REVIEW

The use of machine learning methods in classification of pumpkin seeds (*Cucurbita pepo* L.)

Murat Koklu · **Seyma Sarigil** · **Osman Ozbek**

Abstract Pumpkin seeds are frequently consumed as confection worldwide because of their adequate amount of protein, fat, carbohydrate, and mineral contents. This study was carried out on the two most important and quality types of pumpkin seeds, "Ürgüp Sivrisi" and "Çerçevelik", generally grown in Ürgüp and Karacaören regions in Turkey. However, morphological measurements of 2500 pumpkin seeds of both varieties were made possible by using the gray and binary forms of threshold techniques. Considering morphological features, all the data were modeled with five different machine learning methods: Logistic Regression (LR), Multilayer Perceptrons (MLP), Support Vector Machine (SVM) and Random Forest (RF), and k-Nearest Neighbor (k-NN), which further determined the most successful method for classifying pumpkin seed varieties. However, the performances of the models were determined with the help of the 10 k-fold cross-validation method. The accuracy rates of the classifiers were obtained as LR 87.92 percent, MLP 88.52 percent, SVM 88.64 percent, RF 87.56 percent, and k-NN 87.64 percent.

Keywords Pumpkin seed · Logistic regression · Multilayer peceptrons · Random forest · Classification · Support vector machine · Thresholding

M. Koklu (✉)
Computer Engineering, Faculty of Technology, Selcuk University, Konya, Turkey
e-mail: mkoklu@selcuk.edu.tr

S. Sarigil
Graduate School of Natural and Applied Sciences, Selcuk University, Konya, Turkey
e-mail: seyma.sargil@lisansustu.selcuk.edu.tr

O. Ozbek
Agricultural Machinery and Technologies Engineering, Faculty of Agriculture, Selcuk University, Konya, Turkey
e-mail: ozbek@selcuk.edu.tr

Introduction

The pumpkin plant belongs to the *Cucurbitaceae* family and has seasonal varieties (Düzeltir 2004). Confectionery pumpkins, grown in Turkey, are usually produced from the pumpkin species, *Cucurbita pepo* L and sometimes from the *Cucurbita moschata* Duchesne type pumpkins (Yanmaz and Düzeltir 2003). Pumpkin seeds are considered as important for human health because it contains 37 percent of carbohydrate, 35 percent to 40 percent of fat and protein along with calcium, potassium, phosphorus, magnesium, iron, and zinc (Peričin et al. 2008). Pumpkins are divided into many types, and one of these species is known as "Ürgüp Sivrisi" (Hatun Tırnağı) (Seymen et al. 2019). Ürgüp Sivrisi is a type of pumpkin seed that has a long, white, very bright, thin, and hardly distinguishable shell with a pointed tip (Yavuz et al. 2015; Aktaş et al. 2018; Seymen et al.

2019). One of the other types of pumpkin seeds is "Çerçevelik" (Seymen et al. 2019). It is a particular species grown in Turkey, Nevsehir, Karacaören, and known as "Topak" in Turkey.

With the development of technology, machine learning methods had been started to be used in the agricultural sector as well as other sectors in the world. Many studies had been done on the subject, where attributes, such as size, shape, texture, length, and color of 40 different wheat grains were extracted using the Dense Scale Invariant Feature Transform (DSIFT) algorithm (Olgun et al. 2016). A total of 6400 different images were used by the processing of 160 different images for each of the 40 different varieties. Of those 6400 images, 10 percent was determined as the test set. The success of the DSIFT algorithm for SVM was evaluated. Additionally, the accuracy rate was measured as 88.33 percent for the SVM classification. 5 different types of maize grain (BAINUO 6, NONGDA 86, NONGDA 108, GAOYOU 115, and NONGDA 4967) were examined by Chen et al. (2010). At the first step, they separated maize grains into three types, white, yellow, and mixed maize, with the help of Mahalanobis distance analysis. Later, they extracted a total of 58 different features from different varieties of the same type of maize, and out of them, 17 were responsible for geometry, 13 for shape, and 28 for color features. These models were known as Back-Propagation Neural Network (BPNN) algorithm and examined. At the end of the study on maize species, BAINUO 6, NONGDA 86, NONGDA 108, GAOYOU 115, and NONGDA 4967 were found with the classification accuracies of 100 percent, 94 percent, 92 percent, 88 percent, and 100 percent, respectively. Dheer et al. (2019) used 100 different datasets for seven different wheat varieties and classified their wheat grains with the help of k-NN, LDA, LR, and Navie Bayes (NB) machine learning methods. Then, the accuracy rates were determined by cross-validation with k = 10 for evaluation. The accuracy rates after cross-validation were observed as 97.32 percent for LR, 98.57 percent for LDA, 99.38 percent for k-NN, and 98.78 percent for NB. The observed precision and recall accuracy values of the k-NN model for the training dataset were determined to be 99 percent and 100 percent, respectively. As per the observations, the k-NN algorithm showed higher accuracy than other classification methods. Hua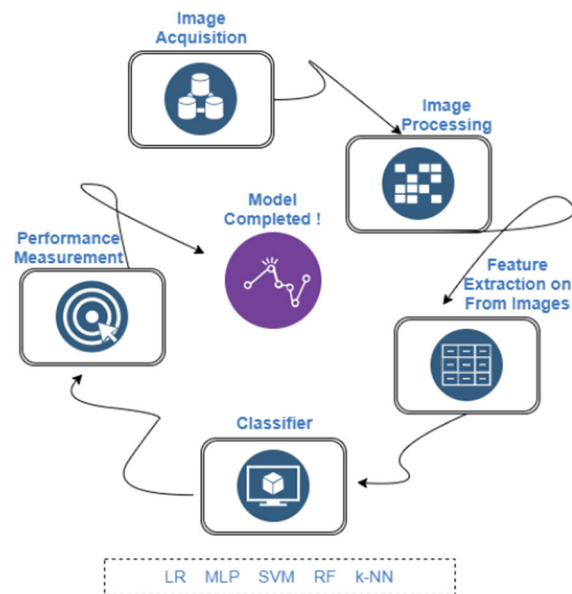ng et al. (2016) obtained images of maize grains with the help of the hyperspectral imaging method during their three years of study using four different corn varieties. Then the areas that were found to be hyperspectral significant in the images were determined with the help of the threshold method. They made classification with SVM using determined areas. At the end of the study, the average accuracy of the model was found to be 94.4 percent. Guevara et al. (2011) used 10 different wheat and 10 different barley images for their research on the classification of wheat and barley grains using LDA and k-NN algorithms. All the images had been tested in three different ways by grouping them only according to the morphological, color, and pattern features. However, the highest accuracy rate was achieved for color and pattern properties. Demir (2017) determined six different characteristics of the three types of pumpkin seeds by using BPNN and RBNN (Radial Basis Neural Network) structures. According to the error metric of root mean squared error (RMSE), the error rate of BPNN was found to be 0.6875, whereas the error rate of RBNN was observed to be 0.0025. As per the study, it was concluded that the BPNN method in the classification of pumpkin seeds was insufficient in predicting the morphological characteristics of the seeds. In this context, it was also inferred that in order to fulfill the research purpose, the RBNN classifier should be used. Additionally, it was also stated in the study that the RBNN classifier produced significant results faster than the BPNN classifier. Jamuna et al. (2010) used NB, J48 (C4.5), and MLP machine learning approaches to classify the quality of seeds belonging to 24 different cotton varieties, which were collected in different growth periods. In order to evaluate the accuracy, the model was cross-validated tenfold. The accuracy of the MLP and J48 (C4.5) model was found to be 98.78 percent, whereas the accuracy of NB was observed as 94.22 percent. However, it was determined that the processing time of MLP took longer than J48, and for this reason, J48 was considered as the most suitable model. Punn and Bhalla (2013) identified five different morphological features with the help of two-dimensional images of wheat grains and modeled those features using SVM and MLP methods. The accuracy rates of these models were determined as 86.8 percent according to the SVM method and 94.5 percent according to the MLP method. As a result, it was found that the model created with MLP was more efficient with the data used in this study. Machine

learning methods used in most of the classification studies with agricultural products were algorithms, such as SVM, MLP, and k-NN.

In this research, some features of the two pumpkin seed varieties, "Ürgüp Sivrisi" and "Çerçevelik", which had never been studied before, were transformed into measurable forms using gray form and binary form threshold methods and modeled with the five different machine learning methods. This research was aimed to pioneer new studies, which would be made using these models.

Materials and methods

Pumpkin seed photographs used in the study were obtained by a specially prepared mechanism because the quality of the images was important for this research. The images obtained were first converted to a gray tone and then to binary images, followed by the removal of noise and application of the feature removal option. At the next step, morphological feature extraction processes were applied to those images. However, the methods used for the classification of pumpkin seeds had been shown in Fig. 1. During the modeling phase, the pumpkin seed classification process was carried out using five machine learning techniques. Moreover, the performances of

those models, which were used at the last step, were evaluated by the cross-validation method.

Image acquisition

In order to obtain quality pumpkin seed images used in the study, a product shooting box was designed as in Fig. 2. The camera, used to take images from the product shooting box, was fixedly placed on the box with a lighting mechanism. The box was designed in such a way so that it could not receive any light from the outside to prevent shadows that might cause damage to the obtained images. The ground color of the box had been chosen as black in order to process the image easily, and the image acquisition setup had been prepared specially for this study.
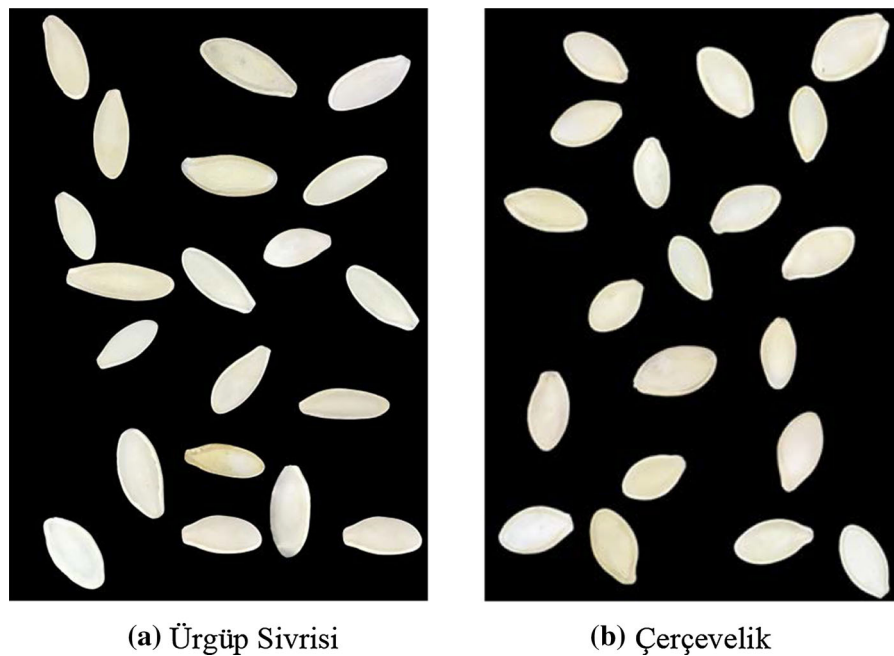
In this research, 2,500 pumpkin seeds were used, including 1200 Ürgüp Sivrisi and 1300 Çerçevelik types. In Fig. 3a and b, examples of the core types had been provided. As seen in the examples, it was very difficult to distinguish those varieties from one another with the help of naked eyes.

Image processing

Image processing was known as the processing of data collected in the form of images by any image capturing device (Eldem et al. 2017; Pandey et al. 2013). Image processing techniques were generally used to obtain meaningful information from the captured images. The reason for using the image processing technique in this study was to extract morphological characteristics of the seeds belonging to each class from the captured pumpkin seed images.

Fig. 1 Steps followed for the classification of pumpkin seeds
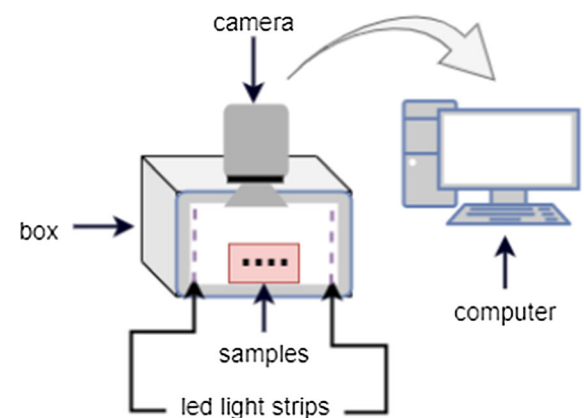
Fig. 2 The image acquisition scheme

(a) Ürgüp Sivrisi   **(b)** Çerçevelik

Fig. 3 Core type of pumpkin seed varieties

These features were discussed in detail in the next section. Taking those features from the clean images played an important role in the classification of those seeds, and MATLAB 18 software was used for the processing of those images. The core photos were first converted to gray level and then to binary forms as in Fig. 4. Threshold levels were determined by the Otsu method (Otsu 1979). Finally, in this way, it was prepared for image analysis and extraction of morphological features.

Morphological features

Twelve morphological features had been extracted for each of the pumpkin seeds, which were transported by image processing. Morphological feature extraction was a wide variety of image processing calculations

that processed images depending on the shapes of the features observed on the image. At this step, each pixel in the image was calculated by considering the values of the other pixels around it.

Considering the sample, which was placed on a two-dimensional plane as in Fig. 5, the extracted features could be highly dependent on the same plane. Therefore, the qualities of the inferred features were considered as determinants in making decisions on the machine learning methods.

Descriptions of the features and characteristics that had been considered as the reference of the two types of pumpkin seeds had been given in Table 1.

The density distribution charts of the two different types of pumpkin seeds, Ürgüp Sivrisi and Çerçevelik, according to the densities of the morphological features, are illustrated in Table 1 and Fig. 6. For the

RGB to Gray Conversion ────────→ Image Binarization
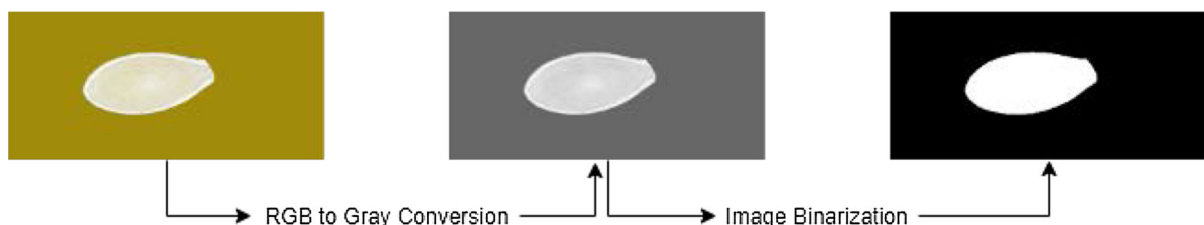
Fig. 4 Steps of image processing
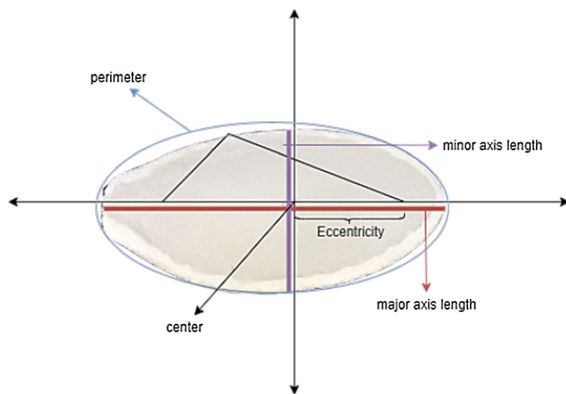
Fig. 5 The sampling of pumpkin seeds on a 2-D plane

9381 pixels provided that the area of the Çerçevelik class also included the density areas measured in pixel range (Fig. 6a). When looking at the graph of the perimeter feature, it could be concluded that Çerçevelik pumpkin seeds were shorter and wider than the Ürgüp Sivrisi pumpkin seeds (Fig. 6b).

Looking at the major axis length feature, it was possible to understand that the Çerçevelik pumpkin seeds could go up to 451 pixels and watch around, while, on the other hand, Ürgüp Sivrisi pumpkin seeds could go up to 526 pixels, but the general distribution was not around (Fig. 6c). In the case of minor axis length, this situation was exactly the opposite because

**Table 1** The most effective morphological features and explanations used in feature extraction

No	Name	Explanation
1	Area (A)	It gave the number of pixels within the borders of a pumpkin seed
2	Perimeter (p)	It gave the circumference in pixels of a pumpkin seed
3	Major Axis Length (Maj.AL)	It gave the circumference in pixels of a pumpkin seed
4	Minor Axis Length (Min.AL)	It gave the small axis distance of a pumpkin seed
5	Eccentricity (e)	It gave the eccentricity of a pumpkin seed
6	Convex Area (CA)	It gave the number of pixels of the smallest convex shell at the region formed by the pumpkin seed
7	Extent (E)	It returned the ratio of a pumpkin seed area to the bounding box pixels
8	Equiv Diameter (ED)	It was formed by multiplying the area of the pumpkin seed by four and dividing by the number pi, and taking the square root
9	Compactness (C)	It proportioned the area of the pumpkin seed relative to the area of the circle with the same circumference
10	Solidity (s)	It considered the convex and convex condition of the pumpkin seeds
11	Roundness (r)	It measured the ovality of pumpkin seeds without considering its distortion of the edges
12	Aspect Ratio (AR)	It gave the aspect ratio of the pumpkin seeds

classification of two species according to this distribution, the five morphological features with the least intersection set were observed as Compactness, Aspect Ratio, Eccentricity, Roundness, and Major Axis Length.

The densities of the results obtained from the morphological features of the Çerçevelik and Ürgüp Sivrisi pumpkin seed varieties were given in Fig. 6 as it could be seen that when the graphs of each feature were examined, the area of the Çerçevelik pumpkin seed was between 6976 and 8628 pixels, the area of the Ürgüp Sivrisi pumpkin seed was between 7248 and

this property of the Ürgüp Sivrisi pumpkin seeds ranged between 199 and 231 pixels and of the Çerçevelik pumpkin seeds ranged between 219 and 247 pixels (Fig. 6d). It was possible to say that the Ürgüp Sivrisi pumpkin seeds had a long and thin structure, although the areas varied, the seeds had a slightly larger area as compared to the Çerçevelik pumpkin seeds.

The Çerçevelik pumpkin seeds were concentrated around 7842 pixels by considering the convex area values (Fig. 6e). It could be seen that the Ürgüp Sivrisi pumpkin seeds were in similar intervals, which were
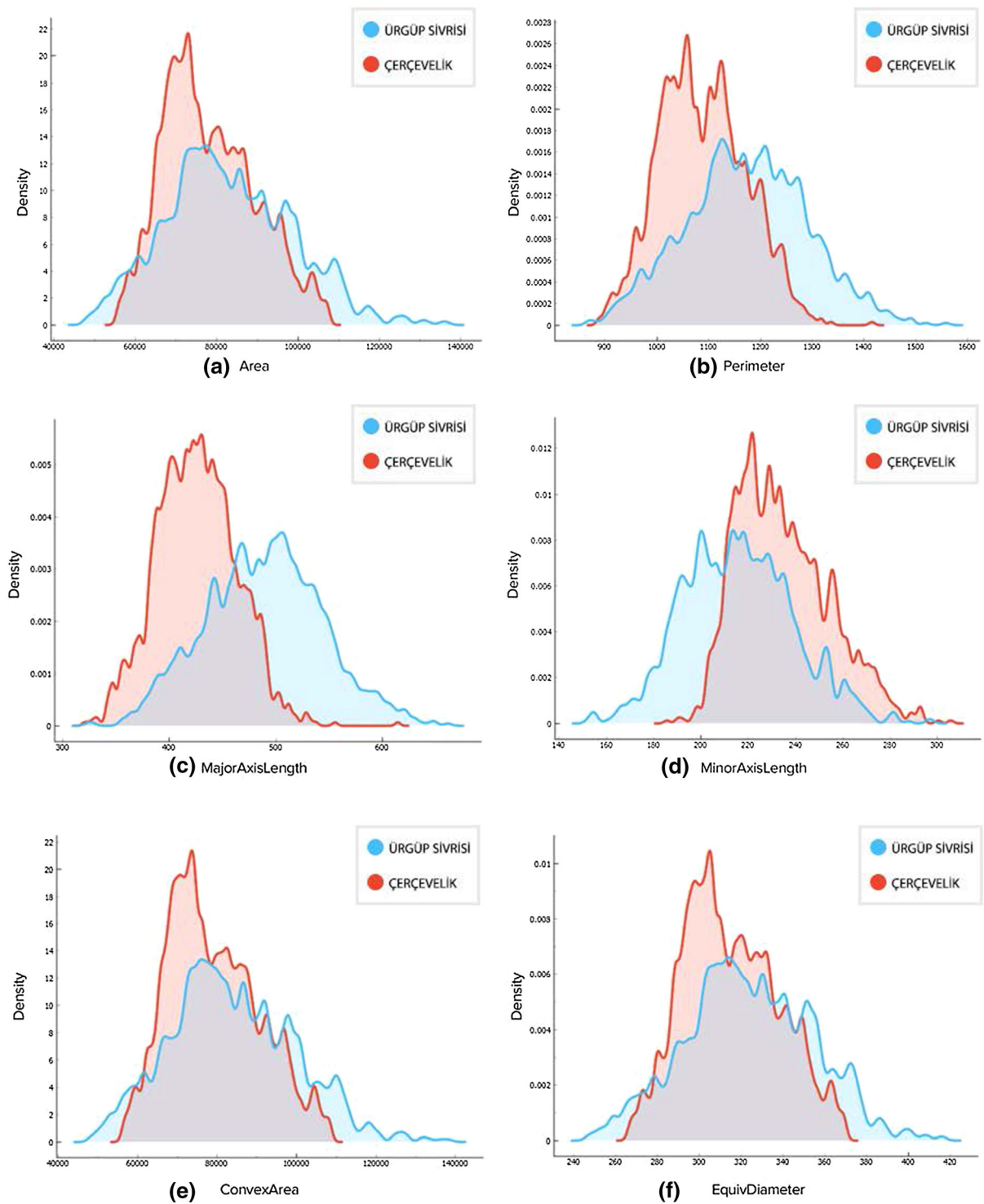
Fig. 6 Density graphs of the (**a**) morphological features-1 (**b**) morphological features-2
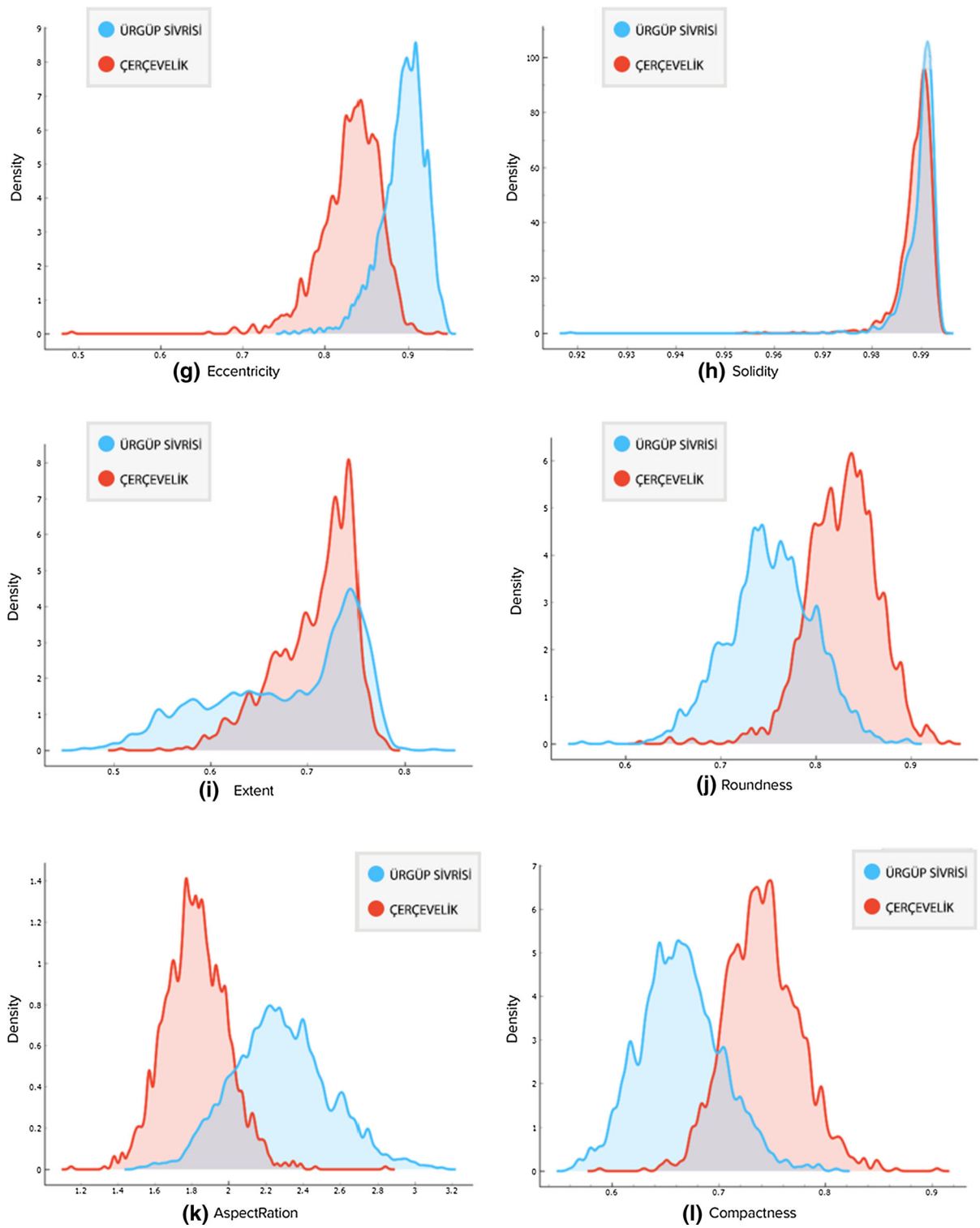
Fig. 6 continued

indistinguishable from the Çerçevelik seeds. The equivalent diameter property also showed similar results with the convex area property, which was predicted because the convex area property had a 100 percent positive correlation with the equivalent diameter property (Fig. 6f).

The small intersection in the density values of both the Çerçevelik and Ürgüp Sivrisi pumpkin seeds could be seen in Fig. 6. The eccentricity feature was noticed as a feature that created the highest classification difference among all other features. It was observed that the values for Çerçevelik class were concentrated around 0.810–0.855 pixels, and on the other hand, for Ürgüp Sivrisi class, this value was between 0.879 and 0.912 (Fig. 6g). The intersection clusters for both the classes were between 0.830 and 0.893, and not very dense. However, the solidity feature controlled both the concave and convex conditions. Overlapping of the density values of solidity property for all the classes under research in the graph occurred due to this situation (Fig. 6h).

The extent feature expressed the values of the red and green squares in Fig. 9. These squares were the boxes themselves, which were called the bounding boxes, drawn on the edges of the beans. While the density value of a box was sparse and not fully evident for Ürgüp Sivrisi pumpkin seeds, it was in the range between 0.680 and 0.739 in Çerçevelik class pumpkin seeds (Fig. 6i). The determinant ratio of the roundness feature varied in the range of between 0.723 and 0.783 for the Ürgüp Sivrisi pumpkin seeds and between 0.803 and –0.852 for Çerçevelik pumpkin seeds (Fig. 6j). The roundness feature was seen as the feature with a high classification capacity similar to eccentricity.

The aspect ratio showed the ratio of the width and length of the pumpkin seeds. The density value for Ürgüp Sivrisi was between 2.099 and 2.444, and for Çerçevelik type pumpkin seeds, it was between 1.7082 and 1.9323 (Fig. 6k). According to the aspect ratio feature, the Çerçevelik pumpkin seeds were remarkable for being shorter and thicker than the Ürgüp Sivrisi type of pumpkin seeds. The compactness property of the Ürgüp Sivrisi variety was concentrated between the values of 0.637 and 0.688, whereas it was between 0.717 and 0.763 for the Çerçevelik pumpkin variety of seeds (Fig. 6l).

Pumpkin seeds dataset

The rate of annual pumpkin production in the world was between 13 and 15.5 million tons, while the rate of annual pumpkin production in Turkey was about 365 thousand tons (Kayak et al., 2018). However, all the pumpkin varieties, which were produced in Turkey, included both the confectionary pumpkin varieties. Pumpkin seeds produced for confectionary purposes were grown in Cappadocia as well as in Ürgüp regions in Nevşehir and Tekirdağ along with the Kırklareli regions in Thrace (Yegul, 2012). In the dataset, two different types of commercial pumpkin seeds from Ürgüp and Karacaören regions of Nevşehir, known as Ürgüp Sivrisi and Çerçevelik, respectively, were used. Çerçevelik and Ürgüp sivrisi are the two most commonly grown pumpkin seeds in our country. Other cultivars grown are subgroups of these two varieties. The classification of these two types of pumpkin seeds is very important, especially for the seed sector. The distributions of these pumpkin seeds in the dataset had been given in Table 2.

The average, standard deviation, maximum, and minimum statistical values of the two types of pumpkin seeds were mentioned in Table 3. A general inference could be made about pumpkin seeds from the table.

Performance measures

Machine learning provided models for the solution to many problems. The success of the model presented on the classifiers was achieved with the use of some performance measures. The success of the algorithm was measured rather than the success of the model by these performance measures. It showed both the estimated and actual classification values, assuming that the values of a confusion matrix of size n × n associated with a classifier, was the number of classes

Table 2 Distribution of Ürgüp Sivrisi and Çerçevelik type pumpkin seeds in the dataset

Ad	Piece
Çerçevelik	1300
Ürgüp Sivrisi	1200
Total	**2500**

Table 3 The statistical distribution of Ürgüp Sivrisi and Çerçevelik pumpkin seed varieties

No	Features	Min	Mean	Max	Std. Dev
1	Area (A)	47,939.0	80,658.220	136,574.0	13,664.510
2	Perimeter (p)	868.485	1,130.279	1,559.45	109.256
3	Major axis length (Maj.AL)	320.844	456.601	661.911	56.235
4	Minor axis length (Min.AL)	152.171	225.794	305.818	23.297
5	Eccentricity (e)	0.492	0.860	0.948	0.045
6	Convex area (CA)	48,366.0	81,508.084	138,384.0	13,764.092
7	Extent (E)	0.467	0.693	0.829	0.060
8	Equiv. diameter (ED)	247.058	319.334	417.002	26.891
9	Solidity (s)	0.918	0.989	0.994	0.003
10	Aspect ratio (AR)	1.148	2.041	3.144	0.315
11	Roundness (r)	0.554	0.791	0.939	0.055
12	Compactness (C)	0.560	0.704	0.904	0.053

Table 4 The confusion matrix used in the classification of pumpkin seed grains

Confusion matrix		Predicted	
		Cercevelik	Urgup Sivrisi
Actual	Cercevelik	*tp*	*fp*
	Urgup sivrisi	*fn*	*tn*

(Townsend 1971). The performance set, used to evaluate the success of the classification models, was called a confusion matrix (Bulut 2016). The structure of the confusion matrix, which was used in the classification of pumpkin seeds in this study, had been mentioned in Table 4.

There are four parameters in the confusion matrix, as seen in Table 4.

These are defined as;

Tp: Çerçevelik was estimated, and the result obtained was Çerçevelik,

Fp: Ürgüp Sivrisi was estimated and the result obtained was Çerçevelik,

Fn: Çerçevelik was predicted, and the result obtained was Ürgüp Sivrisi,

Tn: Ürgüp Sivrisi was estimated and the result obtained was Ürgüp Sivrisi.

Performance criteria for the classification methods used in this study were given in Table 5, together with their formulas and evaluation conditions. The success

of the classification was determined by looking at these criteria (Hossin and Sulaiman 2015).
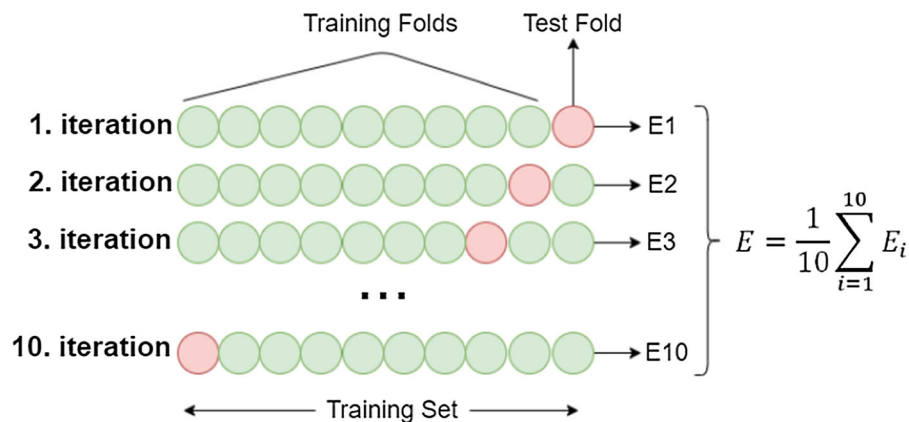
Cross-validation

In the early 1930s (Larson 1931), the best way to evaluate the performance of the statistical relationship between the same data was stated as the testing of the output on new data, and the Cross-Validation method was developed upon this explanation (Arlot and Celisse 2010). In order to create the right model using machine learning methods, training would be done and tested. For this reason, the whole dataset was divided into two, the training set and the test set. It was expected that the data in the training set would realize the learning, whereas the data in the test set would determine the performance of the model against those data that it has never seen. However, the distribution of the dataset also affected the performance of all learning. Therefore, as in the cross-validation method given in Fig. 7, the dataset was divided into k parts, and the k-1 part was used as train data. This process was repeated k times, and the average of the accuracy value found in each iteration was accepted as the performance of cross-validation (Shao 1993). In this study, the tenfold validation method was used.

Development of models

The most important feature of machine learning methods was model creation as different methods used at the time of creating models considered the

Table 5 Performance criteria, formula, and evaluation condition table

No	Performance Measure	Formula	Evaluation
1	Accuracy	$\frac{tp+tn}{tp+fp+tn+fn} \times 100$	The ratio of correct estimates to the total is the number of samples evaluated
2	Precision	$\frac{tp}{tp+fp} \times 100$	It is used to measure the positive patterns that are correctly predicted from the total predictive forms in a positive class
3	Recall	$\frac{tp}{tp+fn} \times 100$	Used to measure the proportion of correctly classified positive patterns
4	Specificity	$\frac{tn}{tn+fp} \times 100$	It is used to measure the proportion of true negatives that have been correctly identified
5	F1-score	$\frac{2tp}{2tp+fp+fn} \times 100$	Represents the harmonic mean between Recall and Precision values

Fig. 7 Cross-validation

characteristics of the dataset and the nature of the problem. These methods included classification, clustering, and regression. Moreover, the dataset used in this study contained numerical input variables and two classes of target variables. Due to these features of the dataset, it was deemed to be appropriate to use classification methods in the study. Classification algorithms trained the model by looking at the patterns of the data in the training set. In this way, it classified the data (test data) so that it could not be seen before in a highly accurate manner.

In this study, pumpkin seed kernels were modeled using LR, MLP, SVM, RF, and k-NN classifiers with the help of Python 3 programming language. They are machine learning methods that are frequently used in classification problems. In addition to these methods, other machine learning methods have also been tried and the methods used in the study have obtained more significant classification results than others.

Logistic regression (LR)

The reason for using logistic regression analysis was to establish a model that could define the fitness between dependent variables and independent variables with the least number of variables (Cruyff 2016). A regression curve was drawn to understand which individual belonged to which population (Kalantar 2018). The curve was calculated by Eq. 1. In this study, the Newton method was used for the optimization during classification with the help of LR.
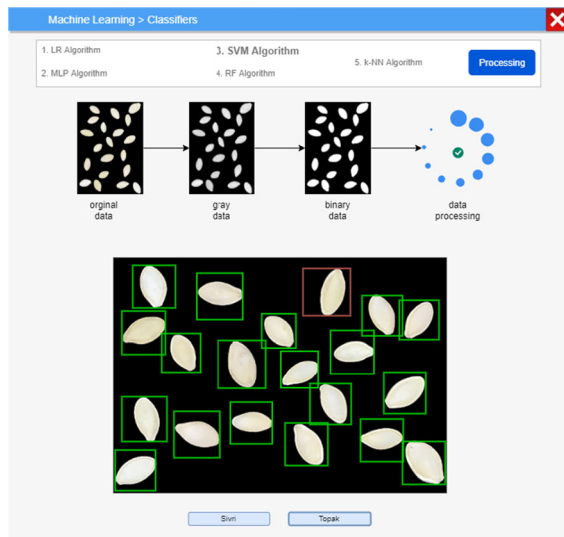
$$\Phi(z) = \frac{1}{e^{-z}} \tag{1}$$

Multilayer perceptron (MLP)

Multi-layered perceptions were formed by many parallel artificial neurons that had non-linear activation functions, which were called MLP (Şen 2004). As seen in Fig. 8, the architecture used for MLP consisted
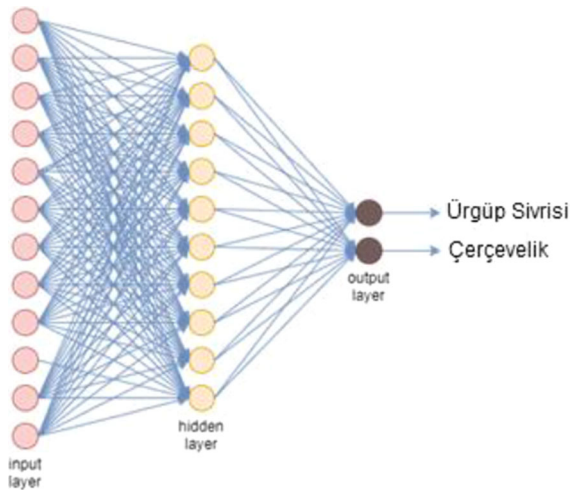
Fig. 8 The MLP classifier architecture used in the study

Fig. 9 The model performance visualization program

of twelve inputs, two outputs, and one hidden layer. There were 10 neurons in parallel throughout the hidden layer. The iteration number was determined as 2500.

Support vector machine (SVM)

Support vector machines predicted a suitable hyperplane function to statistically separate two classes on the multidimensional plane (Kavzoğlu and Çölkesen 2010). In this study, the sigmoid hyperplane function

Table 6 Confusion matrix values of classifiers

Algorithms	Confusion Matrix	
LR	1184	116
	186	1014
MLP	1200	100
	187	1013
SVM	1206	94
	190	1010
RF	1185	115
	196	1004
k-NN	1179	121
	188	1012

Table 7 Performance measurement results of the models

| Measure	LR	MLP	SVM	RF	k-NN
Accuracy	87.92	88.52	**88.64**	87.56	87.64
Precision	91.08	92.31	**92.77**	91.15	90.69
Recall	86.42	**86.51**	86.39	85.80	86.25
Specificity	89.73	91.02	**91.49**	89.72	89.32
F1-Score	88.69	89.32	**89.47**	88.40	88.41

was determined, and the gamma value was accepted as "*1/feature number*".

Random forest (RF)

The random forest classifier classified many random samples, which were sampled independently of the input vector by considering the combination of predictors that received the highest vote from all the tree estimators (Pal 2005). The number of trees in the forest, which was used in the study, was determined to be 100. However, information acquisition was calculated according to the entropy.

K Nearest neighbor (k-NN)

The k-NN or k-nearest neighbor algorithm determined the nearest k points in the same space with each data in the training set, usually by considering the Euclidean distance. The test data entering the model was included in the same class as the lowest one according
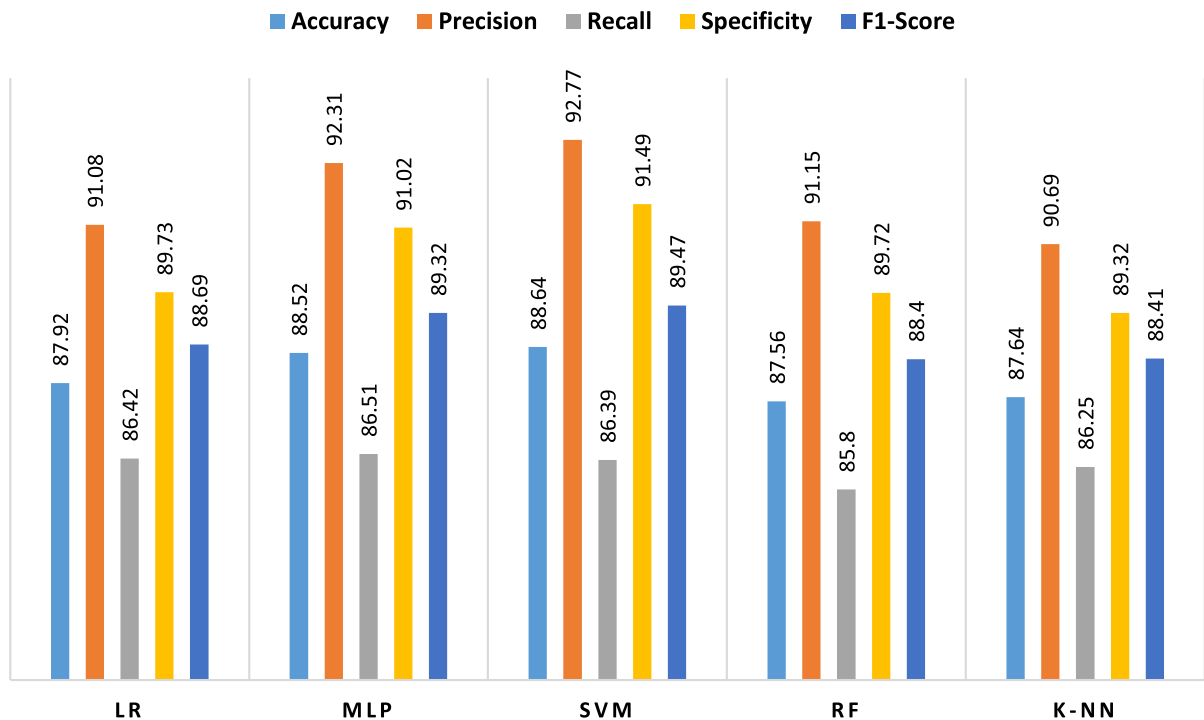
Fig. 10 Performance chart of the models for two types of pumpkin seeds

to the Euclidean distance values (Mahdavinejad et al. 2018). In this study, the k value was considered as five.

Results

The purposes of trade often bred two types of pumpkin seeds in Turkey, which had been studied on a total of 2500 samples. In this study, 12 different morphological features were identified from the images of pumpkin seeds taken in two dimensions. These properties were calculated assuming that each of the nuclei was on a linear plane. The program visual, in which the performances of the models were controlled, had been shown in Fig. 9.

The evaluation results extracted from the confusion matrix of Table 6 had been represented in Table 7. As it could be seen in Table 7 that there were accuracy rates higher than 85 percent in the success of the models made. The accuracy value of the SVM model was found to be 88.64 percent, which proved this model as the most accurate one. Considering the precision value, the success rate was 92.77 percent in the case of the SVM model, whereas 92.31 percent in

the case of the MLP model. According to this evaluation, the SVM model gave more accurate results than the MLP model. As per the Recall evaluation, the rate of correctly defined positive patterns was found to be 86.51 percent in the MLP model, which was higher than that of the SVM. However, concerning the specificity and F1-score evaluations, the results showed that the SVM model was more efficient than other models.

All the performances of the models could be understood by observing the graph in Fig. 10. According to this graph, the models achieved more than 85 percent of success in all the performance values.

Results and discussion

One of the biggest problems of the seed sector is not to remove foreign materials in the seed, but to distinguish different varieties belonging to the same species from the seed. There is no method or machine that can distinguish this. With the machine learning method we use, it will be possible to recognize the varieties and to distinguish the different varieties mixed in the seed

with the machines to be developed. In this way, the recognition of pumpkin seed varieties will be done quickly and accurately by machine learning methods.

The research results were expected to contribute to the quality production of commercial products, such as Çerçevelik and Ürgüp Sivrisi pumpkin seeds used in the study. In this study, 12 morphological features of pumpkin seeds were identified. The identified features were tested using five different machine learning approaches (LR, MLP, SVM, RF, and k-NN) for the two different pumpkin seed classes (Ürgüp Sivrisi and Çerçevelik). The accuracy rates obtained from the models were 87.92 percent, 88.52 percent, 88.64 percent, 87.56 percent, and 87.64 percent, respectively.

Conditional learning was aimed at the investigation on the contribution of 12 morphological features, which had been extracted during the continuation of the study to the model training. However, as per the literature review, the addition of morphological features to the independent ones, such as texture, color, and expert opinion, would ensure the better success of the model. At the same time, this study was aimed to be a source of inspiration, especially for the other exported food and textile products. Quality measurements have been performed with expert opinion for further future. It was predicted that with the introduction of Industry 5.0 into daily lives, the automation of the unmanned factories, whose classifications were dependent on the quality of products, would be increased.

In recent years, the applications of machine learning methods in the field of genetic resources and crop evolution have become increasingly common, bringing solutions to problems in this field and providing an alternative to the methods and techniques used to this day.

The results of the morphological features and technologies related to artificial intelligence, which were used in the study, would be important resources for the development of smart machines for factories. By using these methods with seed data other than pumpkin seeds, seeds can be analyzed. In this way, it will be possible to identify other seed varieties and be used in the solution of yield, disease and species prediction problems.

References

Aktaş N, Uzlaşır T, Tunçil YE (2018) Pre-roasting treatments significantly impact thermal and kinetic characteristics of pumpkin seed oil. Thermochim Acta. https://doi.org/10.1016/j.tca.2018.09.012

Arlot S, Celisse A (2010) A suvey of cross-validation procedures for model selection. Stat Surv. https://doi.org/10.1007/BF02506337

Bulut F (2016) Sınıflandırıcı Topluluklarının Dengesiz Veri Kümeleri Üzerindeki Performans Analizleri. Bilişim Teknolojileri Degisi. https://doi.org/10.17671/btd.81137

Chen X, Xun Y, Li W, Zhang J (2010) Combining discriminant analysis and neural networks for corn variety identification. Computers Electron Agric. https://doi.org/10.1016/j.compag.2009.09.003

Cruyff MJ, Böckenholt U, Van Der Heijden PG, Frank LE (2016) A review of regression procedures for randomized response data, including univariate and multivariate logistic regression, the proportional odds model and item response model, and self-protective responses. Handb Stat 34:287–315. https://doi.org/10.1016/bs.host.2016.01.016

Demir BE (2017) Prediction of physical parameters of pumpkin seeds using neural network. Notulae Botanicae Horti Agrobotanici Cluj-Napoca. https://doi.org/10.15835/nbha45110429

Dheer P, Singh P, Singh V (2019) Classifying wheat vaieties using machine learning model. J Pharmacogn Phytochem. https://doi.org/10.13140/RG.2.2.16338.81600

Düzeltir, B. (2004). *Çekirdek kabağı (Cucurbita pepo L.) hatlarında morfolojik özelliklere göre tanımlama ve seleksiyon çalışmaları.* Master's Thesis, Ankara University Institute of Natural And Applied Sciences. Ankara, Thesis No: 150981

Eldem A, Eldem H, Palalı A (2017) Görüntü İşleme Teknikleriyle Yüz Algılama Sistemi Geliştirme. BEU J Sci. https://doi.org/10.17798/bitlisfen.333984

Guevara F, Gil H, Gomez Gil J (2011) A machine vision system for classification of wheat and barley grain kernels. Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA) 9:672–680. https://doi.org/10.5424/sjar/20110903-140-10

Hossin M, Sulaiman N (2015) A review on evaluation metrics for data classification evaluations. Int J Data Min Knowl Manag Process 5:1–11. https://doi.org/10.5121/ijdkp.2015.52011

Huang M, Tang J, Yang B, Zhu Q (2016) Classification of maize seeds of different years based on hyperspectral imaging and model updating. Computers Electron Agric 122:139–145. https://doi.org/10.1016/j.compag.2016.01.029

Jamuna KS, Kapagavalli S, Vijaya MS, Revathi P, Gokilavani S, Madhiya E. (2010). Classification of seed cotton yield based on the growth stages of cotton crop using machine learning techniques. In: International conference on advances in computer engineering, Bangalore, pp. 312–315. https://doi.org/10.1109/ACE.2010.71

Kalantar B, Pradhan B, Naghibi SA, Motevalli A, Mansor S (2018) Assessment of the effects of training data selection on the landslide susceptibility mapping: a comparison

between support vector machine (SVM), logistic regression (LR) and artificial neural networks (ANN). Geomat Nat Haz Risk 9(1):49–69. https://doi.org/10.1080/19475705.2017.1407368

Kavzoğlu T, Çölkesen İ (2010) Destek Vektör Makineleri ile uydu görüntülerinin sınıflandırılmasında kernel fonksiyonlarının etkilerinin incelenmesi. Harita Dergisi 16:73–82. https://doi.org/10.17475/kastorman.289762

Kayak N, Türkmen Ö, Tevfik A (2018) Çerezlik Kabak (*Cucurbita pepo* L.) Hatlarının SSR (Simple Sequence Repeat) Markörleri ile Karakterizasyonu. Manas J Agric Vet Life Sci. https://doi.org/10.5772/55044

Larson SC (1931) A new formula for predicting the shrinkage of the coefficient of multiple correlation. J Edic Psychol 2:45–55. https://doi.org/10.1214/aoms/1177732951

Mahdavinejad MS, Rezvan M, Barekatain M, Adibi P, Barnaghi P, Sheth AP (2018) Machine learning for internet of things data analysis: a survey. Digit Commun Netw 4:161–175. https://doi.org/10.1016/j.dcan.2017.10.002

Olgun M, Onarcan AO, Özkan K, Işık Ş, Sezer O, Özgişi K et al (2016) Wheat grain classification by using dense SIFT features with SVM classifier. Computers Electron Agric 122:185–190. https://doi.org/10.1016/j.compag.2016.01.033

Otsu N (1979) A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern 9(1):62–66

Pal M (2005) Random forest classifier for remote sensing classification. Int J Remote Sens 26:217–222. https://doi.org/10.1080/01431160412331269698

Pandey N, Krishna S, Sharma S (2013) Automatic Seed classification by shape and color features using machine vision technology. Int J Computer Appl Technol Res 2:208–213. https://doi.org/10.7753/IJCATR0202.1023

Peričin D, Radulović L, Trivić S, Dimić E (2008) Evaluation of solubility of pumpkin seed globulins by response surface method. J Food Eng 84:591–594. https://doi.org/10.1016/j.jfoodeng.2007.07.002

Punn M, Bhalla N (2013) Classification of wheat grains using machine algorithms. Int J Sci Res (IJSR) 2:363–366

Şen Z (2004) Yapay sinir ağları. İstanbul, SU Vakfı. ISBN: 9789756455135

Seymen M, Yavuz D, Dursun A, Kurtar ES, Türkmen Ö (2019) Identification of drought-tolerant pumpkin (*Cucurbita pepo* L.) genotypes associated with certain fruit characteristics, seed yield, and quality. Agric Water Manag 221:150–159

Shao J (1993) Linear model selection by cross-validation. J Am Stat Assoc 88:486–494. https://doi.org/10.1080/01621459.1993.10476299

Townsend JT (1971) Theoretical analysis of an alphabetic confusion matrix. Percept Psychophys 9:40–50. https://doi.org/10.3758/BF03213026

Yanmaz R, Düzeltir B (2003) Çekirdek kabağı yetiştiriciliği. Türk-Koop Ekin, Tarım Kredi Kooparatifi Merkez Bilgi Yayınları 13:22–24. https://doi.org/10.1016/S2095-3119(13)60611-5

Yavuz D, Seymen M, Yavuz N, Türkmen Ö (2015) Effects of irrigation interval and quantity on the yield and quality of confectionary pumpkin grown under field conditions. Agric Water Manag 159:290–298

Yegul M (2012) Seed yield and quality of some inbreed lines in naked seed pumpkin (*Cucurbita pepo var styrica*). Yüzüncü Yıl Üniversitesi Tarım Bilimleri Dergisi. https://doi.org/10.20289/zfdergi.409921