

IDENTIFYING OLIVE OIL FRAUD AND ADULTERATION USING MACHINE LEARNING ALGORITHMS

Yasin Yakar^{a,*} and Kerim Karadağ^b^aFood Processing Department, Organized Industrial Zone Vocational School, Harran University, 63200 Sanliurfa, Türkiye^bDepartment of Electrical and Electronics Engineering, Engineering Faculty, Harran University, 63050 Sanliurfa, Türkiye

Recebido em 24/04/2022; aceito em 28/07/2022; publicado na web em 13/09/2022

As olive oil (OO) is more expensive than other vegetable oils, it is usually adulterated by blending it with more economic edible oils such as cottonseed oil (CSO), canola oil (CO), and soybean oil (SO). This research aimed to determine the fatty acid compositions obtained as a result of blending different proportions of CSO, CO and SO with OO using a gas chromatograph and to reveal OO adulteration by evaluating the obtained data with different machine learning algorithms. The assessment of the OO consisted of two stages. The first step was extraction of the feature vector, while second step was the classification of feature vectors with regard to the data and computing the regression values. Features were extracted using the Relief method, classified with the Support Vector Machine (SVM) and the k-Nearest Neighbor (k-NN) and Decision Tree (DT) algorithms, and the neural network algorithm was used for regression. The highest accuracy values for classification were calculated as 0.946, 0.964 and 0.982 for OO-CO, OO-SO, and OO-CSO mixtures, using the SVM method, respectively. The errors in the regression analysis were computed as 0.005, 0.005 and 0.002 respectively using the neural network algorithm.

Keywords: olive oil; adulteration; gas chromatography; machine learning.

INTRODUCTION

OO is an edible oil extracted as a result of pressing olives without pitting the seeds, and it may be consumed in its natural form without applying any chemical treatment.^{1,2} Owing to its unique flavor, taste, aroma and bio-active components, it is a highly valuable product. OO, which is known for numerous health benefits, has a significant place in our dietary regimens.³ Today, due to the exaggerated consumption of fast food and ready-to-eat products, saturated fats are taken into the body in an unbalanced way. OO is also used in such products to eliminate this negativity in recent studies.⁴

OO is also widely used as a traditional treatment for many diseases, particularly in countries with a Mediterranean coast. Furthermore, experimental studies have demonstrated that it has significant effects on diseases such as hypertension, cardiovascular disorders, diabetes, cancer, Alzheimer's, and Parkinson's disease.^{5,6} The incidence of food imitation and adulteration has recently increased due to various reasons, including technological developments, unfair competition, and economic problems.⁷ Most of the fraud that occurs with regard to foodstuffs involves the intention to maximize profits by blending cheaper products with more expensive ones. OO has recently been the leading foodstuff in terms of the occurrence of food fraud. As OO is more expensive than other vegetable oils, it is usually adulterated by blending with more economic edible oils such as CSO, CO and SO.

80 out of 386 food products listed in the 2020 Food Imitation and Adulteration list by the Ministry of Agriculture and Forestry in Turkey were OO. This figure reveals the extent of the adulteration of this product. A comprehensive study reviewed 1305 articles published in various databases between 1980 and 2010. More than 50% of the scientific records classified under 18 different categories in the database were related to oils, milk, juices, concentrates, jams, purees, canned goods and spices. It was determined that OO, milk, honey, saffron, orange juice, coffee and apple juice occupied the first seven places among food products exposed to adulteration. It was

further determined that OO, with a percentage of 16%, ranked the first among these food products.⁸

Machine learning (ML) is defined as a branch of artificial intelligence that applies different algorithms to find the basic relationships between data and information or as computer algorithms that model any problem in accordance with data obtained with regard to that problem.⁹⁻¹² Machine learning algorithms do not strictly follow a comprehensive list of clear instructions or rules. Instead, they try to develop a model to derive data-based predictions and decisions from sample inputs.¹³ Regression, classification, clustering, and deep learning are among the machine learning algorithms.^{14,15}

The most widely used techniques in previous research conducted to identify OO imitation and adulteration were Gas Chromatography with Flame-Ionization Detection (GC-FID),¹⁶ High-performance liquid chromatography (HPLC),¹⁷ Laser Induced Breakdown Spectroscopy (LIBS),¹⁸ Raman Spectroscopy Analysis,¹⁸⁻²⁰ FT-IR Spectroscopy,²¹ Near Infrared Spectroscopy (NIRS),²² and Nuclear Magnetic Resonance Spectroscopy (NMR).²³ The number of studies combining these techniques with various ML algorithms has been increasing in recent years. This approach makes it possible to obtain more reliable results, both qualitatively and quantitatively. ML has a wide range of applications in food science.²⁴⁻²⁷

This study aimed to use different learning algorithms to identify adulterated OO that had been blended OO with different vegetable oils.

EXPERIMENTAL

Samples

OO, CSO, CO and SO used in the analysis were procured from local manufacturers as 4 different samples of 5 g each.

OO was adulterated using three different models: by blending OO and CSO (OO-CSO), blending OO and canola oil (OO-CO), and blending OO and soybean oil (OO-SO). A total of 19 subgroups were obtained by adding 1%, 5%, 10%, 20% and 40% of the other vegetable oils to the OO in each model. 20 samples from each subgroup were prepared and analyzed (Table 1).

*e-mail: yakar75@gmail.com

Table 1. Blends of olive oil and different vegetable edible oils

	OO-CSO Model (g/g)	OO-CO Model (g/g)	OO-SO Model (g/g)
1	OO	-	-
2	OO+1% CSO	OO+1% CO	OO+1% SO
3	OO+5% CSO	OO+5% CO	OO+5% SO
4	OO+10% CSO	OO+10% CO	OO+10% SO
5	OO+20% CSO	OO+20% CO	OO+20% SO
6	OO+40% CSO	OO+40% CO	OO+40% SO
7	CSO	CO	SO

Chemicals

Isooctane (99.5%), sodium hydrogen sulfate (99%), sodium chloride (99.5%), and potassium hydroxide (2 N) in methanol were purchased from Merck (Darmstadt, Germany). 37-component fatty acid standard mixture was purchased from Sigma-Aldrich (St Louis, MO, USA).

Preparation of samples

Approximately 100 mg of oil sample and 2 mL of isooctane were mixed in a capped test tube. 200 μ L of potassium hydroxide solution was added to the mixture and vortexed for 1 minute. The tube was vortexed after 2 mL of 40% NaCl solution had been added. The supernatant was transferred into a vial and was shaken about 1 g of sodium hydrogen sulfate had been added. Approximately 30 minutes later, 1 μ L of the supernatant was taken and injected into the gas chromatograph.²⁸

Chromatographic conditions

The fatty acid composition of the samples was computed using a Thermo brand, TraceGC Ultra model gas chromatograph with FID (Flame Ion Detector) detector. A 60 m HP-88 column was used for the separation of fatty acids. Detector and injector block temperatures were set at 280 and 260 °C, respectively. A temperature program was applied to the column. After waiting for 2 minutes at 50 °C the temperature was increased to 180 °C with an increase of 20 °C/min. Then, the temperature was further increased to 230 °C with an increase of 5 °C/min and the set-up was kept at this temperature for 5.5 more minutes. The split ratio was set to 1/50 and the injection volume was 1 μ L.

Machine learning algorithms

Experimental data were further analyzed using MATLAB R2018a software run on a computer with the Windows 10 Pro operating system with a 3.00 GHz CPU (Intel Core i5-7400), 8 GB RAM and a 1 TB hard disk. Data were normalized in accordance with the following formula:

$$N = 0.8 \times \frac{(X_i - X_{min})}{(X_{max} - X_{min})} \times 0.1 \quad (1)$$

The Support Vector Machine (SVM), k-Nearest Neighbor (k-NN) and Decision Tree (DT) ML algorithms were used for classification process. SVM is an algorithm that tries to separate any given data sets in such a way that data belonging to different classes are assigned farthest from each other, whereas data belonging to the same class are

assigned closest to each other. k-NN is a machine learning algorithm where the “K” is a parameter that determines the number of nearest neighbors to include in the classification process. DT creates a graph or tree using the branching technique to demonstrate every possible outcome of a decision. Each leaf node represents a class label, whereas each branch corresponds to the result of the master node representing conjunctions of features. To classify a specimen, a top-down approach is applied, starting at the root of the tree. For a given feature or node, the branch that corresponds to the value of that feature’s data point is considered until a leaf is reached or a label is decided on.²⁹ Finally, the Artificial Neural Network (ANN) algorithm was used for regression. ANN is a machine learning estimation algorithm inspired by the nervous system of animals. ANN system is the connection between the layers, the number of neurons as the processing unit of the neural networks in each layer, and the transfer functions between the layers.³⁰

Accuracy, G-Mean and F-Score were used as performance measures. Equations 2-4 were used for these performance measures.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (2)$$

$$F\text{-Score} = 2 \times \frac{\left(\frac{TP}{FP + TP}\right) \left(\frac{TP}{TP + FN}\right)}{\left(\frac{TP}{FP + TP}\right) + \left(\frac{TP}{TP + FN}\right)} \quad (3)$$

$$G\text{-Mean} = \sqrt{\left(\frac{TP}{FP + FN}\right) + \left(\frac{TN}{FP + TN}\right)} \quad (4)$$

RESULTS AND DISCUSSION

The OO industry has a significant social and economic impact in Mediterranean countries, which account for approximately 98% of the world’s OO production.³¹ According to the International Olive Council (IOC) data, OO production in 2019/2020 was approximately 3.26 million tons. The adulteration of OO with low-quality edible oils is major trouble in the industry-leading to economic fraud as well as consumer health issues. Chromatographic techniques supported by ML methods have an important place in the determination of imitation and adulteration in OO.

Grouping and preprocessing data

Fatty acid compositions determined by GC-FID device were included in the research. First recordings were taken from OO, CO, CSO and SO data sets (Figure 1).

Then, groups were formed by adding 1%, 5%, 10%, 20% and 40% CSO, CO and SO to OO in each data set. Mixed oils were then analyzed in the GC instrument. Each of the OO-CSO, OO-CO and OO-SO data sets consisted of 140x13 values. The dataset consisting of 420x13 values in total was used for classification and regression. In the dataset, the value 420 expressed as a row indicates the number of samples, and the value 13, expressed as a column, indicates the properties of the oil. Figures 2a, 2b and 2c graphically show the fatty acid compositions of OO-CSO, OO-CO and OO-SO mixed oils.

Recommended classification system

The classification system recommended in this research is given in Figure 3. First, the data set was obtained and converted into a suitable format to be further analyzed in the computer environment; then, the normalization process was applied. Classification and regression were performed using the features of oils. The data sets obtained for

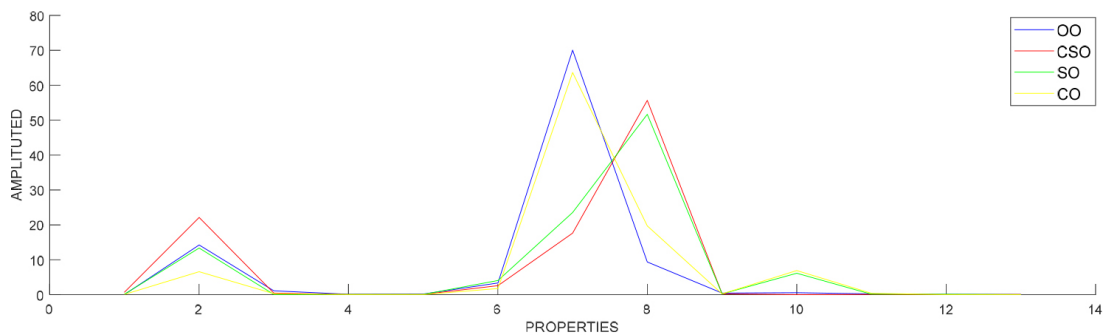


Figure 1. Graphic representation of fatty acid compositions OO, CSO, CO and SO

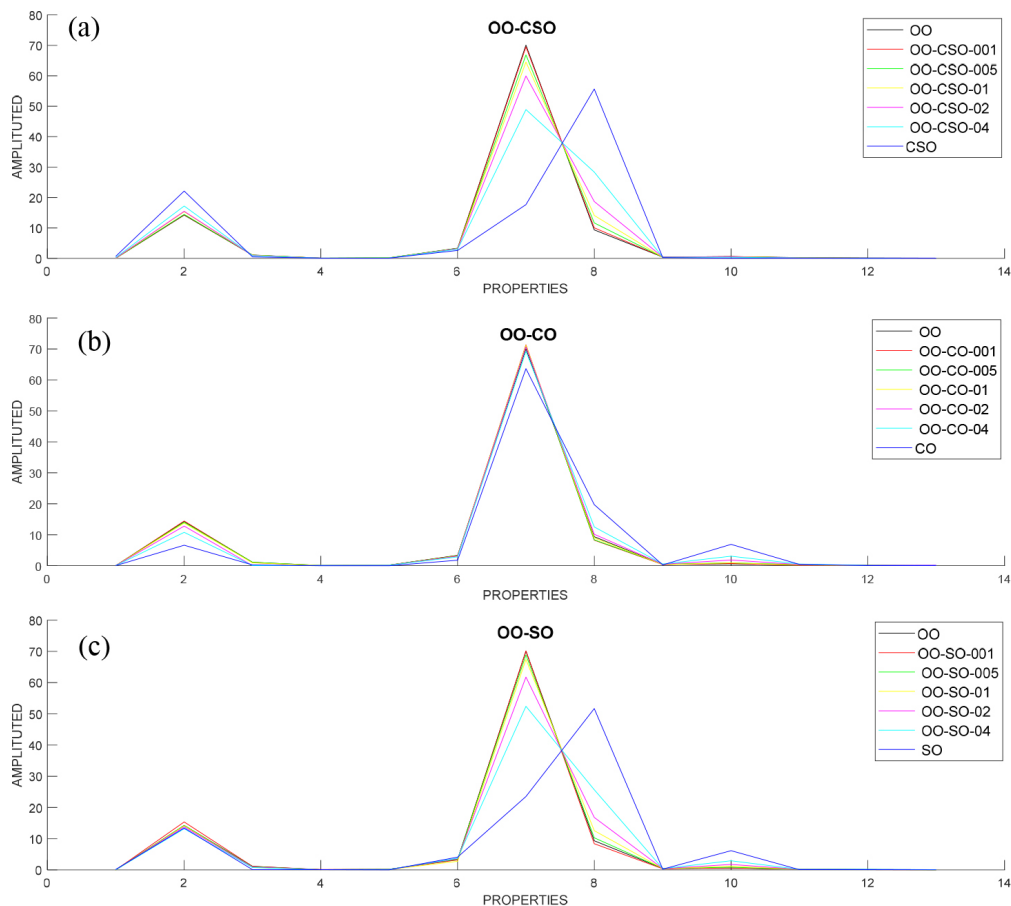


Figure 2. Graphic representation of fatty acid compositions OO-CSO, OO-CO and OO-SO

classification and regression were divided into two groups: training and test data. 60% of the data were used as training data and 40% as test data. Then, cross-validation, which is another decomposition method, was performed. Cross-validated data sets were used in both the training and testing processes. Thus, the software was prevented from memorizing the training and test values. The classification algorithms used were SVM, k-NN and DT. The purity and adulteration of OO were determined using the NN algorithm for regression.

The five most distinctive features of OO adulterated with other vegetable oils are provided in Table 2 for classification and regression purposes. These features were selected using the relief method.

The features were evaluated in order to identify the differences with regard to the mixtures added to the adulterated OO, and it was observed that the most distinctive feature for OO-CSO and OO-SO adulterated oil was linoleic acid, while it was linolenic acid for OO-CO adulterated oil.

Following the regression applied to the data sets consisting of

mixtures OO-CSO, OO-CO and OO-SO, the data set was separated in order to be used in the 60% training and 40% testing stages; the results obtained were converted into graphics and are shown in Figure 4.

Blue lines on the graphics represent actual values, whereas red lines represent estimated regression values. Comparison of the results revealed that the graphics overlapped, and the error rates were found to be 0.005 for OO-CSO, 0.002 for OO-CO and 0.005 for OO-SO.

Then, a final comparison was performed of all groups (Figure 5).

Comparison of the results revealed that the graphics overlapped, and the error rate was 0.0752.

The classification algorithms used were SVM, k-NN and DT. First, obtained data were subjected to the normalization process and then the data were classified so that 60% was allocated to the training phase and 40% to the testing phase, as is frequently encountered in the literature. Accuracy, G-Mean and F-Score were used as classification performance measures. The computed performance measures are shown in Table 3.

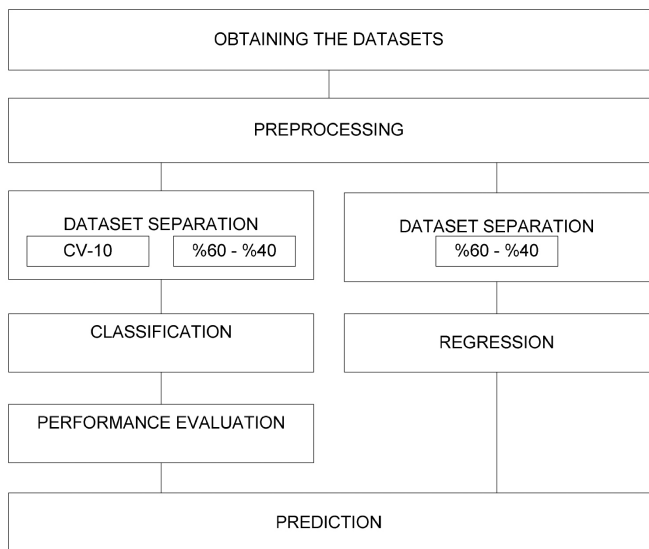


Figure 3. Flow chart of the research

Table 2. Five most effective parameters that distinguish OO from other vegetable oils

Mixtures	Five Most Influential Parameters
OO-CSO	Linoleic Acid, Oleic Acid, Miristic Acid, Palmitic Acid and Arachidic Acid
OO-SO	Linoleic Acid, Oleic Acid, Linolenic Acid, Heptadecenoic Acid and Palmitoleic Acid
OO-CO	Linolenic Acid, Stearic Acid, Palmitic Acid, Eicosenoic Acid and Palmitoleic Acid

Comparison of the performance measures revealed that highest performance was obtained with the SVM algorithm at 0.982 with the accuracy measure in the OO-CO mixed oil. As there was a high number of classification groups, other performance measures were also evaluated; accordingly, the performance rates obtained with the SVM algorithm were 0.941 for F-score and 0.943 for G-mean. The highest performance for the OO-CSO mixed oil was obtained with the SVM algorithm at 0.946 with the accuracy measure, 0.769 with the F-score measure and 0.791 with the G-mean measure. The highest performance for OO-SO mixed oil was obtained with the SVM algorithm at 0.964 with the accuracy measure, 0.875 with the F-score measure and 0.882 with the G-mean measure.

10-fold cross validation (CV-10) was performed in order to transfer all the values used in the research to the training and testing phases. The same performance measures used in the previous application were applied as classification performance measures. The performance measures computed using CV-10 are shown in Table 4.

Comparison of the classification performance measures computed using CV-10 revealed that highest classification performance was obtained with the SVM algorithm in the OO-CSO mixed oil samples. The performance scores for accuracy, F-score and G-mean measures were recorded as 0.993, 0.974 and 0.975 respectively. The highest performance for the OO-CO mixed oil sample was obtained with the SVM algorithm at 0.971 with the accuracy measure, 0.909 with the F-score measure and 0.929 with the G-mean measure. The highest performance for adulterated oil obtained by blending OO and SO was obtained with the SVM algorithm at 0.943 with the accuracy measure, 0.778 with the F-score measure and 0.798 with the G-mean measure.

Comparison of the classification performance measures indicated high performance rates in both cases where the data set was decomposed. Comparison of the classification approaches

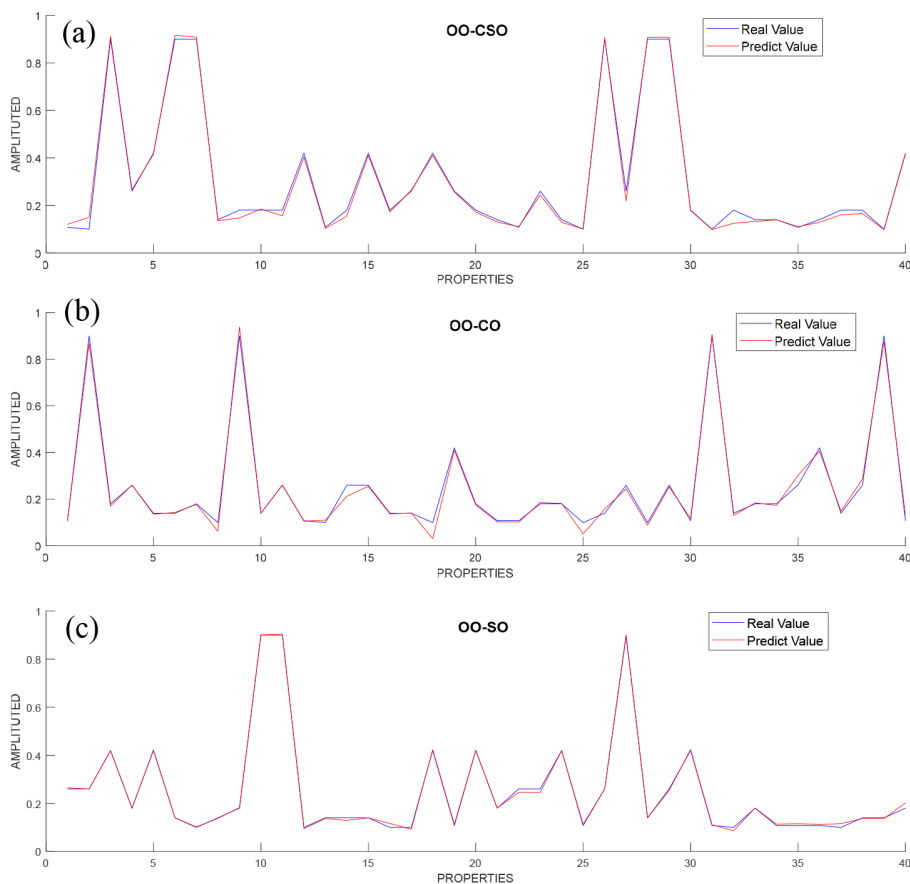


Figure 4. Graphic representation of fatty acid compositions after regression determined for OO-CSO, OO-CO and OO-SO mixed oils

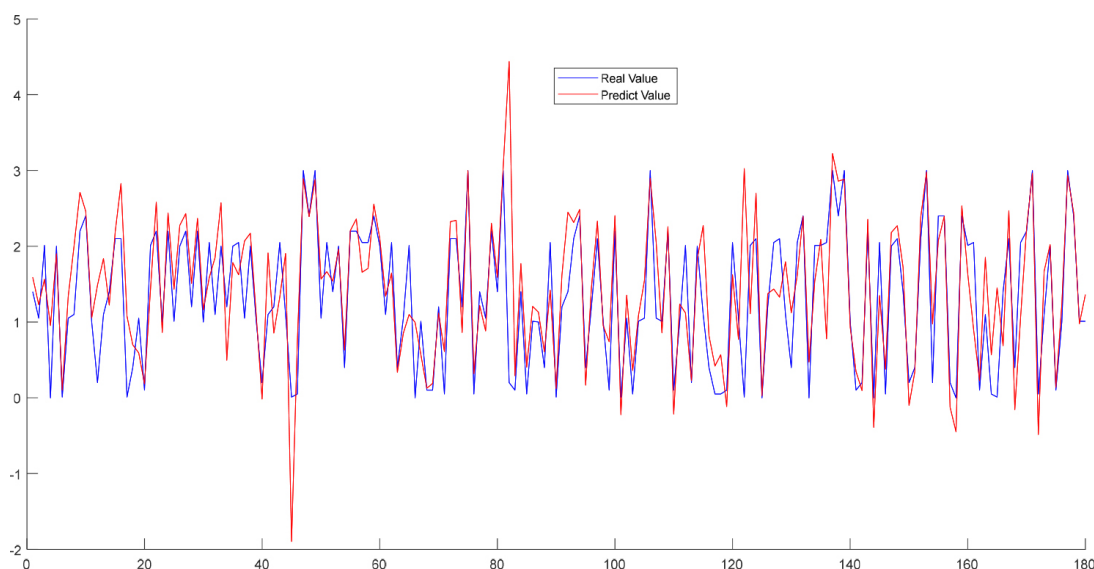


Figure 5. Regression results of the OO and other adulterated mixtures

Table 3. Performance measures with regard to mixed oils

Mixtures	Classification methods	Performance Evaluation		
		F_Score	G_Mean	Accuracy
OO-CSO	SVM	0,769	0,791	0,946
	k-NN	0,526	0,613	0,839
	DT	0,5	0,606	0,804
OO-CO	SVM	0,941	0,943	0,982
	k-NN	0,696	0,788	0,875
	DT	0,5	0,577	0,821
OO-SO	SVM	0,875	0,882	0,964
	k-NN	0,545	0,639	0,821
	DT	0,5	0,606	0,821

Table 4. CV-10 results for blended oil samples

Mixtures	classification methods	Performance Evaluation		
		F-Score	G-Mean	Accuracy
OO-CSO	SVM	0,974	0,975	0,993
	k-NN	0,536	0,624	0,814
	DT	0,654	0,736	0,879
OO-CO	SVM	0,909	0,929	0,971
	k-NN	0,517	0,607	0,80
	DT	0,556	0,642	0,829
OO-SO	SVM	0,778	0,798	0,943
	k-NN	0,481	0,581	0,80
	DT	0,470	0,586	0,814

indicated that the highest performance was obtained using the SVM algorithm and the highest scores were computed using accuracy as the performance criterion.

In another study, low-field nuclear magnetic resonance (LF-NMR) technique was used to detect adulteration from blending hazelnut oil and high-oleic sunflower oil with natural extra virgin OO. The data were evaluated with six different machine learning algorithms. The convolutional neural network algorithm (CNN) was observed to provide the highest accuracy value with a rate of 89.29%.²³

Bellou *et al.* carried out research using LIBS to determine the oil

adulteration obtained by blending Greek virgin OO and pomace oil. They found that the Linear Discriminant Analysis (LDA) algorithm classified OOs in terms of purity and degree of adulteration with an accuracy close to 100%.¹⁸

In an earlier study, a batch of extra virgin OO samples adulterated with soybean oil, corn oil and sunflower seed oil were analyzed by Raman spectroscopy. The mean absolute relative errors were calculated as 7.41%, 7.78% and 9.45% respectively using the external standard method (ESM) and as 5.10, 6.96 and 4.55 respectively using the SVM model. ESM model based on Raman spectroscopy has been shown to be a promising technique for the authentication of extra virgin OO.²⁰

CONCLUSIONS

In this study, classification and regression analysis were carried out using the values computed from the adulterated OO samples by a gas chromatography device with an FID detector. All these processes were carried out with ML algorithms. First, five features that were effective in distinguishing OO were determined, then the SVM, k-NN and DT algorithms were used for classification and the NN algorithm was used for regression. ML algorithms, in particular SVM, were seen to give highly accurate results in the determination of OO imitation and adulteration caused by blending different types of vegetable oils with OO.

REFERENCES

- Armutcu, F.; Namuslu, M.; Kaya, M.; *Konuralp Tıp Dergisi* **2013**, *5*, 60.
- Vitaglione, P.; Savarese, M.; Paduano, A.; Scalfi, L.; Fogliano, V.; Sacchi, R.; *Crit. Rev. Food Sci. Nutr.* **2015**, *55*, 1808. [Crossref]
- Özata, E.; Cömert, M.; *Zeytin Bilimi* **2016**, *6*, 105.
- Los, P. R.; Marson, G. V.; Dutcosky, S. D.; Nogueira, A.; Marinho, M. T.; Simões, D. R. S.; *Food Sci. Technol.* **2020**, *40*, 12. [Crossref]
- Bayram, B.; Özçelik, B.; *Akademik Gıda Dergisi* **2017**, *10*, 77.
- Caramia, G.; Gori, A.; Valli, E.; Cerretani, L.; *Eur. J. Lipid Sci. Technol.* **2012**, *114*, 375. [Crossref]
- Türkmen, S.; Ataseven, Y.; *Tarım Ekonomisi Araştırmaları Dergisi* **2020**, *6*, 65.
- Moore, J. C.; Spink, J.; Lipp, M.; *J. Food Sci.* **2012**, *77*, 4. [Crossref]
- Awad M.; Khanna R.; Apress **2015**, Berkeley, CA. https://doi.org/10.1007/978-1-4302-5990-9_1

10. Atalay, M.; Çelik, E.; *Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi* **2017**, *9*, 155.
11. Boutaba, R.; Salahuddin, M. A.; Limam, N.; Ayoubi, S.; Shahriar, N.; Estrada-Solano, F.; Caicedo, O. M.; *Journal of Internet Services and Applications* **2018**, *9*, 16. [Crossref]
12. Ünal, F.; Birant, D.; Seker, Ö.; *Turkish Journal of Electrical Engineering and Computer Sciences* **2021**, *29*, 1797. [Crossref]
13. Deng, X.; Cao, S.; Horn, A. L.; *Annu. Rev. Food Sci. Technol.* **2021**, *12*, 513. [Crossref]
14. Jiao, Z.; Hu, P.; Xu, H.; Wang, Q.; *J. Chem. Health Saf.* **2020**, *27*, 316. [Crossref]
15. Ahmad, A.; Garhwal, S.; Ray, S. K.; Kumar, G.; Malebary, S. J.; Barukab, O. M.; *Archives of Computational Methods in Engineering* **2020**, *28*, 2645. [Crossref]
16. Siano, F.; Vasca, E.; *J. Chem. Educ.* **2020**, *97*, 4108. [Crossref]
17. Zabarar, D.; Gordon, M.H.; *Food Chem.* **2004**, *84*, 475. [Crossref]
18. Bellou, E.; Gyftokostas, N.; Stefas, D.; Gazeli, O.; Couris, S.; *Spectrochim. Acta, Part B* **2020**, *163*, 105746. [Crossref]
19. Zou, M. Q.; Zhang, X. F.; Qi, X. H.; Ma, H. L.; Dong, Y.; Liu, C.; W.; Guo, X.; Wang, H.; *J. Agric. Food Chem.* **2009**, *57*, 6001. [Crossref]
20. Zhang, X. F.; Zou, M. Q.; Qi, X. H.; Liu, F.; Zhang, C.; Yin, F.; *J. Raman Spectrosc.* **2011**, *42*, 1784. [Crossref]
21. Poiana, M.; Mousdis, G.; Alexa, E.; Moigradean, D.; Negrea, M.; Mateescu, C.; *J. Agroalim. Processes Technol.* **2012**, *18*, 277.
22. Vanstone, N.; Moore, A.; Martos, P.; Neethirajan, S.; *Food Qual. Saf.* **2018**, *2*, 189. [Crossref]
23. Hou, X.; Wang, G.; Wang, X.; Ge, X.; Fan, Y.; Jiang, R.; Nie, S.; *Journal of the Science of Food and Agriculture* **2021**, *101*, 2389. [Crossref]
24. Chen, T.-C.; Yu, S.-Y.; *Food Science and Technology (Campinas)*, **2022**, *42*, e29121.
25. Rocha, R. S.; Calvacanti, R. N.; Silva, R.; Guimaraes, J. T.; Balthazar, C. F.; Pimentel, T. C.; Esmerino, E. A.; Freitas, M. Q.; Granato, D.; Costa, R. G. B.; Silva, M. C.; Cruz, A. G.; *Lebensm.-Wiss. Technol.* **2020**, *126*, 109342.
26. Hou, Y.; Zhao, P.; Zhang, F.; Yang, S.; Rady, A.; Wijewardane, N. K.; Li, M.; *Food Science and Technology (Campinas)* **2022**, *42*, e100821. [Crossref]
27. Farah, J. S.; Cavalcanti, R. N.; Guimaraes, J. T.; Balthazar, C. F.; Coimbra, P. T.; Pimentel, T. C.; Esmerino, E. A.; Duarte, M. C. K. H.; Freitas, M. Q.; Granato, D.; Cucinelli Neto, R. P.; Tavares, M. I. B.; Calado, V.; Silva, M. C.; Cruz, A. G.; *Food Control* **2021**, *121*, 107585. [Crossref]
28. TS EN ISO 12966-2; 2017, *Türk Standartları Enstitüsü*, Ankara.
29. Das, K.; Behera, R. N.; *International Journal of Innovative Research in Computer and Communication Engineering* **2017**, *5*, 1301.
30. Benzer, S.; Benzer, R.; *Savunma Bilimleri Dergisi* **2017**, *16*, 1.
31. Gullon, P.; Gullon, B.; Astray, G.; Carpena, M.; Fraga-Corral, M.; Prieto, M. A.; Simal-Gandara, J.; *Food Res. Int.* **2020**, *137*, 109683.