Ovarian Cancer Prediction Using PCA, K-PCA, ICA and Random Forest

PCA, K-PCA, ICA ve Random Forest Kullanarak Yumurtalık Kanserinin Tahmini

Asiye Sahin[1], Nermin Ozcan[1], Gokhan Nur[1]
[1]Department of Biomedical Engineering, Iskenderun Technical University, Hatay, Turkey
{asiye.sahin, nermin.ozcan, gokhan.nur}@iste.edu.tr

***Abstract*— Ovarian cancer, which is the most common in women and occurs mostly in the post-menopausal period, develops with the uncontrolled proliferation of the cells in the ovaries and the formation of tumors. Early diagnosis is very difficult and in most cases, it is a type of cancer that is in advanced stages when first diagnosed. While it tends to be treated successfully in the early stages where it is confined to the ovary, it is more difficult to treat in the advanced stages and is often fatal. For this reason, it has been focused on studies that predict whether people have ovarian cancer. In our study, we designed a RF-based ovarian cancer prediction model using a data set consisting of 49 features including blood routine tests, general chemistry tests and tumor marker data of 349 real patients. Since the data set containing too many dimensions will increase the time and resources that need to be spent, we reduced the dimension of the data with PCA, K-PCA and ICA methods and examined its effect on the result and time saving. The best result was obtained with a score of 0.895 F1 by using the new smaller-sized data obtained by the PCA method, in which the dimension was reduced from 49 to 6, in the RF method, and the training of the model took 18.191 seconds. This result was both better as a success and more economical in terms of time spent during model training compared to the prediction made over larger data with 49 features, where no dimension reduction method was used. The study has shown that in predictions made with machine learning models over large-scale medical data, dimension reduction methods will provide advantages in terms of time and resources by improving the prediction results.**

Keywords—Dimension reduction, machine learning, ovarian cancer, random forest algorithm

***Özetçe*— Kadınlarda en sık rastlanan ve çoğunlukla menopoz sonrası dönemde ortaya çıkan yumurtalık kanseri, yumurtalıklardaki hücrelerin kontrol dışı çoğalması ve tümör oluşturması ile gelişir. Erken tanısı oldukça zordur ve çoğu durumda ilk tanı konduğunda ileri evrelerde olan bir kanser türüdür. Yumurtalık ile sınırlı olduğu erken evrelerde başarılı bir şekilde tedavi edilmeye yatkınken ileri evrelerde tedavisi daha zordur ve sıklıkla ölümcül olmaktadır. Bu nedenle kişilerin yumurtalık kanseri olup olmadığının tahminini yapan çalışmalar üzerine yoğunlaşılmıştır. Biz de çalışmamızda 349 gerçek hastaya ait kan rutin testi, genel kimya testi ve tümör belirteci verilerini içeren 49 özellikten oluşan veri setini kullanarak Random Forest tabanlı yumurtalık kanseri tahmin modeli tasarladık. Veri setinin çok fazla boyut içermesi harcanması gereken zaman ve kaynakları arttıracağı için PCA, K-PCA ve ICA yöntemleri ile verinin boyutunu azaltıp sonuca ve zaman tasarrufuna etkisini inceledik. Boyutun 49'dan 6'ya düşürüldüğü PCA yöntemi ile elde edilen daha küçük boyutlu yeni verinin RF yönteminde kullanılmasıyla, 0.895 F1 puanı ile en iyi sonuç elde edilmiştir ve modelin eğitimi 18.191 saniye sürmüştür. Bu sonuç, hiçbir boyut azaltma yönteminin kullanılmadığı dolayısıyla 49 özelliğe sahip daha büyük boyutlu veri üzerinden yapılan tahminden hem başarı olarak daha iyi hem de model eğitimi sırasında geçen zaman açısından daha tasarruflu olmuştur. Çalışma büyük boyutlara sahip medikal veriler üzerinden makine öğrenmesi modelleri ile yapılacak tahminlerde, boyut azaltma yöntemlerinin tahmin sonuçlarını iyileştirerek zaman ve kaynaklar açısından avantaj sağlayacağını göstermiştir.**

Anahtar Kelimeler—Boyut azaltma, makine öğrenmesi, yumurtalık kanseri, rastgele orman algoritması.

I. INTRODUCTION

The female reproductive system contains two ovaries, one on each side of the uterus. Ovarian cancer (OC) that starts in the ovaries is one of the most common types of cancer in women. It causes 152,000 deaths worldwide each year [1]. To determine whether the person has OC; physical examination, ultrasound and computed tomography scanning, and blood testing [2], [3]. The overall 5-year survival rate of diagnosed cases is approximately 40%. Most cases are diagnosed in stage 3 and stage 4, with a 5-year survival rate of 3-19% [1]. In addition, the 5-year recurrence rate in these patients with advanced stages is as high as 60–80% [4].

Studies have been conducted to evaluate the effect of various biomarkers such as carbohydrate antigen 125 (CA125), carbohydrate antigen 72-4 (CA72-4), human epididymis protein 4 (HE4)) and indices using them on OC. Moore et al. created a Risk Ovarian Malignancy Algorithm to evaluate the risk of OC according to HE4 and CA125 levels, taking into account the menopausal status in women with pelvic mass. The algorithm has successfully divided patients into high and low risk groups, and 93.8% of OC are correctly classified as high risk [5]. Jacobs et al. evaluated age, ultrasound score, menopausal status, clinical impression score, and CA125 level to see how best to distinguish patients with benign and malignant pelvic masses [6]. Anton et al. showed in their study that HE4 is a parameter with high general sensitivity for the evaluation of ovarian tumor [7]. Zhang et al developed a Linear Multi-Marker odel to differentiate benign ovarian tumors (BOT) and OC patients by combining HE4, CA125, progesterone and estradiol. Multi-marker models showed a significant improvement compared to CA125 or HE4 [8].

Machine Learning (ML), which makes inferences from existing data using mathematical and statistical methods, is a method paradigm that makes predictions about the unknown with these inferences and can evaluate many variables effectively [9]. However, the fact that medical data contains many variables increases the time and resources that should be spent when using ML methods. For this reason, Dimension Reduction (DR) techniques that reduce the dimension by detecting and removing non-essential components of the data or obtaining new features with a smaller size from the data provide a great advantage. Although size reduction sometimes causes a decrease in classification performance, it has been found to be very useful methods in terms of calculation time [10].

In our study, we reduced the dimension of a data set consisting of 49 features, including blood routine test, general chemistry test and tumor marker data collected from 349 patients, with various DR methods. Then we classified BOT and OC using data with reduced dimension with the Random forest algorithm, which is frequently used in the diagnosis of disease in the biomedical field [11], [12], [13], [14].

II. METHOD

A. Dataset

The OC dataset were obtained from Mendeley Data [15]. There are five supplementary datas (i.e. original raw data, list of biomarkers, imputed version of the training data, the raw training data, and the raw test data). The original raw data was selected for this research, the dataset contains 349 records and 50 attributes. There are 49 features that attribute in prediction of OC and one feature performs as the output or the predicted feature for the OC presence in a patient. The OC dataset involves a feature called 'TYPE' to indicate the diagnosis of OC in patients of labels, 0-1. In this case, 0 corresponds to the BOT and 1 represent OC patients. The data used in this study consists of several demographic and clinic measures such as blood routine test, results of tumor markers detection method and the general chemistry tests, collected from 178 patients of BOT and 171 suffering from OC. The details of the database can be found in Mingyang Lu [4].

B. Data Preprocessing

1) Missing Value

Missing data is a frequent issue in nearly all clinical studies, and it can have a considerable impact on the results that can be drawn from the datasets. Most attributes in the OC dataset have a minimal missing value ratio (less than 7%), with the exception of CA72-4 and Neutrophil Ratio percent (NEU), which have more than 65% missing. The features with the low missing value rate were imputed using the mean of theirs. The CA72-4 and NEU features have been removed from the dataset.

2) Data standardization

Medical datasets are made up of several characteristics, which are expressed by various data types. Several of the features are binary in nature, while others may be decimal or fractional. The values of the features can be diverse range, resulting in a bias toward selection of the particular attributes. Data standardization is the method of converting various data forms into a standardized format. The converted single format aids in the comparison and classification of data instances. As a preprocessing step, all of the inputs are standardized in this study.

C. Dimension Reduction

Real life data having too many dimensions (attributes) increases the time and resources we need to spend in all processes from data cleaning to model building. It also makes visualization just as difficult. DR approaches are used to overcome these problems. The main purpose of DR techniques is to minimize the size with the minimum loss in data content. In other words, it is the reduction in dimension by detecting and removing non-essential components of the data. In this study where we used a data set with 49 attributes, we used Principal Component Analysis (PCA), Kernel Principal Component Analysis (K-PCA) and Independent Component Analysis (ICA) methods for DR. These algorithms will be mentioned in subtitles.

1) PCA

PCA is one of the most common unsupervised learning methods used to reduce the dimension of a high dimensional data set. It creates a minimum number of variables holding the maximum variance of the distribution in the data to reduce the dimension with the least information loss. Because if a variable has the same value for each sample, it is an unnecessary variable and variables with the highest variance in the data must be found.

In PCA, the principal components of a data are obtained by calculating the eigenvalue and eigenvector of the covariance matrix after normalizing the data. Let the data be a matrix X of size nxm. X_i $(i\epsilon\{1,2,...,n\})$, represents the ith row of X data of size m. Eq. 1 shows the calculation of the mean value of the data, and Eq. 2 shows the calculation of the covariance matrix. m and C represent the dimension of the data and the covariance matrix, respectively.

$$\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i \tag{1}$$

$$C = \sum_{i=1}^{n}(X - \bar{X})(X - \bar{X})^T \tag{2}$$

The eigenvalues ($\curlywedge$) and eigenvectors (V) of the covariance matrix are calculated using the equations in Eq. 3.

$$\det(\curlywedge I - C) = 0, \quad (\curlywedge_k I - C)xV_k = 0 \tag{3}$$

The eigenvalues are ordered in ascending order and the eigenvectors corresponding to the largest eigenvalues are found. Projection of normalized data onto K eigenvectors produces reduced data [16].

2) K-PCA

Standard PCA allows linear dimension reduction only. However, if the data has more complex structures that may not be well represented in a linear subspace, standard PCA will not be very helpful. K-PCA is the nonlinear form of PCA that makes better use of the complex spatial structure of high dimensional properties [17]. It uses linear, polynomial, radial basis function, sigmoid and cosine kernel structures or multi-kernel structures in which these kernels are used together while doing DR.

3) ICA

ICA is a statistical method that tries to express multivariate data as linear combinations of independent components. Multivariate data is assumed to consist of a linear combination of a set of independent components (factors). The number of factors is generally taken to equal the number of variables. Let us show the data set consisting of p variables, each sampled at n points, with the Z matrix. In this case, the Z matrix in the ICA model is calculated as shown in Eq. 4.

$$Z = AY \tag{4}$$

A is the mixing matrix and Y is the source matrix containing the independent components. Both the mixing matrix and the source matrix are unknown. Both matrices are estimated by maximizing the statistical independence of the predicted components using only the Z data matrix. First, the A mixing matrix is estimated. Then, using the predicted matrix A, the Y matrix is obtained by using Eq. 5.

$$Y = A^{-1}Z \tag{5}$$

Thus, independent components are found and the data is expressed as the linear combination of these independent components [18].

D. Classification

RF is one of the most popular ML models because it can be applied to both regression and classification problems and gives good results. In order to find a solution to the overfitting problem of decision trees, which is a traditional method, it randomly selects 10s and 100s of different subsets from both the data set and the feature set and trains them. Hundreds of subsets (ie 100's of decision trees) that have been created and trained make individual predictions [19]. The GINI index is used to determine the homogeneity of classes [20]. As the GINI index decreases, the homogeneity of classes increases. The GINI index is calculated as in Eq. 6.

$$Gini(T) = 1 - \sum_{j=1}^{n}(p_j)^2 \tag{6}$$

While T refers to the whole data set, p_j represents the division of each data in that row by the sum of all the values in that row, and n represents the selected data. Based on the GINI index, classification or regression is made according to the problem over the test data. If the problem is a regression problem, the estimates of the decision trees are averaged. If the problem is a classification problem, the one with the highest number of votes is chosen among the predictions.

Training on different data sets reduces overfitting, which is one of the biggest problems of decision trees. In addition, it gives results in a very short time. In our study, we used the random forest algorithm for classification to determine whether individuals have OC or not.

E. Performance Evaluation

To analyze the performance of the models' predictions, we need evaluation metrics for example accuracy, recall, precision, F1 score and AUC score. In the most medical datasets, have imbalanced class distribution, F1-score is a crucial metric to evaluate our model on classification problems. F1 Score is used to compare models of diverse precision and recall scores. Recall Score is calculated by dividing the number of true positives by the sum of true positives and false negatives. Precision Score is calculated by dividing the number of true positives by the total number of true positives and false positives. The Area Under the

Curve (AUC) Score computes the area under the ROC curve, which is the Receiver Operator Characteristic Curve. The ROC curve is a probability curve and a schematic representation of the binary classifier system's diagnostic functionality. Simply put, it demonstrates how well the model on a classification problem distinguishes between classes.

III. RESULT

In this study, we evaluated the prediction performance of the RF classifier using an ovarian cancer dataset with 49 real-valued features. The CA72-4 and NEU, the attributes with the high missing value rate dropped from the data. Since our sample is relatively limited, DR methods were needed to boost the prediction model's efficiency. We first standardized the features before implementing DR techniques to our dataset since attributes have different distributions: the minimum, maximum, and average values are -1.5081, 1.4240, and -0.1435, respectively. Furthermore, the variance is 0.4667, skewness is 0.3440, and kurtosis is -0.1367.

A major question that arises is how to determine the optimum number of component that will lead to the highest accuracy. Regarding the optimal component value, we decided the higher the classification performance, the better the data representation so we seeked to maximize the accuracy of prediction in order to acquire an efficient classification model. For this, we used 25 component that was over 90% cumulative variance from 47 features. We prefered the RF algorithm to evaluate the performance of DR due to its widespread use and excellent efficiency. Most of classifier models have tuning hyper-parameters. We optimized the number of estimators, max depth of tree, min samples leaf and min samples split for RF. Hyperparameters range of classification model is shown in Table 1. While the transformed datasets were obtained by using optimal component size, we used 10-fold cross-validation to train RF and the model's final performance was the average Area Under the Curve (AUC) score over ten runs. The AUC is an efficiency metric for classification problems of varying threshold settings. AUC represents the degree or metric of separability, and Receiver Operating Characteristic curve (ROC) is a probability curve. It indicates how well the model can differentiate between classes. By analogy, the higher the AUC, the better the model decides between patients and others.

Hyperparameters	Value
n estimators	[10, 25, 50, 100]
max depth	[5, 10, 25, 50]
min samples leaf	[1, 2, 5, 10]
min samples split	[2, 5, 10, 20]

Table I. Hyperparameters of Random Forest

In Fig. 1, we can see that DR models were compared with number of components. The best AUC score is 6th, 2nd and 15th component for PCA, K-PCA and ICA respectively. Furthermore, in K-PCA with the polynomial kernel, we obtained a higher performance than the other kernel methods. However, we observed the training performance of the original and the three transformed datasets, after we set the optimal n components for DR techniques.

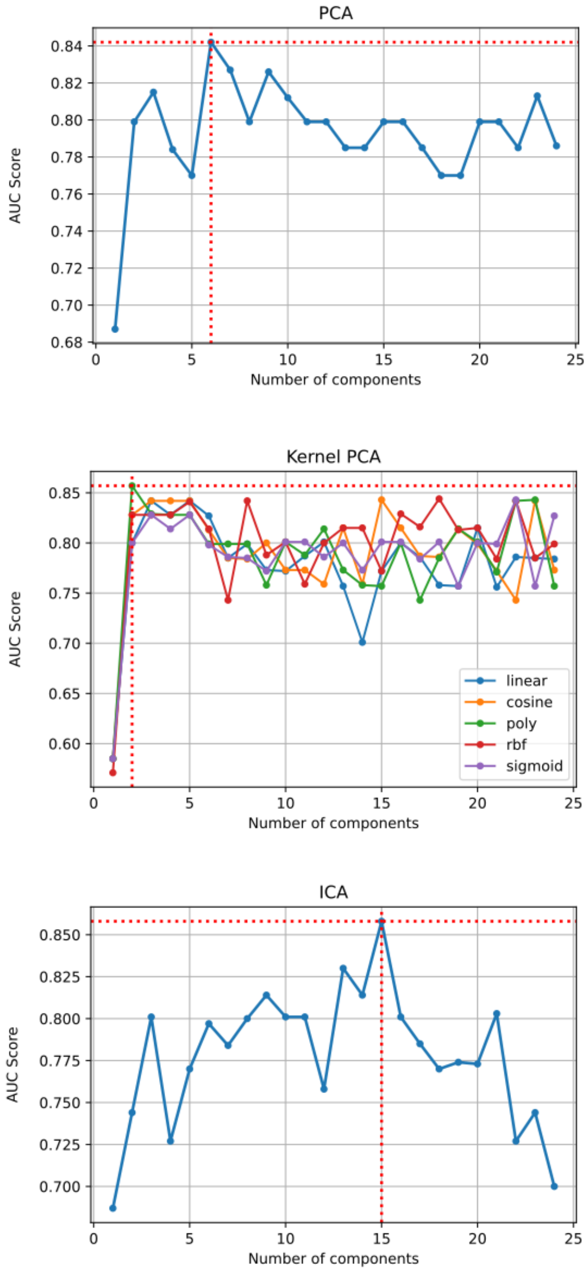

Figure 1. AUC score of dimension reduction methods by component numbers

Table 2 presents the dimension of the transformed datasets, F1-score, and elapsed time in seconds (sec) after DR methods are applied.

Data	Dimension	Random Forest	
		F1-Score	Elapsed Time
Original	47	89.2	24.903
PCA	6	89.5	18.191
Kernel PCA	2	88.3	17.272
ICA	15	84.5	18.594

Table 2. Classification performance of original and reduced datasets.

RF returned the F1-score of 89.20% and a speed of 24.903 seconds on the original data with 47 attributes. The classifier clearly worked better from the current reduced datasets. PCA achieved the best performance, led by K-PCA and ICA. In this way, the best feature space was reduced from 47 to 6, this result is significant for medical datas which is complicated and huge. However, ICA which has the 15 components provided the lowest accuracy. We can also see that K-PCA is the best at classification, taking just 17.272 seconds, followed by PCA at 18.191. The classification of the original dataset takes the longest (24.903 seconds).

IV. CONCLUSION

The medical datasets can contain redundant features and missing values, presenting significant challenges to the prediction model's performance and potentially resulting in pointless predictions. Therefore, proper preprocessing of data sets is necessary before applying ML algorithms. Furthermore, DR strategies such as feature selection and feature extraction have become essential for ML approaches to achieve reasonable classification performance. In this paper, we first defined and compared various dimension reduction methods. The efficiency of DR techniques was then empirically tested and compared to an OC dataset. Using the F1 score and elapsed time, we evaluated the consistency of the different transformed feature spaces. Results showed the feasibility of reduce dimension to enhance classification performance, confirming the gain achieved by performance metrics. Furthermore, it has been observed that the dimesion reduction process visibly reduces processing time on decision-making along performance.

REFERENCES

[1] H. J. Whitwell *vd.*, “Improved early detection of ovarian cancer using longitudinal multimarker models”, *Br. J. Cancer*, c. 122, sayı 6, s. 847—856, 2020, doi: 10.1038/s41416-019-0718-9.

[2] T. Granato, C. Midulla, F. Longo, B. Colaprisca, L. Frati, ve E. Anastasi, “Role of HE4, CA72.4, and CA125 in monitoring ovarian cancer.”, *Tumour Biol. J. Int. Soc. Oncodevelopmental Biol. Med.*, c. 33, sayı 5, ss. 13339, Eki. 2012, doi: 10.1007/s13277-012-0381-8.

[3] K. Aslan, M. A. Onan, C. Yilmaz, N. Bukan, ve M. Erdem, “Comparison of HE 4, CA 125, ROMA score and ultrasound score in the differential diagnosis of ovarian masses.”, *J. Gynecol. Obstet. Hum. Reprod.*, c. 49, sayı 5, s. 101713, May. 2020, doi: 10.1016/j.jogoh.2020.101713.

[4] M. Lu *vd.*, “Using machine learning to predict ovarian cancer.”, *Int. J. Med. Inform.*, c. 141, s. 104195, Eyl. 2020, doi: 10.1016/j.ijmedinf.2020.104195.

[5] R. G. Moore *vd.*, “A novel multiple marker bioassay utilizing HE4 and CA125 for the prediction of ovarian cancer in patients with a pelvic mass.”, *Gynecol. Oncol.*, c. 112, sayı 1, ss. 40–46, Oca. 2009, doi: 10.1016/j.ygyno.2008.08.031.

[6] I. Jacobs, D. Oram, J. Fairbanks, J. Turner, C. Frost, ve J. G. Grudzinskas, “A risk of malignancy index incorporating CA 125, ultrasound and menopausal status for the accurate preoperative diagnosis of ovarian cancer.”, *Br. J. Obstet. Gynaecol.*, c. 97, sayı 10, ss.–9922,Eki. 1990, doi: 10.1111/j.1471-0528.1990.tb02448.x.

[7] C. Anton, F. M. Carvalho, E. I. Oliveira, G. A. R. Maciel, E. C. Baracat, ve J. P. Carvalho, “A comparison of CA125, HE4, risk ovarian malignancy algorithm (ROMA), and risk malignancy index (RMI) for the classification of ovarian masses.”, *Clinics (Sao Paulo).*, c. 67, sayı 5, ss.441,42012, doi: 10.6061/clinics/2012(05)06.

[8] P. Zhang *vd.*, “Development of a multi-marker model combining HE4, CA125, progesterone, and estradiol for distinguishing benign from malignant pelvic masses in postmenopausal women”, *Tumour Biol.*, c. 37, sayı 2, s. 2183-2191, 2016, doi: 10.1007/s13277-015-4037-3.

[9] M. I. Jordan ve T. M. Mitchell, “Machine learning: Trends, perspectives, and prospects.”, *Science*, c. 349, sayı 6245, ss. 255–260, Tem. 2015, doi: 10.1126/science.aaa8415.

[10] E. Yildiz, Y. Sevim, “Comparison of linear dimensionality reduction methods on classification methods”, ELECO 2016, c. 1, sayı 2, ss. 161 –164.

[11] F. Yang, H. Z. Wang, H. Mi, C. De Lin, ve W. W. Cai, “Using random forest for reliable classification and cost-sensitive learning for medical diagnosis”, *BMC Bioinformatics*, c. 10, sayı SUPPL. 1, 2009, doi: 10.1186/1471-2105-10-S1-S22.

[12] C. Nguyen, Y. Wang, ve H. N. Nguyen, “Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic”, *J. Biomed. Sci. Eng.*, c. 06, sayı 05, ss. 551–560, 2013, doi: 10.4236/jbise.2013.65070.

[13] G. Sun, S. Li, Y. Cao, ve F. Lang, “Cervical cancer diagnosis based on random forest”, *Int. J. Performability Eng.*, c. 13, sayı 4, ss. 446–457, 2017, doi: 10.23940/ijpe.17.04.p12.446457.

[14] J. Ramírez *vd.*, “Computer aided diagnosis system for the Alzheimer’s disease based on partial least squares and random forest SPECT image classification.”, *Neurosci. Lett.*, c. 472, sayı 2, ss. 99–103, Mar. 2010, doi: 10.1016/j.neulet.2010.01.056.

[15] M. Mi, Qi; Jiang, Jingting; Znati, Ty; Fan, Zhenjiang; Li,

Jundong; Xu, Bin; Chen, Lujun; Zheng, Xiao; Lu, “Data for: USING MACHINE LEARNING TO PREDICT OVARIAN CANCER”, *Mendeley Data*, 2020. .

[16] M. ÇALIŞAN ve M. F. TALU, “Boyut İndirgeme Yöntemlerinin Karşılaştırmalı Analizi”, *Türk Doğa ve Fen Derg.*, c. 9, sayı 1, ss. 107–113, 2020, doi: 10.46810/tdfd.707200.

[17] Q. Wang, “Kernel Principal Component Analysis and its Applications in Face Recognition and Active Shape Models”, 2012, [Çevrimiçi]. Available at: http://arxiv.org/abs/1207.3538.

[18] B. Sohrab, A. N. Prof, ve E. Tercan, “JEOİSTATİSTİKSEKestiṙiṁ Multivariate Geostatistical Estimation Using Independent Component Analysis”, 2013.

[19] D. S. Palmer, N. M. O’Boyle, R. C. Glen, ve J. B. O. Mitchell, “Random forest models to predict aqueous solubility.”, *J. Chem. Inf. Model.*, c. 47, sayı 1, ss. 150–158, 2007, doi: 10.1021/ci060164k.

[20] M. Pal, “Random forest classifier for remote sensing classification”, *Int. J. Remote Sens.*, c. 26, sayı 1, ss. 217–222, 2005, doi: 10.1080/01431160412331269698.