Computers and Electronics in Agriculture 174 (2020) 105507

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Multiclass classification of dry beans using computer vision and machine learning techniques

Murat Koklu*, Ilker Ali Ozkan

Department of Computer Engineering, Selcuk University, Turkey, Konya, Turkey

ARTICLE INFO

Keywords:
Computer vision system
Image processing
Classification of dry beans
Machine learning techniques

ABSTRACT

There is a wide range of genetic diversity of dry bean which is the most produced one among the edible legume crops in the world. Seed quality is definitely influential in crop production. Therefore, seed classification is essential for both marketing and production to provide the principles of sustainable agricultural systems. The primary objective of this study is to provide a method for obtaining uniform seed varieties from crop production, which is in the form of population, so the seeds are not certified as a sole variety. Thus, a computer vision system was developed to distinguish seven different registered varieties of dry beans with similar features in order to obtain uniform seed classification. For the classification model, images of 13,611 grains of 7 different registered dry beans were taken with a high-resolution camera. A user-friendly interface was designed using the MATLAB graphical user interface (GUI). Bean images obtained by computer vision system (CVS) were subjected to segmentation and feature extraction stages, and a total of 16 features; 12 dimension and 4 shape forms, were obtained from the grains. Multilayer perceptron (MLP), Support Vector Machine (SVM), k-Nearest Neighbors (kNN), Decision Tree (DT) classification models were created with 10-fold cross validation and performance metrics were compared. Overall correct classification rates have been determined as 91.73%, 93.13%, 87.92% and 92.52% for MLP, SVM, kNN and DT, respectively. The SVM classification model, which has the highest accuracy results, has classified the Barbunya, Bombay, Cali, Dermason, Horoz, Seker and Sira bean varieties with 92.36%, 100.00%, 95.03%, 94.36%, 94.92%, 94.67% and 86.84%, respectively. With these results, the demands of the producers and the customers are largely met about obtaining uniform bean varieties.

1. Introduction

Dry bean (*Phaseolus vulgaris L.*) is the most important and the most produced pulse (Fabaceae - Leguminosae) over the world. Similarly, dry bean also takes an important role in agriculture in Turkey. On the other hand, the plant is sensitive to the effect of climatic changes. Resistance and/or tolerance to plant stress factors may be increased by breeding of new seed cultivars and determination of the seed characteristics (Ceyhan et al., 2012) that is one of the most important factors on success of the plant growing.

Ascertainment of the best seed is the main problem for dry bean producers and markets as well (Onder et al., 2012). It is fair that, using of lower quality seed in production will induce to lower quantity even if all the cultivation conditions are provided. The seed is an important input in agricultural expense and dry beans take an important part in food technology. Most of the activities about dry bean seed technology is realized especially by private companies. Dry bean is originally from America, while there is a wide genetic diversity in the world (Gentry, 1969). There are many varieties and local ecotypes of dry beans in Turkey as well (Ceyhan et al., 2014). According to Turkish Standards Institution, dry beans are called as (in Turkish) “Barbunya, Battal, Bombay, Calı, Dermason, Horoz, Tombul, Selanik and Seker” depending on their botanical characteristics (Turkish Standards Institution, 2009). Analysis and classification of dry bean genotypes, which are very common in both Turkey and around the world, constitute one of the main processes in crop production (Granitto et al., 2002).

The presentation of products of dry beans determined by physical features such as appearance, size, color, internal health and diversity increases the market value. Furthermore, the identification of bean varieties helps farmers to use seeds with basic standards for planting and marketing (Chen et al., 2010). Seed quality is the key to bean cultivation in terms of yield and disease. Manual classification and sorting of bean seeds is a difficult process. Furthermore, this method is very time consuming and inefficient, especially when working at high production volumes. For these reasons, automatic methods are required

* Corresponding author at: Department of Computer Engineering, Faculty of Technology, Selcuk University, 42031 Konya, Turkey.
E-mail addresses: mkoklu@selcuk.edu.tr (M. Koklu), ilkerozkan@selcuk.edu.tr (I.A. Ozkan).

https://doi.org/10.1016/j.compag.2020.105507
Received 19 June 2019; Received in revised form 5 May 2020; Accepted 12 May 2020

for grading and classification.

In Turkey, the seeds of beans cultivated are divided into varieties by taking into account the features of the form, shape, type and structure of the Turkish Standards Institute, as well as the market situation. The most well-known of these bean varieties are Barbunya, Bombay, Cali, Dermason, Horoz, Seker and Sira (Onder et al., 2013).

In the last decade, image processing techniques and CVS are used to classify the dry bean seeds in research studies. (Gomes and Leta, 2012; Vibhute and Bodhe, 2012). These systems allow to classify the seed varieties according to parameters such as quality, color and size. In these studies, color, morphological and shape features of seeds are frequently used (Kiratiratanapruk and Sinthupinyo, 2011).

In their study, Sabanci et al. performed a CVS based on ANN for the classification of bread and *durum* wheat grains. In their study, they obtained a total of 21 features over 100 breads and 100 *durum* wheat grains. In order to separate the two types of wheat, they have allocated their data with 90% training and 10% test by Holdout method. They obtained 99.92% classification accuracy as a result of the obtained ANN model (Sabanci et al., 2017).

In their study, Araújo et al. used the multivariate granulometry method based on correlation for quality control in the same bean variety. In this study a computer system for the visual inspection of beans according to their colors was presented. This system consisted of 3 modules: pixel color mapping, grain partitioning and grain sorting. Using the kNN classification algorithm, they achieved 99.88% accuracy (Araújo et al., 2015).

Kilic et al. developed a CVS that takes into account the dimensions and color quantities of the samples for quality control of the beans. They used artificial neural networks for color determination of beans. The samples they obtained were divided into five classes by the system and experts. The ANN was tested with 371 samples. The overall correct classification accuracy of the system is 90.6% (Kılıc et al., 2007).

Kara et al. determined the size and shape of beans by using the image processing method on digital images of 12 different bean varieties. Depending on these features, beans are classified. This study stated that there are differences in the size, shape and morphological features of bean varieties (Kara et al., 2013).

The first step in seed classification is to obtain digital images when evaluating the studies conducted. Different digital image acquisition methods are used according to the developed system and seed. The features of the seeds were extracted by means of image processing techniques on the obtained images. Especially morphological, color and shape features were frequently used in studies. In addition, it is seen that color features have a high effect on classification success in the studies in which the color is important. Due to the difficulties of image acquisition, it is seen that there are few seed images to be classified in many studies. Different machine learning methods were used in the features obtained in order to form the classification models. Generally, Multilayer perceptron (MLP), Support Vector Machine (SVM), K Nearest Neighborhood (kNN) and Decision Tree (DT) models were used for classification in the studies (Kılıc et al., 2007; Sun et al., 2016; Teye et al., 2014).

Dry beans are widely cultivated and consumed in Turkey. Seed color of the bean varieties are very similar. Present studies focused on genetic diversity, cultural practices, agronomic features, physiology and other topics to realize bean classification (Bozoglu and Gulumser, 2000; Onder et al., 2013). When the studies carried out in the classification of beans were examined, a study, which categorizes a large number of standard varieties and in which classification is made by using a large data set containing the features of the sample was not found in the scientific literature.

Using certified dry bean seeds in Turkey is around 10% (Bolat et al., 2017). Dry bean cultivation in Turkey and Asian countries usually in the form of populations containing mixed species of seeds. Also, there is not much certified seed planting area (Varankaya and Ceyhan, 2012). Since different populations which contain different genotypes are cultivated, the final products contain different species of seeds. Thus, when the dry bean seeds obtained from population cultivation are released to the market without being separated by species, the market value decreases immensely (Varankaya and Ceyhan, 2012). Eliminating the disadvantageous situation caused by population cultivation will provide an economic benefit to the producer as a uniform type of bean is obtained. In addition, examination and pricing of products in the market is determined by experts. Because of the human factor, this process is error-prone and not entirely objective. In recent years, the use of data mining and artificial intelligence techniques to solve forecasting and classification problems in agriculture has become widespread.

The primary objective of this study is to provide a method for obtaining uniform seed varieties from the production which is in the form of population. This study will determine the type of products that come to the market and will also provide the product parameters that will determine the price. Also, the parameters obtained in the study will constitute the data set that will enable the use of data mining and artificial intelligence methods. Within this scope, the aim of this study was to develop an artificial intelligence-based CVS for the classification of basic types of dry beans determined by the Turkish Standards Institute, which contains morphologically similar features and has no distinctive color features. In this aspect, Multilayer perceptron (MLP), Support Vector Machine (SVM), k-Nearest Neighbors (kNN), Decision Tree (DT) classification models were created with 10-fold cross validation and performance metrics were compared.

2. Materials and methods

In this study, firstly, the images of dry beans were obtained, and unwanted noise on the image was cleared by image processing methods. Histogram information of gray image information was obtained, and histogram equalization was made. In the image where beans became clear, dry beans were thresholded by Otsu's method (Otsu, 1979). The Otsu's algorithm is based on calculating the optimal threshold value by processing the histogram information of the image. In the next step, after the segmentation of the bean images, dimensional and shape features were extracted. The recommended flow chart for the classification of dry beans is given in Fig. 1. In the modeling stage, the classification of dried beans was carried out with the four most common machine learning techniques. In the last step, the performance evaluation of the models was done, and the best model was selected.

2.1. Image acquisition

In order to obtain images of dry beans, a computer vision system given in Fig. 2 was designed and implemented. The proposed system consists of a camera lens mount and an image capture camera (Prosilica GT2000C) and a special illumination box to prevent shadow formation on the background. The Prosilica GT2000C camera used for the study is 2.2 megapixels, 2048 × 1088 resolution RGB camera, CMOS type sensor, full resolution at a maximum frame rate of 53.7fps and an efficient operating temperature range of −20 °C to +65 °C (Vision, 2018). The camera was placed 15 cm above the samples at the top of the box. To provide a homogeneous lighting environment, the box was illuminated by lightings on the top. The box is completely closed during image capture to have proper lighting and to eliminate environmental noise. The camera has a 12 cm × 10 cm shooting area inside the box. The signals from the dry bean samples were captured by the camera and transferred to the computer via the ethernet port. MATLAB R2016a version was used to capture and save the captured images.

One kilogram of the seven basic dry bean varieties determined by the Turkish Standards Institute were obtained from certified seed producers. Before the image was obtained, dry beans were placed with a gap on a dark background. Thus, they were not touching each other. Photographs were taken until all the samples for each class were finished. In total, 13,611 dry bean samples were obtained from 236

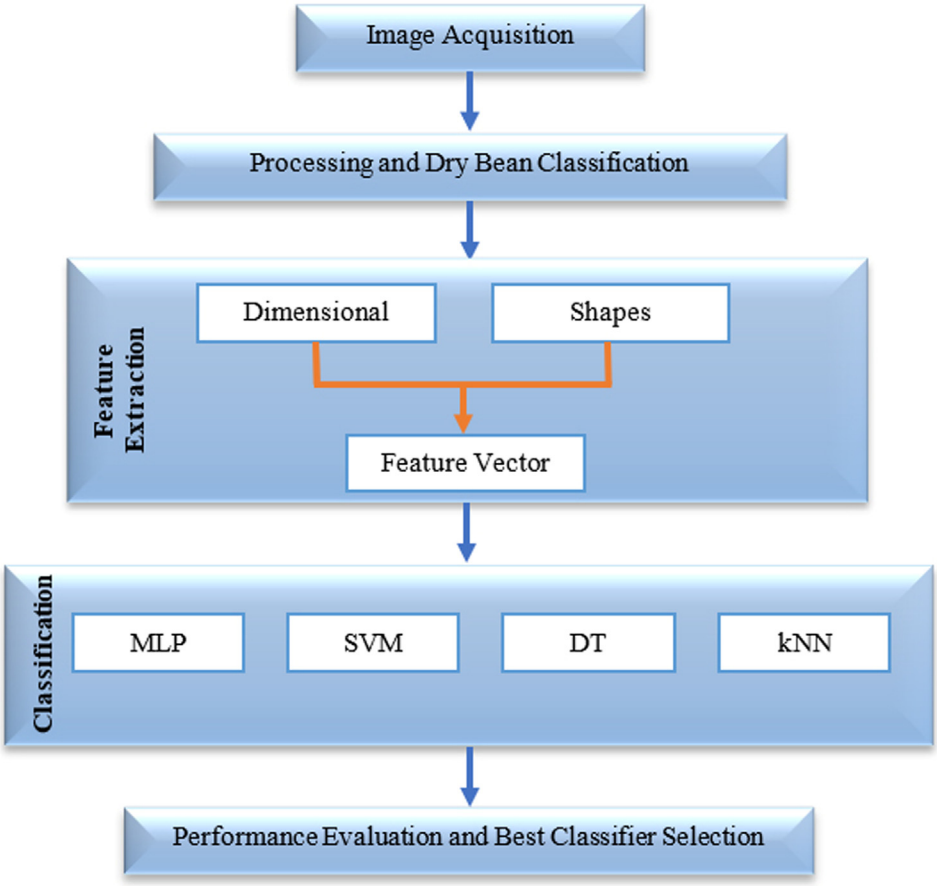

Fig. 1. Flowchart of CVS for bean classification.

Fig. 2. Computer vision system design for dry bean classification.

images. Fig. 3 shows examples of these images.

2.2. Processing and segmentation

This step includes the processes used to prepare images prior to feature extraction. Since the successful implementation of image processing processes directly affects the classification result, it is important that this stage is properly designed. The steps of the pre-processing and segmentation process on the image are as in Fig. 4. The image processing stage consists of removing the dry bean shadows, eliminating background noise and separating each dry bean from the others. Since dimensional and shape features were used in the features, the image was converted to gray. Then, Otsu's global thresholding method was applied in order to select the threshold level, which is used to convert the image into a binary image. Thus, the method of Otsu divides an image into sections, maximizing the separability of the two populations.

2.3. Feature extraction

After the segmentation step, images containing separated dry beans were obtained. A number of features describing each dry bean were extracted. As shown in Fig. 3, dry beans do not have a distinctive color feature. Dimensional and shape features were determined by feature analysis. Geometry related features were obtained with MATLAB software from binary images. In total, 12 dimensional and 4 shape features were obtained for each dry bean. Values found in all features are in pixel count. The most effective dimensional features in the classification of dry bean seeds are as follows (Paliwal et al., 2001):

a) **Area** (A): The area of a bean zone and the number of pixels within its boundaries.

$$A = \sum_{r,c \in R} 1 \tag{1}$$

where r, c is size of region R

b) **Perimeter** (P): Bean circumference is defined as the length of its border.
c) **Major axis length** (L): The distance between the ends of the longest line that can be drawn from a bean.
d) **Minor axis length** (l): The longest line that can be drawn from the bean while standing perpendicular to the main axis.
e) **Aspect ratio** (K): Defines the relationship between L and l.

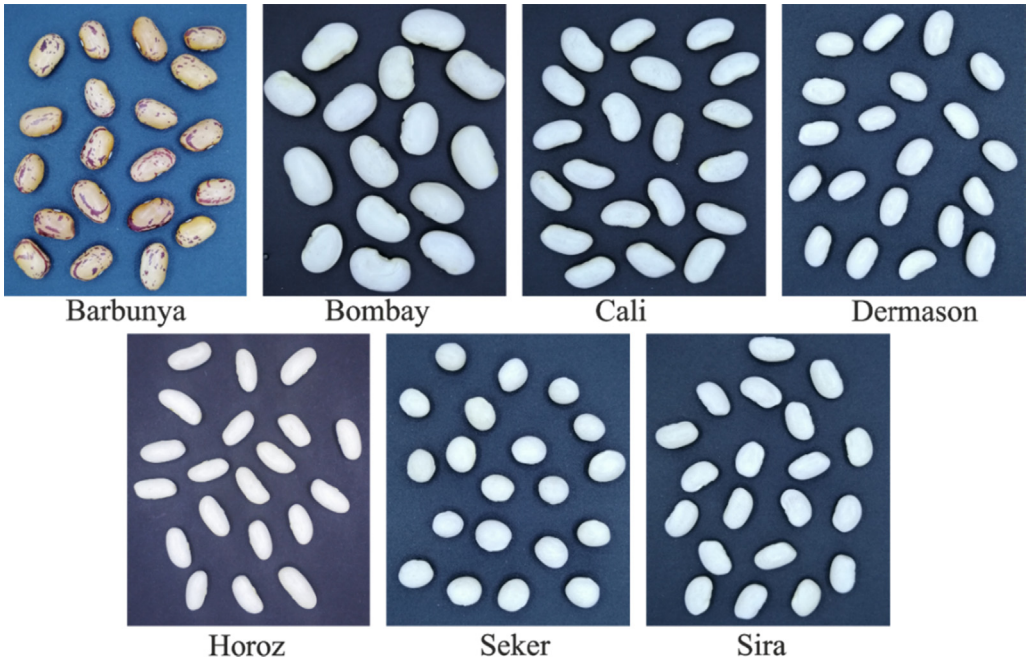

Fig. 3. Sample of taken dry bean images.

$$K = \frac{L}{I} \tag{2}$$

f) **Eccentricity** (Ec): Eccentricity of the ellipse having the same moments as the region.

g) **Convex area** (C): Number of pixels in the smallest convex polygon that can contain the area of a bean seed.

h) **Equivalent diameter** (Ed): The diameter of a circle having the same area as a bean seed area.

$$d = \sqrt{\frac{4 * A}{\pi}} \tag{3}$$

i) **Extent** (Ex): The ratio of the pixels in the bounding box to the bean area.

$$Ex = \frac{A}{A_B} \text{where} A_B = \text{Areaofboundingrectangle} \tag{4}$$

j) **Solidity** (S): Also known as convexity. The ratio of the pixels in the convex shell to those found in beans.

$$S = \frac{A}{C} \tag{5}$$

k) **Roundness** (R): Calculated with the following formula:

$$R = \frac{4\pi A}{P^2} \tag{6}$$

l) **Compactness** (CO): Measures the roundness of an object:

$$CO = \frac{Ed}{L} \tag{7}$$

The shape features used in the classification of the bean seeds are as follows (Pazoki et al., 2014):

$$ShapeFactor1(SF1) = \frac{L}{A} \tag{8}$$

$$ShapeFactor2(SF2) = \frac{l}{A} \tag{9}$$

$$ShapeFactor3(SF3) = \frac{A}{\frac{L}{2} * \frac{L}{2} * \pi} \tag{10}$$

$$ShapeFactor4(SF4) = \frac{A}{\frac{L}{2} * \frac{l}{2} * \pi} \tag{11}$$

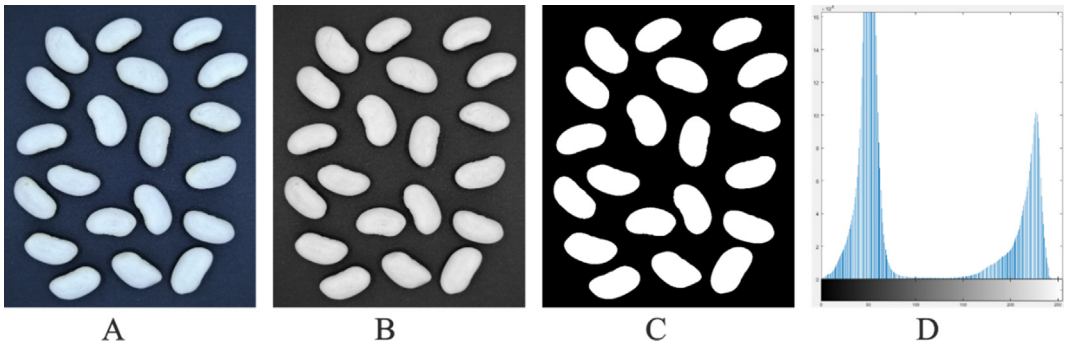

Fig. 4. Image preprocessing and segmentation. (A) Original image in RGB color space. (B) Grayscale image of original image. (C) A segmented beans after preprocessing operations. (D) The channel histogram of grayscale image.

2.4. Dry beans features dataset

Seven different types of dry beans were used in this research, taking into account the features such as form, shape, type and structure by the Turkish Standards Institute as well as the market situation. They are called Seker, Barbunya, Bombay, Cali, Dermosan, Horoz and Sira. The general features of the specified dry beans are as follows (Turkish Standards Institution, 2009).

Cali; It is white in color, its seeds are slightly plump and slightly larger than dry beans and in shape of kidney.
Horoz; Dry beans of this type are long, cylindrical, white in color and generally medium in size.
Dermason; This type of dry beans, which are fuller flat, is white in color and one end is round and the other ends are round.
Seker; Large seeds, white in color, physical shape is round.
Bombay; It is white in color, its seeds are very big and its physical structure is oval and bulging.
Barbunya; Beige-colored background with red stripes or variegated, speckled color, its seeds are large, physical shape is oval close to the round.
Sira; Its seeds are small, white in color, physical structure is flat, one end is flat, and the other end is round.

For the purpose of visual presentation of bean varieties, all bean varieties in the study using CVS are given in Fig. 5.

Table 1 shows the quantity distribution and average weight of 13,611 dry bean samples in the study. The sample weights are different due to the different seed weights of the varieties as the samples are taken from equal weight of each type. This database is publicly available at: https://www.muratkoklu.com/datasets/.

The minimum, maximum, mean and standard deviation data of the features obtained for all dry bean samples are given in Table 2.

2.5. Performance measures

When a model is created for a classification problem or when existing models are used, the success of that model is calculated by the number of accurate estimates from all predictions. However, this information only gives the correctness of the classification. To determine whether a model is good enough, the classification accuracy alone is usually not sufficient. To explain the estimation results of a classifier is to use the confusion matrix. A confusion matrix is a table frequently used to describe the performance of the classification model with a set of known test data and has 4 parameters. These are called true positives, true negatives, false positives and false negatives. These four values constitute the complexity matrix for binary classification in Table 3. Table 4 shows the confusion matrix and its representation for multiple classes (Hossin and Sulaiman, 2015).

Fig. 5. Types of dry beans used in the study.

Table 1
Sample distribution of all types of dry beans.

No	Name	Piece	Seed Weight (average gram per seed)
1	Seker	2027	0.49
2	Barbunya	1322	0.76
3	Bombay	522	1.92
4	Cali	1630	0.61
5	Horoz	1928	0.52
6	Sira	2636	0.38
7	Dermason	3546	0.28
Total		**13,611**	**0.71**

Table 2
Statistical distribution of features of dry bean varieties (in pixels).

No	Features	Min.	Max.	Mean	Std. Deviation
1	Area	20420.000	254616.000	53048.285	29324.096
2	Perimeter	524.736	1985.370	855.283	214.290
3	Major Axis Length	183.601	738.860	320.142	85.694
4	Minor Axis Length	122.513	460.198	202.271	44.970
5	Aspect Ratio	1.025	2.430	1.583	0.247
6	Eccentricity	0.219	0.911	0.751	0.092
7	Convex Area	20684.000	263261.000	53768.200	29774.916
8	Equiv Diameter	161.244	569.374	253.064	59.177
9	Extent	0.555	0.866	0.750	0.049
10	Solidity	0.919	0.995	0.987	0.005
11	Roundness	0.490	0.991	0.873	0.060
12	Compactness	0.641	0.987	0.800	0.062
13	Shape Factor 1	0.003	0.010	0.007	0.001
14	Shape Factor 2	0.001	0.004	0.002	0.001
15	Shape Factor 3	0.410	0.975	0.644	0.099
16	Shape Factor 4	0.948	1.000	0.995	0.004

Table 3
Confusion matrix and representation for binary classes.

		Predicted	
		Positive C^+	Negative C^-
Actual	Positive C^+	True positive **(tp)** *The number of positive predicted positive values.*	False negative **(fn)** *Number of false negative estimates.*
	Negative C^-	False positive **(fp)** *The number of false positive estimates.*	True negative **(tn)** *The number of correctly predicted negative values.*

Table 4
Confusion matrix and representation for multiple classes.

		Predicted				
		C_1	C_2	C_3	…	C_n
Actual	**C_1**	T_1	F_{12}	F_{13}	…	F_{1n}
	C_2	F_{21}	T_2	F_{23}	…	F_{2n}
	C_3	F_{31}	F_{32}	T_3	…	F_{3n}
	…	…	…	…	…	…
	C_n	F_{n1}	F_{n2}	F_{n3}	…	T_n

To be able to evaluate and scale the model, the success criteria such as Accuracy, Error Rate, Sensitivity, Specificity, Precision, Recall and F1-Measure are calculated using Table 3. For binary-class data sets, these calculations are given in Table 5 (Hossin and Sulaiman, 2015; Sokolova and Lapalme, 2009).

The classification metrics and their explanations based on the generalization of the measurements in Table 5 for multi-class data, as in

Table 5
Calculation formulas and explanations of binary class metrics.

Measure	Formula	Evaluation Focus
Accuracy	$\frac{tp+tn}{tp+fp+tn+fn}$	It is used to measure the ratio of accurately estimated samples to the total number of samples. It can be considered that the model is the best if there is high accuracy in the model used.
Error Rate	$\frac{fp+fn}{tp+fp+tn+fn}$	It is used to measure the ratio of the values of incorrectly estimated samples to the total number of samples.
Recall (r)	$\frac{tp}{tp+fn}$	It is used to measure the proportion of positive values classified as true.
Specificity	$\frac{tn}{tn+fp}$	It is used to measure the proportion of negative values classified as true.
Precision (p)	$\frac{tp}{tp+fp}$	The ratio of correctly classified positive samples to estimated total positive samples. This is also called a Positive Predictive Value.
F1-Score	$\frac{2*p*r}{p+r} = \frac{2*\frac{tp}{tp+fp}*\frac{tp}{tp+tn}}{\frac{tp}{tp+fp}+\frac{tp}{tp+tn}}$	It is the harmonic mean of sensitivity. Therefore, it takes into account both false positives and false negatives. Especially in cases of irregular class distribution, looking at the F1-score may be more useful than looking at the accuracy.

Table 6
Calculation formulas and explanations of multiple class metrics.

Measure	Formula	Evaluation Focus
Averaged Accuracy	$\frac{\sum_{i=1}^{l}\frac{tp_i+tn_i}{tp_i+fn_i+fp_i+tn_i}}{l}$	It is used to calculate the mean success of classes.
Averaged Error Rate	$\frac{\sum_{i=1}^{l}\frac{fp_i+fn_i}{tp_i+fn_i+fp_i+tn_i}}{l}$	It is used to calculate the mean error rate of classes.
Averaged Precision (P_M)	$\frac{\sum_{i=1}^{l}\frac{tp_i}{tp_i+fp_i}}{l}$	It is used to calculate the mean of the precision per class.
Averaged Recall (r_M)	$\frac{\sum_{i=1}^{l}\frac{tp_i}{tp_i+fn_i}}{l}$	It is used to calculate the mean of the reminder per class.
F1-Score (Averaged F-Measure)	$\frac{2*p_M*r_M}{p_M+r_M} = \frac{2*\frac{\sum_{i=1}^{l}\frac{tp_i}{tp_i+fp_i}}{l}*\frac{\sum_{i=1}^{l}\frac{tp_i}{tp_i+fn_i}}{l}}{\frac{\sum_{i=1}^{l}\frac{tp_i}{tp_i+fp_i}}{l}+\frac{\sum_{i=1}^{l}\frac{tp_i}{tp_i+fn_i}}{l}}$	It is used to calculate F1-Score per class.

this study, are given in Table 6 (Hossin and Sulaiman, 2015; Sokolova and Lapalme, 2009).

2.6. Cross validation

Cross-validation is a method developed to increase the security of classification. Cross-validation randomly divides the data set into the specified number of identical size sets. It trains the system with the remaining sets by accepting one of the subsets as a test set. This process is repeated until all the number of sets are tested in the system (Arlot and Celisse, 2010). The results obtained from these processes are generalized. As shown in Fig. 6, the number of sub-folds (k) was selected as 10 in this study.

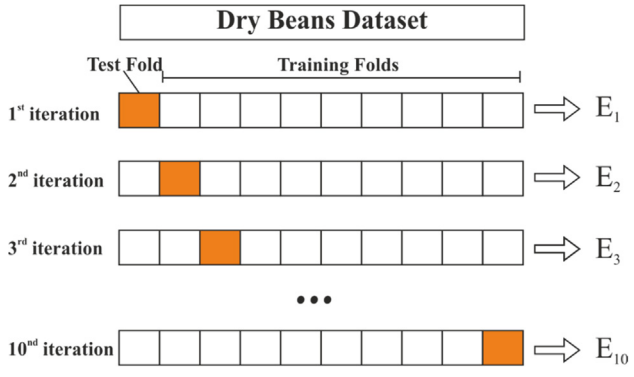

Fig. 6. An example of a 10-fold cross validation.

3. Development of modelling

Classification models have a very important place for automated decision-making system. The model helps to determine which features belongs to which class. There are many algorithms that make the classification process. They can give different results for different data sets. For this, it is important for making a healthy decision to use the most appropriate classifier according to the obtained data. In this study, models were prepared by using MLP, SVM, DT and kNN which are frequently used in literature in order to make classification according to the features obtained from dry beans.

3.1. Multi-layer perceptron (MLP)

MLP are computer systems that can learn events using examples and determine how responses to events from the environment are generated. Similar to the functional characteristics of the human brain, they are successfully applied in areas such as learning, association, classification, generalization, feature identification and optimization. MLP creates its own experiences with the information obtained from the samples and then make similar decisions on similar issues. MLP consists of artificial neurons that are hierarchically connected to each other and capable of working in parallel (Przybył et al., 2018).

As seen in Fig. 7, artificial nerve cells come together to form MLP. Generally, the cells are in the form of three main layers and are parallel in each layer (Castañeda-Miranda and Castaño, 2017). MLP network structure used in this study:

Input layer: 16 parameters (X_1–X_{16}) used in bean data were used as input parameters of MLP.
Hidden Layers: When the results of the study are analyzed, the hidden layer structure with the optimal result is determined as 12-3.

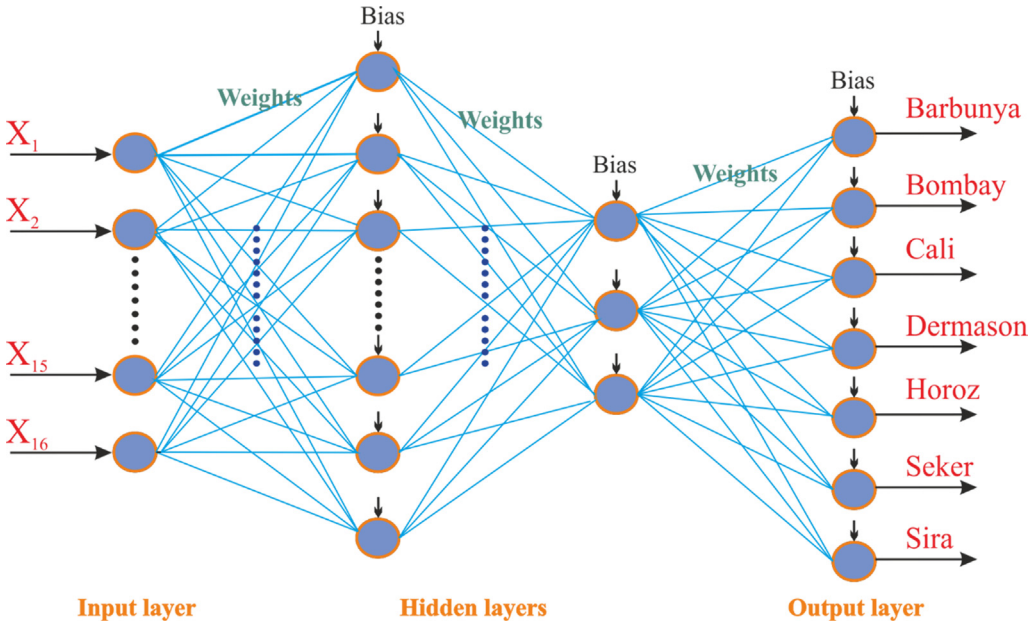

Fig. 7. MLP Architecture used in this study.

Output Layer: The output layer consists of 7 types of dried beans: Seker, Barbunya, Bombay, Cali, Dermosan, Horoz and Sira. They are determined as O_1–O_7.

The MLP network structure with 17-12-3-7 architecture used in this study is shown in Fig. 7 and the parameters of the network used are given in Table 7.

3.2. Support vector machine (SVM)

Support Vector Machines (SVM) is a kernel-based method with high computational power for classification and regression problems (Vapnik, 2013). Compared to other machine learning methods, SVM has better generalization. SVM has a solid theoretical base and provides more accurate results in many applications than other algorithms. As shown in Fig. 8, SVMs use a maximal margin separator (Nakano et al., 2016). This separator represents the most remote control possible to the sample point. SVMs can also classify nonlinear data by moving the data to a higher size by a method called kernel trick. Usually, data that cannot be separated linearly in the original input space can be separated in high dimensional feature space. SVMs are at a level to represent complex problems and are resistant to overfitting. SVM has been developed for binary classifications, and it is possible to obtain accurate classification results with a small number of sampling data (Foody and Mathur, 2004). Originally designed for the classification of binary class linear data, the method was then developed for the classification of multiple-class and non-linear data. Binary classifiers are used to generalize on the multiple class problems. For this generalization, there are two basic schemes: one-vs-one and one-vs.-all. One-vs.-one approach was used in this study. The one-vs-one approach creates a classifier for each class pair. A new instance is applied on binary classifiers, and the decision to estimate is determined by a majority vote technique (Thuraisingham et al., 2017).

The SVM approach can be extended to a non-linear surface using a kernel trick. The Cubic core function used in nonlinear SVM classifiers in this study is as follows:

Polynomial Function (Quadric: d = 2 and Cubic: d = 3)

$$k(x_i, x_j) = (x_i, x_j + 1)^d \tag{12}$$

$$Max \sum_{i=1}^{i} \propto_i - \frac{1}{2} \sum_{j=1}^{i} \propto_i \propto_j y_i y_j k(x_i * y_j) \tag{13}$$

Cubic SVM classifier was used in this study and box constraint level was selected as one in SVM.

Table 7
MLP Parameters.

Parameter Name	Parameter
Hidden Layer Number	2
Hidden Layer Activation Function	Sigmoid
Output Layer Activation Number	Sigmoid
Learning Rate	0.3
Minimum Performance Gradient	1e−5
Performance Goal	1e−3
Maximum Number of Epochs to Train	500

3.3. Decision tree (DT)

A decision tree is a decision support tool that uses a tree-like graphic or model to show its decisions and possible results. DT is used to determine the path to be followed in a decision analysis. Because of their cheap installation, easy interpretation, easy integration with database systems and good reliability, DT is widely used in classification models (Kotsiantis et al., 2007). It has the advantages of using visually and intuitively and being understood.

In this study, the maximum number of partitions for decision trees is 16 at the classification stage. The criterion of partition in decision trees was selected as Ginis diversity index. The decision tree obtained in this study is given in Fig. 9.

3.4. K-nearest neighborhood (kNN)

The K- Nearest Neighborhood (kNN) algorithm is one of the methods of pattern recognition that classifies objects based on the closest educational examples in the attribute space (Alpaydin, 2009). This algorithm makes the classification according to the given k value according to the class of the nearest neighbor. In the kNN algorithm, the classification of a vector is made using known vectors of the class. The sample to be tested is individually processed with each sample in the training set. To determine the class of the sample to be tested, the k

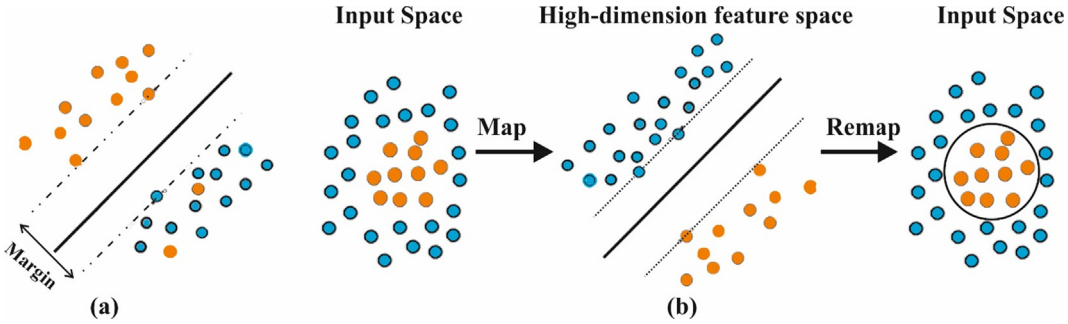

Fig. 8. Feature space of a SVM (a) Linear SVM (b) Non-linear SVM.

samples closest to that sample in the training set are selected. In the cluster consisting of the selected samples, considering which class has the most samples, the sample to be tested belongs to this class. In this study, Eculidean criterion given in Eq. (14) is used.

$$D_{L2}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (14)$$

In this study, for Nearest Neighbor, distance importance is equal, and the number of neighbors is 10.

4. Results

Population cultivation with different types of beans in the production of dry beans prevents the production of uniform crops. Thus, the final product which contains different species of dry beans causes economic losses. In this study, in order to resolve this problem, it was aimed to separate seven different species of dry beans determined by the Turkish Standards Institute (TSE) which are cultivated in Turkey. For this purpose, 13,611 bean sample images of the bean varieties obtained by CVS using image pre-processing and segmentation processes. For each of the samples obtained, 16 features were extracted, and dry beans feature dataset were created. This multiclass dry bean data set is important in terms of showing the relationship between shape and dimensional features of dry bean species. Classification models were created by using MLP, SVM, DT, and kNN machine learning techniques which are widely used for classification, and classification performances were obtained. All models were tested with 10-

Table 8
MLP Confusion Matrix.

Actual	Predict						
	Barbunya	Bombay	Cali	Dermason	Horoz	Seker	Sira
Barbunya	1184	4	87	2	3	10	32
Bombay	1	518	3	0	0	0	0
Cali	44	5	1530	0	34	3	14
Dermason	1	0	0	3293	2	52	198
Horoz	5	0	29	16	1829	0	49
Seker	14	0	0	49	1	1899	64
Sira	9	0	4	325	40	26	2232

Table 9
SVM Confusion Matrix.

Actual	Predict						
	Barbunya	Bombay	Cali	Dermason	Horoz	Seker	Sira
Barbunya	1221	1	60	0	6	11	23
Bombay	0	522	0	0	0	0	0
Cali	42	0	1549	0	24	4	11
Dermason	1	0	0	3346	4	38	157
Horoz	6	0	27	18	1830	0	47
Seker	9	0	1	49	1	1919	48
Sira	9	0	5	278	33	22	2289

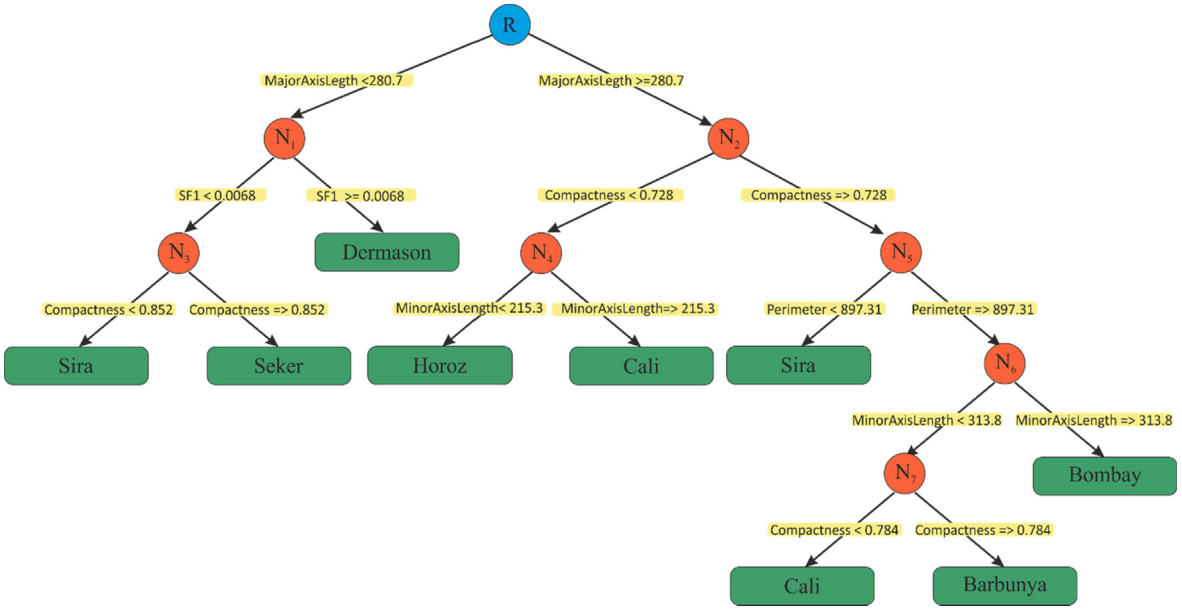

Fig. 9. Decision tree structure used in this study.

Table 10
DT Confusion Matrix.

Actual	Predict						
	Barbunya	Bombay	Cali	Dermason	Horoz	Seker	Sira
Barbunya	904	1	352	0	3	12	50
Bombay	1	517	3	0	1	0	0
Cali	140	0	1455	0	24	1	10
Dermason	0	0	0	3209	1	73	263
Horoz	2	0	102	15	1709	1	99
Seker	7	0	0	62	0	1877	81
Sira	2	0	28	258	7	45	2296

Table 11
kNN Confusion Matrix.

Actual	Predict						
	Barbunya	Bombay	Cali	Dermason	Horoz	Seker	Sira
Barbunya	1199	0	78	0	6	10	29
Bombay	0	522	0	0	0	0	0
Cali	36	0	1546	0	30	2	16
Dermason	0	0	0	3264	6	62	214
Horoz	2	0	25	10	1844	0	47
Seker	8	0	0	41	1	1920	57
Sira	9	0	6	256	39	28	2298

Table 12
Performance values obtained for MLP, SVM, DT ve kNN.

Perfomance Measures	MLP	SVM	DT	kNN
Accuracy (%)	91.73	**93.13**	87.92	92.52
Error Rate (%)	8.27	**6.87**	12.08	7.48
Precision (%)	93.11	**94.45**	89.19	93.93
Recall (%)	92.68	**94.03**	87.93	93.59
Specificity (%)	98.53	**98.77**	97.91	98.67
F1 Score (%)	92.88	**94.23**	88.29	93.75

fold cross validation (Arlot and Celisse, 2010). Confusion matrices of the MLP, SVM, DT, and kNN models developed are given in Tables 8–11, respectively.

The accuracy, error rate, precision, specificity, recall, F1-score classification performance metrics calculated by using the confusion matrix of each model are given in Table 12 (Hossin and Sulaiman, 2015).

When Table 12 is examined, it is seen that all models except DT have classification success with the rate of over 90%. The SVM Model has the best value with a 93.13% of accuracy. In addition, the SVM classification model obtained has the best values for all calculated performance metrics. As a measure of how well the classifier performs, it is significant that the F1-Score value is also high in all the classification models. In addition, it is seen that the Precision value, which is actually positive in all classifiers, determines the ratio of the number of the positive classifieds to the all positives, is proportional to the accuracy.

When the confusion matrix of the SVM model with the best classification value given in Table 9 is examined, accuracy rates of Barbunya, Bombay, Cali, Dermason, Horoz, Seker and Sira were 92.36%, 100.00%, 95.03%, 94.36%, 94.92%, 94.67% and 86.84%, respectively. The accuracy rates of 4 different models obtained for all bean varieties are given in Fig. 10. It can be seen that Bombay variety can be fully classified with 100% accuracy. Sira variety has the lowest classification performance of all kinds. In addition, it is seen that the performance of Dermason and Sira varieties is low in terms of differentiation in confusion matrix.

The DT model given in Fig. 9 can divide all the bean varieties into two groups according to the Major Axis Length feature. In addition, the compactness feature is used in the decision tree model to differentiate the Sira and Seker varieties from each other, Horoz and Cali varieties from Barbunya and Bombay varieties. 87.92% accuracy can be achieved by using only the features of the Major Axis Length, Minor Axis length, Perimeter, Compactness and Shape Factor 1 from the sixteen features available in the DT model.

The SVM model has an accuracy of 93.19% in population planting with seven different dry bean seeds. When the number of classes is

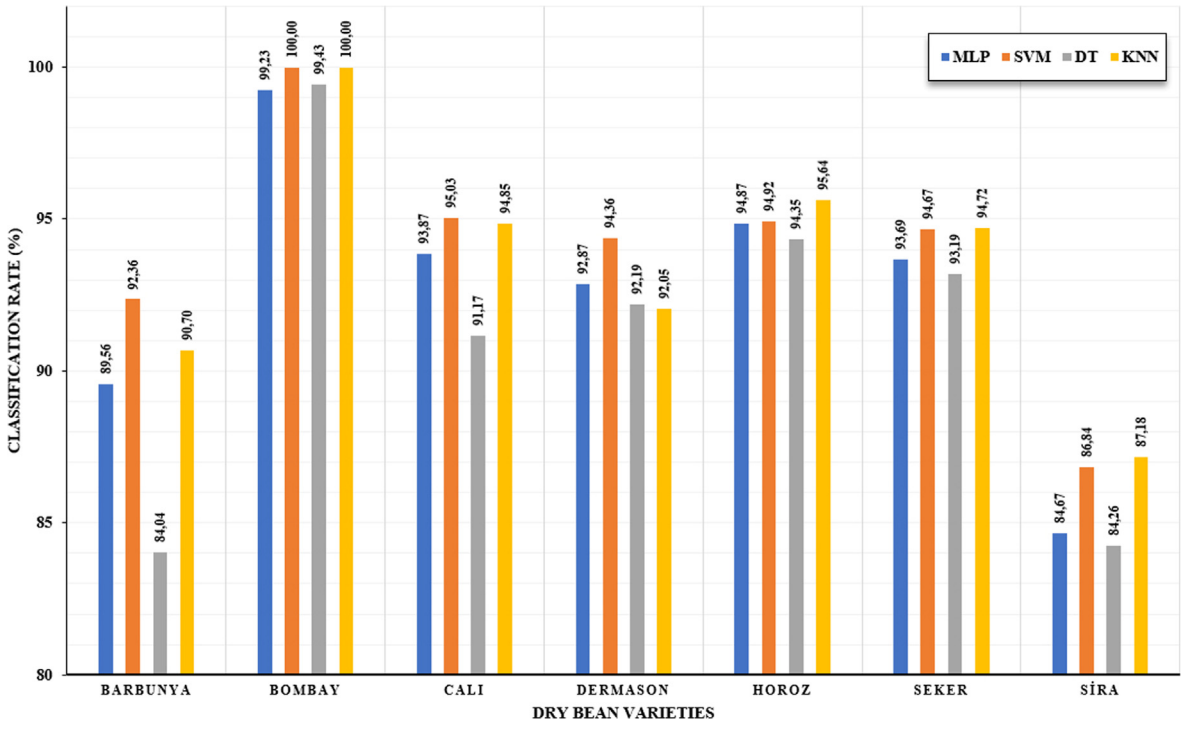

Fig. 10. Accuracy of classification models for all bean varieties.

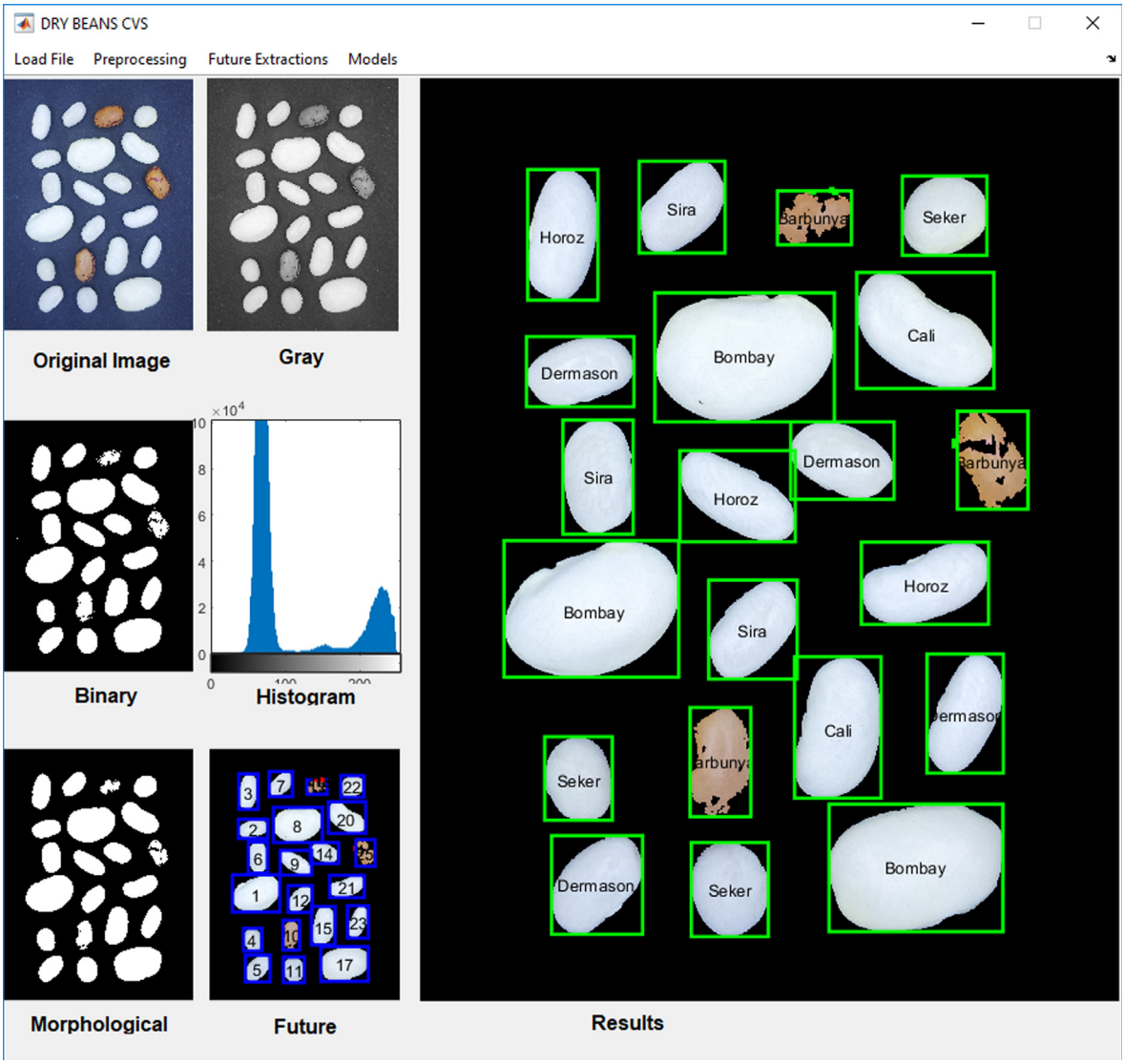

Fig. 11. MATLAB GUI Interface for dry bean classification.

taken into consideration compared to other studies in the literature, it can be seen that the success rate is high. In addition, the success of classification increases in populations containing fewer types of dry bean species.

In order to classify dry bean images outside of this dataset, an interface was developed using MATLAB GUI. As shown in Fig. 11, the images loaded through the interface are firstly subjected to segmentation and the features of each bean obtained are extracted. According to the model selected over the interface, beans are labeled on the image.

5. Conclusions

Regarding the classification of dry bean seeds, dimension and shape features of bean varieties have no external discriminatory features, which causes this classification process to be complex. Classification of bean seed varieties is important in terms of assuring seed uniformity and quality.

In this study, the results obtained with 10-fold cross validation show that the accuracy of the models and the performance scores of the F1 Score are mostly overlapping. This shows that the data distribution is regular. High success rates in all metrics show that the models are successful in the classification. It is seen that all classification models have the lowest sorting performance of the Sira variety due to the low distinguishing of the Sira bean variety with the Dermason variety. The fact that the flatness and roundness features of the Dermason and Sira varieties are similar is effective in this result.

In the study, due to the high number of data in the DT model, a tree was created, which did not include the features of all data. By changing the parameters, there is a high degree of harmony in the different types of tree that use more features and the interpretation of the trees becomes difficult.

The results show that the proposed classifier based on CVS can be successfully used to automatically classify various types of dry beans in this study. In addition, this developed model structure can also be used for the types of dry beans of different regions. The model can be further improved by the hybrid use of machine learning methods, deep learning and new algorithms.

In the study, the variables related to the shape and size characteristics of the bean cultivars were taken from two dimensional images. The third dimension of beans was not included in machine learning. This dimension is the suture axis of the bean. If the suture axis of the bean was included, it would be possible to increase the classification success. However, classification machines in the industry, the seed flows through the orifice quickly. So, the analysis of the third dimension is difficult. Also, the seeds are generally analyzed in two dimensions. In the machine learning technique based on two-dimensional images, the

differences in the shape of each bean variety could be used as a separate variable (e.g., coefficient of variation for the roundness of Sira cultivar). If the coefficient of variance is also included in the shape and size variables of each cultivar, the classification success of bean cultivars may increase. Beside shape and size features, the texture features and statistical features can improve the classification results. Additionally, more studies are needed on this topic to meet the demands for producers and customers.

Another suggestion may be that the resulting MATLAB application can be moved to the mobile environment and turned into a mobile application with more effective and practical usage. Thanks to this mobile application, it can be made easier to use by many people in the agricultural world, and a common image bank of dry beans can be created by uploading images from all around the world.

CRediT authorship contribution statement

Murat Koklu: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - review & editing. **Ilker Ali Ozkan:** Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Alpaydin, E., 2009. Introduction to Machine Learning. MIT Press.

Araújo, S.A. De, Pessota, J.H., Kim, H.Y., 2015. Beans quality inspection using correlation-based granulometry. Eng. Appl. Artif. Intell. 40, 84–94. https://doi.org/10.1016/j.engappai.2015.01.004.

Arlot, S., Celisse, A., 2010. A survey of cross-validation procedures for model selection. Stat. Surv. 4, 40–79. https://doi.org/10.1214/09-SS054.

Bolat, M., Unuvar, F.I., Dellal, I., 2017. The determination of future trends for Turkey's edible grain legumes. J. Agric. Econom. Res. (JAER) 3, 7–18 https://dergipark.org.tr/en/download/article-file/393717.

Bozoglu, H., Gulumser, A., 2000. Determination of genotype x environment interactions of some agronomic characters in dry bean (*Phaseolus vulgaris L.*). Turkish J. Agric. For. 24, 211–220.

Castañeda-Miranda, A., Castaño, V.M., 2017. Smart frost control in greenhouses by neural networks models. Comput. Electron. Agric. 137, 102–114. https://doi.org/10.1016/j.compag.2017.03.024.

Ceyhan, E., Kahraman, A., Onder, M., 2012. The impacts of environment on plant products. Int. J. Biosci. Biochem. Bioinforma. 2, 48–51. https://doi.org/10.7763/IJBBB.2012.V2.68.

Ceyhan, E., Harmankaya, M., Kahraman, A., 2014. Combining ability and heterosis for concentration of mineral elements and protein in common bean (*Phaseolus vulgaris L.*). Turkish J. Agric. For. 38, 581–590. https://doi.org/10.3906/tar-1307-56.

Chen, X., Xun, Y., Li, W., Zhang, J., 2010. Combining discriminant analysis and neural networks for corn variety identification. Comput. Electron. Agric. 71, S48–S53. https://doi.org/10.1016/j.compag.2009.09.003.

Foody, G.M., Mathur, A., 2004. Toward intelligent training of supervised image classifications: directing training data acquisition for SVM classification. Remote Sens. Environ. 93, 107–117. https://doi.org/10.1016/j.rse.2004.06.017.

Gentry, H.S., 1969. Origin of the common bean, Phaseolus vulgaris. Econ. Bot. 23, 55–69 https://www.jstor.org/stable/4253014.

Gomes, J.F.S., Leta, F.R., 2012. Applications of computer vision techniques in the agriculture and food industry: a review. Eur. Food Res. Technol. 235, 989–1000. https://doi.org/10.1007/s00217-012-1844-2.

Granitto, P.M., Navone, H.D., Verdes, P.F., Ceccatto, H.A., 2002. Weed seeds identification by machine vision. Comput. Electron. Agric. 33, 91–103. https://doi.org/10.1016/S0168-1699(02)00004-2.

Hossin, M., Sulaiman, M.N., 2015. A review on evaluation metrics for data classification evaluations. Int. J. Data Min. Knowl. Manag. Process 5, 1.

Kara, M., Sayıncı, B., Elkoca, E., Öztürk, İ., Özmen, T.B., 2013. Seed size and shape analysis of registered common bean (*Phaseolus vulgaris L.*) cultivars in Turkey using digital photography. Tarım Bilim. Derg. 19, 219–234. https://doi.org/10.1501/Tarimbil_0000001247.

Kılıc, K., Boyaci, I.H., Koksel, H., Kusmenoglu, I., 2007. A classification system for beans using computer vision system and artificial neural networks. J. Food Eng. 78, 897–904. https://doi.org/10.1016/j.jfoodeng.2005.11.030.

Kiratiratanapruk, K., Sinthupinyo, W., 2011. Color and texture for corn seed classification by machine vision. In: 2011 International Symposium on Intelligent Signal Processing and Communications Systems (ISPACS). IEEE, pp. 1–5.

Kotsiantis, S.B., Zaharakis, I., Pintelas, P., 2007. Supervised machine learning: a review of classification techniques. Emerg. Artif. Intell. Appl. Comput. Eng. 160, 3–24.

Nakano, T., Nukala, B.T., Zupancic, S., Rodriguez, A., Lie, D.Y.C., Lopez, J., Nguyen, T.Q., 2016. Gaits classification of normal vs. patients by wireless gait sensor and Support Vector Machine (SVM) classifier. In: 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), pp. 1–6.

Onder, M., Ates, M.K., Kahraman, A., Ceyhan, E., 2012. The problems and suggestions to dry bean farming in Konya region. IJANS, Int. J. Agric. Nat. Sci. 5, 143–148.

Onder, M., Kahraman, A., Ceyhan, E., 2013. Correlation and path analysis for yield and yield components in common bean genotypes (*Phaseolus vulgaris L.*). Ratar. i Povrt. 50, 14–19. https://doi.org/10.5937/ratpov50-3958.

Otsu, N., 1979. A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man, Cybernetics 9, 62–66. https://doi.org/10.1109/TSMC.1979.4310076.

Paliwal, J., Visen, N.S., Jayas, D.S., 2001. AE—automation and emerging technologies: evaluation of neural network architectures for cereal grain classification using morphological features. J. Agric. Eng. Res. 79, 361–370. https://doi.org/10.1006/jaer.2001.0724.

Pazoki, A.R., Farokhi, F., Pazoki, Z., 2014. Classification of rice grain varieties using two artificial neural networks (mlp and neuro-fuzzy). J. Anim. Plant Sci. 24, 336–343.

Przybył, K., Gawałek, J., Koszela, K., Wawrzyniak, J., Gierz, L., 2018. Artificial neural networks and electron microscopy to evaluate the quality of fruit and vegetable spray-dried powders. Case study: strawberry powder. Comput. Electron. Agric. 155, 314–323. https://doi.org/10.1016/j.compag.2018.10.033.

Sabanci, K., Kayabasi, A., Toktas, A., 2017. Computer vision-based method for classification of wheat grains using artificial neural network. J. Sci. Food Agric. 97, 2588–2593. https://doi.org/10.1002/jsfa.8080.

Sokolova, M., Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Inf. Process. Manag. 45, 427–437. https://doi.org/10.1016/j.ipm.2009.03.002.

Sun, J., Jiang, S., Mao, H., Wu, X., Li, Q., 2016. Classification of black beans using visible and near infrared hyperspectral imaging. Int. J. Food Prop. 19, 1687–1695. https://doi.org/10.1080/10942912.2015.1055760.

Teye, E., Huang, X., Han, F., Botchway, F., 2014. Discrimination of cocoa beans according to geographical origin by electronic tongue and multivariate algorithms. Food Anal. Methods 7, 360–365. https://doi.org/10.1007/s12161-013-9634-4.

Thuraisingham, B., Khan, L., Parveen, P., Masud, M.M., 2017. Big Data Analytics with Applications in Insider Threat Detection, 1st Editio. ed. Auerbach Publications, New York. Doi:10.1201/9781315119458.

Turkish Standards Institution, 2009. Dry Beans (Kuru Fasulye). Turkish Stand. Inst. Off. Gaz. Repub, Turkey.

Vapnik, V., 2013. The nature of statistical learning theory. Springer-Verlag, New York, NY. Doi:10.1007/978-1-4757-3264-1.

Varankaya, S., Ceyhan, E., 2012. Orta Anadolu Bölgesinde Fasulye Tarımında Karşılaşılan Problemler ve Çözüm Önerileri. Selçuk Tarım Bilim. Derg. 26, 15–26.

Vibhute, A., Bodhe, S.K., 2012. Applications of image processing in agriculture: a survey. Int. J. Comput. Appl. 52, 34–40. https://doi.org/10.5120/8176-1495.

Vision, A., 2018. Prosilica GT2000 URL:https://www.alliedvision.com/en/products/cameras/detail/Prosilica%20GT/2000.html [WWW Document]. URL https://www.alliedvision.com/en/products/cameras/detail/Prosilica GT/2000.html.