Classification of Omani's Dates Varieties Using Artificial Intelligence Techniques

Salima AL-Abri, Lazhar Khriji*, Ahmed Ammari, Medhat Awadalla
Sultan Qaboos University
Muscat, Oman
lazhar@squ.edu.om, s92461@student.squ.edu.om, chiheb@squ.edu.om, medhatha@squ.edu.om

***Abstract*—Date fruits are considered as one of the most popular fruits in the Middle East. Oman is one of the countries that have many varieties of dates and the most well-known are Khalas, Fardh and Khunaizi. Nowadays, the process of classifying different varieties of dates in date's industries is done manually by human workers. The manual process affects the quality of dates, which however is subjective, time consuming, laborious and expensive. The objective of this paper is to classify automatically six popular date varieties, namely, Khalas, Khunaizi, Fardh, Qash, Naghal, and Maan from their images based on color, shape, size, and texture features. Three different artificial intelligence techniques have been used for automatic classification and qualitative comparison; (i) Artificial Neural Network (ANN), (ii) Support Vector Machine (SVM), and (iii) K-Nearest Neighbor (KNN). The Dates' varieties were obtained from AL-Dhahira Governorate. In total, 600 date samples (100 dates/class) were selected. These samples were imaged individually, one date per image. Nineteen features were extracted from each image and used in classification models. Experimental results show that the ANN algorithm outperforms the SVM and KNN based algorithms in all criteria used. The achieved results of ANN using 15 features and seven tan-sigmoid neurons in the hidden layer were 99.2% in classification accuracy, 99.12% in average recall and 99.25% in average precision.**

I. Introduction

Date fruits are of great importance in human diet, due to its high-energy value and nutrient content. North Africa and Middle-East are considered as the largest date producer countries in the world according to Food Agricultural Organization (FAO). As shown in Fig.1, Oman was among the 10 top dates producer countries with annual production of 360,917 tons in 2017 [1]. Oman is one of the countries that depends on dates as a source of income. It has more than 250 varieties of dates, which can be distinguished from each other by their colors, shapes, sizes, and texture. Khunaizi considered as the sweetest variety while Khalas as the most delicious variety [2]. Classification of dates into different classes is very important task in date's industries, which needs careful and hard effort from the workers. If dates classification can be automated by means of an intelligent system, varieties of dates can be classified accurately and very fast which will improve the date's industries [3]. This study aims to propose an automatic date's classification system that classifies dates varieties from their images using Artificial Intelligence Techniques. Most of the related work in the date's classification field classify dates based on grade. In 2012, a method for automatic classification of date fruits using computer vision and pattern recognition is developed. In their system, they tested seven different categories of dates and extracted total of 15 features. They used multiple classifiers in their system for comparison purposes such as Nearest Neighbor, Linear Discriminant Analysis (LDA), and Neural Network (NN) [3]. A new date fruits sorting system using artificial neural networks (ANN) was proposed also in 2012. They used two models of neural networks, (i) multi-layer perceptron (MLP) with (ii) backpropagation and Radial Basis Function (RBF) networks. Their neural networks achieved a recognition rate of 87.5% for MLP and 91.1% for RBF [4]. In 2014, a system of automatically classifying different types of dates from their images was found by Ghulam Muhammad [5]. In this work, texture, color and shape features are extracted from the dates images. For dimensionality reduction of features vector, Fisher discrimination Ratio (FDR) was used and SVM was used as a classifier [5]. A computer vision system with a monochrome camera was proposed in 2016 to classify dates based on hardness. This study used histogram and texture features in their system and LDA and ANN were implemented as classifiers [6,7]. In 2018, an automated system that identifies different dates fruit maturity status and classify their categories is developed. Color, size and skin texture features are extracted. The system counts the number of dates, classify them into different classes and identify the defects [8].

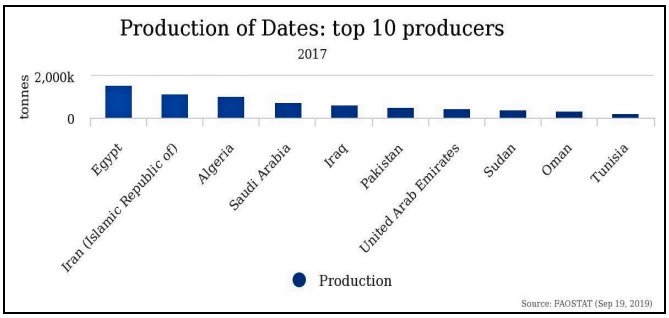

Fig.1. The top 10 dates producer countries in 2017 according to FAO [1]

The aim of this paper is to accurately classify six different varieties of dates: Khalas, Khunaizi, Fardh, Qash, Naghal, and Maan. We will work to extract color, shape, size, and texture features of various date images. Artificial Neural Network (ANN), Support Vector Machine (SVM), and K-Nearest Neighbor (K-NN) classifiers are proposed and comparative performance analysis are conducted.

This paper is organized as follows: Section II describes the materials and methods. Section III gives the basic principles of

classification models. Experimental results and discussions are given in Section IV. Section V concludes the paper.

II. Materials and Methods

The proposed system starts by obtaining colored images of dates with single date per image. Preprocessing and segmentation are then applied to the colored images and are followed by mathematical morphological operations for removing incorrectly segmented pixels. The segmented images are then used to extract color, shape-size and texture features. At the beginning, each type of features was used alone to train the different classifiers. Then the combination of these features was, also, tested in order to examine the effectiveness of each one of them and to tune the parameters of each classifier. The classifiers are then tested by applying new data in order to evaluate their performances. The flowchart of the developed system is shown in Fig.2.

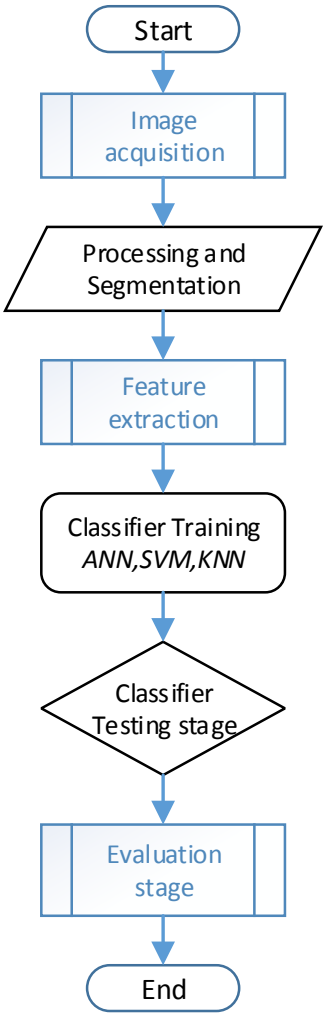

Fig. 2. Flowchart of the proposed system

A. Samples Collection

Khalas, Fardh, Khunaizi, Qash, Naghal and Maan were used in this study since they are the most popular varieties of dates in Oman. All varieties were obtained from AL-Dhahira Governorate. A total of 600 date samples were selected (100 samples for each class). These samples were imaged individually, one date per image, and processed.

B. Image Acquisition System

The image acquisition system used in this study consisted of three main components: a personal computer, RGB color camera (model: EOS 1100D, Canon, Taiwan, resolution of 4272× 2848 pixel) and two fluorescent lights. We used an A4 white paper as image background where we positioned manually each date sample at a distance of 15 cm from the camera. We took images using camera's self-timer mode with one image per sample. Then we downloaded the captured images to the computer for further processing.

C. Preprocessing and Segmentation

An algorithm for processing the acquired images was developed using MATLAB (Version R2014a, The Mathworks Inc., Natick, MA, USA). The steps followed in processing the images are illustrated in Fig.3. First, the colored images are resized to ease and speed up the processing. Then, copies of the colored images are converted into grayscale images. The grayscale images are then segmented into foreground and background regions so that only, the region of interest is used. The segmentation process is implemented using Otsu's method [9] followed by morphological operations, such as "imdilate" and "imopen" (built-in functions in MATLAB software) as well as "hole filling" were employed. Otsu's method is considered as a global thresholding technique, which uses the image histogram to find a global threshold that separates the image into two classes (background and foreground) [10]. Threshold searching process tries to maximize the variance of the two classes [11]. The segmented images are binary images where the foreground is white and the background is black. The morphological operations were performed on images to improve the segmentation process [4]. The steps of segmentation are shown in Fig.3. After the object (date fruit) was segmented from the background, the binary images were used to identify the vertical and horizontal coordinates that correspond to the date. These coordinates then used to crop the original colored gray and binary images so that only the region of interest is used for the feature extraction phase.

D. Features Extraction

Different dates varieties can be distinguished by their colors, size-shape as well as skin texture features. The description for all features is explained in the following sections.

1) Color Features

Since dates varieties are different in colors, the color features provide powerful information in the field of date's classification. Regarding the color, nine features were extracted. First, the cropped RGB images are converted into three-color channels, which are red (R) channel, green (G) channel and blue (B) channel. Then, from each channel, the mean and the standard deviation were computed. In addition, the gray images were used to compute three more features, which are minimum, maximum and mean intensity. The value

of the pixel with the smallest and greatest intensity in the region represents the minimum and maximum intensity respectively while the mean of all the intensity values in the region represents the mean intensity [3].

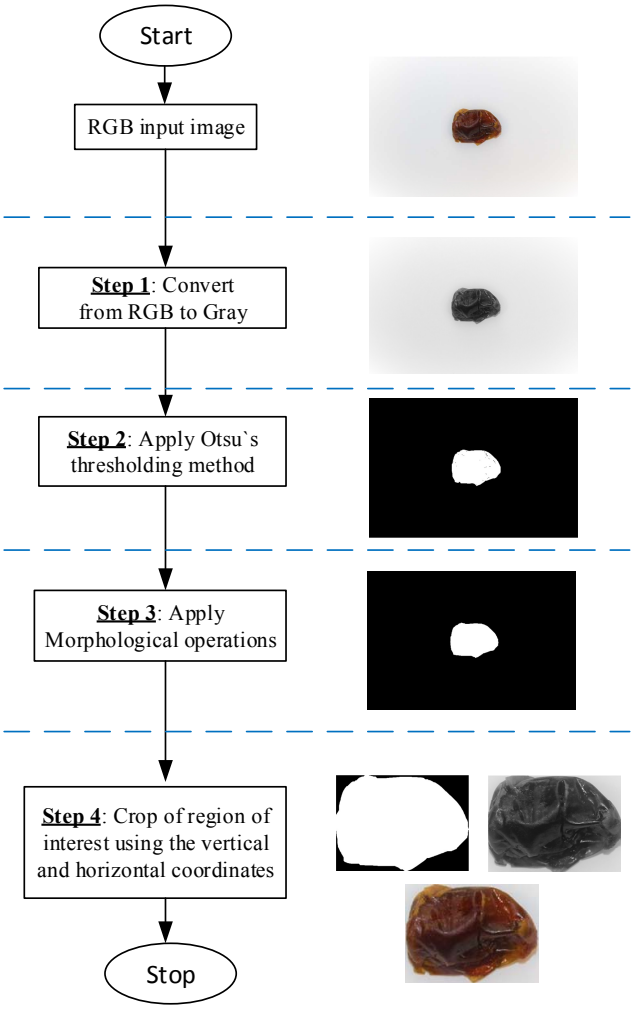

Fig. 3. Segmentation steps applied to the date samples

2) Shape and size Features

The sizes and shape are important features for date's classification because different types of dates have different shapes and sizes. These features can improve the accuracy of classification if they were included in the features vector. As illustrated by Fig.4 the shape and size features were calculated from the segmented images in terms of pixels: Area, Major axis length, Minor axis length, Ellipse eccentricity, solidity and perimeter. The area is obtained by counting the number of pixels in the segmented images. Due to the natural shapes of the dates, ellipse was selected as the best modeling shape [10]. Major axis and minor axis lengths were computed by finding the length of the major and minor axes of the ellipse with the same normalized second central moments as the region. In addition, the eccentricity was also used, which represents the ratio between the major axis length and the distance separating the two foci. The solidity was calculated by finding the proportion of the pixels in the convex hull that are also in the region. The perimeter was computed by counting number of pixels in the boundary of the extracted object [3].

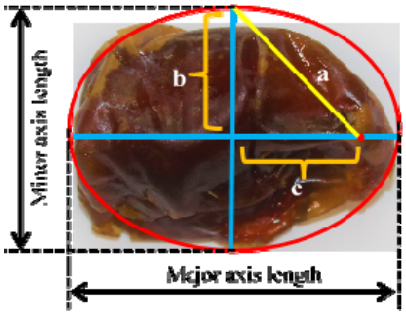

Fig. 4. Major axis, minor axis and eccentricity parameters of the ellipse

Eccentricity (e) is computed as,

$$e = \frac{c}{a} = \frac{\sqrt{a^2 - b^2}}{a} \quad (1)$$

Where c is the distance from the center to the focus of the ellipse and a is the distance from the center to a vertex.

The solidity is calculated by finding the proportion of the pixels in the convex hull that are also in the region. It shows the degree to which shape is concave or convex. It is given by,

$$Solidity = \frac{Area}{Convex\ Area} \quad (2)$$

The perimeter is computed by counting the number of pixels in the boundary of the extracted object.

3) Texture Features

Some types of dates can be separated from each other by their skin texture, so it was necessary to consider dates texture as features. The Gray Level Co-occurrence Matrix (GLCM) is one of the methods that can be used to calculate statistical texture features [13]. GLCM gives an indication of how often different combinations of gray levels co-occur in an image. To calculate the GLCM, the RGB images were first converted into gray scale. Next, the gray-scale range was divided to eight levels [3]. Then, each pixel was substituted with its gray level. After that, the gray level of each pixel and its right neighbor were observed. Finally, GLCM was computed where each entry (i, j) represents the number of occurrence of a pixel j to the right of a pixel i. Once the GLCM is constructed, the texture features are calculated using the content of the GLCM. In this study, only four texture features were used: contrast, energy, correlation and homogeneity. Contrast represents the intensity contrast between a pixel and its neighbor over the whole image. Energy is the sum of squared elements in GCLM. Correlation measures how a pixel is correlated to its neighbor. Homogeneity measures how the distribution of the GLCM's elements is close to the GLCM diagonal [13].

$$Contrast = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(i-j)^2\, p(i,j) \quad (3)$$

$$Correlation = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\frac{p(i,j)-\mu_x\mu_y}{\sigma_x\sigma_y} \quad (4)$$

$$Energy = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}p(i-j)^2 \quad (5)$$

$$Homogeneity = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\frac{p(i,j)}{1+(i-j)^2} \quad (6)$$

Where, *Ng* is the number of distinct gray levels and *p(i,j)* represents the *(i,j)th* entry in the GLCM. μx and μy are the means in the raw and column directions respectively. σx and σy are the standard deviations in the raw and column directions respectively.

III. Classification Models

For classification, three different classifiers were used for comparison purposes: (i) Artificial Neural Network (ANN), (ii) Support Vector Machine (SVM), and (iii) K-Nearest Neighbor (KNN).

A. Artificial Neural Network classifier

Multilayer neural network is one of the common classifiers used in fruit recognition researches. The networks used were two-layer feed-forward networks trained with Levenberg-Marquardt backpropagation with tansig and logsig-hidden neurons and softmax output neurons. The dataset was divided into 3 subsets with 68% for training (408 images), 12% for validation (72 images) and 20% for testing (120 images). In order to test the effect of different features and the architecture of neural network in the classification task, the neural network was trained and tested 30 times with different number of neurons in the hidden layer. For each type of features (color, shape-size and texture), the neurons in the hidden layer was changed over the range (1-10). First, the network is trained and tested 30 times for each number of hidden neurons. Then, over these 30 times, the average accuracy was calculated and the results are shown in the result section.

B. Support Vector Machine

SVM is a set of related supervised learning methods used for regression and classification purposes. It is a binary classifier used for data classification in many applications [11], [14]. SVM takes a set of input data and for each input, it will predict to which class it belongs. SVM training algorithm constructs a model that classifies new sample into either class1 or class2. Specifically, SVM find the optimal boundary that separates the 2 classes with the largest margin between separating boundaries and support vectors (SVs) [5], [15]. As a result, the generalization error of the classifier will be reduced [16]. As the proposed system deals with six types of dates that belong to six different classes, a multi-class SVM was applied for classification. Multiclass SVM reduces the problem to a series of binary problems. One versus All coding design approach was adopted which creates six binary learners. For the kernel function, radial base function (RBF) was used. A 10-fold cross validation approach was used. In this approach, the training subset (80% of images) is divided into 10 folds. In each iteration, 9 out of the 10 folds are used as training, while the remaining fold is used for validating. Then, all the folds are tested after ten iterations. The remaining 20% of the samples were used for testing. The optimal value for the RBF kernel scale was automatically set while the optimization parameter "C" of SVM were changed to 5 different values in the range (1e-1, 1e-3). For each C value, the training and testing were repeated 30 times and the average testing accuracy was calculated.

C. K- Nearest Neighbor classifier

The idea behind KNN classifier is that it tries to find the nearest neighbors to a query data by calculating the distance between the query sample and all other samples which are then used to determine the class of that query [17]. In this study, the Euclidian distance was applied and the number of nearest neighbors is varied in the range from 1 to 10. For each k value, the same procedure as in SVM was applied in order to calculate the average testing accuracy.

IV. Experimental Results and Discussion

A. Setup of the Artificial Neural Network

For the neural network, two activation functions in the hidden layer were used for comparison purposes. Fig.5 and Fig.6 show the Plots of the accuracies of the ANN classifier using different features vs number of hidden neurons for logsig and tansig activation functions respectively. It is clear that as the number of hidden neurons increases from 1 to 3, the increase in the accuracy is very clear. However, the accuracy improvement is very small when the number of neurons is ranging from 4 to 10.

The four texture features computed from GLCM achieves the lowest accuracy among other features. The accuracy ranges between 36.08% and 53.72% for logsig function while it ranges between 35.36% and 54.36% for tansig function. This shows that texture of different dates varieties are very similar so that it cannot be used alone to classify different dates. On the other hand, color and shape features achieve better accuracy when used alone compared to texture features. The color features accuracy range from 58.03% to 80.06% and from 62.58% to 79.67% for logsig and tansig respectively. The shape features achieve comparable accuracy to the color features, which is clear from Fig.5 and Fig.6. Shape accuracy falls in the range from 65.88% to 81.11% for logsig and from 63.67% to 81.19% for tansig. when all features are used to represent the date samples, the classification accuracy improves significantly and can reach up to 96.22% for logsig and 96.21% for tansig. However, the accuracy was improved further (97% and 97.26%) when the texture features were ignored and only color and shape features were used. The highest accuracy achieved in our study was 97.26%, which is reached by tansig neural network with seven hidden neurons as illustrated in TABLE I. . Logsig neural network reached the highest accuracy with more hidden neurons (9 neurons) compared to tansig neural network.

B. *Setup of the Support Vector Machine*

Fig.7 shows the results of the relation between average testing accuracy of different features versus the box constraint parameter C. It is clear that SVM performance is quite similar to the performance achieved by the ANN. The texture features achieved the smallest accuracy, which ranges from 47.06% to 57.14%. However, the combination of color and shape-size features reaches the highest accuracy of 97.1386% when the box constraint is 10. When the box constraint increases more than 10, the accuracy of different features either decreases or the improvement is very small. When all features are used, the accuracy was reduced little bit as compared to color and shape-size combination which shows that texture feature can be ignored.

TABLE I. THE BEST ACHIEVED ACCURACIES (%) USING THE DEVELOPED CLASSIFIERS

Classifier	ANN-logsig	ANN-tansig	SVM	KNN
Best Accuracy	97	97.26	97.14	95.83

C. *Setup of the K- Nearest Neighbor classifier*

KNN classifier performs the worst among the Neural Network and Support Vector Machine, which is obvious from Fig.8. For texture features, the highest accuracy reached was 53.33% with seven nearest neighbors. When color features and shape-size features were used alone, the accuracy improved proportionaly with the number of nearest neighbors and the highest reached at k=10 (70% and 82.5%). On the other hand, the combination of color and shape features attain the highest accuracy (95.83%) with only 5 nearest neighbors.

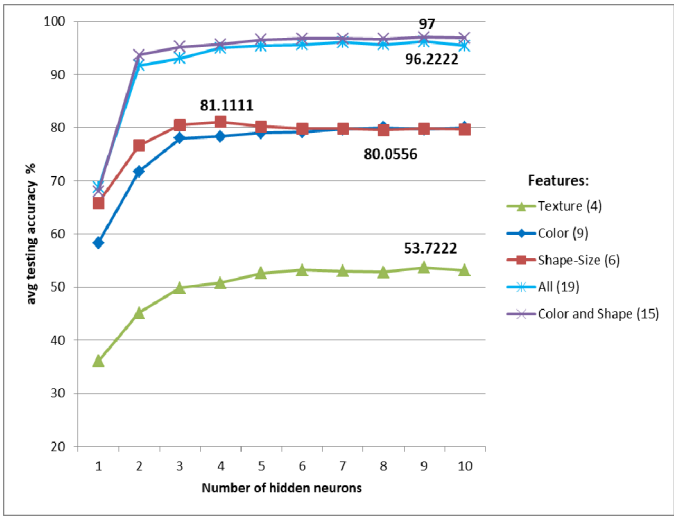

Fig. 5. Plots of the accuracies of the ANN classifier using different features vs number of hidden neurons (logsig-softmax)

D. *Time Complexity analysis*

TABLE II. shows the complexity in terms of testing time (seconds) for the different classifiers. Only the classifiers with the highest performance are considered. It is clear that ANN and SVM classifiers perform the classification process at an approximately similar time. ANN with tan-sigmoid hidden neurons was able to classify the testing samples in about 0.92 seconds/sample, which is considered as the lowest classification time in this paper. Logsig neural network and tansig neural network reach the highest accuracy in very close time. From table 2 we can judge that the times taken by both classifiers (ANN and SVM) are comparable. But simulation results show that SVM takes much more time to achieve the classification. However, the KNN classifier achieves the highest classification time of 2.56 seconds /sample (i.e. the slowest algorithm) since it needs to calculate the distance from each testing sample to all the training samples when a classification is required.

TABLE II. TIME COMPLEXITY OF DIFFERENT CLASSIFIERS FOR THE HIGHEST ACCURACY ACHIEVED

Classifier	ANN-logsig	ANN-tansig	SVM	KNN
Time (sec/sample)	0.939	0.917	0.997	2.56

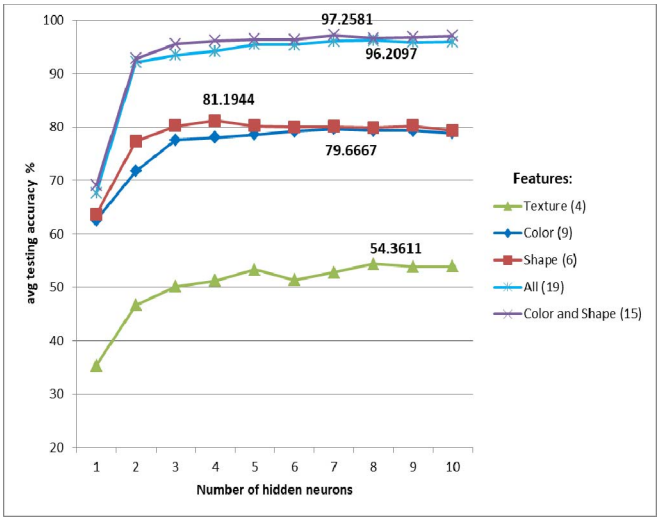

Fig. 6. Plots of the accuracies of the ANN classifier using different features vs number of hidden neurons (tansig-softmax)

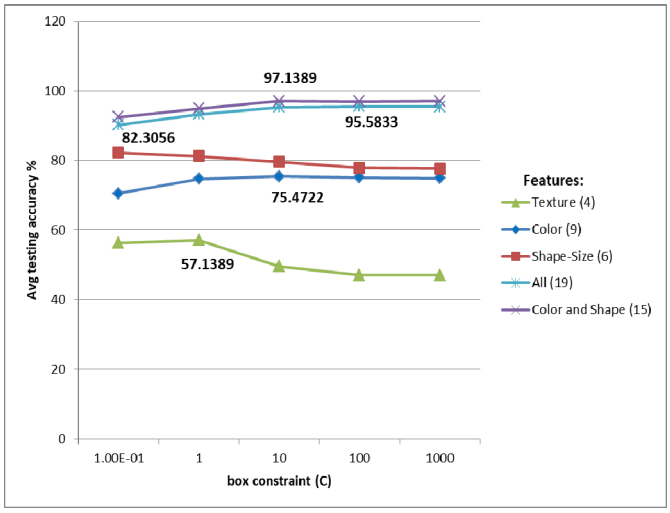

Fig. 7. Plots of the relation between average testing accuracy of different features versus the box constraint parameter C

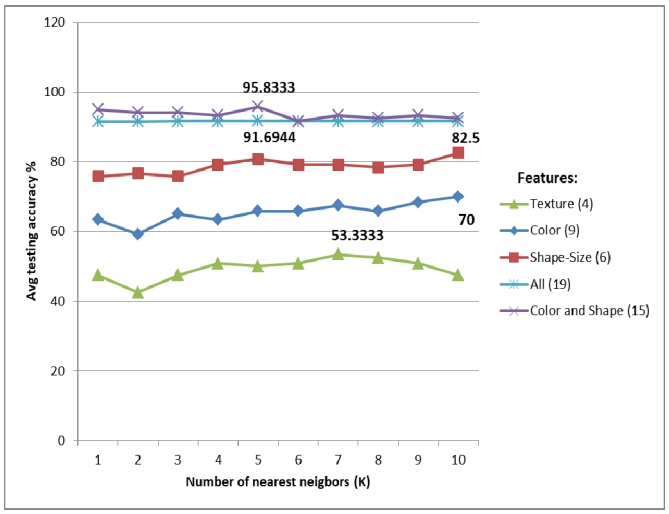

Fig. 8. Plots of the relation between the average testing accuracy of the KNN classifier of different features and the number of nearest neighbors

V. Conclusion

In this project, six types of dates are classified into their varieties by combining image processing and machine-learning techniques. Three techniques have been used and compared to each other in achieving the classification tasks, namely, Artificial Neural Network, Support Vector Machine, and K-nearest neighbor. Different features are extracted from dates images (colour, shape, size, and texture). Intensive experiments and qualitative comparison are conducted among the developed approaches. The impact of these features and the critical parameter of each classifier in the classification task are also addressed. The achieved results show that the combination of colour and shape–size features gives the highest accuracy compared to texture features. The highest classification accuracy attained by ANN, SVM, and KNN classifiers are 97.2581%, 97.1386%, and 95.83%, respectively. In this study, 15 colour and shape-size features are used. This implies that date fruits have significant differences in colors and shape-size rather than textures.

Acknowledgment

The authors want to thank Sultan Qaboos University for the help in providing the necessary tools, equipment and all licensed software's.

References

[1] FAO. Food and Agriculture Organization of the United Nations. Available online: http://www.fao.org/faostat/en/#data/QC (accessed on 19 September 2019).

[2] Popular varieties of dates grown in oman. Available online: https://omaninfo.om/english/module.php?module=topics-showtopic&CatID=35&ID=3793 (accessed on 19 September 2019).

[3] A. Haidar, H. Dong, and N. Mavridis, "Image-based date fruit classification", *4th Int. Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, 2012, pp. 357-363.

[4] K.M. Alrajeh and T.A.A. Alzohairy, "Date fruits classification using MLP and RBF neural networks", *International Journal of Computer Applications*, vol. 41, no.10, 36-41, March 2012

[5] G. Muhammad, "Automatic Date Fruit Classification by Using Local Texture Descriptors and Shape-Size Features", *Modelling Symposium (EMS)*, 2014 European, 2014, pp. 174-179.

[6] A. Manickavasagan, N.H. Al-Shekaili, N.K. Al-Mezeini, M.S. Rahman, and A. Guizani, "Computer vision technique to classify dates based on hardness", *Journal of Agricultural and Marine Sciences*, Vol. 22 (1): 36-41, 2017.

[7] Thomas, G., Manickavasagan, A, Al-Yahyai, R. and L. Khriji, "Contrast Enhancement using Brightness Preserving Histogram Equalization Technique for Classification of Date Varieties", *The Journal of Engineering Research*, 11 (1), 55-63, 2014.

[8] T. Najeeb and M. Safar, "Dates Maturity Status and Classification Using Image Processing", *2018 International Conference on Computing Sciences and Engineering (ICCSE)*, 2018, pp. 1-6.

[9] N. Otsu, "A threshold selection method from gray-level histograms", *IEEE transactions on systems, man, and cybernetics*, vol. 9, pp. 62-66, 1979.

[10] Gonzalez, R.C., R.E. Woods, and S.L. Eddins. 2011. Digital image processing using MATLAB. New Delhi: Tata McGraw Hill Education Private Limited.

[11] K. Hameed, D. Chai, and A. Rassau, "A comprehensive review of fruit and vegetable classification techniques", *Image and Vision Computing*, vol. 80, pp. 24-44, 2018.

[12] Ghulam Muhammad, "Date fruits classification using texture descriptors and shape-size features", *Elsevier, Engineering Applications of Artificial Intelligence*, vol. 37, pp. 361-367, 2015.

[13] P. Mohanaiah, P. Sathyanarayana, and L. GuruKumar, "Image texture feature extraction using GLCM approach", *International journal of scientific research publications*, vol. 3, p. 1, 2013.

[14] G.M. Foody, and A. Mathur, "A relative evaluation of multiclass image classification by support vector machines", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42 (6), pp. 1335-1343, 2004.

[15] Suresha, N. Shilpa, and B. Soumya, "Apples grading based on SVM classifier", *IJCA Proceedings on National Conferecne on Advanced Computing and Communications* 2012 NCACC(1):27-30, August 2012.

[16] A. Kumar and G. Gill, "Automatic fruit grading and classification system using computer vision: a review", *2nd International Conference on Advances in Computing and Communication Engineering*, 2015, pp. 598-603.

[17] P. Cunningham and S. J. Delany, "k-Nearest neighbor classifiers", *Multiple Classifier Systems*, vol. 34, pp. 1-17, 2007.