Bingöl Üniversitesi
İktisadi ve İdari Bilimler Fakültesi Dergisi
Bingöl University
Journal of Economics and Administrative Sciences
Cilt/Volume: 7
Sayı/Issue: Prof. Dr. Muammer ERDOĞAN
Anısına Kongre Özel Sayısı
Yıl/Year: 2023, s. 225-241
DOI: 10.33399/biibfad.1253338
ISSN: 2651-3234/E-ISSN: 2651-3307
Bingöl/Türkiye
Makale Bilgisi /Article Info
Geliş/Received: 19/02/2023 **Kabul/ Accepted:** 25/03/2023

AN APPLICATION OF THE FEATURE SELECTION METHOD BASED ON PAIRWISE CORRELATION FOR DIAGNOSIS OF OVARIAN CANCER WITH MACHINE LEARNING[1]

Makine Öğrenmesi İle Yumurtalık Kanseri Tanısı İçin İkili Korelasyona Dayalı Öznitelik Seçim Yöntemi Uygulaması

Hülya BAŞEĞMEZ*

Abstract

Many machine learning classification problems have high dimensions, and efficient and effective feature selection algorithms are needed to determine the relatively essential features in the dataset. Gene data is often preferred in feature selection applications because it contains many features due to its structure. In addition, it is known from studies in the literature that gene selection plays a significant role in cancer detection. One of the cancer types with very high treatment success in the early period is ovarian cancer. For this purpose, it was aimed to select genes with high descriptiveness in cancer diagnosis by using the ovarian cancer dataset, which is a publicly available dataset. In this study, the feature selection method based on pairwise correlation, which is very new in the literature, was used for classification. Firstly, a feature selection application was made, and 38 genes with the highest cancer

[1] This study is a continuation of the study titled "A Feature Selection Application Using Pairwise Correlation Method in Ovarian Cancer Diagnosis" presented at the 2nd International Congress of Economics and Administrative Sciences held on 19-20 January 2023.

* Asst. Prof. Dr., Beykent University, Faculty of Economics and Administrative Sciences, Department of Management Information Systems, hulyabasegmez@gmail.com, ORCID: https://orcid.org/0000-0001-7768-1666

descriptors were determined. Then, the classification process was carried out using eight different classification algorithms. After the classification process, the lowest success was for the Extra Tree classification algorithm (with 96.44% accuracy), while the highest was for the Multi-Layer Perceptron, Stochastic Gradient Descent, Logistic Regression, and Support Vector Machine (with 100% accuracy). Although there are many studies on feature selection in the literature, this study is the first application of the current method. In this sense, it is thought to contribute to the literature.

Keywords: Feature selection, gene selection, ovarian cancer.

JEL Codes: I10; I19.

Öz

Makine öğrenmesi sınıflandırma problemlerinin birçoğu yüksek boyuta sahip olup, veri kümesindeki özniteliklerden görece önemli olanların belirlenmesi amacıyla verimli ve etkili değişken seçim algoritmalarına ihtiyaç vardır. Gen verileri de yapısı gereği çok sayıda değişken içerdiği için değişken seçim uygulamalarında sıklıkla tercih edilir. Ayrıca gen seçimi kanser tespitinde büyük rol oynadığı literatürde yer alan çalışmalardan bilinmektedir. Erken dönemde tedavi başarısı oldukça yüksek olan kanser türlerinden birisi de yumurtalık (ovarian) kanseridir. Bu amaçla çalışmada erişime açık bir veri kümesi olan yumurtalık kanseri veri kümesi kullanılarak, kanser teşhisinde yüksek tanımlayıcılığa sahip genlerin seçilmesi amaçlanmıştır. Çalışmada, sınıflandırma için literatürde çok yeni olan ikili korelasyona (pairwise correlation) dayalı öznitelik seçim yöntemi kullanılmıştır. Uygulamada, ilk olarak değişken seçim uygulaması yapılmış ve kanser tanımlayıcılığı en yüksek olan 38 gen belirlenmiştir. Daha sonra sekiz farklı sınıflandırma algoritması kullanılarak sınıflandırma işlemi yapılmıştır. Sınıflandırma işlemi sonrası en düşük sınıflandırma başarısı %96.44 doğruluk değeri ile Ekstra Ağaç sınıflandırma algoritması için gerçekleşirken, en yüksek sınıflandırma başarısı ise %100 doğruluk değeri ile Çok Katmanlı Algılayıcı, Stokastik Gradyan İniş, Lojistik Regresyon ve Destek Vektör Makinesi sınıflandırıcıları kullanılarak elde edilmiştir. Literatürde değişken seçimi konusunda yapılan çok sayıda çalışma olmasına rağmen bu çalışma mevcut yöntemle ilgili yapılan ilk uygulama özelliği taşımaktadır. Bu anlamda literatüre katkı sağlayacağı düşünülmektedir.

Anahtar Kelimeler: Öznitelik seçimi, gen seçimi, yumurtalık kanseri.

JEL Kodları: I10; I19.

1. Introduction

In most machine learning classification problems, datasets usually have a high dimension. For this reason, it is necessary to use efficient and effective methods to determine features' relative importance and remove unnecessary features. One of the classification problems in high-dimensional data (especially for cancer classification) is gene expression classification. Gene data is frequently used in feature selection applications because it contains many features due to its structure.

It is well established that microarray datasets are crucial for early cancer diagnosis. However, classification becomes challenging due to these huge datasets' numerous irrelevant and/or redundant features. In addition, due to removing features from the existing structure that is not required, feature (gene) selection has a significant role in this discipline.

Thanks to the correct analysis, the genes responsible for the disease can be easily identified, and thus field experts can act quickly to treat the disease. Analysis of gene expression data is a great opportunity for individuals with the disease, as treatment becomes easier once the genes most likely cause the disease are identified (Ghosh et al., 2019). It is known from the studies in the literature that the remaining genes after the selection process have higher cancer descriptors (Başeğmez et al., 2021).

Features (genes) that are most relevant are called biomarkers, as their expression values can indicate cancer occurrence. Therefore, finding biomarkers is a significant research problem (Ghosh et al., 2019). Irrelevant features increase the accuracy and computation time of the cancer detection system. In other words, not all genetic variations (genes) cause cancer. Only a relatively small portion of all genes are cancer-causing (Özkan ve Erol, 2015). This indicates the importance of feature selection, eliminating irrelevant and/or redundant data in the dataset, and making detection faster and more accurate (Sezer ve Çakir, 2022).

According to results from the American Cancer Society's Center for Cancer Statistics, ovarian cancer is one of the gynecological cancers with the highest mortality rate and the most common among women (The American Cancer Society, 2023). More women die from ovarian cancer than any other cancer in the female reproductive system. The reasons for the high mortality rate of ovarian cancer are the asymptotic and latent growth of the tumor, the delay of symptoms, and the lack of appropriate monitoring tools due to the diagnosis of the disease in advanced stages.

Ovarian cancer risk over a woman's lifetime is about 1/78, and the chance of dying from ovarian cancer in her lifetime is about 1/108. In addition, for any form of ovarian cancer, the 5-year relative survival rate of individuals with this disease is around 49% in developed countries, while this rate drops to 36% in developing countries (Globocan, 2020). Early diagnosis is crucial for ovarian cancer cases and all types of diseases. If the disease can be diagnosed and treated at Stage 1, this rate may increase to 94%. Unfortunately, only 20% of ovarian cancer cases are diagnosed at Stage 1 (The American Cancer Society, 2023).

Therefore, it can be said that increasing the number of women diagnosed with Stage I disease directly impacts the mortality and economy of this cancer without the need to change the approaches used in surgical or chemotherapy treatments. According to a study conducted by (Chin et al., 2020), the average economic burden of an ovarian cancer patient in the treatment and control processes is around 140,000 dollars by 2020. According to the report published by (Globocan, 2020), ovarian cancer is the seventh most common cancer among women and the eighth with the highest mortality rate in the world in 2020. In addition, there are more than 313.00 newly diagnosed patients in 2020, and it is estimated that the number of women diagnosed with ovarian cancer worldwide will increase by 42% in 2040, reaching approximately 442,721 people (Globocan, 2020).

In this study, the pairwise correlation approach, which is very new in the literature for classification and recommended by (Jimenez et al., 2022), is applied for gene selection in ovarian cancer detection. The Ovarian cancer dataset obtained by (Zhu et al., 2007) was used in this study. As a result, the results obtained will be compared with other studies in the literature, and the current method's success will be measured. Although there are many studies on gene selection to cancer detection in the literature, it is thought that this study is the first application of the current method and will contribute to the literature.

The structure of this study is as follows. In section II, some basic concepts and approaches related to feature selection and classification, which are widely used in the literature and will be included in our study, are explained in detail. In section III, the obtained analysis results are given. In section IV, we draw some final conclusions.

2. Material and Method

2.1. Ovarian Dataset

The ovarian cancer dataset was produced as a result of research conducted by (Zhu et al., 2007), a faculty member of the Department of Computer Science and Software Engineering at Shenzhen University of China, and is a dataset open to researchers. The dataset has two classes, 253 observations and 15154 genes (features). The current observation group involves 162 diseased and 91 healthy people.

2.2. Feature Selection for Classification

Feature selection algorithms for classification are examined in two main groups: wrappers and filters, according to their operation. It is a simple approach to use the accuracy of the classifier as a performance measure in feature selection algorithms. In this case, a classifier should be created to achieve the highest possible classifier accuracy, and the most suitable features for the classifier should be selected. This method is called the "wrapper model" (Liu ve Motoda, 1998). Wrapper models use the classification error rate as the evaluation criterion (Blum ve Langley, 1997). Due to the high workload and longtime requirement in wrapper models, studies on indirect performance measurements have been directed. In general, models in which measures such as distance, information, and correlation are used for feature selection are called "filtering models". Filtering models use an evaluation measure other than classification accuracy, which allows measurement indirectly. Filtering models do not guarantee that they will achieve the best subset of features, while wrapper methods have the risk of overfitting (Ladha ve Deepa, 2011).

The main filtering feature selection methods are described below:

- Feature selection based on information gain
- Feature selection based on gain ratio
- Feature selection based on symmetric uncertainty
- Correlation-based feature selection
- Consistency-based feature selection
- Feature selection based on pairwise correlation

2.3. Multivariate Feature Ranking Based on Pairwise Correlation

The pairwise Correlation method is suggested by (Jimenez et al., 2022). This method is inspired by the correlation-based feature selection (CFS) method. The CFS method was developed by (Hall, 1999). This method uses a correlation-based evaluation function to rank the features (Hall, 1999). The CFS algorithm has the advantage of creating more accurate models and, in most cases, halving the number of selected features over other feature selection methods.

The CFS method uses the $\Phi_D(S)$ function to measure the quality of an S subset of k attributes in the D dataset, with $1 \leq k \leq n$. $\Phi_D(S)$ function is defined as follows:

$$\Phi_D\ (S) = \frac{\left(k.\sigma_D^C\right)}{\sqrt{k + k.(k-1).\sigma_D^f}} \tag{1}$$

Here, σ_D^C is the mean of the correlation values between the class attribute and each feature in S, σ_D^f is the mean correlation between each of the possible pairs of features $\binom{k}{2}$ in S.

In other words, the denominator represents the excess between features, and the numerator represents the predictive power of a group of features. Since the CFS method works with categorical features, it is ensured that the features are categorized as a result. Here, the approach generally used in the literature for the discretizing process was proposed by (Fayyad ve Irani, 1993).

The pairwise correlation method proposed by Jimenez et al. (2022) is defined as follows for a feature $i, i \in \{1,2,\dots,n\}$:

$$\Phi_A^D(i) = \sum_{\substack{j \in \{1,\dots,n\} \\ j \neq i}} \Phi_D\ (\{i,j\}) \tag{2}$$

$\Phi_D(\{i,j\})$ is the merit value of the subset formed by the features i and j for all $j = 1,\dots,n$values such that $i \neq j$ (Eq. (1)). That is, the $\Phi_A^D(i)$ merit value of the $i-$th attribute is the total value of the attribute subsets created by the i and other attributes. Here, features with a high correlation with class and a low correlation with other attributes are selected. Since the evolution of each feature considers all other features along with the class, the dual correlation method is a multivariate feature ranking method. Therefore, it also considers the intersections between the features.

Jimenez et al. (2022) have officially added the pairwise correlation method as a package called PairwiseCorrelationAttributeEval to the WEKA platform.

2.4. Classification

In the literature, a variety of classification methods are employed. In this study, eight different classification algorithms were used. These algorithms are briefly described below.

2.4.1. Stochastic Gradient Descent

Gradient descent (GD) is a widely used technique in deep learning and machine learning as an optimization tool. The gradient of a given function is its slope, which is its derivative. GD is an iterative process for determining parameter values to minimize the cost function. Instead of using all samples collectively, the Stochastic Gradient Descent (SGD) technique uses a small number of randomly chosen samples from the dataset to minimize the cost function (Bottou, 1991).

2.4.2. Extra Trees Classifier

The Extra Trees algorithm creates an ensemble of the unpruned decision or regression trees in accordance with the traditional top-down procedure. However, it has two main differences over other tree-based ensemble methods. The first of these variations is that it allocates the nodes by choosing cut points completely randomly, and the second is that it uses the entire learning instance (rather than a bootstrap copy) to grow the trees (Geurts et al., 2006).

2.4.3. Multi-layer Perceptron (MLP)

Multi-layer Perceptron, which is one of the fundamental approaches used in machine learning, has three layers: the input layer, the hidden layer, and the output layer. Interconnected adaptive processing units called neurons make up each layer. A neuron is a general computational unit (Ozer et al., 2004). This unit has a structure that takes m inputs and produces a single output. Parameter connection weights distinguish the outputs of neurons. Each neuron in a layer is coupled to every neuron in the top layer with varying weights. After the data coming to the input layer is multiplied by the weight values, it is transmitted to the hidden layer (Baxter et al., 2011). A transfer function is used to gather the multiplication results in the hidden layer (Yesilkaya et al., 2022).

2.4.4. C4.5 Decision Trees

Decision trees are a classification algorithm consisting of roots, nodes, branches, and leaves. Each attribute in the dataset represents a node. In the literature, there are many different application areas depending on the structure of the tree. The reason why decision trees are frequently preferred in applications is the faster training and test process. Also, its outputs are easily interpreted (Hart et al., 2000). C4.5 is a decision tree classification algorithm proposed by Ross Quinlan and is performed in two steps. These steps are creating and pruning the tree (Quinlan, 1987).

2.4.5. Random Forest

Random Forests (RF), as it is used today, was proposed by Breiman in 2001 (Breiman, 2001). This classifier creates a model consisting of multiple decision trees and trains this model (Biau ve Scornet, 2016). In the classifier model, each internal node represents the feature in the corresponding instance, each branch represents the test result, and the leaf node represents the class label. Each decision tree is built using values drawn at random from the input data. The random forest method produces a model made up of several trees. Decision trees (a forest) in this classifier are permitted to expand without going over their limit.

2.4.6. Logistic Regression

Logistic regression is a method for estimating possible classes from categorically distributed variables (Hart et al., 2000). LR is actually a curve fitting, that is, a regression method, but the desired values are not real-valued numbers. They are discrete numbers (i.e., categories) (Yesilkaya et al., 2022).

2.4.7. Support Vector Machine

The main advantage of the support vector machine classifier is that it transforms into an optimization problem to solve the classification problem. Thanks to this transformation, it is possible to obtain a faster solution by reducing the calculation process (Yesilkaya et al., 2022). It is a machine learning-based classification technique developed by Vapnik, built on support vector machines, statistical learning theory, and structural risk minimization techniques (Boser et al., 1992). The main purpose of this classifier is to create the optimal decision boundary, the hyperplane, which can distinguish the data points labeled in different ways and maximize the distance between the support vectors.

2.4.8. Naive Bayes

The naive Bayes algorithm is an application of Bayes' theorem, a family of simple probability-based "classifiers” (Bayes ve Price, 1763; Hart et al., 2000). This algorithm assumes that all features are conditionally independent and contribute equally to the output when the class label is given (Murphy, 2006).

3. Findings

In this section, we have discussed the classification stage and feature selection studies. The ovarian cancer dataset used in this study was downloaded as an Arff file. The feature types in this dataset are as follows: 15154 numeric (continuous) attributes and one categorical attribute named "Class". When the Class feature was investigated at, it was discovered that there were 253 observations overall (162 diseased people and 93 healthy people) and two classes. 38 genes (features) are selected by using feature selection based on the pairwise correlation method. Table 1 displays selected genes.

Table 1: Selected genes using feature selection based on the pairwise correlation method

Number	Gene	Number	Gene	Number	Gene
1	MZ0.022435711	14	MZ246.12233	27	MZ435.46452
2	MZ2.8234234	15	MZ246.41524	28	MZ435.85411
3	MZ2.9824141	16	MZ246.70832	29	MZ466.77814
4	MZ11.165473	17	MZ247.00158	30	MZ555.74254
5	MZ25.49556	18	MZ247.295	31	MZ674.57738
6	MZ222.41828	19	MZ261.58446	32	MZ2665.3973
7	MZ243.49401	20	MZ261.88643	33	MZ4003.6449
8	MZ244.07686	21	MZ262.18857	34	MZ4906.9617
9	MZ244.66041	22	MZ288.82415	35	MZ5270.3916
10	MZ244.95245	23	MZ417.73207	36	MZ6803.0344
11	MZ245.24466	24	MZ434.29682	37	MZ15097.351
12	MZ245.53704	25	MZ434.68588	38	MZ17012.128
13	MZ245.8296	26	MZ435.07512		

In this study, a classification study was conducted using the Weka 3.8.6 program. First, we applied eight different classification methods, including 10-fold cross-validation. These classifiers are Extra Tree, Multi-LayerPerceptron, Stochastic Gradient Descent, C4.5, Random Forest, Logistic Regression, Support Vector Machine, and Naive Bayes. The results obtained after the classification processes using these classification algorithms on the Ovarian dataset are shown in Table 2.

When Table 2 is examined, it can be seen that the highest accuracy value (100%) is obtained for MLP, SGD, LogReg, and SVM classifiers. In addition, it is seen from this table that the accuracy value obtained for the other four classifiers is close to 100%. As well as it was observed that the F-measure values were around 1.

Table 2: Classifier performances

	Classifier	Accuracy	ROC Area	F-Measure	Kappa	MAE
1	ExtraTree	96.44%	0.965	0.965	0.9233	0.0356
2	MLP	100%	1	1	1	0
3	SGD	100%	1	1	1	0
4	C4.5	97.63%	0.983	0.976	0.9488	0.0244
5	RF	98.81%	1	0.988	0.9743	0.0308
6	LogReg	100%	1	1	1	0
7	SVM	100%	1	1	1	0
8	NB	98.81%	0.999	0.988	0.9742	0.0119

Since the number of samples was only 15154 and the error distribution between the classes was not uneven, it was unnecessary to perform data balancing before the classification study. This data set, containing 15154 columns and 253 rows, was classified faster using pairwise correlation, a feature selection method. A higher success was attained using only 38 features compared to the success acquired from the complete data set.

Eight different classification prediction algorithms were applied with 10-fold cross-validation on the reduced dataset of selected genes. 100% predictive accuracy was obtained in four of the eight estimation algorithms applied.

Finally, different classifier performances in classification and diagnosis problems were compared with the studies in the literature. The accuracy values of both this study and other studies are summarized in Table 3. While making these comparisons, the comparison classifier performances table created by Yesilkaya et al. (2022) was developed. In comparison, it was seen that the results obtained were better than many studies in the literature.

Table 3: Comparison of classifier performance results

Study	Classifier	Optimisation	Validation	ACC
Ubaidillah et al. (2013)	Support Vector Machine	–	70%+30%	0.64
Belciug ve Gorunescu, (2018)	Adaptive Single-HiddenLayer Feedforward Neural Network	–	10-fold	0.72
Ubaidillah et al. (2013)	Multi-Layer Perceptron	–	70%+30%	0.78
Belciug ve Ivanescu, (2019)	Bayesian Initialization of Extreme Learning Machine	–	10-fold	0.80
Q. Liu et al. (2017)	Support Vector Machine	Procrustes	10-fold	0.95
Elhoseny et al. (2019)	Optimal Recurrent Neural Networks + Self Organizing Map	–	90%+10%	0.96
Kilicarslan et al. (2020)	Convolutional Neural Network	–	70%+30%	0.98
Rahman et al. (2019)	Multi-Layer Perceptron	Taguchi	70%+30%	0.98
Talbi et al. (2008)	Support Vector Machine	Particle Swarm Optimization	10-fold	0.99
Al-Murad ve Hossain, (2021)	Multi-Layer Perceptron	Integrated Feature Selection	70%+30%	0.99
Yesilkaya et al. (2022)	Logistic Regression	Multi-Dimensional Scaling	70%+30%	0.99
Yesilkaya et al. (2022)	Stochastic Gradient Descent	Locally Linear Embedding	70%+30%	0.99
This study	Multi-Layer Perceptron	Pairwise Correlation	10-fold	1.00
This study	Support Vector Machine	Pairwise Correlation	10-fold	1.00
This study	Stochastic Gradient Descent	Pairwise Correlation	10-fold	1.00
This study	Logistic Regression	Pairwise Correlation	10-fold	1.00
Başeğmez et al. (2021)	Support Vector Machine	Correlation Based Feature Selection	5-fold	1.00
Demircioğlu ve Bilge, (2015)	Support Vector Machine	FisherScore	40%+60%	1.00
Y. Liu (2012)	Support Vector Machine	Principal Component Analysis + Linear Discriminant Analysis	3-fold	1.00

Yeşilbaş ve Güven, (2021)	Multi-Layer Perceptron	Principal Component Analysis	90%+10%	1.00

4. Results and Discussion

Studies have been conducted to benefit from gene data in cancer diagnosis for many years. Early diagnosis is of great importance in reducing the rate of death due to ovarian cancer, one of the most common types of cancer in women. Machine learning and data mining techniques in computer science are used to help doctors diagnose this type of cancer early.

In this study, the pairwise correlation-based feature selection approach proposed by (Jimenez et al., 2022) was applied to the ovarian cancer dataset, which has been predicted with many different approaches in the literature. 38 genes were selected with pairwise correlation-based feature selection from the Ovarian dataset consisting of 15154 features.

As a result, it has been seen that the pairwise correlation-based feature selection approach proposed can give highly successful results in gene selection studies. It has been said that MLP, SGD, LogReg, and SVM classifiers stand out compared to others with 100% accuracy.

As seen from Table 3, the results obtained are better than the other studies in the literature that we examined within the scope of this study. Based on these results, it is thought that using the pairwise correlation-based feature selection approach in classification studies on different gene datasets would be helpful.

Ethics Statement: The author declare that ethical rules are followed in all preparation processes of this study. In case of detection of a contrary situation, BİİBFAD Journal does not have any responsibility, and all responsibility belongs to the authors of the study.

Acknowledgement: I would like to thank the Editorial Board of BİİBFAD Journal for their intense interest and efforts and the referees for their contribution.

Reference

Al-Murad, A., & Hossain, M. F. (2021). An integrated feature selection method for neural network to classify ovarian cancer. *In 2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI)*, 1–6.

Başeğmez, H., Sezer, E., & Erol, Ç. S. (2021). *Optimization for Gene Selection and Cancer Classification*. 21. https://doi.org/10.3390/proceedings2021074021

Baxter, C. W., Zhang, Q., Stanley, S. J., Shariff, R., Tupas, R.-R., & Stark, H. L. (2011). Drinking water quality and treatment: the use of artificial neural networks. *Canadian Journal of Civil Engineering*, *28*(S1), 26–35. https://doi.org/10.1139/L00-053

Bayes, T., and Price, R. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, *53*, 370-418.

Belciug, S., & Gorunescu, F. (2018). Learning a single-hidden layer feed forward neural network using a rank correlation-based strategy with application to high dimensional gene expression and proteomic spectra datasets in cancer detection. *Journal of Biomedical Informatics*, *83*, 159–166.

Belciug, S., & Ivanescu, R. C. (2019). A Bayesian framework for extreme learning machine with application for automated cancer detection. *Annals of the University of Craiova, Mathematics and Computer Science Series*, *46*(1), 189–202.

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*(2), 197–227. https://doi.org/10.1007/S11749-016-0481-7/FIGURES/4

Blum, A. L., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, *97*, 245–271.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A Training Algorithm for Optimal Margin Classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*.

Bottou, L. (1991). Stochastic gradient learning in neural networks.

Proceedings of Neuro-Nimes, *91*(8), 1–12.

Breiman, L. (2001). Random Forests. *Machine learning*, 45, 5-32.

Chin, L., Hansen, R. N., & Carlson, J. J. (2020). Economic burden of metastatic ovarian cancer in a commercially insured population: A retrospective cohort analysis. *Journal of Managed Care and Specialty Pharmacy*, *26*(8), 962–970. https://doi.org/10.18553/JMCP.2020.26.8.962/ASSET/IMAGES/SMALL/FIG1.GIF

Demircioğlu, H., & Bilge, H. (2015). Yumurtalık kanseri veri kümesindeki gen ifadelerinin veri madenciliği ile analizi. *Marmara Fen Bilimleri Dergisi*, *27*(4), 125–134.

Elhoseny, M., Bian, G.-B., Lakshmanaprabu, S. K., Shankar, K., Singh, K. A. K., & Wu, W. (2019). Effective features to classify ovarian cancer data in internet of medical things. *Computer Networks*, *159*, 147–156.

Fayyad, U. M., & Irani, K. B. (1993). Multi-lnterval Discretization of Continuous-Valued Attributes for Classification Learning. *Thirteenth International Joint Conference on Artificial Intelligence*, 1022–1027.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*, 3–42.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*(1), 3–42. https://doi.org/10.1007/s10994-006-6226-1

Ghosh, M., Adhikary, S., Ghosh, K. K., Sardar, A., Begum, S., & Sarkar, R. (2019). Genetic algorithm based cancerous gene identification from microarray data using ensemble of filter methods. *Medical and Biological Engineering and Computing*, *57*(1), 159–176. https://doi.org/10.1007/s11517-018-1874-4

Globocan. (2020). *World Ovarian Cancer Coalition*. Ovarian Cancer Key Stats. https://worldovariancancercoalition.org/about-ovarian-cancer/key-stats/ Date accessed: 24/01/2023.

Hall, M. A. (1999). *Correlation-based Feature Selection for Machine*

Learning [The University of Waikato]. https://www.cs.waikato.ac.nz/~mhall/thesis.pdf

Hart, P. E., Stork, D. G., & Duda, R. O. (2000). *Pattern classification*. Hoboken: Wiley.

Jiménez, F., Sánchez, G., Palma, J., Miralles-Pechuán, L., & Botía, J. A. (2022). Multivariate feature ranking with high-dimensional data for classification tasks. *IEEE Access*, 10, 60421-60437.

Kilicarslan, S., Adem, K., & Celik, M. (2020). Diagnosis and classification of cancer using hybrid model based on relieff and convolutional neural network. *Medical Hypotheses, 109577.*

Ladha, L., & Deepa, T. (2011). Feature selection methods and algorithms. *International Journal on Computer Science and Engineering*, *3*, 1787–1797.

Liu, H., ve Motoda, H. (1998). Feature Selection for Knowledge Discovery and Data Mining. Içinde H. Liu ve H. Motoda (Ed.), *Feature Selection for Knowledge Discovery and Data Mining*. Springer, New York. https://doi.org/10.1007/978-1-4615-5689-3

Liu, Q., Gu, Q., & Wu, Z. (2017). Feature selection method based on support vector machine and shape analysis for high-throughput medical data. *Computers in Biology and Medicine*, *91*, 103–111.

Liu, Y. (2012). Dimensionality reduction and main component extraction of mass spectrometry cancer data. *Knowledge-Based Systems*, *26*, 207–215.

Murphy, K. P. (2006). Naive bayes classifiers. *University of British Columbia*, *18*(60), 1–8.

Ozer, M., Isler, Y., & Ozer, H. (2004). A computer software for simulating single-compartmental model of neurons. *Computer Methods and Programs in Biomedicine*, *75*(1), 51–57. https://doi.org/10.1016/J.CMPB.2003.08.002

Özkan, Y., & Erol, Ç. (2015). *Biyoenformatik DNA mikrodizi: veri madenciliği*. Papatya Yayıncılık Eğitim.

Quinlan, J. R. (1987). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

Rahman, M. A., Muniyandi, R. C., Islam, K. T., & Rahman, M. M. (2019). Ovarian cancer classification accuracy analysis using 15-neuron artificial neural networks model. *2019 IEEE Student Conference on Research and Development (SCOReD)*, 33–38.

Sezer, E., & Çakir, Ö. (2022). A Feature Selection Application for Classification: A Banking Application. *Dicle University Journal of Economics and Administrative Sciences*, *12*(24), 480–498.

Talbi, E. G., Jourdan, L., Garcia-Nieto, J., & Alba, E. (2008). Comparison of population based meta heuristics for feature selection: Application to microarray data classification. *In 2008 IEEE/ACS International Conference on Computer Systems and Applications*, 45–52.

The American Cancer Society. (2023). Key Statistics for Ovarian Cancer. https://www.cancer.org/cancer/ovarian-cancer/about/key-statistics.html

Ubaidillah, S. H. S. A., Sallehuddin, R., & Ali, N. A. (2013). Cancer detection using artifical neural network and support vector machine: A comparative study. *Jurnal Teknologi*, *65*(1), 73–81.

Yeşilbaş, D., & Güven, A. (2021). Kütle Spektrometresi Verileri Kullanılarak Yumurtalık Kanserinin Yapay Sinir Ağlarıyla Sınıflandırılması. *Çukurova Üniversitesi Mühendislik Fakültesi Dergisi*, *36*(3), 781–790.

Yesilkaya, B., Perc, M., & Isler, Y. (2022). Manifold learning methods for the diagnosis of ovarian cancer. *Journal of Computational Science*, *63*. https://doi.org/10.1016/j.jocs.2022.101775

Zhu, Z., Ong, Y.-S., & Dash, M. (2007). Markov Blanket-Embedded Genetic Algorithm for Gene Selection. *Pattern Recognition*, *40*(11), 3236–3248. https://csse.szu.edu.cn/staff/zhuzx/Datasets.html