Food Control 130 (2021) 108318

Contents lists available at ScienceDirect

Food Control

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

Discrimination of olive oils based on the olive cultivar origin by machine learning employing the fusion of emission and absorption spectroscopic data

Dimitrios Stefas [a,b], Nikolaos Gyftokostas [a,b], Panagiotis Kourelias [a,b], Eleni Nanou [a,b], Vasileios Kokkinos [c], Christos Bouras [c], Stelios Couris [a,b,*]

[a] Department of Physics, University of Patras, 26504, Patras, Greece
[b] Institute of Chemical Engineering Sciences (ICE-HT), Foundation for Research and Technology-Hellas (FORTH), 26504, Patras, Greece
[c] Department of Computer Engineering & Informatics, University of Patras, 26504, Patras, Greece

ARTICLE INFO

Keywords:
Laser-induced breakdown spectroscopy (LIBS)
Absorption spectroscopy
Olive oil
Classification
Olive cultivar discrimination
Data fusion

ABSTRACT

In this work Laser-Induced Breakdown Spectroscopy (LIBS) and absorption spectroscopy aided by machine learning are employed for discriminating some extra virgin Greek olive oils of different olive cultivars for the first time. LIBS and absorption spectra of extra virgin olive oils belonging to Kolovi and Koroneiki cultivars, as well as mixtures of them, were collected, analyzed, and used to develop classification schemes employing Linear Discriminant Analysis and Gradient Boosting, the latter allowing the determination of the most important spectral features. Both algorithms were found to provide efficient classification of the olive oil spectra with accuracies exceeding 90%. Furthermore, for the first time, the emission spectra of LIBS were fused with the absorption spectra to create predictive models and their accuracies were found to be significantly improved. This work demonstrates the enhanced capabilities of LIBS and absorption spectroscopy and the potential of their combination for olive oil quality monitoring and control.

1. Introduction

The quality of extra virgin olive oil (EVOO) is determined by its composition and can be influenced by several factors that affect both the fruit physiology and the obtained oil. Factors like the olive cultivar, the climatic conditions, the type of soil, the harvesting and fruit ripening influence the distributions of various olive oil constituents, as e.g., fatty acids and triglycerides (Boskou, 2015). A common practice for producing olive oils with unique quality and characteristics and allowing a brand name in the market is the use of monovarietal olive fruits. In that way, the resulting olive oil is highly dependent on the olive cultivar, geographical origin and climatic conditions. The knowledge of cultivar and geographical origin can be of high commercial interest, especially for premium and high-quality olive oils that can bear marks such as protected designation of origin (PDO), protected geographical indication (PGI) and traditional specialty guaranteed (TSG) (Kosma et al., 2017). However, such extra virgin and premium olive oils are often adulterated using anonymous or less expensive oils as well as other types of vegetable oils. For these reasons, there is an emerging need for developing appropriate methodologies to guarantee oil traceability and identification of geographical origin and/or cultivar (Conte et al., 2020).

During the last decades, several analytical methods have been developed for the determination of authenticity of olive oil such as Gas and High-Performance chromatography (Aparicio, Morales, Aparicio-Ruiz, Tena, & García-González, 2013), Fourier Transform IR spectroscopy (Valand, Tanna, Lawson, & Bengtström, 2019) and Raman spectroscopy (Berghian-Grosan & Magdas, 2020). For the classification of olive oils based on their cultivar/variety, a common strategy is the combination of analytical techniques with chemometric and machine learning methods. For example, Casale, Sinelli, Oliveri, Di Egidio, and Lanteri (2010) have shown the potential of Near- and Mid-Infrared spectroscopy combined with Linear Discriminant Analysis (LDA) for cultivar identification of extra virgin olive oils, while Maléchaux, Laroussi-Mezghani, Le Dréau, Artaud, and Dupuy (2020) applied Partial Least Squares Discriminant Analysis (PLS-DA) on both gas chromatographic and mid-infrared spectroscopic data, for classification of some Tunisian olive oils' varieties (i.e., Chemlali, Chetoui and Oueslati). In another study, Binetti et al. (2017), have applied multilayer perceptron

* Corresponding author. Department of Physics, University of Patras, 26504, Patras, Greece.
E-mail address: couris@upatras.gr (S. Couris).

https://doi.org/10.1016/j.foodcont.2021.108318
Received 26 March 2021; Received in revised form 3 June 2021; Accepted 4 June 2021
Available online 7 June 2021

neural networks for the classification of NMR and NIR data of olive oils from four different Italian cultivars.

In the present work, two spectroscopic techniques, namely Laser-Induced Breakdown Spectroscopy (LIBS) and Absorption Spectroscopy are employed for the classification of olive oils based on their olive cultivar origin, with the aid of machine learning. LIBS is an emerging analytical technique in food analysis (Peng et al., 2019; Senesi, Cabral, Menegatti, Marangoni, & Nicolodelli, 2019; Velásquez-Ferrín, Babos, Marina-Montes, & Anzano, 2020) and, very recently, has been applied for olive oil analysis and classification (Bellou, Gyftokostas, Stefas, ; Caceres et al., 2013; Gazeli, Bellou, Stefas, & Couris, 2020; Gyftokostas, Stefas, & Couris, 2020). LIBS is a laser-based technique capable to provide very rapidly the elemental analysis of a sample, independently of its state of matter and physical properties (e.g., solid, gas, liquid, dielectric or conductive), by creating a plasma on the sample's surface using a focused laser beam. The plasma plume contains the ablated/vaporized material, which is partially atomized, and finally excited and/or partially ionized. By monitoring the emission spectrum of the radiation emitted from the plasma, the elements present in the ablated material can be identified and, eventually, quantified (Fortes, Moros, Lucena, Cabalín, & Laserna, 2012). This well-established technique does not require any sample preparation or pre-treatment and only a very small quantity of the sample is ablated. Moreover, it can be performed in situ, on-line and remotely (François et al., 2020). LIBS has been combined with machine learning and chemometrics for various applications (Yang, Hao, Zhang, & Ren, 2020; Yu, Ren, & Zhao, 2020) because it can quickly create large datasets containing a large number of variables depending on the spectral resolution (Képeš, Vrábel, Střítežská, ; Vrábel et al., 2020). On the other hand, absorption spectroscopy is among the most common spectroscopic techniques used routinely in lab providing information on the absorption characteristics of a sample. Absorption spectroscopy can also produce rapid results and has often been used to fingerprint the authenticity of a material or food. The officially authorized method to characterize olive oils by absorption spectroscopy (and the most common one) is the determination of the extinction coefficients in the UV region, namely K232 and K270, at 232 and 270 nm, respectively, which determine the proportion of oxidized constituents. This method is highly recommended by the European Commission, the International Olive Council and the American Oil Chemists' Society (Aparicio et al., 2013; Conte et al., 2020). Among the applications of absorption spectroscopy for olive oils, some approaches are focusing on the multivariate statistical analysis of olive oil absorbance spectra in the visible and near–infrared spectral regions. For instance, Kružlicová, Mocak, Katsoyannos, and Lankmayr (2008) combined UV-VIS absorption spectra with machine learning techniques such as quadratic discriminant analysis, logistic regression and neural networks for the classification of olive oils according to their geographical origin. Violino et al. (2020) succeeded in discriminating various olive oils in terms of their geographical origin, manipulating their VIS-NIR absorption spectra using neural networks and multivariate analysis of variance. In other works, UV-VIS spectroscopy is combined with other spectroscopies. In particular, Milanez et al. (2017) employed chemometric algorithms such as partial least squares on both fluorescence and UV-VIS absorption spectra to predict the adulteration of extra virgin olive oils, whereas Kontzedaki et al. (2020) applied partial least squares discriminant analysis to Raman, UV-VIS-NIR absorption and fluorescence spectroscopic data for EVOOs classification based on their geographical origin. In the same spirit, LIBS has been combined with other spectroscopic methods, such as Raman spectroscopy for identifying bacterial species and strains (Prochazka et al., 2018), wavelength dispersive X-ray fluorescence for predicting the contents of various inorganic elements in bean seed (Gamela, Costa, Sperança, & Pereira-Filho, 2020) and Ft-IR and Raman for quantifying calcium content in infant formula (Zhao et al., 2020).

The present work presents a significant extension of a recent work (Gyftokostas et al., 2021), where LIBS emission data and data from absorption measurements were used for the geographical discrimination of Greek olive oils employing some machine learning algorithmic approaches. Specifically, in the present work, for the first time, to the best of our knowledge, LIBS emission data and absorption data separately and also in fused form, assisted by machine learning algorithms are used for the discrimination/classification of olive oils in terms of their olive cultivar. In that view, LDA and Gradient Boosting algorithms were used to classify the olive oil spectra in terms of their olive cultivars, while the most important features of the spectra were identified and selected to create a dataset with much fewer features. Then, LDA and Gradient Boosting were used to create new predictive models, which are subsequently assessed and compared with the former models. Finally, a hybrid machine learning model is proposed, for the first time for olive oils classification/discrimination, whereas data fusion LIBS and absorption spectra are used to determine the olive cultivar origin of some Greek extra virgin olive oils.

2. Materials and methods

2.1. The olive oil samples

A total of 41 monovarietal olive oil samples (i.e., 38 extra virgin olive oils (EVOOs) and 3 virgin olive oils (VOOs) samples) were collected from producers, from different areas of the island of Lesvos, Greece, belonging to two types of olive cultivars, i.e., Kolovi and Koroneiki. In addition, 10 extra virgin olive oil commercial samples were purchased from the local market. More information about the olive oil samples is presented in Table S1, including sample names, type of olive farming, altitude of the olive trees' locations and kind of olive fruit ripening. Two of the EVOOs samples were used to prepare mixtures/blends of the two different types of olive cultivars studied here, i.e., the Kolovi and Koroneiki, resulting to 9 more samples corresponding to different mixing ratios of each variety, ranging from 10 to 90% v/v (e.g., 10:90, 20:80, …, 90:10).

The spectra obtained from the samples were split into two sets; one containing spectra from 44 olive oil samples used for the algorithmic training, and the other one comprising the spectra from 16 samples (i.e., 14 monovarietal olive oil samples out of 51, and 2 mixture samples out of 9) used for the external validation of the algorithmic models (see also in Table S1 the samples used for training and external validation). It should be emphasized at this point that the external validation procedure is quite important for a realistic assessment of the predictive models' accuracies. All samples, after their collection, were stored in dark-colored glass bottles and were kept at a temperature of 2–4 °C, protected from light and humidity. Prior to the laser measurements, the olive oil samples were left at room temperature for several hours.

2.2. LIBS experimental setup

LIBS is essentially an emission spectroscopy based technique, where a laser beam is used to create a plasma, emitting the characteristic emissions of its constituents. For the experiments, about 2 ml of each olive oil sample were placed in small Petri dish-like recipient, allowing the access of the laser beam on the sample free surface to induce a spark, i.e., plasma. For the creation of the plasma, the focused laser beam from a 5 ns Q-switched Nd: YAG laser (Quanta-Ray INDI, Spectra Physics) operating at its fundamental wavelength at 1064 nm and with a repetition rate of up to 10 Hz, was used. The laser beam was focused perpendicular to the sample's surface by means of a 150 mm focal length quartz lens. The energy of the laser pulses was about 90 mJ. The laser focusing conditions and the laser energy were optimized to provide a good signal-to-noise-ratio (SNR) and minimizing splashing. The plasma emission was collected via a quartz lens and fed into a quartz optical fiber bundle, being coupled to the entrance slit of a 75 mm focal length spectrograph (Avantes, AvaSpec-ULS4096CL-EVO). The spectrograph was equipped with a 300 lines/mm grating and a 4096-pixel detector

(CMOS) covering the spectral region from 185 to 1347 nm. From these pixels, 2754 ones were used corresponding to the 200–1000 nm spectral region. The measurements were performed using a time delay (t_d) of 1.28 μs and an integration window (t_w) of 1.05 ms for the detector. For the measurements, every laser shot was inducing a plasma, while ten successive laser shots were averaged corresponding to one LIBS measurement. Then, up to 30 such LIBS spectra were collected, from different places on the free surface, and were used for the algorithmic training. More detailed information about the LIBS experimental setup and the spectra acquisition conditions can be found elsewhere (Bellou et al., 2020).

2.3. Absorption spectroscopy experimental setup

The absorption spectra of the olive oil samples were obtained using a spectrophotometer (Jasco V-670), employing 1 mm optical pathlength glass cells. The spectral range used was extending from 350 to 750 nm. Each olive oil sample was pipetted in a 1 mm glass cuvette and for every sample 20 absorption spectra were acquired and used for the algorithmic training. Each spectrum consisted of 801 data points (i.e., pixels).

2.4. Data Analysis

For the analysis of the collected spectroscopic data, two machine learning techniques were selected, employed and assessed for each kind of input dataset, i.e., the LIBS spectral data, the absorption data and the combined LIBS-absorption data using the Python library Scikit-learn (Pedregosa et al., 2011). The different machine learning algorithms applied were Linear Discriminant Analysis (LDA) and Gradient Boosting Classifier. Three classes were used for the classification procedure, namely Kolovi and Koroneiki (corresponding to the samples from Kolovi and Koroneiki cultivars respectively) and Mixtures (corresponding to the Kolovi-Koroneiki blended samples). Data analysis was performed separately for the LIBS and the absorption spectra. Finally, the emission and absorption spectra, from the two techniques, were combined to create machine learning models that were simultaneously taking into account information about the samples' elemental composition (provided by LIBS) and the absorption features (provided from the absorption measurements). This procedure, where two or more types of datasets are combined to create machine learning models, is widely known as data fusion (Andrade, De Almeida, De Carvalho, Pereira-Filho, & Amarasiriwardena, 2021). According to Borràs et al. (2015) data fusion procedures can be categorized in three types, namely low-, mid- and high-level of data fusion. Low-level data fusion is the simple, sample-wise, concatenation of the independent datasets into a single matrix having as many rows as samples analyzed and as many columns as the features measured by the different analytical methods. Mid-level data fusion is performed when feature extraction or selection is performed on each one of the independent datasets prior to their concatenation into a single one. High-level data fusion is performed when machine learning models resulting from each independent dataset are combined. In this work, the low-level data fusion approach was followed on the LIBS and absorption spectroscopic datasets. So, the fused dataset consisted of 3555 features, 2754 of them originating from the LIBS spectral features and the rest 801 from the absorption spectra of the same sample. For the analysis, some data preprocessing was also performed, i.e., the LIBS spectra were normalized by max, while the absorbance data were used as is.

LDA was used for dimensionality reduction, as a pre-processing step, as it is often used for different machine learning applications and for visualization purposes as well. LDA being a supervised learning technique, was used for classification purposes as well. As it is known, LDA maximizes the ratio of the between-class variance over the within-class variance, through a classifier with a linear decision boundary which is generated by fitting class conditional densities to the data using Bayes' rule (Hastie, Tibshirani & Friedman, 2009). The LDA model fits a Gaussian probability density to each class with the assumption that all classes share the same variance-covariance matrix. Gradient Boosting classification algorithm (also known as Gradient Boosted Trees classification) being a supervised ensemble learning algorithm, it produces a predictive model in the form of an ensemble of weak learners, i.e., decision trees, in order for the model to be a strong learner. Similarly to other tree-based ensemble methods, such as Decision Trees and Random Forests, each feature's importance can be evaluated. In this work, Gradient Boosting is used to identify the spectral features that are mostly contributing to the efficient classification of the acquired LIBS and absorption spectroscopic data (Huffman, Sobral, & Terán-Hinojosa, 2019).

Both supervised algorithms (i.e., LDA and Gradient Boosting) were applied on the raw and the reduced data as well. The latter ones were obtained by using the Gradient Boosting algorithm on the raw data and using the resulting important features next, as inputs for both classification algorithms. In general, such pre-treatment procedures can result in effective reduction of the dataset's size, thus using much less inputs compared to the raw data (i.e., the un-treated ones) and having a direct impact on the computational time.

For the evaluation of the model, a train-test split method was applied to the data while the training data was split further during cross validation. In more details, the k-fold cross validation technique was implemented to the algorithmic training data to ensure the stability of the algorithm and obtain the prediction accuracy with $k = 10$. In this way, the dataset is shuffled and split into k groups, where one group is used as test and the remaining k-1 are used as training samples. This procedure is performed k times. In this manner, the classification accuracies can be obtained, allowing for the better and more accurate assessment of the classification procedure. Finally, to ensure the stability of the machine learning models, an external validation procedure with the test data was performed using spectra from olive oil samples left out from the algorithmic training, i.e., previously un-seen by the algorithms. So, the machine learning models are tested and evaluated by checking their capacity to accurately predict new spectra from new unknown (i.e., un-seen) olive oil samples (Hastie, Tibshirani & Friedman, 2009).

3. Results and discussion

3.1. LIBS spectral features

A representative LIBS spectrum of an olive oil sample is presented as an example in Fig. 1a, showing the most characteristic emission features (i.e., spectral lines) related to olive oil's elemental composition. As olive oil consists basically of organic constituents, it has to expected that the plasma emission spectrum will be dominated by spectral lines of Carbon, Hydrogen, Nitrogen and Oxygen. Indeed, the most prevailing of them are the atomic carbon line at 247.9 nm, the Hydrogen's Balmer lines, H_α and H_β, at 656.3 and 486.1 nm respectively, the atomic nitrogen's triplet at 742.4, 744.2 and 746.8 nm, as well as the oxygen's triplet centered at about 777 nm. Moreover, the progression of the molecular emission bands of CN and C_2, arising from the fragmentation of the different olive oil's constituents and subsequently formed under plasma conditions are also clearly visible (Acquaviva et al., 1997; Parigger, 2013). The spectral features were assigned and identified using the NIST atomic spectra database (Kramida, A., Ralchenko, Y., Reader, J. & NIST ASD Team) and other LIBS-related studies of olive oil and other organic materials (Caceres et al., 2013; Gyftokostas et al., 2020; Moros & Laserna, 2019; Stefas, Gyftokostas, & Couris, 2020). Following previous works (Bellou et al., 2020; Caceres et al., 2013; Gazeli et al., 2020; Gyftokostas et al., 2020), it should be noted that the classification of different olive oil LIBS spectra is a rather challenging task, due to their spectral similarity arising from the very similar elemental composition of olive oils. Fig. 1b, presents some representative LIBS spectra of olive oil of the Kolovi and Koroneiki cultivars, as well as a mixture of them. In this figure, the similarities between the spectra are evident while their differences lie

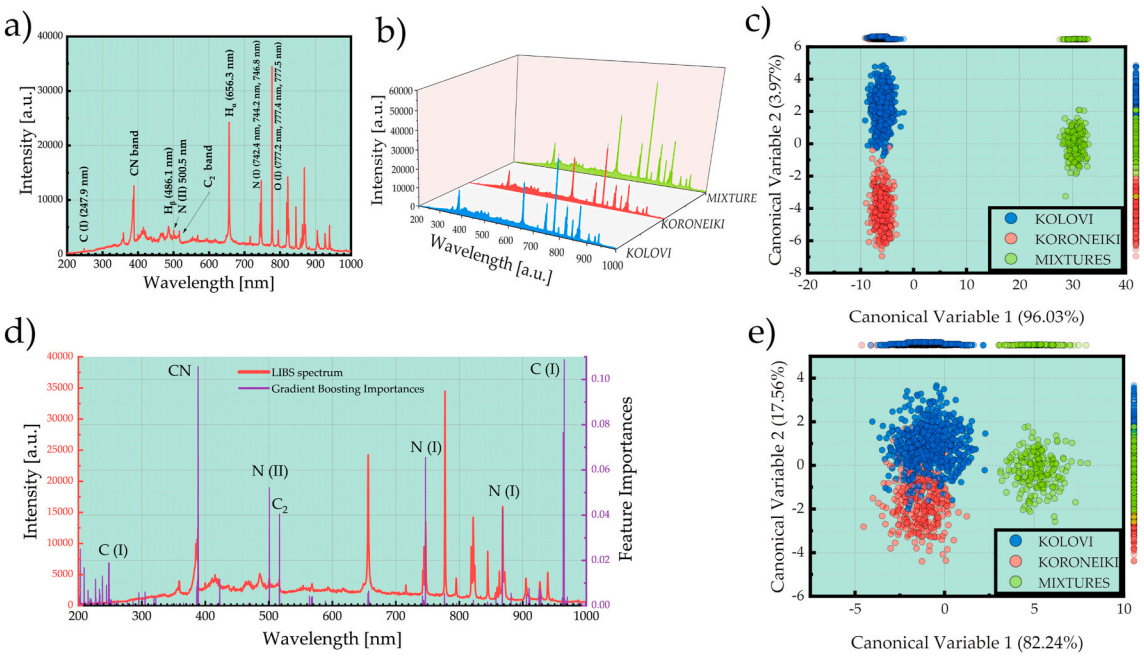

Fig. 1. a) LIBS spectrum of an olive oil, showing the most important emission spectral lines and molecular bands. b) Comparison of LIBS spectra originating from olive oils of the Kolovi and Koroneiki cultivars, and a Kolovi-Koroneiki olive oil mixture. c) Canonical Variable plot resulting from the LDA algorithm showing the distribution of LIBS spectra after dimensionality reduction. d) Feature importances resulting from the Gradient Boosting algorithm and their comparison with an olive oil LIBS spectrum. e) Canonical Variable plot resulting from the LDA algorithm after applying the feature selection procedure, with a threshold of 0.018.

basically on their relative intensities.

3.2. Linear Discriminant Analysis and Gradient Boosting for olive cultivar discrimination

Fig. 1c shows the canonical variable plot of LDA resulting from the training dataset. As can be seen, three distinct clusters of data points are formed, indicating their successful classification bearing a training accuracy of (92.6 ± 2.3)%. A small overlap can be noticed between the Kolovi and Koroneiki samples while their mixtures are clearly discriminated. The first canonical variable describes 96.03% of the dataset's variance and the second one describes 3.97% of it, in such a way that class separability is maximized. In that view, the Mixtures samples are discriminated from both the Kolovi and Koroneiki samples along the first canonical variable axis. For the second canonical variable (describing the 3.97% of the total variance), the Mixtures are placed in between the Kolovi and Koroneiki samples. This is visualized on Fig. 1c, where each data point is projected on the canonical variable's axes.

The LDA model achieved a high classification accuracy after the training procedure, i.e., (92.6 ± 2.3)%. All misclassifications occurred within samples from the Koroneiki and Kolovi cultivars, something which can be observed from the two overlapping clusters of Fig. 1c. This model was used to identify 16 unknown samples and attained a 90.6% accuracy of prediction. Similarly, Gradient Boosting algorithm was used to classify the same LIBS spectra and attained (96.0 ± 1.7)% training accuracy and 93.8% accuracy on predicting the test data.

In Figure S1 the confusion matrices for both LDA and Gradient boosting predictions of the test data are depicted, outlining how accurately each instance from the unknown samples' spectra is predicted. The confusion matrix of LDA on Figure S1a, shows that only 45 of 330 instances (i.e., spectra) of Kolovi origin, are falsely predicted as of Koroneiki origin, corresponding to the overlap of the two classes, noticed previously in Fig. 1c. The instances of Koroneiki origin and the mixtures, in total, were accurately predicted within their class. Figure S1c, shows that merely 30 out of 330 instances of Kolovi origin, are confused with Koroneiki origin but only 1 of 90 instances of Koroneiki is predicted as Kolovi origin. The 60 instances of the mixture, in total, were accurately predicted.

3.3. Feature importances of the LIBS spectra

Gradient Boosting algorithm can estimate the important features from a trained predictive model, as it has been discussed above. In principle, importances are computed as scores, that indicate how useful each feature is for constructing the model's boosted decision trees. Thus, for each feature (i.e., wavelength) of the dataset (which has a total of 2754 features), the feature importances were calculated and are associated with the LIBS spectra. In Fig. 1d, an olive oil LIBS spectrum and the resulting important features after the implementation of the Gradient Boosting algorithm are shown. As it can be seen, the most important features (i.e., those over a threshold of about 0.018) are coinciding with some spectral lines. These are the band-head of the CN progression, lying at 388.3 nm, the ionic nitrogen line at 500.5 nm, the C_2 progression's band-head at 516.5 nm, as well as, the atomic nitrogen lines at 746.8 and 868.0 nm. Two more features correspond to the atomic lines of carbon at 962.1 and 965.8 nm. Other spectral lines seem to have weak or negligible importance, as for instance the H_α and H_β lines and some lines of oxygen, carbon and nitrogen. By using these seven spectral features, the creation of LDA and Gradient Boosting models were performed, and their performances are assessed. Fig. 1e shows the canonical variable plot of LDA resulting from the training dataset.

As it can be observed from Fig. 1e, along the first Canonical Variable's axis, which explains 82.24% of the total variance, the Kolovi and Koroneiki samples are completely separated from their intermediate mixtures. However, along the second Canonical Variable's axis, explaining 17.56% of the variance, the samples seem not to be well-separated. This model's classification accuracy is (90.2 ± 2.9)% and in comparison, with the LDA model prior to the feature selection (i.e., (92.6 ± 2.3)%) it performed quite well, considering that only 7 features were used. Concerning the prediction of the unknown samples, the algorithm predicted 91.5% of the unknown instances.

The Gradient Boosting algorithm, also, performed quite successfully, with the trained model presenting an accuracy of (92.7 ± 1.8)% and a 90.8% prediction accuracy of the unknown data. Despite the accuracies of these feature-reduced models being slightly lower than the original models (i.e., (96.0 ± 1.7)% training accuracy and 93.8% testing accuracy), the obtained results are quite satisfactory taking into account the massive features' reduction, namely from 2754 features to only 7 of them.

3.4. Olive oil absorption spectra and Linear Discriminant Analysis/ Gradient Boosting for olive cultivar discrimination

Some representative absorption spectra of olive oil corresponding to the visible spectral region 350–750 nm, are depicted in Fig. 2a. The observed absorption bands are due to pigments, such as carotenoids and chlorophylls (Domenici et al., 2014; Jimenez-Lopez et al., 2020). As can be seen, Kolovi, Koroneiki and their mixtures absorption spectra exhibit significant differences among them, due to the different amounts of these pigments.

In Fig. 2b the canonical variable plot of LDA is shown, resulting from the training dataset. As can be seen, three distinct clusters of data points are formed, indicating a remarkably successful classification attaining a training accuracy of (100.0 ± 0.0)%. The first canonical variable describes 59.62% of the dataset's variance while the second one describes the rest 40.38% of it. However, when the LDA model was put to the test, only 86.3% of the unknown data were correctly predicted (see, i.e., the confusion matrix in Figure S2a). Next, Gradient Boosting algorithm was used to classify the same absorption spectra. It has attained a training accuracy of (99.8 ± 0.5)% and 80.6% accuracy on predicting the test data (see, i.e., the confusion matrix in Figure S2b). The features that Gradient Boosting algorithm which were found to be important were determined to be in the spectral range 400–430 nm and 650–660 nm, as can be seen in Fig. 2c. The performance of both LDA and Gradient Boosting were investigated for various threshold values. So, the optimum threshold value was determined to 0.01, resulting to only 16 features (i.e., wavelengths) having the highest importances. LDA attained an accuracy of (91.6 ± 3.9)% for training and 84.4% for testing. Correspondingly, the Gradient Boosting attained an accuracy of (99.5 ± 0.6)% for training and 87.2% for testing. The confusion matrices for both LDA and Gradient Boosting can be found in Figure S2c and Figure S2d, respectively. However, these results are rather affected by overfitting of the training data, as the LDA and Gradient Boosting models correspond exactly on the training data and therefore they lose their ability to generalize when new previously unseen data are used for prediction. Actually, overfitting is a very common situation in machine learning and chemometrics, and the most efficient way to check for it, is through the control of the efficiency of a model to predict new data. In the present case, the observed overfitting can have two origins. Either the spectral information of the absorption spectra within the 350–750 nm range is not satisfactory for predicting the cultivar origin of olive oils, or to the low spectral variances of the absorption spectra, the latter not allowing the algorithms to "learn" efficiently from the observed absorption patterns and perform successful classification. On the contrary, the intensities of the LIBS spectral data exhibiting large variances allow for the efficient training of the predictive model. In general, LIBS can provide large data sets in very short time (i.e., hundreds of spectra per second), the corresponding spectra consisting of thousands of variables with high spectral resolution and large variances. The high variance of the LIBS spectra is commonly addressed by averaging the emission collected from several shots, or the comparable approach of numerically averaging the spectra after the acquisition (Képeš, Pořízka, & Kaiser, 2019). The latter is especially effective in the analysis of homogenous samples, such as olive oils.

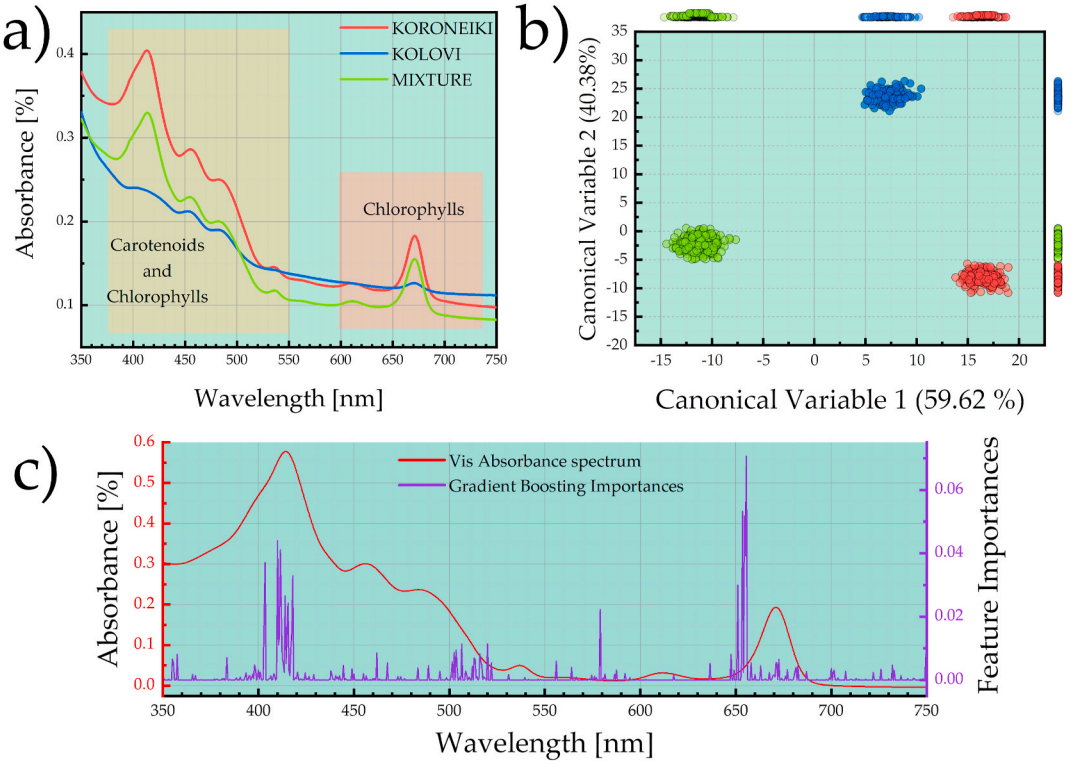

Fig. 2. a) Comparison of olive oil absorption spectra from olive oils of Kolovi and Koroneiki cultivars and a Kolovi-Koroneiki mixture. b) Canonical Variable plot resulting from the LDA algorithm. c) Feature importances resulting from the Gradient Boosting algorithm in comparison with an indicative olive oil absorption spectrum.

3.5. Use of fused emission (LIBS data) and absorption data for the creation of machine learning models

In this section, the performance of the fusion of LIBS and absorption data is examined for the classification of olive oils in terms of their cultivar origins. In that view, machine learning models are created using both LIBS and absorption data as a single data file (i.e., fusing the emission and the absorption data) and it is shown that high classification accuracies of the olive oils in terms of the olive cultivar origin can be obtained. In fact, the determined classification accuracies were found to be significantly improved compared with the classification accuracies obtained using only each one of the two types of spectroscopic data separately. In addition, the overfitting issues encountered, e.g. when the absorption spectroscopic data were used as inputs to the machine learning models, seem to be avoided in this case. A fused LIBS - Absorption spectrum is presented in Fig. 3a, where the two types of spectra have been merged. The x-axis of the plot indicates the number of features of the combined spectra (3555 features, 2754 stemming from the LIBS spectrum and 801 from the corresponding absorption spectrum). The y-axis indicates the numerical value of each feature (which originates from the normalized intensities and the absorbance value of the LIBS and the absorption spectra, respectively). The presented results were obtained after preprocessing of the spectroscopic data (as mentioned in Section 2.4 Data Analysis). More information can be found in the supplementary material, i.e., see Figure S4. When using the LDA algorithm, the training accuracy was (96.0 ± 1.7)% and the testing accuracy was 82.5%. The resulting canonical variable plot is shown in Fig. 3b. However, when using the Gradient Boosting algorithm both high training and testing accuracies were attained, ca. (99.4 ± 0.9)% and 100.0%, respectively. The three most important features in this case correspond to the oxygen atom triplet at ~777 nm and nitrogen line at 868.0 nm for the LIBS-related part of the data, and some part of the chlorophylls and carotenoids at about 370–390 nm for the absorption-part of the data.

The obtained results are summarized in Fig. 3c, where the training and testing accuracies are plotted against the threshold value for the important features. The best results have been obtained for a threshold value of 0.02 and the training and testing accuracies were (99.3 ± 1.0)% and 99.7%, respectively. For this threshold value 9 features were selected. Thus, from a total of 3555 features, with only 9 spectral features a nearly 100% classification accuracy can be attained. The confusion matrices for the predictive models are shown in Figure S3. As can be seen, the results are quite high and the validation accuracy falls within the standard deviation of the training accuracy, avoiding possible overfitting of the predictive model. It is concluded that the accuracy of discrimination compared to each method separately is higher and improved.

Such approaches, i.e., data fusion from different spectroscopies and analytical techniques, have been widely used for the quality assessment and authentication of foodstuff. A detailed overview of this issue can be found in the review of Borràs et al. (2015) and the references within. In the present work, data obtained from LIBS and absorption spectroscopies, were fused together for strengthening of the predictive models' capabilities. With this approach, the weaknesses that may occurs in each technique can be eliminated resulting to excellent discrimination results.

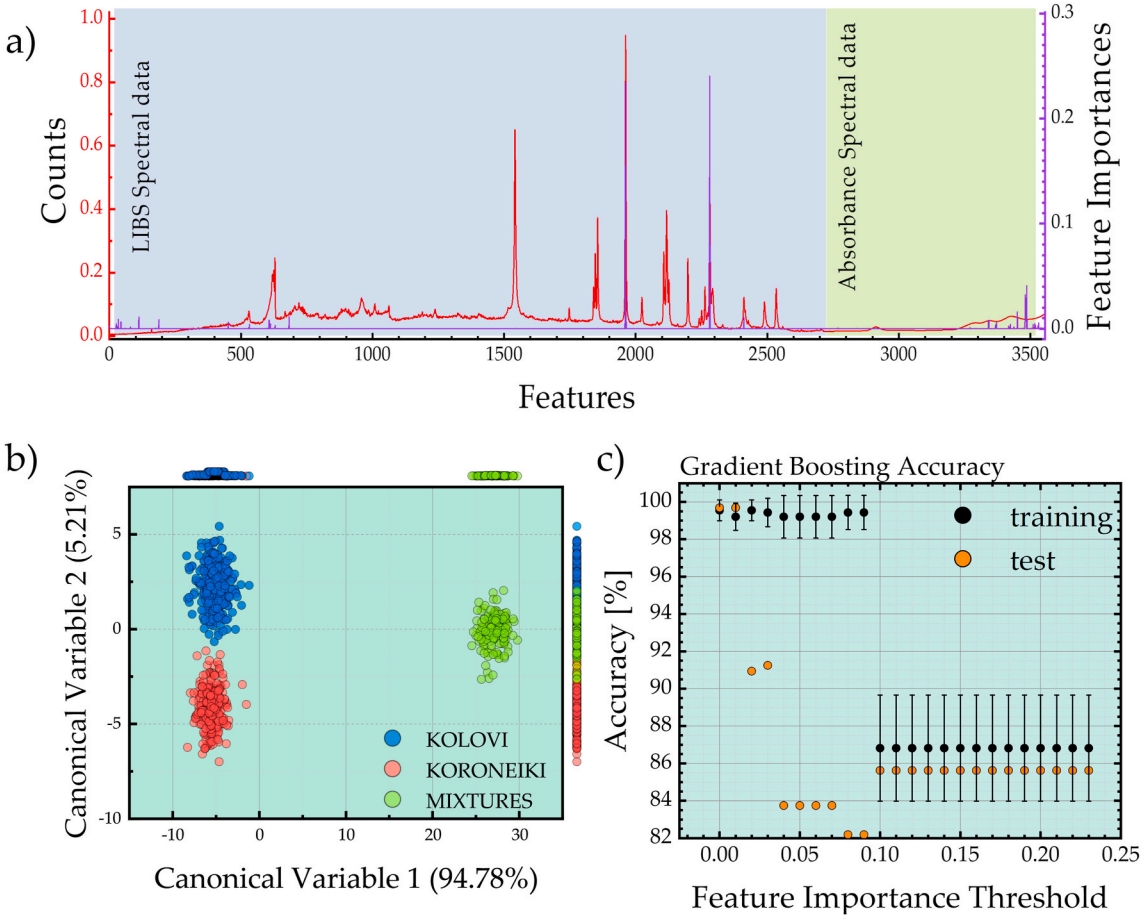

Fig. 3. a) Fused spectrum consisting of LIBS and absorbance data. Feature importances are also indicated. b) Canonical Variable plot resulting from the LDA algorithm. c) Accuracies for training and predicting unknown data for varying feature importance thresholds by using Gradient Boosting algorithm.

4. Conclusion

In this work, two spectroscopic techniques, an emission spectroscopy based one (ca., Laser-Induced Breakdown Spectroscopy) and absorption spectroscopy were used for the discrimination of some Greek olive oils based on their cultivar origin. LDA and Gradient Boosting algorithms were used for the classification and were found to provide excellent classification accuracies, with the best results obtained using the LIBS data, attaining (96.0 ± 1.7)% for training accuracy and 93.8% for external validation, by using the Gradient Boosting algorithm. In the case of absorption spectroscopy, the best results were 100% for training and 86.3% for external validation and were obtained by implementing the LDA algorithm. It was also shown that the classification results can be efficiently improved when using the important features recognized by the Gradient Boosting algorithm, as a feature selection method. Furthermore, the fusion of the two different origins spectroscopic data, i.e., the emission and the absorption spectra, and their use by the predictive models, can lead to an efficient strategy for predicting the cultivar origin of olive oils. By the data fusion methodology employed in this work, the discrimination of olive oils in terms of their cultivar origin, by using Gradient Boosting and LDA, resulted to external validation accuracies of 100% and 82.5%, respectively. The obtained results demonstrate the high potential of such spectroscopic data when assisted by machine learning algorithms for olive oil authentication and classification. In addition, the use and/or the combination of these well established and experimentally mature techniques, such as LIBS and absorption spectroscopy, which can acquire data very rapidly (e.g., in few seconds for LIBS and few minutes for absorption) can be potentially an efficient tool of great importance for food safety and food industry.

Author contributions

D.S, N.G, P.K., E.N.: data curation; S.C.: funding acquisition; D. S, N. G., P.K., E.N.: investigation; C.B, S.C.: methodology; S.C.: project administration; S.C.: resources; D.S., N.G., P.K., V.K: software; C.B., S.C.: supervision; D.S., N.G., P.K., E.N., V.K., C.B., S.C.: writing–original draft; D.S., N.G., V.K., C.B., S.C.: review and editing.

Declaration of competing interest

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

Acknowledgements.

This research has been partially supported by Greek national funds through the Public Investments Program (PIP) of General Secretariat for Research and Technology (GSRT), under the Emblematic Action “The Olive Road” (project code: 2018ΣΕ01300000). D.S. acknowledges support from the «Andreas Mentzelopoulos Foundation».

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodcont.2021.108318.

References

Acquaviva, S., Caricato, A. P., Giorgi, M. L., Dinescu, G., Luches, A., & Perrone, A. (1997). Evidence for CN in spectroscopic studies of laser-induced plasma during pulsed irradiation of graphite targets in nitrogen and ammonia. *Journal of Physics B: Atomic, Molecular and Optical Physics, 30*(19), 4405–4414. https://doi.org/10.1088/0953-4075/30/19/026

Andrade, D. F., De Almeida, E., De Carvalho, H. W., Pereira-Filho, E. R., & Amarasiriwardena, D. (2021). Chemical inspection and elemental analysis of electronic waste using data fusion - application of complementary spectroanalytical techniques. *Talanta, 225*, 122025. https://doi.org/10.1016/j.talanta.2020.122025

Aparicio, R., Morales, M. T., Aparicio-Ruiz, R., Tena, N., & García-González, D. L. (2013). Authenticity of olive oil: Mapping and comparing official methods and promising alternatives. *Food Research International, 54*(2), 2025–2038. https://doi.org/10.1016/j.foodres.2013.07.039

Bellou, E., Gyftokostas, N., Stefas, D., Gazeli, O., & Couris, S. (2020). Laser-induced breakdown spectroscopy assisted by machine learning for olive oils classification: The effect of the experimental parameters. *Spectrochimica Acta Part B: Atomic Spectroscopy, 163*, 105746. https://doi.org/10.1016/j.sab.2019.105746

Berghian-Grosan, C., & Magdas, D. A. (2020). Raman spectroscopy and machine-learning for edible oils evaluation. *Talanta, 218*, 121176. https://doi.org/10.1016/j.talanta.2020.121176

Binetti, G., Del Coco, L., Ragone, R., Zelasco, S., Perri, E., Montemurro, C., et al. (2017). Cultivar classification of Apulian olive oils: Use of artificial neural networks for comparing NMR, NIR and merceological data. *Food Chemistry, 219*, 131–138. https://doi.org/10.1016/j.foodchem.2016.09.041

Borràs, E., Ferré, J., Boqué, R., Mestres, M., Aceña, L., & Busto, O. (2015). Data fusion methodologies for food and beverage authentication and quality assessment – a review. *Analytica Chimica Acta, 891*, 1–14. https://doi.org/10.1016/j.aca.2015.04.042

Boskou, D. (2015). *Olive oil: Chemistry and Technology*. Elsevier Science.

Caceres, J. O., Moncayo, S., Rosales, J. D., de Villena, F. J., Alvira, F. C., & Bilmes, G. M. (2013). Application of laser-induced breakdown spectroscopy (LIBS) and neural networks to olive oils analysis. *Applied Spectroscopy, 67*(9), 1064–1072. https://doi.org/10.1366/12-06916

Casale, M., Sinelli, N., Oliveri, P., Di Egidio, V., & Lanteri, S. (2010). Chemometrical strategies for feature selection and data compression applied to NIR and MIR spectra of extra virgin olive oils for cultivar identification. *Talanta, 80*(5), 1832–1837. https://doi.org/10.1016/j.talanta.2009.10.030

Conte, L., Bendini, A., Valli, E., Lucci, P., Moret, S., Maquet, A., et al. (2020). Olive oil quality and authenticity: A review of current eu legislation, standards, relevant methods of analyses, their drawbacks and recommendations for the future. *Trends in Food Science & Technology, 105*, 483–493. https://doi.org/10.1016/j.tifs.2019.02.025

Domenici, V., Ancora, D., Cifelli, M., Serani, A., Veracini, C. A., & Zandomeneghi, M. (2014). Extraction of pigment information from near-UV vis absorption spectra of extra virgin olive oils. *Journal of Agricultural and Food Chemistry, 62*(38), 9317–9325. https://doi.org/10.1021/jf503818k

Fortes, F. J., Moros, J., Lucena, P., Cabalín, L. M., & Laserna, J. J. (2012). Laser-Induced breakdown spectroscopy. *Analytical Chemistry, 85*(2), 640–669. https://doi.org/10.1021/ac303220r

François, E., Gazeli, O., Couris, S., Angelopoulos, G. N., Blanpain, B., & Malfliet, A. (2020). Laser-induced breakdown spectroscopy analysis of the free surface of liquid secondary copper slag. *Spectrochimica Acta Part B: Atomic Spectroscopy, 170*, 105921. https://doi.org/10.1016/j.sab.2020.105921

Gamela, R. R., Costa, V. C., Sperança, M. A., & Pereira-Filho, E. R. (2020). Laser-induced breakdown spectroscopy (LIBS) and wavelength dispersive X-ray fluorescence (WDXRF) data fusion to predict the concentration of K, Mg and P in bean seed samples. *Food Research International, 132*, 109037. https://doi.org/10.1016/j.foodres.2020.109037

Gazeli, O., Bellou, E., Stefas, D., & Couris, S. (2020). Laser-based classification of olive oils assisted by machine learning. *Food Chemistry, 302*, 125329. https://doi.org/10.1016/j.foodchem.2019.125329

Gyftokostas, N., Nanou, E., Stefas, D., Kokkinos, V., Bouras, C., & Couris, S. (2021). Classification of Greek olive oils from different regions by machine learning-aided laser-induced breakdown spectroscopy and absorption spectroscopy. *Molecules, 26*, 1241. https://doi.org/10.3390/molecules26051241

Gyftokostas, N., Stefas, D., & Couris, S. (2020). Olive oils classification via laser-induced breakdown spectroscopy. *Applied Sciences, 10*(10), 3462. https://doi.org/10.3390/app10103462

Hastie, T., Friedman, J., & Tisbshirani, R. (2017). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.

Huffman, C., Sobral, H., & Terán-Hinojosa, E. (2019). Laser-induced breakdown spectroscopy spectral feature selection to enhance classification capabilities: A t-test filter approach. *Spectrochimica Acta Part B: Atomic Spectroscopy, 162*, 105721. https://doi.org/10.1016/j.sab.2019.105721

Jimenez-Lopez, C., Carpena, M., Lourenço-Lopes, C., Gallardo-Gomez, M., Lorenzo, J. M., Barba, F. J., et al. (2020). Bioactive compounds and quality of extra virgin olive oil. *Foods, 9*(8), 1014. https://doi.org/10.3390/foods9081014

Képeš, E., Pořízka, P., & Kaiser, J. (2019). On the application of bootstrapping to laser-induced breakdown spectroscopy data. *Journal of Analytical Atomic Spectrometry, 34*(12), 2411–2419. https://doi.org/10.1039/c9ja00304e

Képeš, E., Vrábel, J., Střítežská, S., Pořízka, P., & Kaiser, J. (2020). Benchmark classification dataset for laser-induced breakdown spectroscopy. *Scientific Data, 7*(1). https://doi.org/10.1038/s41597-020-0396-8

Kontzedaki, R., Orfanakis, E., Sofra-Karanti, G., Stamataki, K., Philippidis, A., Zoumi, A., et al. (2020). Verifying the geographical origin and authenticity of Greek olive oils by means of optical spectroscopy and multivariate analysis. *Molecules, 25*(18), 4180. https://doi.org/10.3390/molecules25184180

Kosma, I., Vatavali, K., Kontakos, S., Kontominas, M., Kiritsakis, A., & Badeka, A. (2017). Geographical differentiation of Greek extra virgin olive oil from late-harvested Koroneiki cultivar fruits. *JAOCS. Journal of the American Oil Chemists' Society, 94*(11), 1373–1384. https://doi.org/10.1007/s11746-017-3036-5

Kramida, A., Ralchenko, Y., Reader, J., & NIST ASD Team. (2019). *NIST atomic spectra database (version 5.7.1)*. Retrieved from https://www.nist.gov/pml/atomic-spectra-database.

Kružlicová, D., Mocak, J., Katsoyannos, E., & Lankmayr, E. (2008). Classification and characterization of olive oils by UV-Vis absorption spectrometry and sensorial analysis. *Journal of Food & Nutrition Research, 47*(4), 181–188.

Maléchaux, A., Laroussi-Mezghani, S., Le Dréau, Y., Artaud, J., & Dupuy, N. (2020). Multiblock chemometrics for the discrimination of three extra virgin olive oil varieties. *Food Chemistry, 309*, 125588. https://doi.org/10.1016/j.foodchem.2019.125588

Milanez, K. D. T. M., Nóbrega, T. C. A., Nascimento, D. S., Insausti, M., Band, B. S. F., & Pontes, M. J. C. (2017). Multivariate modeling for detecting adulteration of extra virgin olive oil with soybean oil using fluorescence and UV–vis spectroscopies: A preliminary approach. *Lebensmittel-Wissenschaft & Technologie, 85*, 9–15. https://doi.org/10.1016/j.lwt.2017.06.060

Moros, J., & Laserna, J. (2019). Laser-Induced breakdown spectroscopy (LIBS) of organic compounds: A review. *Applied Spectroscopy, 73*(9), 963–1011. https://doi.org/10.1177/0003702819853252

Parigger, C. G. (2013). Atomic and molecular emissions in laser-induced breakdown spectroscopy. *Spectrochimica Acta Part B: Atomic Spectroscopy, 79–80*, 4–16. https://doi.org/10.1016/j.sab.2012.11.012

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Peng, J., He, Y., Jiang, J., Zhao, Z., Zhou, F., & Liu, F. (2019). High-accuracy and fast determination of chromium content in rice leaves based on collinear dual-pulse laser-induced breakdown spectroscopy and chemometric methods. *Food Chemistry, 295*, 327–333. https://doi.org/10.1016/j.foodchem.2019.05.119

Prochazka, D., Mazura, M., Samek, O., Rebrošová, K., Pořízka, P., Klus, J., et al. (2018). Combination of laser-induced breakdown spectroscopy and Raman spectroscopy for multivariate classification of bacteria. *Spectrochimica Acta Part B: Atomic Spectroscopy, 139*, 6–12. https://doi.org/10.1016/j.sab.2017.11.004

Senesi, G. S., Cabral, J., Menegatti, C. R., Marangoni, B., & Nicolodelli, G. (2019). Recent advances and future trends in LIBS applications to agricultural materials and their FOOD DERIVATIVES: An overview of developments in the last DECADE (2010–2019). Part II. crop plants and their food derivatives. *TRAC Trends in Analytical Chemistry, 118*, 453–469. https://doi.org/10.1016/j.trac.2019.05.052

Stefas, D., Gyftokostas, N., & Couris, S. (2020). Laser induced breakdown spectroscopy for elemental analysis and discrimination of honey samples. *Spectrochimica Acta Part B: Atomic Spectroscopy, 172*, Article 105969. https://doi.org/10.1016/j.sab.2020.105969

Valand, R., Tanna, S., Lawson, G., & Bengtström, L. (2019). A review of Fourier Transform Infrared (FTIR) spectroscopy used in food adulteration and authenticity investigations. *Food Additives & Contaminants: Part A, 37*(1), 19–38. https://doi.org/10.1080/19440049.2019.1675909

Velásquez-Ferrín, A., Babos, D. V., Marina-Montes, C., & Anzano, J. (2020). Rapidly growing trends in laser-induced breakdown spectroscopy for food analysis. *Applied Spectroscopy Reviews*, 1–21. https://doi.org/10.1080/05704928.2020.1810060

Violino, S., Ortenzi, L., Antonucci, F., Pallottino, F., Benincasa, C., Figorilli, S., et al. (2020). An artificial intelligence approach for Italian EVOO origin traceability through an open source IoT spectrometer. *Foods, 9*(6), 834. https://doi.org/10.3390/foods9060834

Vrábel, J., Képeš, E., Duponchel, L., Motto-Ros, V., Fabre, C., Connemann, S., et al. (2020). Classification of challenging Laser-Induced Breakdown Spectroscopy soil sample data - EMSLIBS contest. *Spectrochimica Acta Part B: Atomic Spectroscopy, 169*, 105872. https://doi.org/10.1016/j.sab.2020.105872

Yang, Y., Hao, X., Zhang, L., & Ren, L. (2020). Application of scikit and keras libraries for the classification of iron ore data acquired by laser-induced breakdown spectroscopy (LIBS). *Sensors, 20*(5), 1393. https://doi.org/10.3390/s20051393

Yu, K., Ren, J., & Zhao, Y. (2020). Principles, developments and applications of laser-induced breakdown spectroscopy in agriculture: A review. *Artificial Intelligence in Agriculture, 4*, 127–139. https://doi.org/10.1016/j.aiia.2020.07.001

Zhao, M., Markiewicz-Keszycka, M., Beattie, R. J., Casado-Gavalda, M. P., Cama-Moncunill, X., O'Donnell, C. P., et al. (2020). Quantification of calcium in infant formula using laser-induced breakdown spectroscopy (LIBS), Fourier transform mid-infrared (FT-IR) and Raman spectroscopy combined with chemometrics including data fusion. *Food Chemistry, 320*, 126639. https://doi.org/10.1016/j.foodchem.2020.126639