Original article

Classification of Pistachio Species Using Improved k-NN Classifier

İlker Ali Özkan[1], *Murat Köklü*[1], *Rıdvan Saraçoğlu*[2,*]

[1] Selcuk University, Faculty of Technology, Department of Computuer Engineering 42100 Konya, Turkey
[2] Van Yuzuncu Yil University, Faculty of Engineering, Department of Electrical and Electronics Engineering, 65080 Van, Turkey

Abstract: In order to keep the economic value of pistachio nuts which have an important place in the agricultural economy, the efficiency of post-harvest industrial processes is very important. To provide this efficiency, new methods and technologies are needed for the separation and classification of pistachios. Different pistachio species address different markets, which increases the need for the classification of pistachio species. In this study, it is aimed to develop a classification model different from traditional separation methods, based on image processing and artificial intelligence which are capable to provide the required classification. A computer vision system has been developed to distinguish two different species of pistachios with different characteristics that address different market types. 2148 sample image for these two kinds of pistachios were taken with a high-resolution camera. The image processing techniques, segmentation and feature extraction were applied on the obtained images of the pistachio samples. A pistachio dataset that has sixteen attributes was created. An advanced classifier based on k-NN method, which is a simple and successful classifier, and principal component analysis was designed on the obtained dataset. In this study; a multi-level system including feature extraction, dimension reduction and dimension weighting stages has been proposed. Experimental results showed that the proposed approach achieved a classification success of 94.18%. The presented high-performance classification model provides an important need for the separation of pistachio species and increases the economic value of species. In addition, the developed model is important in terms of its application to similar studies.

Keywords: Classification, Image processing, k nearest neighbor classifier, Pistachio species.

Introduction

Pistachio is one of the most nutritious products. It provides 560 calories in 100 gr and is a rich source of protein, dietary fiber, various dietary minerals and vitamins B, thiamine and vitamin B6(1). It is known that pistachio has many benefits especially heart health. (2,3).

Turkey is ranked in the top three in World Pistachio production and pistachios are grown in 56 provinces of Turkey. Fig. 1 shows a pistachio grain. Kirmizi and Siirt species that have more prominent fruits and fewer tendencies to periodicity are often the preferred species to increase production of pistachio in Turkey (1).

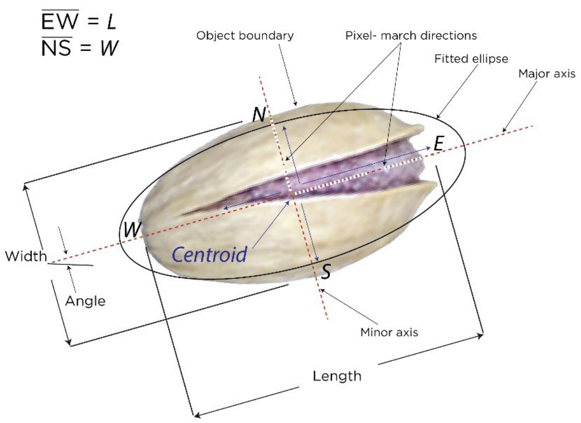

Figure 1. Morphological characteristics and appearance of pistachio grain

In addition, production is concentrated on Kirmizi and Siirt species because of economic value. Kirmizi species are preferred to use in sweet confectionery and pastry industry due to their dark green color, taste and distinct aroma; Siirt species are preferred as a snack because of its high cracking rate and round shape (4). Therefore, these two types address different markets. The majority of the operations on pistachios are post-harvest processes.

There are different studies in the literature in order to evaluate the obtained product in an economic and high-quality way. Cetin et al. has developed an algorithm to distinguish between the close and open state of pistachio shell (5). In their study, they took into account that after the impact on a steel plate, the pistachios in the closed state has a different sound than the pistachios in the open state. In the classification of audio signals, a linear combination of mel-cepstrum and PCA feature vectors were used. They have achieved 99% success in determining the open and close states of pistachios shells.

Casasent et al., performed classification using x-ray images of pistachios. They used piecewise quadratic neural network (PQNN) in their classification process (6). They indicated that X-ray imaging had the potential to provide real-time inspection of similar agricultural products. They have achieved a success rate of 88% in determining the quality of products.

Omid et al. applied an algorithm based on combined image processing and machine learning techniques, including artificial neural networks (ANN) and support vector machine (SVM) to classify peeled pistachio beans into five classes (7). They first segmented the images and obtained 72 chromatic and four shape features from each sample. After feature selection and PCS operations, they obtained the input vector size as 7. They achieved 99.4% accuracy with ANN classifier and 99.88% accuracy with SVM classifier.

In order to increase the market value of pistachio, automatic separation and classification is needed for different types of products that come from different suppliers. Therefore, new techniques and technologies should be used to improve the methods used in post-harvest mechanization processes of pistachio and to increase the yield of the product which has high economic value (8).

In this study, it is aimed to develop a system that can classify the pistachio species which are needed to increase the efficiency of post harvest processes and address different market types. For this purpose, a computer vision system (CVS), image processing techniques and appropriate artificial intelligence methods were used together.

With the proposed system and similar classification studies (9), improvements can be achieved in production and marketing processes. This can contribute to an increase in product quality and a decrease in product price. Facilitating access to pistachio and pistachio products will provide many nutritional benefits.

Material and Methods

Image acquisition and forming pistachio dataset

In this study, firstly images of licensed pistachio species were obtained in order to form dataset. A CVS was used to obtain the pistachio images. The developed CVS consists of a lens socket and a Prosilica GT2000C image capture camera. In addition, a special image shooting box was used to prevent shade formation in pistachios and to eliminate outdoor light differences. The distance between the camera and pistachio samples is 15 cm. Furthermore, black is used as background surface to prevent unwanted noise.

The noise on obtained pistachio images were cleaned by using image processing methods. Histogram information was obtained for contrast enhancement and histogram equalization was performed. With Otsu's method, a threshold detection method that can be applied on grayscale images, pistachio images are partitioned into two parts: background and foreground. Otsu algorithm, computes the optimal threshold value by processing the histogram information of the image (10). After segmentation, from each pistachio sample morphological and shape features were extracted.

MATLAB was used for feature extraction and sixteen features were obtained for each pistachio sample, twelve of them morphological and four of them shape information. The values found in all features are from pixel type. The morphological and shape characteristics used are given in Table 1(11-14): Kirmizi and Siirt pistachio species were used in this study. Of the total 2148 pistachio samples, 1232 belong to Kirmizi and 916 belong to Siirt specie. Samples were made by taking one kilogram of each species. Descriptive Statistics of Kirmizi and Siirt species are given in Table 2.

When the statistical values, given in Table 2, are examined, the differences between the species in terms of type are seen on the morphological features. Between two pistachio species, morphological feature values of Siirt species are higher than Kirmizi species. Despite this difference, morphological features of the two species overlap in many value ranges. This conclusion suggests that no feature alone will be sufficient for decision-making or statistical classification.

Performance Criteria.

The success of the model used in the study is determined by the number of samples assigned to the right class and the number of samples assigned to the wrong class.

The basic concepts used to evaluate model performance are accuracy, error rate, precision, sensitivity and F-score. Performance results obtained by the test can be expressed by confusion matrix.

In order to evaluate and scale the model, success criteria such as accuracy, sensitivity, specificity, precision and F1-Score are calculated (15-17). Detail of the formulas of success criteria is given in Table 3. In here *tp, fp, tn,* and *fn* are true positive, false positive, true negative and false negative respectively.

To prevent high bias and high variance while testing pistachio classification algorithms, "10-fold cross validation" method was used (18). In the 10-fold cross-validation method, the dataset is divided into 10-folds and each fold is used once as a test set and the remaining 9-fold used as training set. After this process is repeated 10 times, the result is reached by taking the average of the performance criteria of the test sets (18, 19).

Principial Component Analysis

Principal Component Analysis (PCA) is a useful statistical technique (20). The main purpose of this technique is to reduce the feature vector size. PCA is now widely used in signal and image processing techniques. The PCA feature reduction method can be described as follows:

Let *M*, be a *t* dimensional feature set. *n* basic axes *G1, G2*, . . . , *Gn* and $1 \leq n \leq t$. These axes are also called orthonormal axes. Where *xk*∈ *M*, m is the average of these samples and L the number of samples, the covariance matrix *U* is calculated as follows:

$$U = (\frac{1}{L}) + \sum_{k=1}^{L} (x_k - m)^T (x_k - m)$$

according to this:

$UGk = v_k Gk,\ k \in 1,\ldots,n,$

where v_k is the largest eigen value of *U*. According to this, n basic components of $x_k \in M$ are given as follows:

$$q = [q_1, q_2, \ldots\ q_n, = [G_1^T x, G_2^T x, \ldots, G_n^T] = G^T x$$

here, q is n basic component vector of x input feature vector.

Table 1. Attributes and formulas used in this study

Type	Feature	Formula
Morphological	Area (A):	$A = \sum_{r,c\,\in R} 1$
	Perimeter (P):	Bean circumference is defined as the length of its border.
	Major axis length (L):	The distance between the ends of the longest line that can be drawn from a bean.
	Minor axis length (l):	The longest line that can be drawn from the bean while standing perpendicular to the main axis.
	Eccentricity (Ec):	Eccentricity of the ellipse having the same moments as the region.
	Equivalent diameter (Ed):	$Ed = \sqrt{\frac{4 * A}{\pi}}$
	Solidity (S):	$S = \frac{A}{C}$
	Convex area (C):	Number of pixels in the smallest convex polygon that can contain the area of a bean grain.
	Extent (Ex):	$Ex = \frac{A}{A_B}$ where A_B = Area of bounding rectangle
	Aspect ratio (K):	$K = \frac{L}{I}$
	Roundness (R):	$R = \frac{4\pi A}{P^2}$
	Compactness (CO):	$CO = \frac{Ed}{L}$
Shapes	Shape Factor 1 (SF1):	$SF1 = \frac{L}{A}$
	Shape Factor 2 (SF2):	$SF2 = \frac{l}{A}$
	Shape Factor 3 (SF3):	$SF3 = \frac{A}{\frac{L}{2} * \frac{L}{2} * \pi}$
	Shape Factor 4 (SF4):	$SF4 = \frac{A}{\frac{L}{2} * \frac{l}{2} * \pi}$

Table 2. Descriptive Statistics of pistachio species

No	Features	Kirmizi				Siirt			
		Min	Max	Mean	Std. Dev.	Min	Max	Mean	Std. Dev.
1	Area	29.808	96.582	73.285,43	11.049,99	55.000	124.008	88.915,24	9.973,60
2	Perimeter	858,363	2.755,05	1.378,93	341,5137	1.018,38	2.746,77	1.489,24	408,7394
3	Major axis length	320,3445	541,9661	442,4214	34,8984	336,944	535,6422	451,397	28,0239
4	Minor axis length	133,5096	305,8938	220,8538	22,4381	171,062	383,0461	261,7925	22,718
5	Eccentricity	0,6249	0,946	0,8632	0,0339	0,5049	0,9247	0,8093	0,0486
6	Equivalent diameter	194,8146	350,6737	304,5275	23,9449	264,6284	397,3561	335,9306	19,0103
7	Solidity	0,588	0,9936	0,9346	0,0501	0,6975	0,9951	0,9474	0,05
8	Convex area	37.935	109.071	78.386,02	11.079,02	59.754	132.478	93.932,81	10.106,46
9	Extent	0,4272	0,8123	0,7115	0,0542	0,5035	0,8204	0,7222	0,0496
10	Aspect ratio	1,281	3,0858	2,0185	0,2086	1,1585	2,6263	1,7363	0,1761
11	Roundness	0,0628	0,874	0,5498	0,1952	0,122	0,9336	0,5953	0,2319
12	Compactness	0,476	0,8082	0,6891	0,0333	0,6067	0,8779	0,7454	0,0366
13	Shape Factor 1	0,0047	0,0131	0,0061	0,0008	0,004	0,0077	0,0051	0,0004
14	Shape Factor 2	0,0024	0,0053	0,0031	0,0004	0,0024	0,0047	0,003	0,0003
15	Shape Factor 3	0,2266	0,6532	0,4759	0,0455	0,3681	0,7706	0,5569	0,0551
16	Shape Factor 4	0,6204	0,9987	0,9524	0,0489	0,6598	0,999	0,959	0,0553

Table 3. Calculation formulas and descriptions of two class metrics

Measure	Formula
Accuracy (acc)	$\frac{tp + tn}{tp + fp + tn + fn}$
Sensivitiy (se)	$\frac{tp}{tp + fn}$
Specificity (sp)	$\frac{tn}{tn + fp}$
Precision (p)	$\frac{tp}{tp + fp}$
F1-Score	$\frac{2 * p * se}{p + se}$

K-NearestNeighborAlgorithm

The K-NN algorithm is among the most basic example-based learning algorithms. In sample-based learning algorithms, the learning process is based on the data in the training set. A new sample is classified according to the similarity between the samples in the training set (21). In the K-NN algorithm, the samples in the training set have n-dimensional numerical attributes. All training samples are held in an n-dimensional sample space, with each sample representing a point in the n-dimensional space. When an unknown sample is encountered, the class label of the new sample is assigned by determining the k samples closest to the relevant sample from the training set and deciding by majority of the class labels of the nearest neighbor k (22). The K value is usually set as the square root of the number of training samples (23).

In the basic k-NN algorithm, determining the class label with majority voting leads frequent classes in datasets with asymmetric distribution to have more dominant role in idenfiying new samples class labels (24). Therefore, there are methods such as weighted k-NN that assign different weight values to the the distance criterion of the basic k-NN algorithm (25).

In the K-NN algorithm, when an unlabeled sample comes, various feature calculation criteria are used to find the nearest sample. In this study Euclidean criterion is used. Euclidean distance is the most commonly used distance measure in classification and clustering algorithms. Euclidean distance is the linear distance between two points (26).

In Weighted k-NN, by assigning weight values to neighbors, neighboring samples closer to the sample being classified are intended to contribute more to the majority vote. The most commonly used methods of weight value assignment are that the weight of each neighbor is taken as 1/d or $1/d^2$, where d, the distance between neighbors (27). Actually, this weight assignment can be explained as a conversion of distances to weight value. For each class, the weight values of the samples belonging to these classes are summed and a value of membership to the class is determined. When these values are considered, the class with the largest value is determined as the result class (28).

Proposed Classification Model

The feature extraction and modeling steps used for pistachio classification are given in Fig. 2. First of all, images of pistachios were obtained using CVS developed in the image preparation step. An image acquisition set was used to provide the same distance, light, etc. environment characteristics in order to avoid differences during image acquisition. The pistachio images obtained in the feature extraction stage were subjected to segmentation. In this stage, for each pistachio sample morphological and shape features that are detailed in Table 1, were extracted.

Features of pistachio dataset are closely related to each other. To find the basic structure and size of the dataset with this relationship, PCA was performed on the dataset in the dimension reduction stage. Dimension reduction process eliminates the high dimensional data disadvantage of the classification algorithm used in this study.

In the classification stage, one of the most popular machine learning techniques, k-NN algorithm, was used. The k-NN algorithm is easy to implement with respect to the size of the obtained dataset and does not require a training process. Since it is a sample-based algorithm, new samples can be added to the dataset without any problems. It does not require too many parameters which provides ease of application. It is also preferred because it is robust against noisy data. In addition, weighted k-NN structure was employed in order to use the contribution of neighboring samples of k-NN model to voting.

In the last stage, the models developed for the classification of pistachio varieties were evaluated based on the confusion matrix and the performance criteria given in Table 3 and the model selection was made.

Results

In the study, MATLAB was used for model development process. Modeling process was performed by using standard k-NN algorithm on the obtained pistachio dataset. In the classification study, k-NN models were developed by using all sixteen attributes. For the k-NN model, the Euclidean criterion was used as the distance criteria. In addition, the best k value was determined using trial and error method starting from 1. As a result of the experiments, the best k value was chosen as 43, equal to the square root of the training dataset, which is generally one of the k selection method (29).

Two different models have been developed on the dataset, namely the k-NN model and the weighted k-NN model. The confusion matrices are given in Table 4 and 5 have been obtained as the result of the 10-fold cross validation kNN models. In addition, the average values of the performance criteria obtained after 30 runs for the k-NN models are given in Table 6.

The k-NN model has 83.38% accuracy and the weighted k-NN model has 87.38% accuracy. Developed weighted k-NN model can estimate the class of

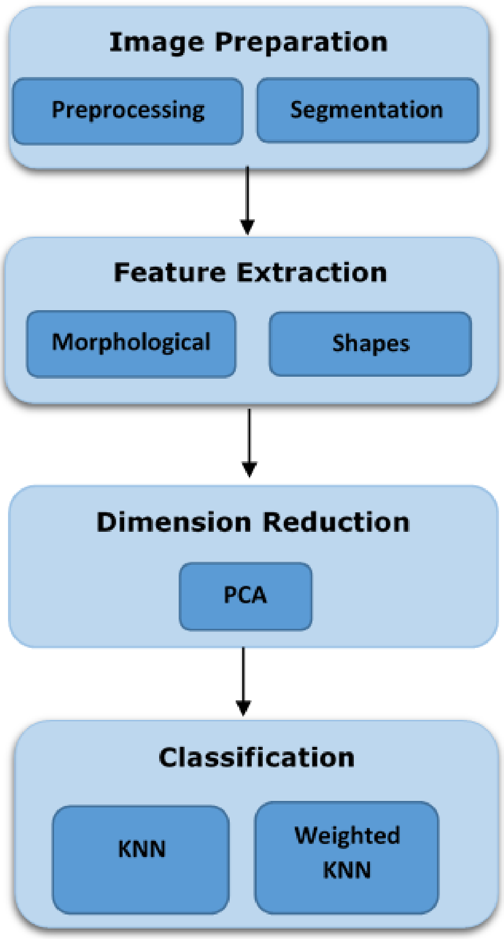

Figure 2. Flow chart of the proposed system for pistachio classification

Table 4. Confusion matrix of k-NN model

		Predicted Classes Weighted k-NN Value	
		Kirmizi	Siirt
True Classes	Kirmizi	1047	185
	Siirt	172	744

Table 5. Confusion matrix of Weighted k-NN model

		Predicted Classes Weighted k-NN Value	
		Kirmizi	Siirt
True Classes	Kirmizi	1106	126
	Siirt	145	771

Table 6. Performance criteria for developed K-NN models

Measure	k-NN Value	Weighted k-NN Value
Sensivitiy	0.8589	0.8841
Specificity	0.8009	0.8595
Precision	0.8498	0.8977
F1 Score	0.8543	0.8909
Accuracy	0.8338	0.8738

being Siirt subspecies with the accuracy of 89.77%. On the otherhand the model estimation success rate is 85.95% for Kirmizi subspecies.

PCA was used to increase accuracy performance on pistachio dataset, to reduce the number of features sizes and to improve the distribution of samples. PCA yielded 4 features with 95% variance.

Weigthed k-NN model, which gave successful results before, was used for classification on the data set obtained by PCA. The confusion matrice is given in Table 7. As the result of the classification, 94.18% accuracy rate was obtained. In addition, performance criteria for PCA based k-NN model (using 10-fold cross validation) are given in Table 8.

The PCA-based weighted k-NN has of 94.18% accuracy. Developed PCA based weighted k-NN model can estimate the class of being Siirt subspecies with the accuracy of 95.13%, while the same model estimation success rate is 93.41% for Kirmizi subspecies.

The PCA method is known to be successful for independent variables that fit the normal distribution. The method was successful because the features in this study also fit the normal distribution. In addition, PCA is used to overcome the noise problem in the data. In this study, it can be interpreted that PCA method reduces the noise in the data.

Table 7. Confusion matrix of PCA based weighted k-NN model

		Predicted Classes Weighted k-NN Value	
		Kirmizi	Siirt
True Classes	Kirmizi	1172	60
	Siirt	65	851

Table 8. Performance criteria for developed PCA based weighted k-NN performance

Measure	Value
Sensivitiy	0.9475
Specificity	0.9341
Precision	0.9513
F1 Score	0.9494
Accuracy	0.9418

Conclusion

There is not any fundamental decisive feature in the classification of pistachio varieties. This study proposes an image-based approach to pistachios classification process. According to our experience in image obtainment and processing, a CVS is needed to prevent the noised data as much as possible. Properties for obtained images should be similar for all samples. Dataset obtained with the image-based approach and the classification model can be considered successful witout any optimization. This result shows that this image-based approach can be used to identify the characteristics of pistachio species.

In this study, it was seen that the PCA applied on pistachio dataset increased the classification performance. With this process, it is shown that none of its attributes alone is sufficient and PCA process gives good results in expressing the basic factors in the dataset containing high dimensional structure.

The dimension size reduction by PCA also resulted in improved performance in the k-NN model. This system with k-NN algorithm, can be applied to different species easly because it is sample-based. It is possible to include different species in the model if necessary. In this respect, the proposed system has an expandable structure.

In addition, with the optimization of the k-NN algorithm in according to the model, the accuracy was increased from 83.38% to 94.18%. The preprocessing for the dataset in the modeling stage and the use of the weighted k-NN algorithm resulted in a 10% increase in the accuracy.

With the developed system, it will be easier to store and process the same kind of products and the time and energy expenses of the enterprises will be less. By means of the system, it will be ensured that the nuts to be classified to the standard specifications. As the result of all these factors, they will have price and sales advantages as the pistachio nuts are standard at the marketing stage.

The success of the developed system can be further enhanced by the hybrid use of machine learning methods. The resulting system can be combined with ultrasonic sounds to make a complicated classifier that considers open and close conditions of pistachios shells.

REFERENCES

1. Ertürk YE. , Geçer MK., Gülsoy E, and Yalçın S. Production and Marketing of Pistachio, Journal of the Institute of Science and Technology of Igdir University 2011;5: 43–62.
2. Dreher ML. Pistachio nuts: composition and potential health benefits, Nutrition Reviews 2012, 70(4): 234–240.
3. Kay CD, Gebauer SK, West SG, Kris-Etherton PM. Pistachios Increase Serum Antioxidants and Lower Serum Oxidized-LDL in Hypercholesterolemic Adults, The Journal of Nutrition 2010, 140(6): 1093–1098
4. Tunalıoğlu R, and Taşkaya B. Antepfıstığı. Tarımsal Ekonomi Araştırma Enstitüsü Dergisi, 2003.
5. Cetin AE, Pearson TC, and Tewfik AH. Classification of closed and open shell pistachio nuts using principal component analysis of impact acoustics. In 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing 2004; V–677.
6. Casasent DA, Sipe MA, Schatzki TF, Keagy PM, Lee LC. Neural net classification of X-ray pistachio nut data. LWT-Food Science and Technology 1998, 31(2): 122–128.
7. Omid M, Firouz MS, Nouri-Ahmadabadi H, Mohtasebi SS. Classification of peeled pistachio kernels using computer vision and color features, Engineering in Agriculture, Environment and Food 2017,10: 259–265.
8. Atay Ü. The investigation of classification systems used for pistahio and construction of an alternetive classification system, PhD Thesis Harran University, Sanliurfa, 2007.
9. Sabanci K, Kayabasi A, and Toktas A. Computer vision-based method for classification of wheat grains using artificial neural network, Journal of The Science of Food and Agriculture 2016, 97(8): 2588–2593.
10. Otsu N. A threshold selection method from gray-level histograms, IEEE Transactions on Systems, Man, and Cybernetics 1979, 9: 62–66.
11. Paliwal J, Visen NS, and Jayas DS, AE—Automation and Emerging Technologies: Evaluation of Neural Network Architectures for Cereal Grain Classification using Morphological Features, Journal of Agricultural Engineering Research 2001, 79: 361–370.

12. Liu ZY, Cheng F, Ying YB, and Rao XQ, Identification of rice seed varieties using neural network, Journal of Zhejiang University Science. B 2005, 6: 1095–1100.
13. Symons SJ, and Fulcher RG. Determination of wheat kernel morphological variation by digital image analysis: I. Variation in Eastern Canadian milling quality wheats, Journal of Cereal Science 1988, 8: 211–218.
14. Pazoki A, Farokhi F, and Pazoki Z. Classification of rice grain varieties using two Artificial Neural Networks (MLP and Neuro-Fuzzy), The Journal of Animal & Plant Sciences 2014, 24: 336–343.
15. Hossin M, Sulaiman MN, Mustapha A, and Mustapha N. A Novel Performance Metric for Building an Optimized Classifier, Journal of Computer Science 2011, 7(4):582–590.
16. Hossin M, and Sulaiman M. A review on evaluation metrics for data classification evaluations, International Journal of Data Mining & Knowledge Management Process 2015, 5: 1.
17. Sokolova M, and Lapalme G. A systematic analysis of performance measures for classification tasks, Information Processing & Management 2009, 45: 427–437.
18. James G, Witten D, Hastie T, and Tibshirani R. An Introduction To Statistical Learning, Springer, 2013.
19. Kuhn M, and Johnson K. Applied predictive modeling, Springer, 2013.
20. Chatterjee C, Roychowdhury VP, Chong EKP. On Relative Convergence Properties of Principal Component Analysis Algorithms, IEEE Transactions on Neural Networks 1998, 9(2): 319–329.
21. Mitchell TM. Machine learning, McGraw Hill, 1997.
22. Han J, Pei J, and Kamber M. Data mining: concepts and techniques, Elsevier, 2011.
23. Duda RO, Hart PE, and Stork DG. Pattern classification, John Wiley & Sons, 2012.
24. Coomans D, and Massart DL. Alternative k-nearest neighbour rules in supervised pattern recognition: Part 1. k-Nearest neighbour classification by using alternative voting rules, Analytica Chimica Acta 1982,136: 15–27.
25. Gärtner T, Lloyd JW, and Flach PA. Kernels and distances for structured data, Machine Learning 2004, 57: 205–232.
26. Mohamed TM. Pulsar selection using fuzzy kNN classifier, Future Computing and Informatics Journal 2018, 3: 1–6.
27. Dudani SA. The distance-weighted k-nearest-neighbor rule, IEEE Transactions on Systems, Man, and Cybernetics 1976, 325–327.
28. Srivastava A., Singh MP, and Kumar P. Supervised semantic analysis of product reviews using weighted k-NN classifier. 11th International Conference on Information Technology: New Generations 2014, 502–507.
29. Hassanat AB, Abbadi MA, Altarawneh GA, Alhasanat AA. Solving the Problem of the K Parameter in the KNN Classifier Using an Ensemble Learning Approach. International Journal of Computer Science and Information Security, 2014,. 12(8): 33–39.

Corresponding author: ridvansaracoglu@yyu.edu.tr