 tumor biomarkers thought to increase the outside OC. This screening is recommended only for high risk groups because of the fact that routine screening and mass screening is unlikely that conditions in Turkey [11].

There are some researches for finding the best biomarkers and detecting between stage and grade of OC. While Su [12] summarized the most recent serum biomarkers and clinical applications of biomarkers for the early detection and treatment monitoring of OC and discussed the algorithms for predicting the risk of OC, Colak [13] showed that Human Epididymis Protein 4 (HE4) might be a better tumor marker than CA125 in the diagnosis of postmenopausal endometrial cancer. On the other hand, Einhorn [14] in a study carried out on patients who underwent screening over age 40 because of adnexal mass CA125 is not sufficient diagnosis with alone because of low sensitivity and if possible, it should be complemented by using Doppler measurements. Moore [15] identified a panel of complementary biomarkers in order to construct a multiple marker panel that could be used to aid in the triage of patients with a pelvic mass to appropriate centers for surgery.

The above-mentioned studies in the early diagnosis of OC and cannot be said to definitive and conclusive in the determination of appropriate biomarkers. According to data that morbidity and mortality rates are very closer of IARC in 2012, it is expected to consider a major contribution for new early OC diagnosis studies in this area. There is no doubt that early detection and diagnosis of OC is very important because the diagnosis of OC is usually made at late stage. Therefore, it is very important to determine the identifier(s) that will use for diagnosis and early detection. In this study, we searched the best identifier(s).

The rest of this research is organized as follows: the

related materials and methods are described in Section 2. The performed experiments and obtained experimental results has been explained in Section 3. Finally, we summarized the most relevant conclusions and discussion of this work in Section 4, which is then followed by the future work has been explained in Section 5.

MATERIALS and METHODS

Tumor markers

Performed studies for the determination various risk factors that might play a role in OC were performed and continues to be performed. Risk factors and protective factors in OC are given in Table 1.

Table 1. Risk factors and protective factors in ovarian cancer [16].

Risk factors	Protective factors
Age	
Family history	Multiparity
BRCA1 mutations	Use of oral contraceptives
	Hysterectomy
BRCA2 mutations	Tubal ligation
	Lactation
LYNCH II/HNPCC	
Infertility	
Nulliparity	
Late menopause	
Early menarche	
Increased CA125 level	
Smoking	
Asbestosiz and talc	

Tumor biomarkers that are molecules can be measured in blood or body fluids for diagnosis of cancer that are generated by the cancer and its environment, screening or treatment of monitored. In addition, other information about the patient (age, menopause and family history) can be used as a marker for diagnosis [17]. Over time, a large number of markers have been investigated in terms of its role in OC [15]. Diagnosis of OC is largely based on symptoms, imaging, and laboratory biomarkers. Overall, more than 200 potential biomarkers differentially expressed in OC have been identified [18]. However, no single marker has been found useful for the diagnosis of OC. Increased sensitivity (SEN) and specificity (SPE) for the diagnosis of OC are observed when multiple markers are used in combination [12].

CA125 is the most commonly used markers between these tumor markers in cancer screening and diagnosis. CA125 is a high molecular weight glycoprotein and it has identified by Bast et al. using a mouse monoclonal antibody in 1981[19]. The SEN and the SPE levels of CA125 is 95% and 43.3% when used alone[15]. Therefore, the CA125 combination studies are carried out with other tumor markers.

Data selection

In this study, the 39 female subjects who suffer because of a mass in the abdomen and admitted to Selcuk University, Faculty of Medicine, Department of Obstetrics and Gynecology (Non-Invasive Clinical Research Ethics Committee No. 2012/238) were dealt as retrospectively. Some demographic characteristics such as age, menopausal status, blood group of patients, CA125 values before surgery, imaging parameters such as computed tomography were obtained from the files of patients. Taken from abdominal masses that are taken after surgeries belonging to all the patients were sent pathological examination, it has been reported to be precisely benign or malign of the mass after some results that include performed pathological examination. Fifteen markers and its features form this data are given in Table 2.

Encoding is the process of converting data into a format required for a number of information processing needs. Therefore, we have converted some markers such as location mural thickness etc. into numbers. However, the some markers that include numerical expressions such as age, CA125, lesions, mural lesion and gravid weren't encoded because of already numerical. According to pathological results, 0 value is defined benign tumor, 1 value is defined malign tumor to classify. These markers given in Table 2 are used by means of some classification methods to determine the most appropriate identifier(s) for diagnosis of OC. Thus, markers and relationships between them that are important for early detection of OC were obtained.

Data Mining Methods

Support Vector Machine

In today's machine learning applications, SVM [20] are considered a must try—it offers one of the most robust and accurate methods among all well-known algorithms. It has a sound theoretical foundation, requires only a dozen examples for training, and is insensitive to the number of dimensions. In addition, efficient methods for training SVM are also being developed at a fast pace.

In a two-class learning task, the aim of SVM is to find the best classification function to distinguish between members of the two classes in the training data. The metric for the concept of the "best" classification function can be realized geometrically. For a linearly separable dataset, a linear classification function corresponds to a separating hyperplane $f(x)$ that passes through the middle of the two classes, separating the two. Once this function is determined, new data instance x can be classified by simply testing the sign of the function $f(x_n)$; x_n belongs to the positive class if $f(x_n)>0$. Because there are many such linear hyperplanes, what SVM additionally guarantee is that n the best such function is found by maximizing the margin between the two classes. Intuitively, the margin is defined as the amount of space, or separation between the two classes as defined by the hyperplane. Geometrically, the margin corresponds to the shortest distance between the closest data points to a point on the hyperplane. Having this geometric definition allows us to explore how to maximize the margin, so that even though there are an infinite number of hyperplanes, only a few qualify as the solution to SVM [21].

Classification Tree

The TREES module computes classification and regression trees. Classification trees include those models in which the dependent variable (the predicted variable) is categorical. Classification trees are parallel to discriminant analysis and algebraic classification methods [22]. Classification trees are an increasingly popular from of multistage or sequential decision rules. Tree classifiers provide some significant advantages over more traditional nonparametric classifiers. Classification trees easily accommodate data from all measurement scales (i.e., nominal, ordinal, interval, and ratio scales) and make no distributional assumptions [23]. As a class of machine

Table 2. Some of markers that play a role in determining of OC and the properties of these markers for this study

No	Clinical Markers	Attributes Definition and Encoding
1	Age	Among 21 and 80 years of age
2	CA125	Among 5.4 and 5000 U/mL
3	Location	0 The mass on the right side 1 The mass on the left side 2 The mass on both sides. (Refers to the area where the lesions)
4 5 6	The size of lesion 1 The size of lesion 2 The size of lesion 3	The value of the three-dimensional size of the lesion was made into a one-dimensional number by multiply with each other.
7	Mural thickness	0 Thin wall thickness 1 Thick wall thickness
8	Septum	0 Without septum 1 No septum contrast enhancement 2 There septum contrast enhancement
9	Contour status of the lesion	0 The plain 1 Lobular
10	Distribution of lesion	0 Homogeneous distribution 1 Heterogeneous distribution
11	Mural Lesion	0 1
12	Mural characteristics	0 No solid contrast involvement 1 There are solid contrast involvement 2 Cystic 3 Solid cystic
13	Blood Type	1 ARh+ 2 BRh+ 3 AB+ 4 ORh+ 0 Indefinite
14	Gravid	Among 0 and 11
15	Menopause	0 No menopausal status 1 There are menopausal status

learning algorithms, classification trees automatically select variables and their hierarchical structure is capable of detecting non-additive interactions between variables without explicit specification [24].

Naïve Bayes

Given a set of objects, each of which belongs to a known class, and each of which has a known vector of variables, our aim is to construct a rule which will allow us to assign future objects to a class, given only the vectors of variables describing the future objects. Problems of this kind, called problems of supervised classification, are ubiquitous, and many methods for constructing such rules have been developed. One very important one is the naive Bayes method—also called idiot's Bayes, simple Bayes, and independence Bayes. This method is important for several reasons. It is very easy to construct, not needing any complicated iterative parameter estimation schemes. This means it may be readily applied to huge data sets. It is easy to interpret, so users unskilled in classifier technology can understand why it is making the classification it makes. And finally, it often does surprisingly well: it may not be the best possible classifier in any particular application, but it can usually be relied on to be robust and to do quite well. General discussion of the naive Bayes method and its merits are given in [25,26].

Logistic Regression

Logistic regression is used to classify cases into the most likely category [27]. It is a standard for predicting binary, binomial and multinomial outcomes. Since the response variable is discrete, linear regression cannot be directly used for modeling. Instead, rather than predicting the point estimate of the event, it predicts the odds of its occurrence. In a two-class problem, odds greater than 50% would assign the case to the desired event (designated as "1") and to non-event (designated as "0") otherwise.

While a powerful modeling tool, logistic regression assumes that the log odds of the response variable are linearly related to the predictor variables. This might render the explanation of predictor coefficients difficult. High-Performance (HP) logistic regression completes model selection in seconds or minutes. This allows the user to include more variables, explore their effects, and finally, and build better models. Some features of HP logistic regression include variable selection, weighted and group analysis, and modeling capabilities for unordered multinomial data.

RESULTS

According to International Agency for Research on Cancer (IARC) that is connected World Health Organization (WHO), 2400 incidence and 1588 mortality was determined in Turkey, in 2012. The value of mortality is high despite value incidence of OC is lower than in other diseases as shown in Table 3. Although OC occurrence less than the probability of breast cancer is often results in death. Estimated incidence, mortality and 5-year prevalence of some diseases rates that belong to female patients in Turkey

Table 3. Estimated incidence, mortality and 5-year prevalence: women (GLOBOCAN 2012 (IARC) Section of Cancer Surveillance)

Cancer	Incidence			Mortality			5-year prevalence		
	Number	(%)	ASR (W)	Number	(%)	ASR (W)	Number	(%)	ASR (W)
Stomach	4182	6.7	10.9	3577	10.8	9.3	5349	3.5	19.1
Breast	15230	24.5	39.1	5199	15.7	13.4	52360	34.0	186.9
Ovary	2400	3.9	6.3	1588	4.8	4.2	5816	3.8	20.8
Thyroid	7076	11.4	17.8	682	2.1	1.9	26739	17.4	95.5

Incidence and mortality data for all ages. 5-year prevalence for adult population only. An age-standardized rate (ASR) (W) and proportion per 100.000.

is shown in Figure 1.

Incidence and mortality rates to getting OC that is closer are fifth rank, although it has the highest mortality rate for breast cancer. In this case, it can be said that OC carries a significant risk and must be resolved through early diagnosis and treatment.

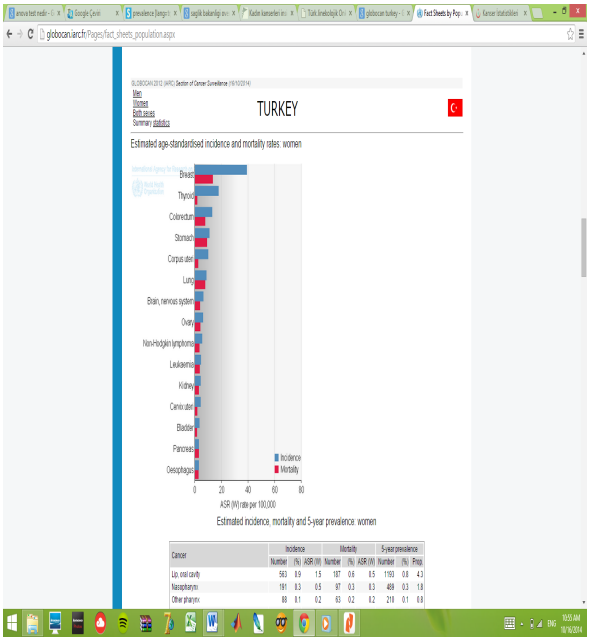


Fig. 1. Estimated age-standardized incidence and mortality rates: women

CA125 is the most commonly used serum biomarkers in patients with pelvic mass and it rises in 80% of patients with epithelial OC. CA125 value can also increase in benign gynecologic cases such as menstruation, endometriosis, cirrhosis and heart failure and other diseases [28], [29]. Moreover, serum CA125 values are changing with age, while it was detected high values in premenopausal women; it is seen to decrease with age in postmenopausal women [28–32]. Although CA125 is the most common tumor identifier in the diagnosis of OC, it is insufficient for the recognition of OC in early.

In this study, while whether benign or malign of a mass before preoperative is determined result of some investigations that include blood markers, demographic data and pathologic consequences, place on disease of the designated specified markers place and relationship between these markers were studied by using the data mining methods. Therefore, the data that include 39 women patients in the hospital were retrospectively obtained by scanning the recorded files of these patients. In that context, the data was transferred to a computer in some respects by summarizing, then data mining methods was carried out for classify. 15 identifiers that include serum CA125 before surgery and some demographic features such as septum, age, menopause and location were used as input parameters in the study. For the best model selection in the classification problems described, 10-fold cross-validation technique, which 10-fold cross-validation is the most common in data mining and machine learning, was used [33–35]. Cross-validation technique provides to seen and employ all of data. In this study, data was subjected to 10-fold cross-validation before running classification. After preprocessing data and cross-validation techniques we performed to classify on dataset.

SVM Regression, CT, NB and LR methods are accomplished, respectively. Firstly, the classification methods were carried out by using default parameters. Afterwards, optimum parameter values were explored by trial and error method. The optimum parameter values that obtained as a result of experiments are given shown in Table 4. The SEN, SPE, and ACC that ca be defined as (Equations 1, 2 and 3) are the commonly used parameters to evaluate the performance of the classification methods [36], [37].

$$SEN = \frac{P}{(P + N)} x100 \quad (1)$$

$$SPE = \frac{N}{(N + P)} x100 \quad (2)$$

$$ACC = \frac{P + N}{(N + N + P + N)} x100 \quad (3)$$

where *TP* and *TN* represent the total number of correctly detected true positive patterns and true negative patterns, respectively. The *FP* and *FN* represent the total number of erroneously positive patterns and erroneously negative patterns, respectively. The positive and negative patterns represent detected benign and malign OC, respectively.

The confusion matrix that includes TP, TN, FP and FN of each classification method for determining whether benign or malign tumor are numerically given in Table 5. While 0 value is defined benign tumor, 1 value is defined malign tumor result of pathological consequences in abdomen patients. In addition to these, the columns indicate the value of predicted condition positive (0) and negative (1); the rows indicate the condition positive (0) and negative (1).

Table 4. Classification methods and their optimum parameters

Classification Methods	Learning parameters / Settings
SVM	Type: Nu SVM regression Cost (C): 0.500 Complexity bound (nu): 0.500 Kernel: RBF, $e^{-0.0000*(x-y).(x-y)}$ Tolerance: 0.001 Normalize data: Yes
Classification Tree	Attribute selection: Information Gain Binarization: No binarization Pruning: 2 instances in leaves Recursively merge leaves with same majority class: Yes Pruning with m-estimate: m=2
Naïve Bayes	Probability estimation: Relative Frequency LOESS window size: 0.5 Number of points in LOESS: 100 Adjust classification threshold: No
Logistic Regression	Stepwise attribute selection: add at 10%, remove at 10% Imputation of known values: Classification/Regression trees

Table 5. Confusion Matrix for all classification methods

	Naïve Bayes		Logistic Regression		SVM Regression		Classification Tree	
	0	1	0	1	0	1	0	1
0	21	1	16	6	21	1	19	3
1	3	14	4	13	3	14	0	17
	24	15	20	19	24	15	19	20

According to the above table, while NB and SVM methods classified 21 of 22 cases that is not take OC diagnosis, it classified 14 of 17 cases that take OC diagnosis in a correct way that is named as TP and TN, respectively. On the other hand, while 16 cases that are TP of 22 non-cancer mass were classified, 6 cases that are FP of them were not classified as correctly after carrying out LR. In addition, LR is not realized with high classify for detection of OC patients. It can be said that LR is the worst classifier in all of classification methods as shown in Table 5. The CT method identified 19 of 22 cases that is not take OC diagnosis as TP, while it estimated 17 of 17 cases that take OC diagnosis as FP. We obtained that CT method is more successful than the other methods when it was compared with the other classification methods as shown in Table 5.

Statistical performances of all classification methods that are carried out in this study was computed. This statistical performance shown in Table 6 and 7, respectively, for benign and malignant tumors. According to the results in Table 6, CT method is the highest classification with ACC value that is 92.31% in identifying patients without OC.

Table 6. The result of test learners for benign tumors (Target class: 0)

	Classification Performance		
Classification Methods	**ACC (%)**	**SEN (%)**	**SPE (%)**
Naïve Bayes	89.74	95.45	82.35
Logistic Regression	74.36	72.73	76.47
SVM Regression	89.74	95.45	82.35
Classification Tree	**92.31**	86.36	100

Table 7. The result of test learners for malignant tumors (Target class: 1)

	Classification Performance		
Classification Methods	**ACC (%)**	**SEN (%)**	**SPE (%)**
Naïve Bayes	89.74	82.35	95.45
Logistic Regression	74.36	76.47	72.73
SVM Regression	89.74	82.35	95.45
Classification Tree	**92.31**	100	86.36

According to the results in Table 7, CT method is higher ACC (92.31%) value than the other methods as well as it is more successful than the other methods in the detection of non-cancer cases. On the other hand, the ACC of SVM regression and NB classifier are the same value with 89.74% values. At the same time, all classification methods performed the same values for patients who include both non-OC and OC. This study is a special operation on data that were obtained some patients in Turkey. Therefore, it can be considered that our study is a unique study in this area because of belongs to a community of Turkey. The aim of this research identify to effective markers because early diagnosis is very important. In that context, we investigated the relationship of all markers to each other's by performing various experiments.

First of all, we investigated the relationship of each marker alone to distinguish normal from cancer. The best of fifteen markers was found as a CA125, lesion 2, 3 and mural lesion. The one on one relationship between some markers and the disease is shown in Figure 2. While, x coordinate shows the OC, y coordinate shows markers. Furthermore, 0 value (blue) shows the distribution of benign mass, while 1 value (red) show the distribution of OC in these figures except the relationship between disease and markers. We can say that these markers play very important role to distinguish normal from cancer. According to results of classification, especially CA125, lesion 1, lesion 2, lesion 3, mural lesion and location markers that are belong to the

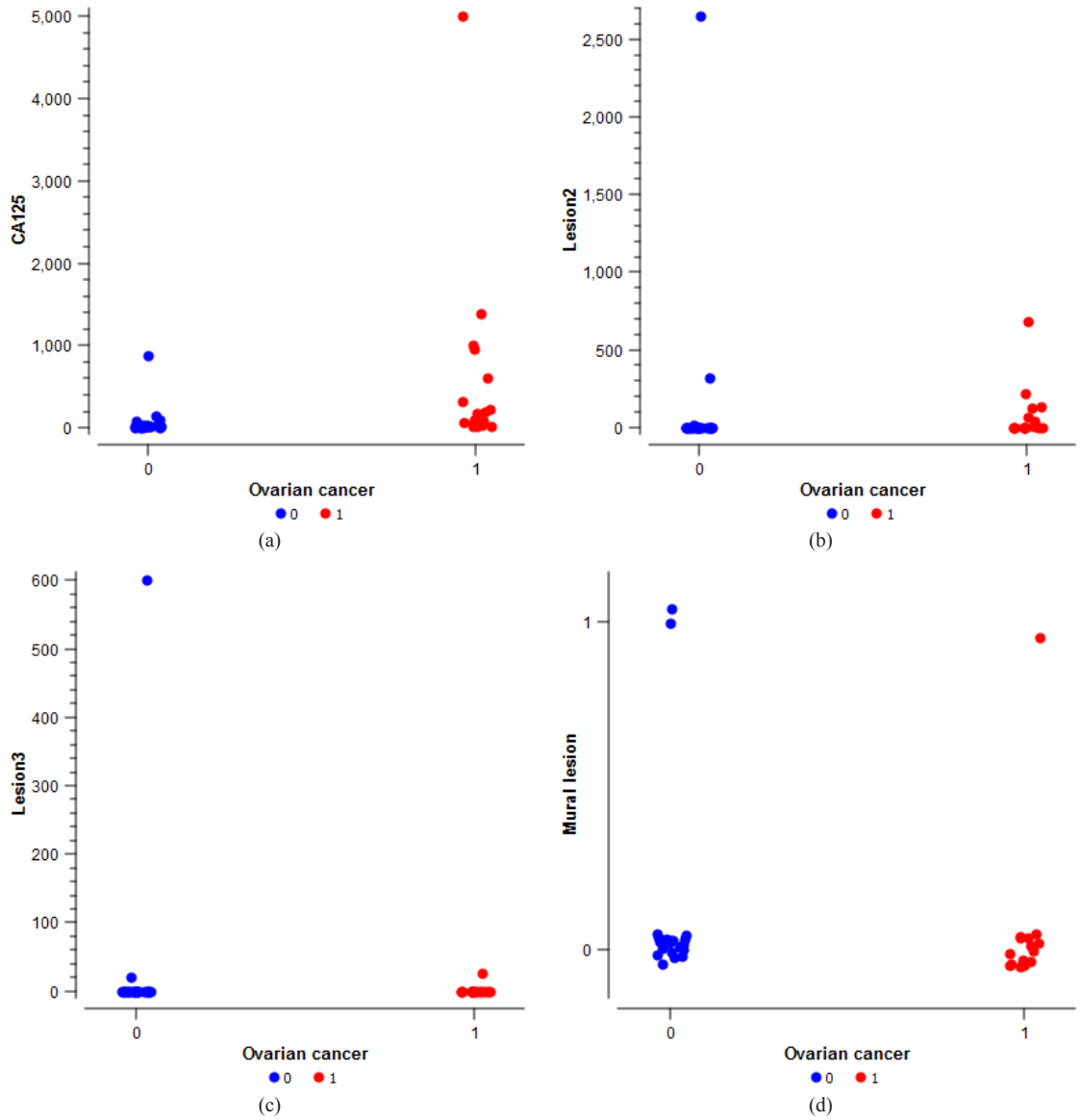


Fig. 2. The relationship of the best markers to distinguish benign tumor from cancer (a) CA125 and OC (b) lesion 2 and OC (c) lesion 3 and OC (d) mural lesion and OC.

15 markers in this dataset may predict the OC in a correct way. CA125 was seen as a major success on this data in the diagnosis of cancer cases. CA125 was found to be the marker of the highest association with cancer. What is more is that the direct relationship between lesion2, lesion 3 and mural lesion with the disease is seen high achievement. The all relationship with each other of markers (15) was studied individually. The relationship with each other of some markers is shown in Figure 3.

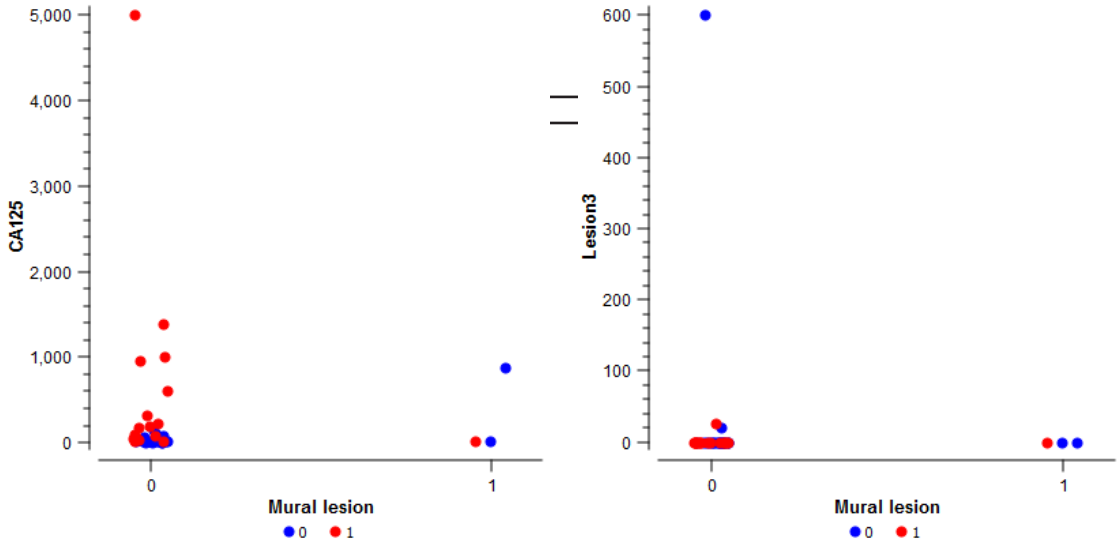
Some studies indicated that while OC is rare before the age of 30, it increases with increasing age [38] and is seen among furthermore the most common age 45-79 [39]. On the other hand, OC can be seen at any age although the incidence of cancer is frequently in the 57-64 for this data as shown in Figure 4.

It is also observed relationship that is seen as shown in Table 8 with each other among markers may use and four attributes are investigated for this purpose. While the relationship among CA125, lesion 1 and lesion is a significant, CA125, lesion 1 and lesion 2 is a significant after evaluating. As a result of the classification performed on the dataset, the highest correlation markers are given in Table 8.

Table 8. The multiple relationship among markers. (Finished evaluation: evaluated 1.247 projections in 0 min, 26 sec)

Rank	Score	Most Interesting Projections
1	75.01	CA125, Lesion 1, Lesion 3
2	74.40	CA125, Lesion 1, Lesion 2
3	70.99	Location, Mural thickness, contour status of the lesion, structure
4	69.63	CA125, lesion 2, lesion 1, distribution of the lesion
5	69.63	Location, CA125, Mural lesion, Gravid

In this study, CT method is carried out as the best



(a) (b)

Fig. 3. The relationship of the best markers combination of two markers to distinguish benign tumor from cancer (a) CA125 and mural lesion (b) mural lesion and lesion 3.

classifier method. Meanwhile, a CT is easier to read and understand by using flow chart symbols. The structure of CT is shown by using Classification Tree Graph (CTG) in Figure 5.

There are 14 nodes and 8 leaves in this tree structure. The values of majority class, majority class probability, and target class probability and number of instances are also shown in this figure in order to monitor the detail impact of these markers in OC. We can see easily the most important markers that are effective in the diagnosis of OC and at what they settle and at what they point. It can be said that CA125 biomarker has a major impact on the classification because of that the highest information gain parameter is settled to the top position in tree structure. Then, location and lesion 1 markers are settled to the second node of the tree because of information gain. While lesion 1 shows 14 OC of 19 cases, it shows 5 benign cases. In cases where CA125 small and equal 39.0, 10 subjects for left abdominal mass, 8 subjects for right abdominal mass and 2 subjects for bilateral mass are seen for location parameter. Consequently, the dataset was indicated graphically in the most optimum way by means of CTG, thus it can be observed most appropriate markers and what these markers are effective for diagnosing disease and cancer detection because CT method provides optimum performance in different classification methods.

According to finding in the left branch of CTG, except OC cases that CA125 is higher from 39.0, there is no OC in situation of CA125 value is greater than 39.0, lesion 1 value is greater than 1290.0 and smaller than 2106.0 (IF CA125 > 39.000 AND Lesion 1= in (1290.0, 2106.0 THEN 0). In this case, we say that lesion 1 is not carrying any cancer risk for range from 1290.0 to 2106.0. On the other side of branch, OC is not observed in some situation that include small and equal to 39.0 of the CA125 and the mass of bilateral. It is also observed that among 12.53 and 39.0 of the CA125 value, small and equal to 4.5000 of the gravid and left side of the abdominal mass (IF CA125 = in [12.53, 39.0] AND Gravid =<=4.500 AND Location = 1 THEN 1) for disease is except for right side of tree.

CONCLUSIONS

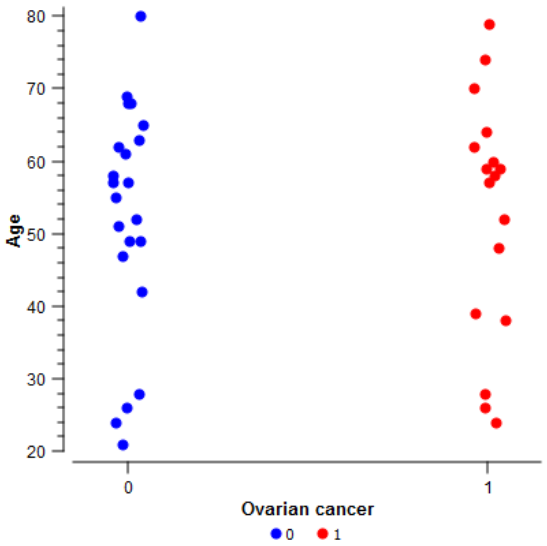


Fig. 4. The effect of age to distinguish benign tumor from cancer.

OC is one of the most dangerous type of cancer in gynecological cancers. Complete and definitive treatment cannot be still carried out because of absence of appropriate clinical indicators and diagnosed in advanced stages of cancer. Therefore, the early detection of OC studies are still continuing. According to the results of research and literature reviews based on information obtained, many parameters, biomarkers and drug combination were investigated whether appropriate or not for diagnosis and treatment of disease. However, tumor biomarkers and value of ultrasonography has not yet been determined in screening for OC [40]. Although one of the CA125 tumor biomarker, the actual purpose of CA125 diagnosed as determining follow-up to tumor and the response to treatment. Essentially an ideal tumor marker must the original of tumor and the level of marker must increase with tumor size [41]. CA125 that is serum biomarker may be higher in some patients who are

OC may higher the other patients who are not OC. CA125 is not a token used alone in the diagnosis of the disease in result of many performed studies. In addition to these study, our study that supports the other studies showed CA125 is not sufficient for cancer diagnosis by itself. Furthermore, in this work were found to be effective clinical indicators and relationship between them for diagnosis in addition to this biomarker. Accordingly, it is shown that the using CA125 with some markers that include mural lesion, lesions and location was found to be more effective than used alone in order to diagnosis OC subjects who are in Turkey.

We carried out different data mining methods for determining the best classifier. First of all, all classifiers were executed by using their default parameters. Afterwards, we found the optimum parameter values by trial and error method. While the highest ACC value (92.31%) was achieved CT method, the same ACC value (89.74%) was achieved SVM and NB classifier methods. There is no doubt that CT method is the best classifier for this dataset.

Furthermore, the all of blood type was observed positive values that include ARH+, BRH+ AB+ and 0RH+ when the entire dataset is analyzed. In this case, it can be search whether negative blood group play or not play a role in the early diagnosis of cancer with further studies that will be carried out by using new dataset. Risk factors that causes cancer was considered, the blood types are expected to contribute to investigate for identifying of OC subjects in Turkey.

ACKNOWLEDGEMENTS

We are grateful to Selcuk University, Faculty of Medicine, Department of Obstetrics and Gynecology (Non-Invasive Clinical Research Ethics Committee No. 2012/238) and Dr. Nasuh Utku Doğan for their support and contributions. The authors declare that they have no competing interests.

REFERENCES

[1] Agarwal R, Kaye SB. Ovarian cancer: strategies for overcoming resistance to chemotherapy. Nature Reviews Cancer. 2003;3(7):502–16.

[2] Ozols RF. Recurrent ovarian cancer: evidence-based treatment. Journal of clinical oncology. 2002;20(5):1161–3.

[3] Parkin DM, Pisani P, Ferlay J. Estimates of the worldwide incidence of 25 major cancers in 1990. International journal of cancer. 1999;80(6):827–41.

[4] Harries M, Gore M. Part I: Chemotherapy for epithelial ovarian cancer–treatment at first diagnosis. The lancet oncology. 2002;3(9):529–36.

[5] Jemal A, Siegel R, Ward E, Hao Y, Xu J, Murray T, et al. Cancer statistics, 2008. CA: a cancer journal for clinicians. 2008;58(2):71–96.

[6] Morrow CP, Curtin JP, Townsend DE, Blessing JA. Synopsis of gynecologic oncology. Vol. 282. Churchill Livingstone Philadelphia; 1998.

[7] Yarbro C, Wujcik D, Gobel BH. Cancer nursing: principles and practice. Jones & Bartlett Learning; 2010.

[8] Gershenson DM, McGuire WP, Gore M, Quinn MJ, Thomas G. Gynecologic cancer: controversies in management. Churchill Livingstone; 2004.

[9] Berek JS. Epithelial ovarian cancer. Practical gynecologic oncology. 2000;3:457–522.

[10] Arvas M, Göker B. Germ hücreli over tümörleri. Jinekolojik Onkoloji. 2002;3:245–55.

[11] Ayhan A, Başaran M. Epitelyal over kanserleri. Jinekolojik Onkoloji. 2002;3:201–43.

[12] Su Z, Graybill WS, Zhu Y. Detection and monitoring of ovarian cancer. Clinica Chimica Acta. 2013;415:341–5.

[13] Abdullah Çolak, Over Kanserinin Erken Tanı ve Takibinde CA125 ve HE4'ün Sensitivitesi ve Spesifitesinin Karşılaştırılması, Hacettepe Üniversitesi, Tıp Fakültesi, Uzmanlık Tezi, 2013. 2013;2013.

[14] Einhorn N, Knapp RC, Bast RC, Zurawski VR. CA 125 assay used in conjunction with CA 15-3 and TAG-72 assays for discrimination between malignant and non-malignant diseases of the ovary. Acta Oncologica. 1989;28(5):655–7.

[15] Moore RG, Brown AK, Miller MC, Skates S, Allard WJ, Verch T, et al. The use of multiple novel tumor biomarkers for the detection of ovarian carcinoma in patients with a pelvic mass. Gynecologic oncology. 2008;108(2):402–8.

[16] Edmondson RJ, Monaghan JM. The epidemiology of ovarian cancer. International Journal of Gynecological Cancer. 2001;11(6):423–9.

[17] Burtis CA, Ashwood ER, Bruns DE. Tietz fundamentals of clinical chemistry. 2001;

[18] Lokshin AE. The quest for ovarian cancer screening biomarkers: are we on the right road? International Journal of Gynecological Cancer. 2012;22:S35–40.

[19] Bast RC, Feeney M, Lazarus H, Nadler LM, Colvin RB, Knapp RC. Reactivity of a monoclonal antibody with human ovarian carcinoma. The Journal of clinical investigation [Internet]. 1981;68(5):1331–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/7028788%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC370929

[20] Vladimir VN, Vapnik V. The nature of statistical learning theory. Springer Heidelberg; 1995.

[21] Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, et al. Top 10 algorithms in data mining. Knowledge and Information Systems. 2008;14(1):1–37.

[22] Wilkinson L. Classification and regression trees. Systat. 2004;11:35–56.

[23] Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. CRC press; 1984.

[24] Clark LA, Pregibon D, Chambers JM, Hastie TJ. Statistical models in S. chapter Tree-Based Models. 1992;377–419.

[25] Domingos P, Pazzani M. On the optimality of the simple Bayesian classifier under zero-one loss. Machine learning. 1997;29(2–3):103–30.

[26] Fix E, Hodges Jr JL. Discriminatory analysis-nonparametric discrimination: consistency properties. DTIC Document; 1951.

[27] Olson DL, Chae BK. Direct marketing decision support through predictive customer response modeling. Decision Support Systems. 2012;54(1):443–51.

[28] Alagoz T, Buller RE, Berman M, Anderson B, Manetta A, DiSaia P. What is a normal CA125 level? Gynecologic oncology. 1994;53(1):93–7.

[29] Bon GG, Kenemans P, Verstraeten R, van Kamp GJ, Hilgers J. Serum tumor marker immunoassays

in gynecologic oncology: establishment of reference values. American journal of obstetrics and gynecology. 1996;174(1):107–14.

[30] Maggino T, Gadducci A, D'addario V, Pecorelli S, Lissoni A, Stella M, et al. Prospective multicenter study on CA 125 in postmenopausal pelvic masses. Gynecologic oncology. 1994;54(2):117–23.

[31] Malkasian Jr GD, Knapp RC, Lavin PT, Zurawski Jr VR, Podratz KC, Stanhope CR, et al. Preoperative evaluation of serum CA 125 levels in premenopausal and postmenopausal patients with pelvic masses. Discrimination of benign from malignant disease. American journal of obstetrics and gynecology. 1988;159(2):341–6.

[32] Einhorn N, Bast Jr RC, Knapp RC, Tjernberg B, Zurawski Jr VR. Preoperative evaluation of serum CA 125 levels in patients with primary epithelial ovarian cancer. Obstetrics & Gynecology. 1986;67(3):414–6.

[33] Acharya UR, Sree SV, Alvin APC, Suri JS. Use of principal component analysis for automatic classification of epileptic EEG activities in wavelet framework. Expert Systems with Applications. 2012;39(10):9072–8.

[34] Kaya Y, Uyar M, Tekin R, Yıldırım S. 1D-local binary pattern based feature extraction for classification of epileptic EEG signals. Applied Mathematics and Computation. 2014;243:209–19.

[35] Wang D, Miao D, Xie C. Best basis-based wavelet packet entropy feature extraction and hierarchical EEG classification for epileptic detection. Expert Systems with Applications. 2011;38(11):14314–20.

[36] Li S, Zhou W, Yuan Q, Geng S, Cai D. Feature extraction and recognition of ictal EEG using EMD and SVM. Computers in biology and medicine. 2013;43(7):807–16.

[37] Bajaj V, Pachori RB. Classification of seizure and nonseizure EEG signals using empirical mode decomposition. Information Technology in Biomedicine, IEEE Transactions on. 2012;16(6):1135–42.

[38] Altchek A, Deligdisch L, Kase N. Diagnosis and management of ovarian disorders. Academic Press; 2003.

[39] Morris CR, Sands MT, Smith LH. Ovarian cancer: predictors of early-stage diagnosis. Cancer Causes & Control. 2010;21(8):1203–11.

[40] Berek JS, Hacker NF, Hengst TC. Practical gynecologic oncology. Vol. 204. Lippincott Williams & Wilkins Philadelphia; 2000.

[41] MONTAG TW. Tumor markers in gynecologic oncology. Obstetrical & gynecological survey. 1990;45(2):94–105.