Food Chemistry 373 (2022) 131471

Contents lists available at ScienceDirect

Food Chemistry

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

Analytical Methods

# The application of machine-learning and Raman spectroscopy for the rapid detection of edible oils type and adulteration

Hefei Zhao [a], Yinglun Zhan [b], Zheng Xu [c], Joshua John Nduwamungu [a], Yuzhen Zhou [b], Robert Powers [d,e], Changmou Xu [a,*]

[a] Food Processing Center, Department of Food Science and Technology, University of Nebraska-Lincoln, Lincoln NE 68588, USA
[b] Department of Statistics, University of Nebraska-Lincoln, Lincoln, NE 68583, USA
[c] Department of Mathematics and Statistics, Wright State University, Dayton, OH 45435, USA
[d] Department of Chemistry, University of Nebraska-Lincoln, Lincoln, NE 68588, USA
[e] Nebraska Center for Integrated Biomolecular Communication, University of Nebraska-Lincoln, Lincoln, NE 68588, USA

ARTICLE INFO

*Keywords:*
Raman spectroscopy
Machine learning
Edible oil quality
Food adulteration

ABSTRACT

Raman spectroscopy is an emerging technique for the rapid detection of oil qualities. But the spectral analysis is time-consuming and low-throughput, which has limited the broad adoption. To address this issue, nine supervised machine learning (ML) algorithms were integrated into a Raman spectroscopy protocol for achieving the rapid analysis. Raman spectra were obtained for ten commercial edible oils from a variety of brands and the resulting spectral dataset was analyzed with supervised ML algorithms and compared against a principal component analysis (PCA) model. A ML-derived model obtained an accuracy of 96.7% in detecting oil type and an adulteration prediction of 0.984 ($R^2$). Several ML algorithms also were superior than PCA in classifying edible oils based on fatty acid compositions by gas chromatography, with a faster readout and 100% accuracy. This study provided an exemplar for combining conventional Raman spectroscopy or gas chromatography with ML for the rapid food analysis.

## 1. Introduction

Edible oils are an indispensable source of nutrition and, accordingly, are widely present in food. Oil adulteration has been a chronic issue for many years (Zhang et al., 2012) because of the large differences between oil prices. Simply, higher priced quality oils are mixed with lower quality oils to enhance profits and deceive the consumer. Currently, the authentication of edible oils mainly depends on the analysis of fatty acid composition via gas chromatography (GC) that requires pretreatment to achieve methyl esterification. Methyl esterification is a time-consuming chemical reaction that produces toxic solvent waste and is impractical for high throughput measurements. A relatively fast, one-hour detection approach has been recently developed to authenticate extra virgin olive oils based on the direct analysis of triacylglycerols (TAGs) using ultra-high-performance liquid chromatography (UHPLC) coupled with charged aerosol detection (CAD) (Green et al., 2020). This UHPLC-CAD method required minimal sample preparation, which took an important step forward to achieving a rapid high-throughput screen for the olive oil industry. However, for undertaking large amount of sample determination workload, chromatography is still a low throughput measurement (~0.5 to 1 h per sample) that needs organic solvents as mobile phase. Therefore, alternative technologies that are organic solvent-free and high throughput are urgently needed to enable rapid oil quality determination, especially for on-site measurements.

Raman spectroscopy does not require any chemical reagents for sample pretreatment. Notably, Raman spectroscopy has been used to characterize the chemical composition of bulk lipids, to determine the free fatty acid content and the degree of unsaturation of oils, and to discriminate between and authenticate different edible oils and fats (Baeten et al., 2005; Jiménez-Sanchidrián & Ruiz, 2016; Yang, Irudayaraj, & Paradkar, 2005). However, the difference in Raman spectra between oils is subtle; therefore, it is necessary to apply statistical analysis to accurately and efficiently identify these unique spectral differences. Currently, the interpretation of Raman spectra requires manual or semi-manual data processing and technical expertise to compare an unknown spectrum with known spectra in database, and in many cases, elaborative comparison of specific Raman bands is needed. The Raman spectrum does not provide a simple direct readout that

* Corresponding author.
*E-mail address:* cxu13@unl.edu (C. Xu).

https://doi.org/10.1016/j.foodchem.2021.131471
Received 30 June 2021; Received in revised form 17 September 2021; Accepted 22 October 2021
Available online 26 October 2021

outputs chemical or compound names or concentrations. Instead, a statistical analysis is needed. One recent example was the application of an unsupervised principal component analysis (PCA)-assisted surface-enhanced Raman spectroscopy (SERS) for the discrimination of edible oils (Du et al., 2019). Although the PCA method could differentiate between the six types of edible oils, the approach still required manual intervention to match each data cluster with the oil types instead of providing a direct readout. Although SERS significantly increased the sensitivity of detection, the use of organic solvents and gold nanoparticles for sample pretreatment greatly increased the cost and analysis time. Accordingly, a rapid and reliable spectral data processing method may enhance the efficiency and wide-acceptance of Raman spectroscopy for the characterization of edible oils.

Machine learning algorithms have facilitated numerous breakthroughs in the processing of complicated data sets, such as medical images (Ardila et al., 2019). Recently, they have been coupled with Raman spectroscopy or surface-enhanced Raman spectroscopy for the rapid analysis of a diversity of samples that have included medicines and microbes (Lussier, Thibault, Charron, Wallace, & Masson, 2020). Unlike unsupervised PCA, a trained machine learning algorithm can rapidly classify a new analyte from a data set of raw Raman spectra and provide a direct readout. Nevertheless, machine learning algorithms has seen limited applications in solving food science related problems, coupled with advanced food analysis equipment. To the best of our knowledge, the rapid validation of a variety of edible oil quality by coupling Raman spectroscopy with machine learning has not been previously demonstrated.

Nine supervised machine learning algorithms were integrated into a Raman spectroscopy protocol for achieving the rapid classification of oil type and the quick detection of adulterated edible oils. Raman spectra were obtained for ten commercial edible oils from a variety of brands and the resulting spectral data set was analyzed with supervised machine learning algorithms. The results were compared against a PCA model. The fatty acid composition of the edible oils were also analyzed using the same data processing protocol. The correlation between fatty acid composition and the Raman spectra of various edible oils was also examined. Our study provides an exemplar for the application of machine learning for the rapid analysis of Raman spectra in the field of food science.

## 2. Materials and methods

### 2.1. Chemicals and supplies

Heptane, an alkane standard solution (C8 to C20), glyceryl triheptadecanoate, and Supelco 37 component fatty acid methyl esters (FAME) mixed in dichloromethane were purchased from Sigma-Aldrich (St. Louis, USA). Hexane was bought from Fisher Chemical (Fair Lawn, USA). The four inch gold (99.99%) coated silicon wafer was purchased from Sigma-Aldrich (St. Louis, USA). The gold coated silicon wafer was sliced into 10 mm × 6 mm pieces by a diamond cutter and attached to the center of regular glass microscopy slides (25 mm × 75 mm × 1 mm) by adhesive tapes for further use.

### 2.2. Edible oils

Forty-seven edible oils from forty-six different brands and comprising twelve different oil types from at least seventeen countries of origin were purchased from local markets in Lincoln, Nebraska, USA, between 2019 and 2020 (S. Table 1). The oil types include avocado, canola, coconut, liquid coconut, corn, grapeseed, olive, peanut, soybean, sunflower, algae, hemp, and safflower oils. To minimize deteriorations and changes in fatty acid composition before analysis, 10 mL of each oil was transferred into a 30 mL GC headspace bottle with a tight silicon cap and stored in a dark refrigerator at 4 °C. All samples were analyzed within a week of collection.

### 2.3. Preparation of oil samples for adulteration study

Two adulteration models, avocado oils adulterated by canola oils and olive oils adulterated by soybean oils, were prepared for this study. Specifically, two randomly selected avocado oils (# 3 and # 43) from S. Table 1 were blended with two randomly selected canola oils (#8 and #5) for the training data set for the adulterated avocado oils. Two other avocado oils (#1 and #42) were blended with two other canola oils (#6 and #7) for the testing data set for the adulterated avocado oils. Similarly, two randomly selected olive oils (#26 and #27) were blended with two soybean oils (#34 and #35) for the training data set for the adulterated olive oils. Two other olive oils (#25 and #28) were blended with two other soybean oils (#33 and #36) for the testing data set for the adulterated olive oils. The inexpensive oil (canola or soybean oil) was mixed with the target oil (avocado or olive oil) in glass vials. The adulterated mixtures were prepared with a range of inexpensive oil compositions (mass/mass) consisting of: 0%, 1%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, and 100% of inexpensive oil in a total of 5 g of oil. Overall, in each of the two adulterated models (avocado oil canola oil mixtures, olive oil and soybean oil mixtures), the adulterated mixtures consisted of 4 different oil blends, 12 different percent compositions per oil blend, and 4 replicates per mixture group for a total of 48 oil samples.

### 2.4. Determination of fatty acid composition by GC-FID

The fatty acid composition of the 47 edible oil samples was determined with a Hewlett Packard 6890 series GC equipped with a 30 m × 250 μm × 0.25 μm DB-WAX bonded-phase fused-silica capillary column (J & W Scientific, Folsom, CA) and a flame ionization detector (FID). Methyl esterification of each oil sample followed a previously published protocol (Monfreda, Gobbi, & Grippa, 2012). Triplicates of each samples were injected into the GC with the following experimental parameters: injection volume of 1.0 μL, a split injection with 30:1 at 280 °C, a constant helium flow rate of 30 mL $min^{-1}$, and a detector temperature of 280 °C. The oven temperature was initially set to 50 °C for 1 min and then ramped to 200 °C with a gradient of 25 °C $min^{-1}$. The oven temperature was held for 1 min at 200 °C and then ramped to 230 °C with a gradient of 2 °C $min^{-1}$. The oven temperature was then held for 5 mins at 230 °C for a total run time of 28 mins. The fatty acid composition of each oil was identified using alkane standards (C8 to C20) and 37 fatty acid methyl esters (FAMEs). The fatty acid composition for each of the four biological replicates from each of the 47 oil types listed in S. Table 1 was determined in triplicate for a total of 564 GC spectra. The composition of each fatty acid was expressed as a percentage of the total fatty acid composition by the peak area ratio derived from the GC spectrum.

### 2.5. Raman spectra of edible oils

Raman spectra were acquired on an XploRA ONE™ Raman spectrometer system (HORIBA, Ltd., Kyoto, Japan) with a 785 nm near-infrared diode laser. The Raman spectra of the edible oil samples were collected as previously described (Du et al., 2019; Zhao, Shen, Wu, Zhang, & Xu, 2020) with the following modifications. Specifically, a tiny portion (~0.1 μL) of a single edible oil or oil mixture was placed on a pre-prepared gold film silicon (10 mm × 6 mm) wafer. The 50X lens was used to focus and then observe the edible oil samples. Each Raman spectrum was collected within 5 min over a wavenumber range of 670 $cm^{-1}$ to 3435 $cm^{-1}$ with a resolution of approximately 3.4 $cm^{-1}$. A Raman spectrum was acquired for five different spots for each edible oil sample. The Raman spectra were collected in triplicate for a total of 15 spectra per edible oil sample. MATLAB® R2020a software (MathWorks®, Natick, USA) software was used for baseline correction, normalization, and Raman shift alignment. The –CH-($CH_3$) asymmetric stretch at an average wavenumber of 1437.95 ± 1.87 $cm^{-1}$ was used to align and normalize the set of Raman spectra to correct for any

displacement along the *x*-axis and intensity differences at *y*-axis. To refine the spectral alignment because of sampling location difference of each data points, Raman spectra with a 1 $cm^{-1}$ resolution between 670 $cm^{-1}$ and 3435 $cm^{-1}$ were generated by a linear interpolation to a standardized reference spectrum by using the MATLAB® software. Detail information and schematic diagrams of Raman signal processing can be found in S. Figs. 1-3 of Supplementary materials.

### 2.6. Data processing and machine learning algorithms

Hierarchical cluster analysis of fatty acid composition was applied by using the MATLAB® R2020a software (MathWorks®, Natick, USA). PCA models were produced using R version 3.5.2. The heatmap and hierarchical clustering of pairwise Pearson correlation coefficients (*r*) to correlate fatty acids with Raman bands were also generated with the R software. Machine learning algorithms were implemented in the Python 3.5.7 programming environment.

To create an equally distributed training and test data set for each edible oil type, four brands from each of the ten oil types listed in S. Table 1 (only #1 to #40) were used for the machine learning and deep learning study. For each edible oil type, there were four biological replicates. The first two oil brands were randomly assigned and selected as training data sets. The remaining two brands were used as a test data set. In this regard, the data was equally and independently partitioned between technical data and biological samples. Overall, there were 60 (3 replicates × 2 brands × 10 oil types) GC fatty acid compositions (FACs) in the training data set and 60 FACs in the testing data set, respectively. Similarly, there were 300 (15 replicates × 2 brands × 10 oil types) Raman spectra in the training data set and 300 Raman spectra (15 replicates × 2 brands × 10 oil types) in testing data set, respectively. For the oil adulteration study, the training and test datasets were derived from different set of edible oil samples as described above. For each of the avocado-canola and olive-soybean mixture systems, 144 Raman spectra (6 replicates × 2 independent oil mixtures × 12 oil at different percent compositions) were used in the training data set and 144 Raman spectra (6 replicates × 2 independent oil mixture × 12 oil at different percent compositions) were used in the testing data set, respectively.

Supervised machine learning algorithms included PCA for dimension reduction with multinomial logistic regression (MLR), MLR with L1 penalty, MLR with L2 penalty, MLR with elastic net penalty, PCA with RF, RF, PCA with boosting, boosting, and one-dimensional convolutional neural network (1D-CNN), which were used for classification of the GC fatty acids and edible oils spectral data sets. PCA with linear regression (LNR), LNR with L1 penalty, LNR with L2 penalty, LNR with elastic net penalty, partial least squares (PLS) regression, PCA with RF, RF, PCA with boosting, and boosting were applied for regression analysis of the adulterated oil data sets. Prediction accuracy and coefficient of determination ($R^2$) were used to evaluate the performances of each machine learning model in regard to edible oil type classification and detection of adulterated oils. The predictive models were implemented in the Python 3.5.7 programming environment. The RF model provided variable importance in addition to oil classification. All the machine learning and 1D-CNN models were computed on a Windows 10 x64 system with an Intel® Core™ i5-6300HQ 2.30 GHz*2 CPU and 16 GB DDR3 ram.

## 3. Results and discussion

### 3.1. Fatty acid composition and classification of edible oils by PCA

GC techniques are routinely employed to obtain fatty acid compositions (FACs) of oils to authenticate vegetable oils (Aparicio & Aparicio-Ruíz, 2000; Lim, Pan, Yu, & Xiao, 2020). Therefore, GC was used to obtain the FACs for the 47 edible oil samples in S. Table 1 to establish a standard reference data set in order to evaluate the performance of the machine learning or deep learning models. As can be seen from the heatmap and hierarchical cluster analysis in Fig. 1, different types of edible oils have different FAC profiles while the same type of oils has a similar profile.

PCA is an unsupervised learning method that visualizes data by dimensional reduction and cluster analysis (Çam, Hişil, & Durmaz, 2009; Liu, Liu, Hu, Yang, & Zheng, 2016). PCA was used for the classification and comparison of the FAC profiles for the edible oils. A biplot of the resulting PCA model is shown in S. Figure 5. The PCA biplot distributed the edible oils into separate clusters in PC-space, which enabled a different approach to visualize relative similarities in the FAC profiles. Importantly, the key fatty acid classes that contribute to group separation are readily apparent from the biplot. For instance, the C8:0 (caprylic acid/octanoic acid), C10:0, C12:0 and C14:0 vectors point to coconut oils, which indicates the coconut oils contain more medium-chain fatty acids (Kinsella, Maher, & Clegg, 2017) relative to the other oils. This is further evident by the heatmap and hierarchical cluster analysis in Fig. 1. In general, different brands of the same type of oil, especially staple commodities such as canola, soybean, and olive, clustered together in the PCA biplot (S. Figure 5). A PCA biplot has previously been shown to be an effective approach to differentiate between six types of edible oils by cluster analysis (Green et al., 2020). However, as revealed by our PCA model, the edible oils cannot be fully differentiated based on only the first two principal components because the PC1 and PC2 explained 36.9 % and 17.2 % of variance respectively, which can be regarded as low explanation variance by the first two principal components (PC), and indicates more PCs, such as PC3 and PC4 should be added for differentiating oil types and to increase the explained variance. The FAC of the avocado oils was found to be similar to the olive oils and, as a result, clustered together in the PCA biplot. A similar outcome was obtained when comparing peanut oil to olive oil, or when comparing hemp oil, grapeseed oil, and corn oil. Incorporating additional principal components into the PCA model may increase the ability of the PCA biplot to differentiate between the edible oils; nevertheless, it is important to understand that principal component analysis (PCA) is generally applied to feasibly observe an original multidimension dataset on reduced dimensions (Townes, Hicks, Aryee, & Irizarry, 2019), thereby to increase data interpretability. For example, hemp oil overlapped with the corn and grapeseed oils in the PC1 vs. PC2 plot (S. Figure 5A), but were well separated in the PC2 vs. PC3 plot (S. Figure 5B); however, using multiple PCs is time-consuming and not an effective way to classify the multidimension data of fatty acid composition of edible oils.

### 3.2. Fatty acid composition and classification of edible oils with machine learning algorithms

Nine supervised machine learning algorithms were employed to classify the edible oils based on FAC (Table 1A) and to compare to the PCA model. The predictive models based on MLR with L1 penalty, MLR with L2 penalty, or MLR with elastic net penalty obtained a 100% testing accuracy, which indicated that these algorithms were very effective in the classification of edible oils. The PCA-RF model is an attractive alternative since training was completed in 0.085 s while achieving a testing accuracy of 95.0%. MLRs with L1 Penalty, L2 Penalty and Elastic net Penalty all achieved 100% test accuracy; however, the training times were 5.458 s, 13.524 s and 4.848 s respectively, which were about 2 to 3 order of magnitudes longer than PCA-RF. The remaining algorithms failed to provide an accurate classification, which included the 1D-CNN deep learning method. The failure of 1D-CNN was not surprising considering the limited-size of the dataset. It is important to note, a deep learning neural network model was previously successful in classifying 19,583 oil samples collected over 5 years (Lim et al., 2020).

Overall, we demonstrated that at least three different machine learning algorithms were highly effective in accurately classifying edible oils based on fatty acid compositions. Importantly, the machine learning algorithms provided a significant improvement over PCA in the

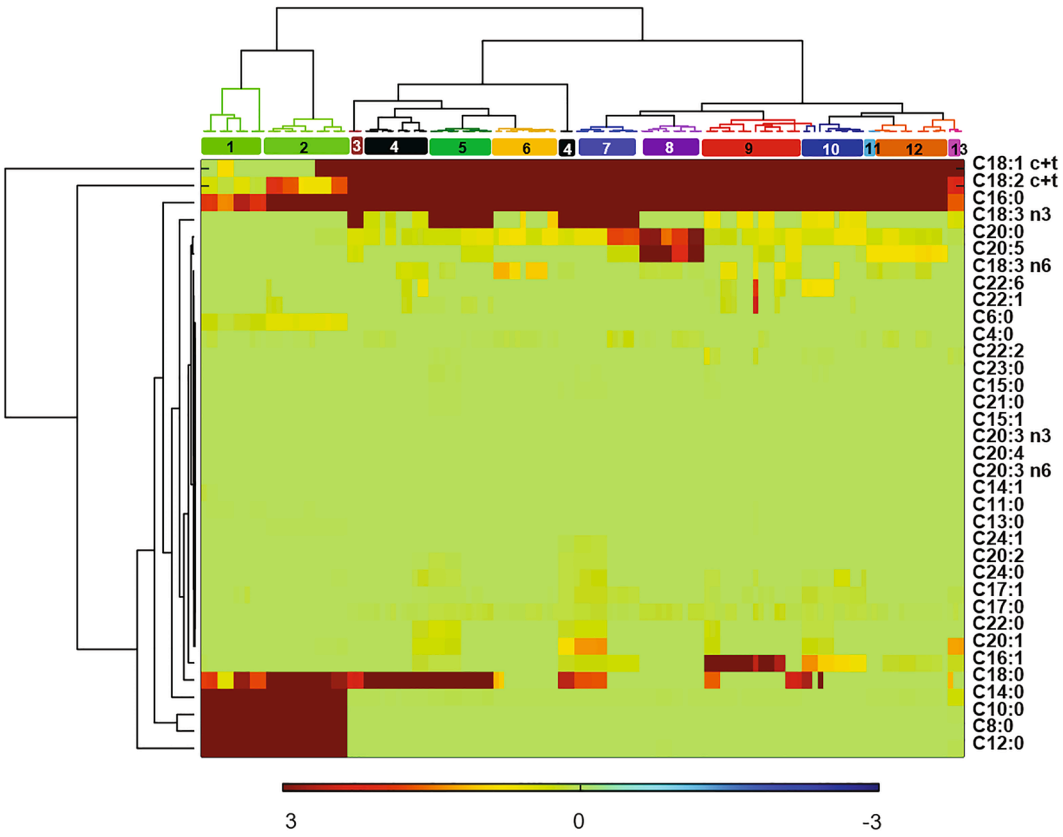


**Fig. 1.** Hierarchical cluster analysis of fatty acid composition of 47 edible oils with different types and brands. 1 LiquidCoconut, 2 Coconut, 3 Hemp, 4 GrapeSeed, 5 Soybean, 6 Corn, 7 , Canola, 8 Peanut, 9 Avocado, 10 Olive, 11 Safflower, 12 Sunflower, 13 Algae. The dendrogram (hierarchical cluster) on the top of columns indicates the similarity of fatty acid composition of oils, whereas the dendrogram on left side of rows indicates the similarity of the distribution of specific fatty acids (biomarkers) among oils.

**Table 1A**
Classification of 10 types of edible oils with different brands by machine learning algorithms based on the fatty acid composition.

| Methods | PCA + MLR | MLR with L1 Penalty | MLR with L2 Penalty | MLR with Elastic net Penalty | PCA + RF | RF | PCA + Boosting | Boosting | 1D-CNN |
|---|---|---|---|---|---|---|---|---|---|
| | Machine learning | | | | | | | | Deep learning |
| Training time (s) | 0.004 | 5.458 | 13.524 | 4.848 | 0.085 | 0.085 | 0.591 | 0.652 | 15.912 |
| Training accuracy | 0.450 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.600 |
| Testing accuracy | 0.450 | 1.000 | 1.000 | 1.000 | 0.950 | 0.817 | 0.650 | 0.700 | 0.600 |

*Note*: PCA (principal component analysis), MLR (multinomial logistic regression), RF (random forest), 1D-CNN (one-dimensional convolutional neural network). 10 types of oils included avocado, canola, coconut, liquid coconut, corn, grapeseed, olive, peanut, soybean, and sunflower oils. Accuracy, 1 = 100%.

classification of edible oil types. Machine learning was faster, more accurate, and provided a direct readout of each oil's classification.

### 3.3. Classification of edible oils by combining Raman spectroscopy with PCA

The Raman spectra of 47 edible oils were collected and 10 major peaks with relative intensities from ~0.2 to ~1 were detected in the range of 500 to1800 $cm^{-1}$ and 2700 to 3010 $cm^{-1}$ (Fig. 2A). The major peaks in the Raman spectrum depicted in Fig. 2B were annotated with chemical functional groups. As expected, different types of oils have unique Raman spectra. For example, the ratio of peak intensities ($I_{1266/1300}$) comparing fatty acid unsaturation (1266 $cm^{-1}$, =CH-H deformation (def)) to saturation (1300 $cm^{-1}$, $-CH_3$) ranged from 0 to 1.5 for the edible oils in this study. Specifically, saturated coconut oil had a $I_{1266/1300}$ of 0.30 ± 0.01, unsaturated olive oil a value of 0.60 ± 0.02, unsaturated canola oil a value of 0.84 ± 0.02, and unsaturated soybean oil a value of 1.0 ± 0.1. The increase in $I_{1266/1300}$ may reflect the high content of C18:3 n3 (linolenic acid), where it was 18.15% in the polyunsaturated hemp oil samples based on the GC analysis results. This is consistent with a previous report by Jiménez-Sanchidrián and Ruiz (Jiménez-Sanchidrián & Ruiz, 2016), which identified a $I_{1266/1300}$ value of 1.8 as being characteristic of a polyunsaturated linseed oil. However, it was difficult to differentiate all the oils based solely on the $I_{1266/1300}$ ratio since algae, avocado, grapeseed, olive, and soybean oil had nearly identical ratios. Compared to the fatty acid profiles showed in Fig. 1, the observed differences in the Raman spectra of edible oils (Fig. 2) were modest, at best, and difficult to visually identify. Accordingly, a PCA was conducted to better differentiation oil types based on their Raman spectrum.

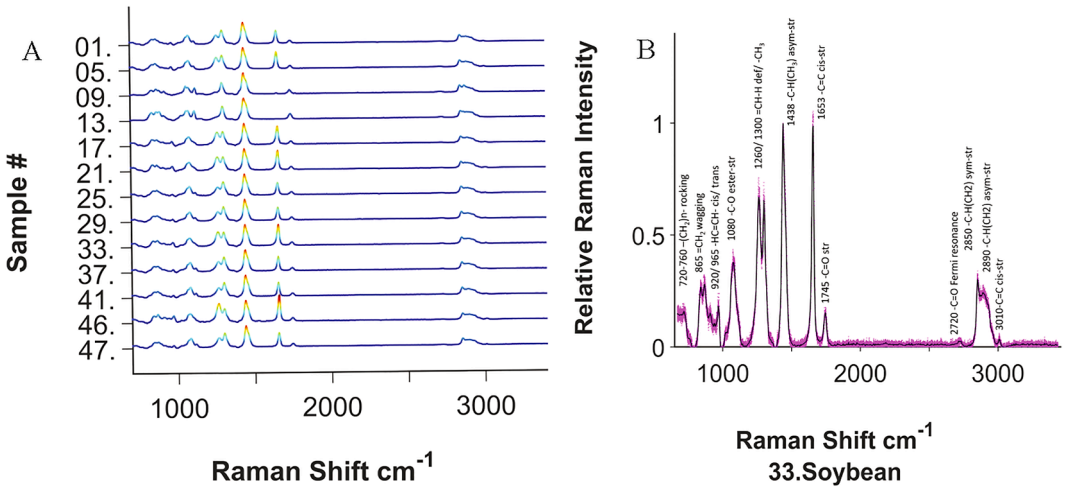


**Fig. 2.** Raman spectra (heat map) of 47 edible oils with different types and brands (A) and representative Raman spectra of #33. soybean oil with marked functional groups (B), purple dots represent 15 replicates of spectra and center black line represent the average. *Note*: numbers of sample codes refer to S.Table 1. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The PCA scores plot generated from the Raman spectral data set is shown in S. Figure 6. The PC1 and PC2 of Raman spectra explained 17.0 % and 9.8 % of variance, respectively, which were lower than the explanations of the top two PCs (PC1 = 36.9%, PC2 = 17.2%) on FACs data (S. Figure 5). The same types of oils were clustered together in the scores plot regardless of brand. This was true for coconut, liquid coconut, canola, olive, soybean, and hemp oils. In general, different oil types formed distinct clusters in the PCA scores plot. However, some oil types were not completely separated from each other and were not differentiated by the PCA model. Specifically, grapeseed, soybean, and corn oil clustered together, while avocado, olive, sunflower, and peanut oil formed a second large cluster. Overall, the PCA model generated from the Raman spectral data set was less effective in classifying edible oil types than the PCA model produced from the fatty acid composition data set. Thus, an advanced classification method was employed to improve the utility of Raman spectral analysis in differentiating various edible oils.

### *3.4. Classification of edible oils by combining Raman spectroscopy with machine learning*

Similar to the fatty acid composition analysis, nine supervised machine learning algorithms were employed to classify edible oils based on Raman spectra (Table 1B). The testing accuracies from these machine learning models ranged from 57% (1D-CNN) to 84.7% (RF) and were generally lower than the accuracies obtained with the fatty acid composition models (Table 1A). However, the 84.7% classification accuracy by RF may qualify as a rapid and reliable test. The machine learning algorithms commonly misassigned avocado oil as either olive oil or peanut oil. Similarly, grapeseed oil was misassigned as either soybean oil or corn oil in the confusion matrix (data not shown). These incorrect assignments were a primary cause of the reduced testing accuracies. Simply, the oils exhibited comparable fatty acid compositions (Fig. 1) and similar Raman spectra (Fig. 2).

A higher testing accuracy was obtained by excluding avocado oil and grapeseed oil from the data set (Table 1C). In this regard, the testing accuracies of PCA with RF, RF, MLR with L1 penalty, MLR with L2 penalty increased to 96.7%, 92.9%, 90.0%, and 89.2%, respectively, although some specific oils had relatively high classification errors. For instance, 46.67 % of olive oils were categorized as the sunflower oils in Fig. 3B due to the similarity of fatty acid composition as can be seen from the clusters 10 olive oils and 12 sunflower oils in Fig. 1; however, the machine learning models classified most of the individual oils at high test accuracy > 90%. The major vegetable oils that are internationally traded include coconut, cotton, olive, palm, peanut, rapeseed (canola), soybean, and sunflower oils (Sharma, Gupta, & Mondal, 2012). Notably, most of these edible oils were included in our investigation. Also, the biological replicates for each oil type were randomly selected from local markets and used as both training and testing samples. Thus, the high overall testing accuracies strongly validated the effectiveness of combining machine learning with Raman spectroscopy to authenticate the major classes of internationally traded vegetable oils.

The classification confusion matrices for the best-performing machine learning models are shown in Fig. 3. As an illustration, the top ten Raman bands from the RF model that were used for the classification of edible oils are shown in Fig. 3E. The Raman bands at 1262 $cm^{-1}$ (=CH–H def) and 1654 $cm^{-1}$ (–C=C *cis*-stretching, *cis*-str) were the top important variables in the RF model. Accordingly, the RF classification

**Table 1B**
Classification of 10 types of edible oils with different brands by machine learning algorithms based on the Raman spectra.

| Methods | PCA + MLR | MLR with L1 Penalty | MLR with L2 Penalty | MLR with Elastic net Penalty | PCA + RF | RF | PCA + Boosting | Boosting | 1D-CNN |
|---|---|---|---|---|---|---|---|---|---|
| | Machine learning | | | | | | | | Deep learning |
| Training time (s) | 0.022 | 1350.886 | 122.772 | 1171.726 | 2.357 | 3.176 | 2.070 | 64.326 | 849.498 |
| Training accuracy | 0.847 | 0.997 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.600 |
| Testing accuracy | 0.713 | 0.747 | 0.803 | 0.780 | 0.817 | 0.847 | 0.680 | 0.663 | 0.570 |

Note: PCA (principal component analysis), MLR (multinomial logistic regression), RF (random forest), 1D-CNN (one-dimensional convolutional neural networks). 10 types of oils included avocado, canola, coconut, liquid coconut, corn, grapeseed, olive, peanut, soybean, and sunflower oils. Accuracy, 1 = 100%.

**Table 1C**
Classification of 8 types of edible oils with different brands by machine learning algorithms based on the Raman spectra.

| Methods | PCA + MLR | MLR with L1 Penalty | MLR with L2 Penalty | MLR with Elastic net Penalty | PCA + RF | RF | PCA + Boosting | Boosting | 1D-CNN |
|---|---|---|---|---|---|---|---|---|---|
| | Machine learning | | | | | | | | Deep learning |
| Training time (s) | 0.016 | 972.72 | 73.24 | 917.64 | 0.68 | 2.340 | 0.660 | 40.87 | 548.60 |
| Training accuracy | 0.904 | 0.996 | 1.000 | 0.954 | 1.000 | 1.000 | 1.000 | 1.000 | 0.625 |
| Testing accuracy | 0.829 | 0.900 | 0.892 | 0.863 | 0.967 | 0.929 | 0.858 | 0.758 | 0.621 |

Note: 8 types of oils included canola, coconut, liquid coconut, corn, olive, peanut, soybean, and sunflower oils, but excluded avocado and grapeseed oils. Accuracy, 1 = 100%.

identified these spectral features as a potential chemical fingerprint of edible oil types. It should be noted that our study did not use the SERS technique, which may provide a higher sensitivity and further improve the predictive model. Liquid interfacial SERS with gold nanoparticles has been previously reported to discriminate between edible oil types, oxidation state, and adulteration using a PCA model (Du et al., 2019). Another SERS study was able to quickly differentiae six types of edible oils (Vander Ende et al., 2019). However, the preparation and maintenance of surface-enhanced nanoparticles significantly reduced the throughput and increased the cost of the Raman analysis. Although the PCA model showed a difference in how oil types clustered in the resulting scores plots, no biological replicates were used for validating the accuracy of the developed models.

Overall, our findings demonstrated the general utility of combining Raman spectroscopy with machine learning for the classification of edible oils. The machine learning models performed better than PCA in the classification of edible oil types by being faster, more accurate, and by providing a direct readout of group membership. Our Raman-machine learning method exhibited a comparable accuracy with the previously reported SERS-PCA model and with the machine learning model of fatty acid compositions described herein. Importantly, the Raman-machine learning method is faster and cheaper, and could be used to develop a rapid on-line or off-line analysis platform.

### 3.5. Predicting adulterated edible oils by combining Raman spectroscopy with machine learning

The high classification accuracies which were achieved by combining Raman spectroscopy with machine learning suggested the same approach would be amenable to detecting adulterated oils. Two adulteration models were selected to evaluate the utility of the Raman-machine learning approach to detect adulterated oils. Specifically, avocado oil was adulterated with canola oil; and olive oil was adulterated with soybean oil. The results of the machine learning models are summarized in Table 2A, which indicates that the LNR with L2 penalty was the best performing model for predicting avocado oil adulterated with canola oil with an $R^2$ of 0.910. LNR with L2 penalty was the best model for predicting olive oil adulterated with soybean oil with an $R^2$ of 0.984 (Table 2B). Interestingly, the testing accuracies were higher for all models when predicting olive oil adulterated by soybean oil compared to the models predicting avocado oil adulterated with canola oil. Overall, the LNR with L2 penalty was identified as the best-performing machine learning algorithm for predicting the adulteration of edible oils based on Raman spectra. The regression for true values versus predicted values for the LNR with L2 penalty model is shown in S. Figure 7. Simply, a better convergence or smaller variance in the data was apparent when predicting olive oil adulterated by soybean oil. Furthermore, the mean squared prediction errors (MSPE) as shown in Tables 2A and B suggests the LNR with L2 penalty model converged better for the olive oil adulterated by soybean oil (*i.e.,* 14.851) compared to the avocado oil adulterated by canola oil (*i.e.,* 83.029). The incremental improvement in performance was likely attributed to a larger difference in the Raman spectra of the edible oils. For example, the difference (*e.g.,* $\Delta I_{1260} \approx 0.28$) between the olive oil ($I_{1260} \approx 0.40$) and soybean oil ($I_{1260} \approx 0.68$) spectra was almost twice the difference (*e.g.,* $\Delta I_{1260} \approx 0.15$) observed between the avocado oil ($I_{1260} \approx 0.40$) and canola oil ($I_{1260} \approx 0.55$) spectra.

Recently, the identification of rainbow trout meat adulterated with Atlantic salmon meat was accomplished by combining Raman spectroscopy with machine learning techniques (Chen, Wu, Xiang, Xu, & Tian, 2019). The mean squared prediction errors (MSPE) of the test dataset was 107.95 and the prediction accuracy ($R^2$) was 0.87. These metrics are comparable to the outcomes from our predictive models, but we observed a higher prediction accuracy ($R^2 = 0.984$) and lower MSPE (14.851). Also, the identification of oil adulterations has been achieved by a deep-learning coupling with GC-FID technique based on a 2-, 3- and 4-way oil mixture model (Lim et al., 2020). For 3-way adulterated mixtures of groundnut oil, the authors observed a median absolute error between 1.2 and 0.95% for predicting both the major groundnut oil and the minor adulterant oil. The GC-FIDs were collected with an approximate 50 min protocol, but in our approach both the Raman spectra and the predictions from the machine learning models are obtained within seconds (Tables 2A and B). Thus, our Raman-machine learning approach may greatly reduce the time and cost of an analysis of adulterated oils.

### 3.6. Correlation between fatty acid composition and the Raman spectra of edible oils

A correlation between fatty acid composition and the Raman spectra of various edible oils was also examined. The intensity of Raman spectral bands corresponding to specific fatty acid functional groups should be consistent with the fatty acid composition observed for each edible oil. Simply, as the fatty acid composition changes between the different oil types, a proportional change in the intensity of the corresponding Raman band should occur. The fatty acid composition of each oil type was accurately determined using GC-FID, thus, it should be feasible to correlate the known variation in fatty acid composition with the corresponding Raman spectrum. Pearson correlation coefficients ($r$) were calculated between each detected fatty acid and observed Raman band and then plotted as a heatmap with hierarchical clustering (Fig. 4). In this regard, highly variable fatty acids would be expected to correlate with highly variable Raman band intensities.

As shown in Fig. 4, Raman bands associated with carbon double bonds (C═C) at 920, 965, 1260, 1653, and 3010 $cm^{-1}$ were found to be positively correlated with unsaturated fatty acids such as C18:1 ($r \approx 0.3$), C18:2 ($r \approx 0.8$), and C18:3 ($r \approx 0.5$). Conversely, the C═C bond vibrations at 920, 965, 1260, 1653, and 3010 $cm^{-1}$ were negatively correlated ($r \approx -0.5$) with saturated fatty acids (C6:0 to C18:0). Presumably, an increase in the proportion of saturated fatty acids led to a corresponding decrease in unsaturated fatty acids (C18:2 and C18:3), which, in turn, resulted in a reduction in the total amount of C═C bonds in the oils and a decrease in the intensity of the corresponding Raman

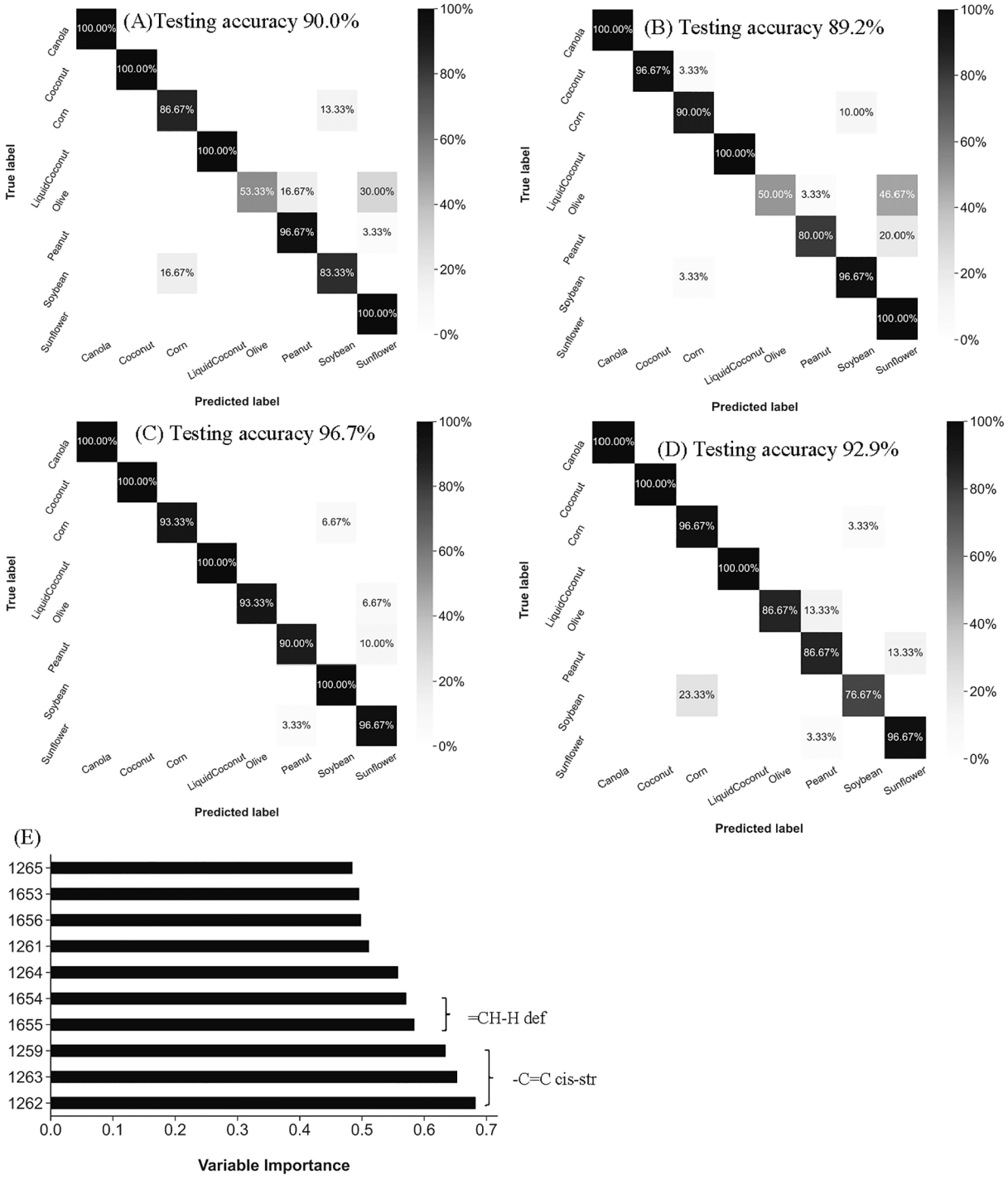


**Fig. 3.** Classification confusion matrix of Raman spectra of 8 types of edible oils by machine learning models: (A) MLR with L1 penalty, (B) MLR with L2 penalty, (C) PCA + random forest (RF), (D) RF, and (E) representative top 10 variable (Raman shift $cm^{-1}$) importance from RF.

bands. Similarly, ester bands, including -C—O- at 1080 $cm^{-1}$ ($r \approx 0.8$) and -C═O at 1745 $cm^{-1}$ ($r \approx 0.8$), were mainly correlated with short to medium chain fatty acids in coconut oils, such as C8:0, C10:0, and C11:0. Presumably, the ester bands were more pronounced in the short to medium chain fatty acids because of the lower molecular mass, which simply led to a relative increment in the ester vibrations. The observed and expected correlation between Raman bands and functional groups within fatty acid molecules provided further evidence that a Raman spectrum can explain differences in fatty acid composition between edible oils. To the best of our knowledge, this is the first reported correlation between fatty acid composition and Raman spectra for a variety of edible oils.

## 4. Conclusion

We described a protocol that combined machine learning algorithms with Raman spectroscopy or fatty acid composition to characterize edible oils. Our method yielded a high accuracy in classifying edible oil types and, accordingly, is an effective means of detecting adulterated oils. Our approach is faster, more accurate, and provides a clear oil classification compared to standard PCA methods. The PCA with RF model was found to be the best performing machine learning algorithm

**Table 2A**
Machine learning for regression of avocado oil adulterated by canola oil.

| Methods | PCA + LNR | LNR with L1 Penalty | LNR with L2 Penalty | LNR with Elastic net Penalty | PLS Regression | PCA + RF | RF | PCA + Boosting | Boosting |
|---|---|---|---|---|---|---|---|---|---|
| Training time (s) | 0.005 | 4.775 | 0.022 | 4.145 | 4.854 | 0.831 | 0.252 | 0.097 | 4.762 |
| $R^2$ | 0.993 | 0.997 | 0.997 | 0.990 | 1.000 | 0.988 | 0.986 | 1.000 | 1.000 |
| MSE | 6.961 | 3.113 | 3.105 | 9.428 | 0.001 | 11.353 | 13.083 | 0.059 | 0.001 |
| Predicted $R^2$ | 0.862 | 0.873 | 0.910 | 0.903 | 0.827 | 0.858 | 0.814 | 0.879 | 0.827 |
| MSPE | 127.55 | 117.677 | 83.029 | 89.480 | 159.654 | 131.649 | 171.917 | 111.707 | 159.570 |

*Note*: LNR (linear regression), PLS (partial least square), RF (random forest), MSE (mean squared error), MSPE (mean squared prediction error), $R^2$ (coefficient of determination).

**Table 2B**
Machine learning for regression of olive oil adulterated by soybean oil.

| Methods | PCA + LNR | LNR with L1 Penalty | LNR with L2 Penalty | LNR with Elastic net Penalty | PLS Regression | PCA + RF | RF | PCA + Boosting | Boosting |
|---|---|---|---|---|---|---|---|---|---|
| Training time (s) | 0.002 | 3.394 | 0.022 | 1.972 | 2.943 | 0.476 | 1.298 | 0.060 | 2.987 |
| $R^2$ | 0.997 | 0.997 | 0.999 | 0.995 | 1.000 | 0.997 | 0.996 | 1.000 | 1.000 |
| MSE | 2.357 | 2.448 | 0.883 | 4.474 | 0.001 | 3.147 | 3.393 | 0.056 | 0.001 |
| Predicted $R^2$ | 0.984 | 0.975 | 0.984 | 0.974 | 0.954 | 0.963 | 0.959 | 0.966 | 0.954 |
| MSPE | 15.089 | 22.722 | 14.851 | 24.237 | 42.986 | 34.535 | 38.021 | 31.535 | 42.666 |

*Note*: LNR (linear regression), PLS (partial least square), RF (random forest), MSE (mean squared error), $R^2$ (coefficient of determination).

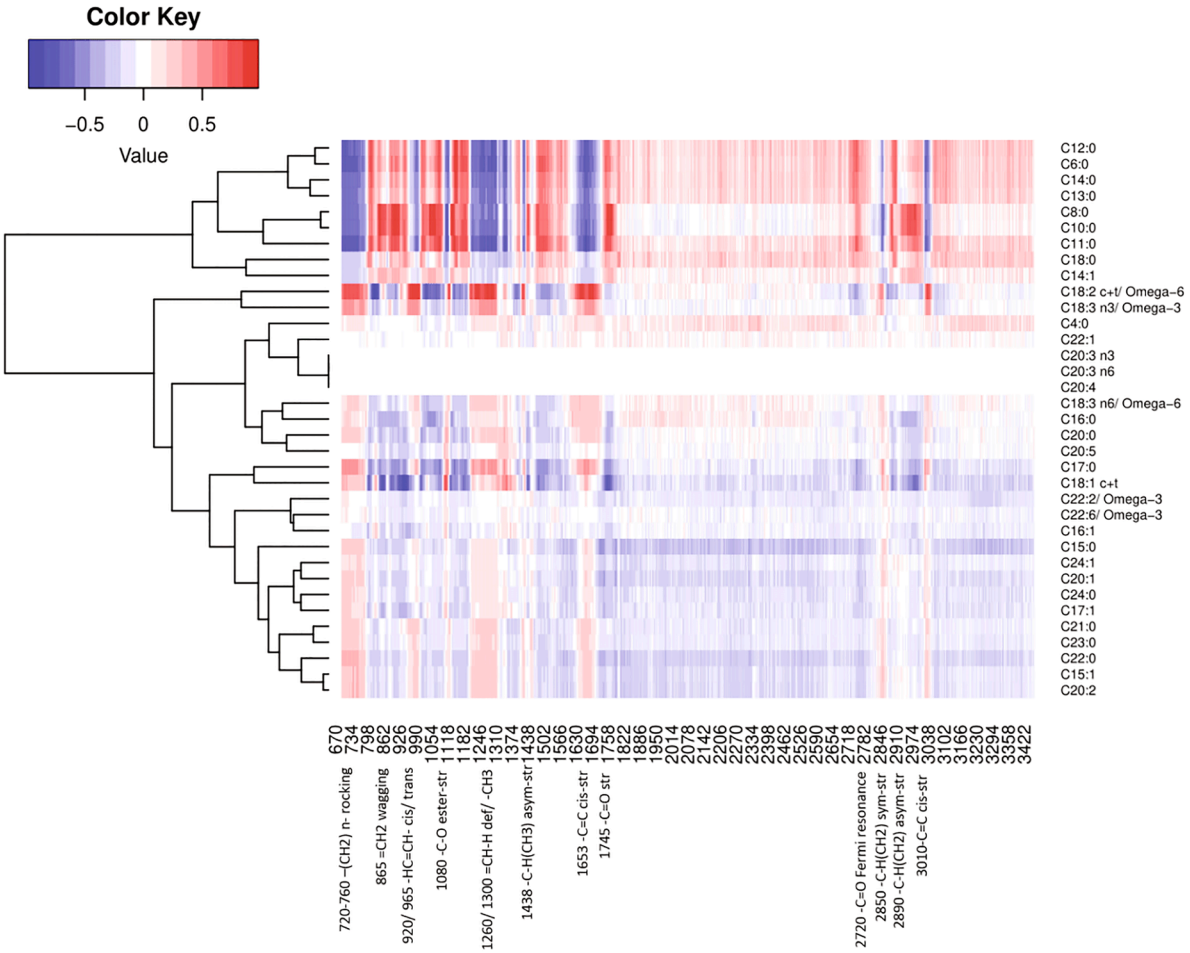


**Fig. 4.** Heatmap of pairwise Pearson correlation coefficients ($r$) between the proportion of fatty acid compositions and Raman spectra of edible oils.

for the classification of edible oils based on Raman spectra. Alternatively, the LNR with L2 penalty model was determined to be the best performing machine learning algorithm for predicting adulterated edible oils. Our approach may be used to establish rapid on-line or off-line platforms for the analysis of edible oils or other food contaminants. Overall, our study demonstrated the potential and value of machine learning assisted Raman spectra analysis for the rapid authentication and detection of contaminants in food products, or identification of origin of agricultural products based on their chemical compositions.

*CRediT authorship contribution statement*

**Hefei Zhao:** Conceptualization, Investigation, Writing – original draft, Data curation, Formal analysis, Methodology, Validation. **Yinglun Zhan:** Formal analysis, Methodology, Validation. **Zheng Xu:** Formal analysis, Methodology, Validation. **Joshua John Nduwamungu:** Investigation. **Yuzhen Zhou:** Conceptualization. **Robert Powers:** Conceptualization, Writing – review & editing. **Changmou Xu:** Conceptualization, Writing – review & editing, Funding acquisition.

## 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.

## Acknowledgments

This work was supported by the Layman Seed Award program of the University of Nebraska Foundation for developing rapid detection methods of food components using Raman Spectroscopy (No. 1024460). This work was supported in part by the National Science Foundation under Grant Number (1660921) to RP and by funding from the Nebraska Center for Integrated Biomolecular Communication (P20 GM113126, NIGMS).

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodchem.2021.131471.

## References

Aparicio, R., & Aparicio-Ruíz, R. (2000). Authentication of vegetable oils by chromatographic techniques. *Journal of Chromatography A, 881*(1–2), 93–104. https://doi.org/10.1016/S0021-9673(00)00355-1

Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., … Shetty, S. (2019). End-to-End Lung Cancer Screening with Three-Dimensional Deep Learning on Low-Dose Chest Computed Tomography. *Nature Medicine, 25*(6), 954–961. https://doi.org/10.1038/s41591-019-0447-x

Baeten, V., Fernández Pierna, J. A., Dardenne, P., Meurens, M., García-González, D. L., & Aparicio-Ruiz, R. (2005). Detection of the presence of hazelnut oil in olive oil by FT-Raman and FT-MIR spectroscopy. *Journal of Agricultural and Food Chemistry, 53*(16), 6201–6206.

Çam, M., Hişil, Y., & Durmaz, G. (2009). Classification of eight pomegranate juices based on antioxidant capacity measured by four methods. *Food Chemistry, 112*(3), 721–726. https://doi.org/10.1016/j.foodchem.2008.06.009

Chen, Z., Wu, T., Xiang, C., Xu, X., & Tian, X. (2019). Rapid Identification of Rainbow Trout Adulteration in Atlantic Salmon by Raman Spectroscopy Combined with Machine Learning. *Molecules (Basel, Switzerland), 24*(15), 2851. https://doi.org/10.3390/molecules24152851

Du, S., Su, M., Jiang, Y., Yu, F., Xu, Y., Lou, X., … Liu, H. (2019). Direct Discrimination of Edible Oil Type, Oxidation, and Adulteration by Liquid Interfacial Surface-Enhanced Raman Spectroscopy. *ACS Sensors, 4*(7), 1798–1805. https://doi.org/10.1021/acssensors.9b00354

Green, H. S., Li, X., De Pra, M., Lovejoy, K. S., Steiner, F., Acworth, I. N., & Wang, S. C. (2020). A rapid method for the detection of extra virgin olive oil adulteration using UHPLC-CAD profiling of triacylglycerols and PCA. *Food Control, 107*, Article 106773. https://doi.org/10.1016/j.foodcont.2019.106773

Jiménez-Sanchidrián, C., & Ruiz, J. R. (2016). Use of Raman spectroscopy for analyzing edible vegetable oils. *Applied Spectroscopy Reviews, 51*(5), 417–430. https://doi.org/10.1080/05704928.2016.1141292

Kinsella, R., Maher, T., & Clegg, M. E. (2017). Coconut oil has less satiating properties than medium chain triglyceride oil. *Physiology & Behavior, 179*, 422–426. https://doi.org/10.1016/J.PHYSBEH.2017.07.007

Lim, K., Pan, K., Yu, Z., & Xiao, R. H. (2020). Pattern recognition based on machine learning identifies oil adulteration and edible oil mixtures. *Nature Communications, 11*(1), 5353. https://doi.org/10.1038/s41467-020-19137-6

Liu, W., Liu, C., Hu, X., Yang, J., & Zheng, L. (2016). Application of terahertz spectroscopy imaging for discrimination of transgenic rice seeds with chemometrics. *Food Chemistry, 210*, 415–421. https://doi.org/10.1016/j.foodchem.2016.04.117

Lussier, F., Thibault, V., Charron, B., Wallace, G. Q., & Masson, J.-F. (2020). Deep learning and artificial intelligence methods for Raman and surface-enhanced Raman scattering. *TrAC - Trends in Analytical Chemistry, 124*, 115796. https://doi.org/10.1016/j.trac.2019.115796

Monfreda, M., Gobbi, L., & Grippa, A. (2012). Blends of olive oil and sunflower oil: Characterisation and olive oil quantification using fatty acid composition and chemometric tools. *Food Chemistry, 134*(4), 2283–2290. https://doi.org/10.1016/J.FOODCHEM.2012.03.122

Sharma, M., Gupta, S. K., & Mondal, A. K. (2012). Production and Trade of Major World Oil Crops Volume 1: Breeding. In S. K. Gupta (Ed.), *Technological Innovations in Major World Oil Crops* (pp. 1–15). https://doi.org/10.1007/978-1-4614-0356-2_1.

Townes, F. W., Hicks, S. C., Aryee, M. J., & Irizarry, R. A. (2019). Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biology, 20*(1), 295. https://doi.org/10.1186/s13059-019-1861-6

Vander Ende, E., Bourgeois, M. R., Henry, A.-I., Chávez, J. L., Krabacher, R., Schatz, G. C., & Van Duyne, R. P. (2019). Physicochemical trapping of neurotransmitters in polymer-mediated gold nanoparticle aggregates for surface-enhanced raman spectroscopy. *Analytical Chemistry, 91*(15), 9554–9562. https://doi.org/10.1021/acs.analchem.9b00773

Yang, H., Irudayaraj, J., & Paradkar, M. M. (2005). Discriminant analysis of edible oils and fats by FTIR. *FT-NIR and FT-Raman spectroscopy. Food Chemistry, 93*(1), 25–32. https://doi.org/10.1016/J.FOODCHEM.2004.08.039

Zhang, Q., Liu, C., Sun, Z., Hu, X., Shen, Q., & Wu, J. (2012). Authentication of edible vegetable oils adulterated with used frying oil by Fourier Transform Infrared Spectroscopy. *Food Chemistry, 132*(3), 1607–1613. https://doi.org/10.1016/J.FOODCHEM.2011.11.129

Zhao, H., Shen, C., Wu, Z., Zhang, Z., & Xu, C. (2020). Comparison of wheat, soybean, rice, and pea protein properties for effective applications in food products. *Journal of Food Biochemistry, 44*(4), Article e13157. https://doi.org/10.1111/jfbc.13157