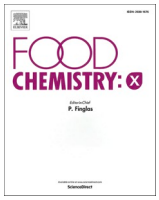
Food Chemistry: X 20 (2023) 100902

Contents lists available at ScienceDirect

Food Chemistry: X

journal homepage: www.sciencedirect.com/journal/food-chemistry-x



Geographical and varietal origin differentiation of alcoholic beverages through the association between FT-Raman spectroscopy and advanced data processing strategies

Ariana Raluca Hategan [a,b], Maria David [a,b], Camelia Berghian-Grosan [a], Dana Alina Magdas [a,b,*]

[a] National Institute for Research and Development of Isotopic and Molecular Technologies, 67-103 Donat Street, 400293 Cluj-Napoca, Romania
[b] Faculty of Physics, Babeș-Bolyai University, Kogălniceanu 1, 400084 Cluj-Napoca, Romania

ARTICLE INFO

Keywords:
Fruit spirits
SVM
PLS-DA
Raman spectroscopy
Recognition models

ABSTRACT

The present work aimed to test the efficiency of FT-Raman spectroscopy for fruit spirits discrimination by developing differentiation models based on two approaches, namely a supervised statistical method (Partial Least Squares Discriminant Analysis), and a Machine Learning technique (Support Vector Machines). For this purpose, a data set comprising 86 Romanian distillate samples was used, which aimed to be differentiated in terms of the raw material used for production (plum, apple, pear and grape) and county of origin (Cluj, Satu Mare and Salaj). Eight distinct preprocessing methods (autoscale, mean center, variance scaling, smoothing, 1st derivative, 2nd derivative, standard normal variate and Pareto) followed by a feature selection step were applied to identify the meaningful input data based on which the most efficient classification models can be constructed. Both types of models led to accuracy scores greater than 90% in differentiating the distillate samples in terms of geographical and botanical origin.

1. Introduction

Fruit spirits represent, especially in Eastern Europe, a highly appreciated product, being related to tradition, as their production 'secret' is left as a legacy from father to son. The manufacturing process is directly reflected in the final personality of this type of alcoholic beverage and therefore, brand protection became a very important issue to be addressed (Bauer-Christoph, Wachter, Christoph, Roßmann & Adam, 1997). Authentic fruit spirits are alcoholic drinks with a high commercial value because of the important costs that are involved in their production process. Because of the high market price, such alcoholic beverages are more prone to falsification with a view to making an illicit economical gain. In this regard, common falsification techniques include the wrong declaration of the raw material's botanical or geographical origin. For this reason, the development of analytical tools able to differentiate the fruit distillates, according to some predefined criteria as a function of the classification purpose, is necessary.

The chemical composition of fruit spirits consists mainly of water and ethanol (Dolenko et al., 2015). However, an important component of fruit distillates is represented by volatile compounds, which offer to spirits the taste and aroma and consist of tannic and polyphenolic substances, aromatic acids, nitrogen- and sulphur-containing compounds, hydrocarbons, unfermented sugars, di- and tribasic carboxylic acids (Coldea, Socaciu, Fetea, Ranga, Pop & Florea, 2013; Mangas, Rodríguez, Moreno, Suárez & Blanco, 1996).

For a standard spirit drink quality analysis, higher alcohols and other volatile compounds are determined using gas chromatography (Ledauphin et al., 2004). Other analytical approaches used and reported in the literature for checking the authenticity of fruit spirits were: i) the determination of $^{13}C/^{12}C$ isotope ratios (Winterova, Mikulikova, Mazáč & Havelec, 2008); ii) stable isotopes of light elements and mineral content (Magdas, Cristea, Pîrnau, Feher, Hategan & Dehelean, 2021), or iii) HPLC study of the phenolic acids (Coldea et al., 2013; Mangas et al., 1996). However, these methods are relatively expensive, time-consuming and require highly skilled operators.

In this light, vibrational spectroscopy became more and more applied in food and beverage differentiation studies, being a nondestructive technique, with a fast response time. Vibrational methods were also

* Corresponding author at: National Institute for Research and Development of Isotopic and Molecular Technologies, 67-103 Donat Street, 400293 Cluj-Napoca, Romania.
E-mail addresses: ariana.hategan@itim-cj.ro (A.R. Hategan), maria.david@itim-cj.ro (M. David), camelia.grosan@itim-cj.ro (C. Berghian-Grosan), alina.magdas@itim-cj.ro (D.A. Magdas).

https://doi.org/10.1016/j.fochx.2023.100902
Received 20 February 2023; Received in revised form 7 September 2023; Accepted 23 September 2023
Available online 25 September 2023

conducted for quantitative studies of some components such as the ethanol and methanol of the fruit distillates (Ellis, Muhamadali, Xu, Eccles, Goodall & Goodacre, 2019; Mendes, Oliveira, Suarez & Rubim, 2003; Šrámek, Švancara & Sýs, 2019; Vaskova, 2014), but a very limited number of studies aimed the botanical or geographical differentiation of these beverages based on metabolomics (Berghian-Grosan & Magdas, 2020; Magdas, David & Berghian-Grosan, 2020). Raman spectroscopy is a more suitable technique as compared to infrared (IR) spectroscopy for matrices having a high water content, like fruit spirits, mainly due to the weak water bending mode in the fingerprint region (Berghian-Grosan & Magdas, 2020; Magdas, David, & Berghian-Grosan, 2020). For the last purpose, vibrational spectroscopy in corroboration with supervised statistical methods, such as Partial Least Squares Discriminant Analysis (PLS-DA), Linear Discriminant Analysis or Machine Learning tools proved to be very effective for both food and beverage products differentiation with respect to the geographical origin of the raw materials and to the fruit or botanical variety (David, Hategan, Berghian-Grosan, & Magdas, 2022; Magdas, Guyon, Feher & Cinta Pinzaru, 2018; Magdas, Cozar, Feher, Guyon, Dehelean & Cinta Pinzaru, 2019; Magdas et al., 2020; Magdas, Guyon, Berghian-Grosan, & Muller Molnar, 2021; Parastar, van Kollenburg, Weesepoel, van den Doel, Buydens, & Jansen, 2020; Tsakanikas, Karnavas, Panagou & Nychas, 2020).

This study reports the development of new recognition models capable of discriminating fruit spirits with respect to the geographical origin and the type of raw material used for distillate production, based on statistical methods and Artificial Intelligence algorithms. With the aim of performing a step forward in relation to our previous work (Berghian-Grosan & Magdas, 2020; Magdas, David & Berghian-Grosan, 2020), special attention was given to identifying reliable ways to improve the classification ability of the prediction models through effective preprocessing methods. The assumption was that an adequate data preprocessing strategy allowing the construction of recognition models based on the variables having statistical significance leads to the obtainment of differentiation models with a considerably improved performance. The practical importance of the proposed approach is related to the application of a rapid, cost-effective, and easy-to-use tool for assuring quality standards in the case of alcoholic beverages. Therefore, the development of reliable data processing approaches that can be incorporated into the software system of already existing portable equipment can allow the on-site control of beverages even by non-specialists.

2. Materials and methods

2.1. Sample description

The study was based on a sample set consisting of 86 distillate samples, whose distribution in terms of the raw material used for production included the following varieties: plum (48), apple (11), pear (6), grape (5), fruit mixture (5), cherry (3), quince (3), sour cherry (2), beer (1), apricot (1) and blackcurrant (1). All samples were produced in Romania: 79 distillates originated from Transylvania, while 7 samples had as geographical origin other Romanian regions. The highest representativeness among the Transylvanian samples corresponded to the following counties: Satu Mare (23), Cluj (21) and Salaj (19).

2.2. FT-Raman measurements

The samples were investigated using a Bruker Equinox 55 Fourier transform Raman (FT-Raman) spectrometer, equipped with an integrated FRA 106S Raman module. The laser was employed to emit at 1064 nm, with an output power of 350 mW. The Ge detector used was cooled with liquid nitrogen. The spectral range was chosen to comprise both Stokes and anti-Stokes regions and was situated between − 1000 and 3600 cm^{-1}. 450 scans were collected for each sample.

2.3. Model development

2.3.1. Data preprocessing

The data preprocessing workflow consisted of two main steps. The first one was represented by the application of eight preprocessing methods for determining the most suitable way of transforming the raw data acquired through Raman spectroscopy, namely autoscale, mean center, variance scaling, Savitzky–Golay smoothing using first order polynomial approximation, Savitzky–Golay 1^{st} derivative using second order polynomial approximation, Savitzky–Golay 2^{nd} derivative using second order polynomial approximation, standard normal variate and Pareto. For all Savitzky–Golay filters, a filter width of 15 was applied. The second step illustrated the use of a variable selection technique for identifying the spectral points that have the highest discrimination power with respect to a certain classification criterion (i.e. botanical or geographical). In this, regard, a PLS-based feature selection method that takes into account variable importance in projection (VIP) and selectivity ratio (SR) scores was applied. This method determines, in an iterative manner, the subgroup of variables that conduct to the lowest RMSECV (root mean square error of cross-validation) value (Eigenvector Research Inc, 2023, https://wiki.eigenvector.com/index.php?title=Selectvars).

2.3.2. Partial Least Squares Discriminant Analysis (PLS-DA)

Partial Least Squares Discriminant Analysis (PLS-DA) is a supervised statistical method that relies on the Partial Least Squares (PLS) regression and attempts the maximization of the covariance between the matrix X representing the independent variables (e.g. the acquired spectra) and matrix Y containing the corresponding dependent variable (e.g. the classes aimed to be differentiated) of multidimensional data. This is achieved by identifying a linear subspace of variables that allows predicting Y based on a reduced number of components, also called latent variables (LVs) (Gromski et al., 2015). In this study, the number of LVs was chosen such that the cross-validation classification error average score was minimized.

All statistical treatments were conducted under the software SOLO 8.9.1 (2021) (Eigenvector Research Inc., 2022 Manson, WA, USA).

2.3.3. Support Vector Machines (SVM)

The second category of fruit distillates classifiers was represented by Support Vector Machines (SVM) models, which were implemented by means of the *svm* module available in the scikit-learn library (Pedregosa et al., 2011). Three types of SVM kernels were investigated, namely linear, polynomial and radial basis function (RBF). For each of these kernels, the tested values for the C parameter were: 2^{-5}, 2^{-4}, …, 2^{15}. Moreover, the experimented values for the degree of the polynomial kernel were: 1, 2, …, 10, and the search space of the gamma parameter associated with the RBF kernel was: 2^{-15}, 2^{-14}, …, 2^{3}. All these hyperparameters were optimized using the *model_selection.GridSearchCV* class. In this regard, 10-fold cross-validation was applied and the accuracy score was used to evaluate the performance of the models.

2.3.4. Evaluation

For evaluating and comparing the developed models, constructed either through PLS-DA or SVM, the Venetian Blinds cross-validation method (Eigenvector Research, Inc, 2023, https://wiki.eigenvector.com/index.php?title=Using_Cross-Validation) was applied, and the number of data splits was set to ten.

3. Results

The present study prospected the potential given by Raman spectroscopy for the development of alcoholic beverage recognition models aimed to classify fruit distillates with respect to their geographical and botanical origin. In order to improve the classification abilities of the constructed models, eight distinct preprocessing methods (i.e. autoscale,

mean center, variance scaling, smoothing, 1st derivative, 2nd derivative, standard normal variate and Pareto) were applied to the experimental FT-Raman spectra and further compared by means of PLS-DA in order to identify the most efficient approach for transforming the raw spectra into input data for model development. In this regard, Figures S1 and S2 illustrate an example of a raw FT-Raman spectrum and the corresponding preprocessed spectra, obtained by applying each of the eight pretreatment techniques. Corresponding to each type of classification and preprocessing technique, a feature selection step was conducted to identify the spectral points that have a higher discrimination power. Therefore, the input data for constructing the final differentiation models (i.e. based on either PLS-DA or SVM) corresponded to the variables that were transformed through the most suitable preprocessing method and that were subsequently selected as relevant markers for a certain classification.

The main signals that dominate the Raman sample spectra, as it can be seen in Fig. S1, are located at 884, 1056, 1456, 2713–2780, 2883, 2932, 2976 and 3100–3400 cm^{-1} and can be attributed to the presence of the main fruit spirit distillates componence: ethanol and water (Magdas et al., 2020; Spaho, 2017; da Silva et. al., 2019). The bands at 884 and 1056 cm^{-1} are present due to the stretching vibration of C–C and C–O, respectively. The band at 1456 cm^{-1} can be assigned to the bending vibration of CH_2 and CH_3 from ethanol. The broad band between 2713 and 2780 cm^{-1} comes from a combination of frequencies, while the bands at 2883, 2932, 2976 cm^{-1} are attributed to the stretching of the symmetric vibration of CH_2 and CH_3. The broad band at 3100–3400 cm^{-1} is assigned to the stretching vibrations of the O–H groups (Magdas et al., 2020; Spaho, 2017; da Silva et al., 2019; Socrates, 2001).

There are a few weaker bands that can be observed between 1560 and 1770 cm^{-1} that are due to the C=C and C=O stretching vibration from the minor concentration components such as esters. These compounds are mostly responsible for the flowery and fruity aroma of the distillates. Ethyl acetate is the most common ester present in fruit distillates, while isoamyl acetate, isobutyl acetate and hexanoate (present especially in apple distillates) are in lower concentrations (Spaho, 2017). Other minor components that can present a signal in this region are: carbonyl compounds such as acetaldehyde, or low concentrations of isobutyraldehyde, 2-propenal (acrolein); benzaldehyde (especially present in the plum, cherry and apricot spirits samples) or acetic acid (Spaho, 2017).

Weaker signals at 200–700 cm^{-1} can be assigned to the metal oxide vibrations, Cu-O, Fe-O, Zn-O, Mn-O, Co-O, whose concentrations depend on the floral origin (Barai, Banerjee, & Joo, 2017; Magdas, David, & Berghian-Grosan, 2020; Rashad, Rüsing, Berth, Lischka, & Pawlis, 2013; Santillan et al., 2017)).

3.1. Geographical differentiation

The geographical recognition models aimed the differentiation of the fruit distillates according to three Transylvanian counties of origin: Cluj (21 samples), Salaj (19 samples) and Satu Mare (23 samples).

The first phase of the model development workflow corresponded to the application of PLS-DA for the assessment of distinct preprocessing techniques. As shown in Table 1, when the entire Raman spectra were used as input data for the development of the PLS-DA models, the highest accuracy score corresponded to smoothing preprocessing, namely 61 % of the samples were correctly attributed to the geographical class in the cross-validation procedure. The model identified the right county of origin for 78 % of the samples originating from Satu Mare, while modest true positive rates were obtained for the group of distillates produced in Cluj and Salaj. In order to improve the classification abilities of the constructed differentiation models, a feature selection step was further performed. The main aim of this step was to reduce the variables based on which the model is further constructed only to those meaningful for the geographical origin differentiation.

Table 1
Performance of the PLS-DA geographical discrimination models, before and after feature selection, as function of the type of preprocessing applied.

Preprocessing	Number of variables	True Positive Rate (Cross-validation)			Cross-validation accuracy
		Cluj (21)	Salaj (19)	Satu Mare (23)	
Autoscale	2386	0.66	0.15	0.60	0.49
Mean Center	2386	0.47	0.36	0.60	0.49
Variance Scaling	2386	0.61	0.05	0.60	0.44
Smoothing	2386	0.57	0.47	0.78	0.61
1st derivative	2386	0.47	0.57	0.65	0.57
2nd derivative	2386	0.66	0.47	0.34	0.49
Standard Normal Variate	2386	0.42	0.42	0.56	0.47
Pareto	2386	0.42	0.73	0.47	0.53
Feature Selection					
Autoscale	257	0.90	0.89	1.00	0.93
Mean Center	150	1.00	1.00	1.00	1.00
Variance Scaling	309	0.85	0.84	0.95	0.88
Smoothing	782	0.71	0.73	0.91	0.79
1st derivative	1312	0.47	0.73	0.86	0.69
2nd derivative	469	0.61	0.78	0.78	0.73
Standard Normal Variate	321	0.95	1.00	0.95	0.96
Pareto	245	0.95	1.00	1.00	0.98

Thus, as it can be seen in Table 1, conducting a model-based feature selection step prior to model development substantially improved the performance of the PLS-DA classifiers. The application of different preprocessing techniques implied the obtainment of distinct sets of significant variables and, therefore, all classification models built on a reduced data set, as function of the type of preprocessing, illustrated an increase in accuracy, between 12 % (i.e. in the case when the 1st derivative was used to preprocess the data) and 51 % (i.e. when the data was mean centered). The highest accuracy was acquired when the lowest number of features selected for the geographical differentiation was employed as input data (i.e. 150 spectral points, corresponding to mean center preprocessing), while the lowest performance resulted from the model built on the data with the highest dimensionality (i.e. 1312 attributes, corresponding to the 1st derivative preprocessing). Thus, the utilization of smoothing, 1st derivative and 2nd derivative as preprocessing methods generated the lowest accuracy improvement after the feature selection step (i.e. between 12 % and 24 %), while this dimensionality reduction method proved to be the most effective when it was applied on the mean cantered data (100 % accuracy).

In accordance with the obtained results, the input data for the development of the final geographical differentiation models corresponded to 150 Raman variables (Fig. 1) that were preprocessed through mean centering. As can be seen in Table 1, the PLS-DA model constructed based on this input data allowed a perfect discrimination of the distillate samples with respect to the geographical origin. Thus, the model, defined by means of the first 17 LVs, conducted to a 100 % accuracy in the cross-validation evaluation procedure (Fig. 2).

The potential of these 150 Raman features preprocessed through mean centering for the geographical discrimination of the fruit distillates was also highlighted by the performance of the SVM model. In this regard, the ML classifier correctly identified the county of origin for 57 of the distillate samples, leading to an accuracy score of 90 %. This performance resulted from the fact that during cross-validation, one sample from Salaj was predicted as being from Cluj, two samples from Satu Mare were attributed to the Salaj group, and three samples from Cluj were wrongly assigned to Salaj (one sample) and Satu Mare (two samples) counties. An interesting aspect is reflected by the fact that among these wrongly classified instances were the samples indexed 38, 39 and 59, which presented the lowest Y predicted values in cross-validation for their actual class (Fig. 2). The SVM model was characterized by a RBF kernel, a C parameter of 2^{13}, and a gamma parameter of

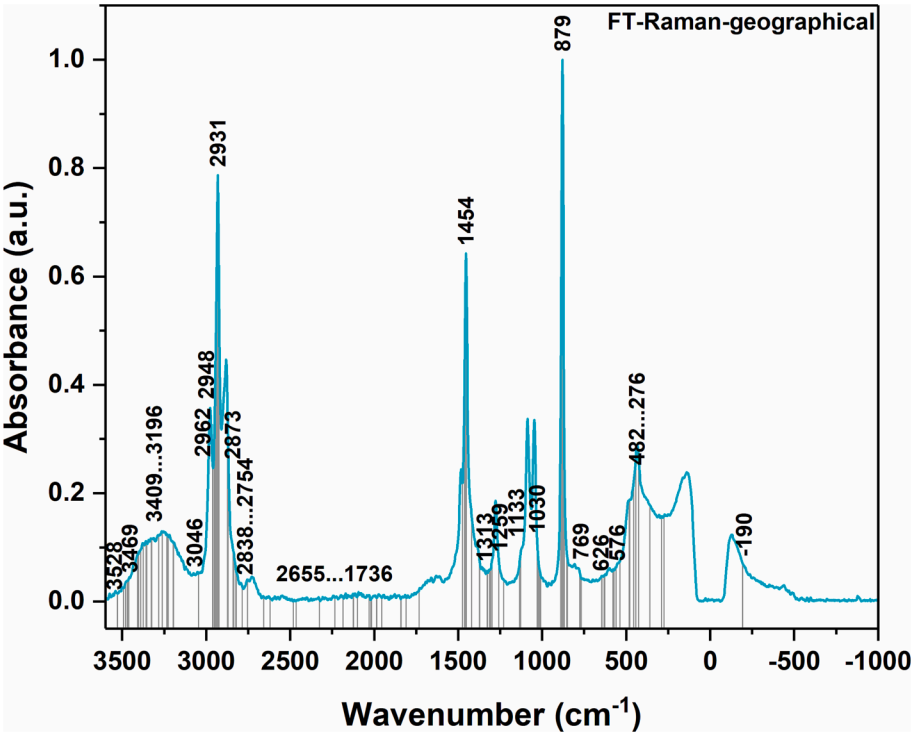


Fig. 1. Raman spectrum containing the most important predictors used for geographical differentiation of fruit distillates.

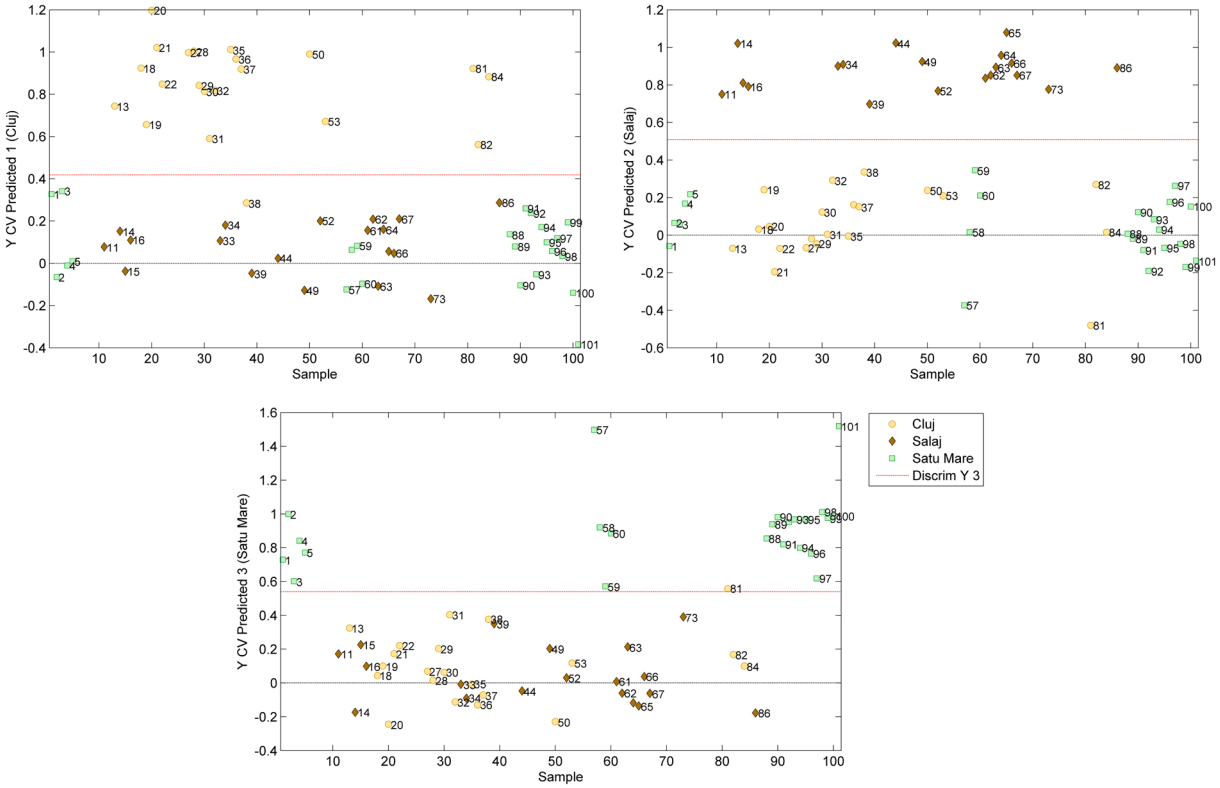


Fig. 2. Y cross-validation predicted values by the PLS-DA model developed for the geographical classification of the fruit distillates. The input data is represented by 150 Raman variables preprocessed through the mean center.

2^{-4}.

Based on the obtained results, it can be stated that the set of attributes having the highest relevance for the geographical differentiation of the distillates corresponds to the 150 spectral points obtained through the feature selection method applied to the mean centered data (Fig. 1). The differentiation markers that allowed the geographical discrimination of fruit distillates samples are located all over the spectral domain; their assignments are identified in accord with the data from the work of

Socrates, 2001. Thus, while the bands from 3528 and 3469 cm^{-1} are associated with the O–H stretching vibrations specific to *ortho*-substituted phenols, those from 3409 to 3196 cm^{-1} could be mainly due to the hydrogen-bonded O–H intermolecular vibrations from water, alcohols, wood-derived carbohydrates – originated from the cask-aged fruit spirits (Buglass, 2014) – or phenols; 3046 cm^{-1} could be the result of the asymmetric vibration of $–CH_3$ from methyl esters; the peaks from 2962, 2948 and 2931 cm^{-1}, which are characteristic to $–CH_3$ and $–CH_2$ stretching modes could be correlated with the ethyl esters, as well as with the $–CH_3$ stretching vibration from 2873 cm^{-1}; the bands from 2838 to 2754 cm^{-1} could be associated to the –C–H stretching vibration of aldehydes and those from 2655 to 1736 cm^{-1} to the stretching modes of some compounds containing the nitrile group or the hydrogen cyanide derivatives. In the region below 1500 cm^{-1}, some significant contributions of the bands around 1454 cm^{-1}, due to the $–CH_2$ symmetric deformation vibrations of acyclic esters, can be pointed out; the O-CH_3 stretching vibrations of the alkyl ester generally appear in the region 1315–1195 cm^{-1}, so the bands from about 1313 and 1259 cm^{-1} could be associated with various alkyl esters, whereas the peak from 1313 cm^{-1} could be also due to the combination between the O–H deformation and C–O stretching modes of phenols that can be detected at about 1313 cm^{-1}; in the same time, the bands from 1259 and 1133 cm^{-1} could be the result of the asymmetric and symmetric stretching vibrations of the C–O–C group from aromatic esters, formed during the maturation process. The peaks from around 1030 cm^{-1} could be associated with the C–O stretching vibrations of aromatic alcohols or of the carbohydrates formed during the aging in wooden barrels, while those from around 879 cm^{-1} to the C–C–O group stretching vibrations of primary or secondary alcohols. The aromatic =C–H out-of-plane deformation vibrations appeared around 769 cm^{-1}, whereas the 626 and 576 cm^{-1} peaks are characteristic of the aromatic ring deformation vibrations; different interaction between organic molecules and metals and their appropriate M−C, M−N or M−O stretching vibrations could be revealed in the region 482–276 cm^{-1}.

3.2. Varietal differentiation

The varietal discrimination models aimed to classify the fruit distillates according to the type of raw material used for production. In this regard, the fruit spirits having as botanical variety plum (48 samples), apple (11 samples), pear (6 samples) and grape (5 samples) were included for constructing and validating PLS-DA models. Similar to the data processing workflow conducted for the geographical differentiation, the first step consisted in the application of eight distinct methods for preprocessing the Raman spectra. The impact of each pretreatment was illustrated through the classification performance of the PLS-DA models having as input data the Raman spectra preprocessed through that preprocessing method (Table 2). It can be observed that the highest accuracy score corresponds to the data transformed through the 2nd derivative, namely 67 % of the samples were correctly predicted. However, in this case, the model had a very low ability in determining the botanical source of the distillates that belong to classes represented by a smaller number of samples (i.e. apple, pear and grape). Based on these results, it can be stated that no matter what preprocessing approach was applied, the models developed on the entire Raman spectral range did not lead to reliable results. When the input space was reduced to the variables having the highest discrimination power, determined with respect to each pretreatment, the performance of the PLS-DA models constructed based on these sets of significant attributes increased. Therefore, accuracy scores ranging between 67 % and 95 % were obtained, and the most efficient model proved to be the one developed based on the lowest number of variables, namely 133 spectral points. This is a total agreement with the results obtained for the geographical differentiation of the fruit distillates when a 100 % model accuracy was achieved only for the input space having the lowest dimensionality (i.e. 150 data points). Moreover, the preprocessing methods that proved to be the most efficient ones for the geographical discrimination corresponded to Pareto, standard normal variate, mean centering and variance scaling, leading to accuracies of 95 %, 94 %, 91 % and 90 % respectively. On the contrary, the feature selection step led to a moderate effect in terms of model accuracy when this treatment was applied to the experimental data set preprocessed prior to model development through autoscaling, smoothing, 1st derivative and 2nd derivative.

Table 2
PLS-DA model performance in predicting the type of raw material used for distillate production, as function of the preprocessing method and the type of input data (i.e. entire spectra / significant variables).

Preprocessing	Number of variables	True Positive Rate (Cross-validation)				Cross-validation accuracy
		plum (48)	apple (11)	pear (6)	grape (5)	
Autoscale	2386	0.89	0.00	0.16	0.00	0.62
Mean Center	2386	0.50	0.18	0.33	0.00	0.40
Variance Scaling	2386	0.93	0.00	0.00	0.00	0.64
Smoothing	2386	0.81	0.18	0.00	0.20	0.60
1st derivative	2386	0.75	0.00	0.00	0.20	0.52
2nd derivative	2386	0.91	0.18	0.16	0.00	0.67
Standard Normal Variate	2386	0.75	0.18	0.66	0.00	0.60
Pareto	2386	0.58	0.27	0.16	0.00	0.45
Feature Selection						
Autoscale	397	0.89	0.54	0.33	0.00	0.72
Mean Center	281	0.97	0.63	1.00	0.80	0.91
Variance Scaling	218	0.95	0.81	0.66	0.80	0.90
Smoothing	218	0.75	0.54	0.50	0.40	0.67
1st derivative	722	0.91	0.54	0.33	0.40	0.77
2nd derivative	782	0.93	0.45	0.33	0.00	0.74
Standard Normal Variate	178	1.00	0.63	1.00	1.00	0.94
Pareto	133	0.97	0.90	1.00	0.80	0.95

As opposed to the botanical differentiation models built on the entire spectral range, the PLS-DA classifiers that had as input data only the relevant markers obtained when (i) Pareto, (ii) standard normal variate, (iii) mean centering and (iv) variance scaling were used as preprocessing methods proved to have a higher ability in correctly predicting the samples produced from apples, pears and grapes. Therefore, even though the number of fruit distillates produced from these raw materials was not as high as the one of plum spirits, these PLS-DA models succeeded in offering reliable predictions for the botanical origin of the samples, leading to true positive rates greater than 63 %.

Based on these results, the chosen input data for constructing the final varietal prediction models corresponded to 133 Raman features (Fig. 3) preprocessed through Pareto. Therefore, as previously presented in Table 2, the PLS-DA model developed on this data led to a 95 % accuracy score during cross-validation, namely 67 out of 70 samples were correctly classified. The true positive rates associated with the plum, apple, pear, and grape classes were 97 %, 90 %, 100 %, and 80 %, respectively. Nonetheless, the number of LVs used for this model was set to 12, as it corresponded to the lowest cross-validation classification error average. The PLS-DA plots illustrating the Y predicted values for each sample and each varietal class are presented in Fig. 4.

The SVM model built on the chosen input data (i.e. the Raman spectral points depicted in Fig. 3 and preprocessed through Pareto) was able to correctly classify 90 % of the fruit distillates with respect to the botanical origin during cross-validation. This performance proved the differentiation ability of the chosen markers, as well as the efficiency of the Pareto preprocessing for the varietal discrimination of fruit spirits based on Raman measurements. Therefore, the varietal origin of 81 %, 66 %, 97 %, and 60 % of the apple, pear, plum, and grape samples

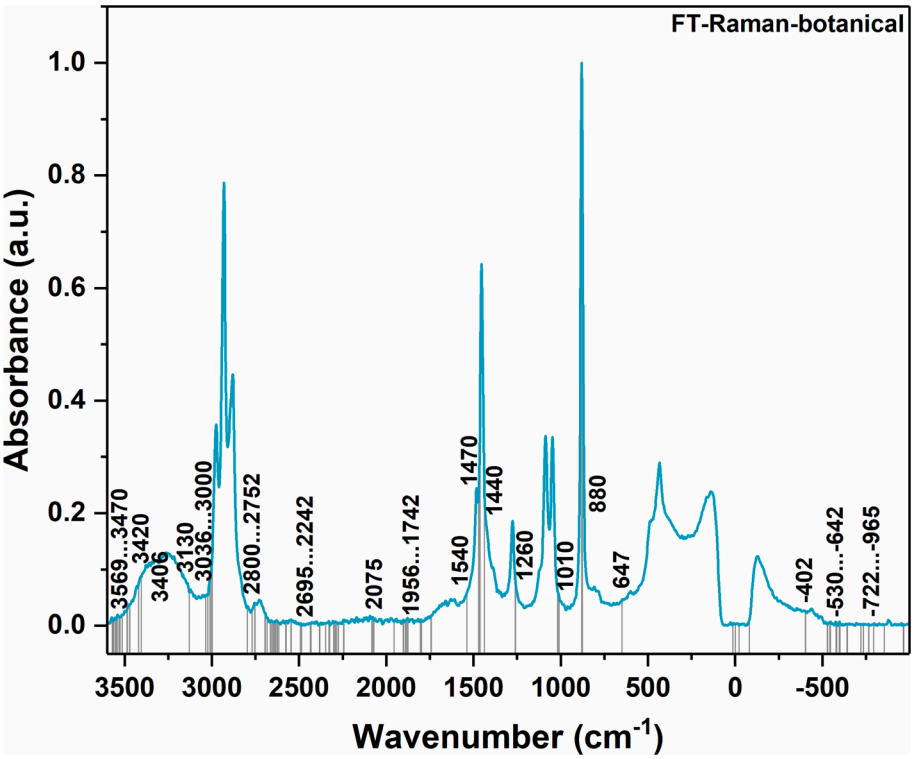


Fig. 3. Raman spectrum containing the most important predictors used for the botanical differentiation of fruit distillates.

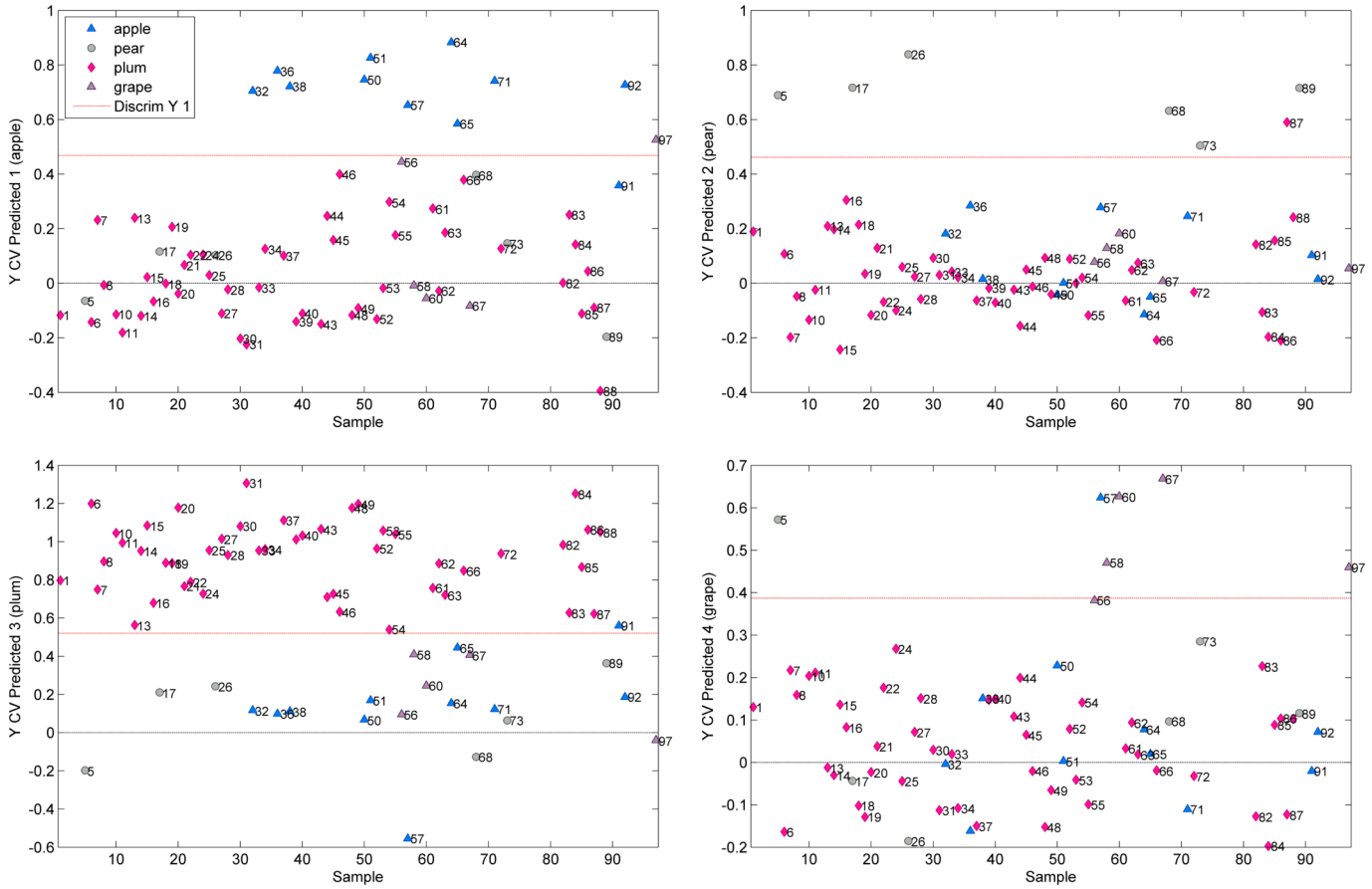


Fig. 4. Y cross-validation predicted values by the PLS-DA model developed for the varietal discrimination of the fruit distillates. The input data is represented by 133 Raman variables preprocessed through Pareto.

respectively was successfully predicted. Among the seven wrongly attributes instances, three were also misclassified by the PLS-DA model, namely samples indexed 56, 87 and 91 (Fig. 4). The other four wrongly classified samples corresponded to one grape (sample 58), one apple (sample 92) and two pear (samples 26 and 73) distillates that were all predicted as plum samples. The incorrect attribution of sample indexed 73 can be related to the fact that, as can be seen in Fig. 4, the PLS-DA Y cross-validation predicted value associated with it was the lowest one observed among all six pear samples. The configuration of the SVM model, determined by means of a grid search approach, corresponded to the use of a RBF kernel, a gamma parameter of 2^{-5} and a C parameter of 2^{8}.

The 133 Raman spectral points based on which the best botanical differentiation model was constructed can be visualized in Fig. 3. According to Socrates (2001), the bands situated in the region 3569–3470 cm^{-1} or around the 3420, 3406, 3130 cm^{-1} are characteristic to the vibrations of the hydroxyl group from water, alcohols, phenols or carbohydrates; the latter compounds are dependent of the wooden barrel type and although there is a similarity with the markers from the geographical discrimination, it seems that their contribution is smaller in the botanical differentiation model (Figs. 1 and 3). Contrary to the previous situation, here, the region between 3036 and 3000 cm^{-1} (assigned to the asymmetric CH_3 stretching vibrations from the saturated or unsaturated methyl esters) becomes more important. The aldehydes, through the –C–H group stretching or overtone CH in-plane deformation vibrations appearing in the domain 2800–2752 cm^{-1}, and the carboxylic acids or phenols, whose the O–H stretching vibrations occur in the frequency domain 2695–2242 cm^{-1}, seems to be relevant for the botanical origin recognition too. The signals from about 2075 cm-1 could be related to the stretching vibration of the cyanide ion from some cyanide derivatives that could be found in the distilled products (IARC Working Group, 1988). The discriminants from the frequency region between 1956 and 1742 cm^{-1} could be due to the presence of aromatic compounds, while the peak from 1540 cm^{-1} could be from the asymmetric stretch vibration of CO_2^- group or from the CC in-plane/ring vibrations of some heterocyclic constituents of fruit spirits. The C=C ring stretching vibrations of some furan derivatives could have occurred at 1470 cm^{-1}, whereas the 1440 cm^{-1} band could be related to asymmetric CH_3 deformation vibrations of the aliphatic ketones or methyl esters, to the in-plane C–H rocking vibrations of aliphatic aldehydes, to the O–H deformation vibrations from alcohols, to the symmetric vibration of the CO_2^- group from carboxylic acids salts or acetate salts. At 1260 cm^{-1} are detectable the O–H deformation vibrations of some alcohols, around 1010 cm^{-1}, the aromatic =C–H in-plane deformation vibrations, around 880 cm^{-1}, the CH_2 out-of-plane deformation vibrations from esters, while at 647 cm^{-1}, the rocking or in-plane deformation vibrations of CO_2 group of the aromatic esters. In the anti-Stokes region, the -402 cm^{-1} could be associated with the M−O stretching vibrations, whereas the bands originated from the domain -530–642 cm^{-1} with the O–C–O group vibrations (bending or out-of-plane deformation), which indicates the presence of aliphatic esters or acetates. The domain from -722 to -965 cm^{-1} is specific to the C–H out-of-plane deformation vibrations from the aromatic or heterocyclic compounds, and also to the C–C–O stretching modes of various alcohols.

4. Conclusion

The present work proposes the development of new recognition tools for the fast, non-destructive and cost-effective assessment of fruit spirits' botanical and geographical origin based on the association between FT-Raman spectroscopy and advanced data processing strategies.. In this regard, strategies based on either supervised statistical methods (i.e. Partial Least Squares Discriminant Analysis) or Machine Learning approaches (i.e. Support Vector Machines) were applied for the development of the prediction models. With the aim of constructing reliable classifiers, a special attention was given to the utilization of an appropriate preprocessing method for transforming the raw Raman spectra, along with the selection of the most significant variables, identified individually for each differentiation criterion. The obtained results illustrated that by introducing a data reduction step before model development, with the aim of keeping only the attributes that have the highest discrimination power, a significant improvement in the prediction accuracy is obtained as compared to the utilization of the entire variable set. Based on this approach, recognition models having prediction accuracy higher than 90 % were developed with respect to the geographical origin and the type of raw material used for distillate production, highlighting the effectiveness of the proposed approach for alcoholic beverage origin control.

CRediT authorship contribution statement

Ariana Raluca Hategan: Conceptualization, Methodology, Software, Validation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Maria David:** Methodology, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Camelia Berghian-Grosan:** Software, Validation, Writing – original draft, Writing – review & editing. **Dana Alina Magdas:** Conceptualization, Resources, Writing – original draft, Writing – review & editing, Supervision, Project administration, 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.

Data availability

Data will be made available on request.

Acknowledgements

This work was supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI – UEFISCDI, project number PN-III-P2-2.1-PED-2021-1095 (contract no. 651PED/ 2022) within PNCDI III.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.fochx.2023.100902.

References

Barai, H. R., Banerjee, A. N., & Joo, S. W. (2017). Improved electrochemical properties of highly porous amorphous manganese oxide nanoparticles with crystalline edges for superior supercapacitors. *Journal of Industrial and Engineering Chemistry, 56*, 212–224. https://doi.org/10.1016/j.jiec.2017.07.014

Bauer-Christoph, C., Wachter, H., Christoph, N., Roßmann, A., & Adam, L. (1997). Assignment of raw material and authentication of spirits by gas chromatography, hydrogen- and carbon-isotope ratio measurements I. Analytical methods and results of a study of commercial products. *Zeitschrift für Lebensmitteluntersuchung und -Forschung A, 204*, 445–452. https://doi.org/10.1007/s002170050111

Berghian-Grosan, C., & Magdas, D. A. (2020). Application of Raman spectroscopy and Machine Learning algorithms for fruit distillates discrimination. *Scientific Reports, 10* (1), 21152. https://doi.org/10.1038/s41598-020-78159-8

Buglass, A. J. (2014). Chemical Composition of Beverages and Drinks. In P. Cheung (Ed.), *Handbook of Food Chemistry* (pp. 1–62). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-41609-5_29-1.

Coldea, T. E., Socaciu, C., Fetea, F., Ranga, F., Pop, R. M., & Florea, M. (2013). Rapid quantitative analysis of ethanol and prediction of methanol content in traditional fruit brandies from Romania, using FTIR spectroscopy and chemometrics. *Notulae Botanicae Horti Agrobotanici, 41*, 143–149. https://doi.org/10.15835/nbha4119000

David, M., Hategan, A. R., Berghian-Grosan, C., & Magdas, D. A. (2022). The Development of Honey Recognition Models Based on the Association between ATR-

IR Spectroscopy and Advanced Statistical Tools. *International Journal of Molecular Sciences, 23*(17), 9977. https://doi.org/10.3390/ijms23179977

Dolenko, T. A., Burikov, S. A., Dolenko, S. A., Efitorov, A. O., Plastinin, I. V., Yuzhakov, V. I., et al. (2015). Raman spectroscopy of water–ethanol solutions: The estimation of hydrogen bonding energy and the appearance of clathrate-like structures in solutions. *The Journal of Physical Chemistry A, 119*(44), 10806–10815. https://doi.org/10.1021/acs.jpca.5b06678

Eigenvector Research, Inc. Eigenvector Research Wiki. *Selectvars*. Retrieved from https://wiki.eigenvector.com/index.php?title=Selectvars. Accessed February 9, 2023.

Eigenvector Research, Inc. Eigenvector Research Wiki. *Using Cross-Validation*. Retrieved from https://wiki.eigenvector.com/index.php?title=Using_Cross-Validation. Accessed February 9, 2023.

Ellis, D. I., Muhamadali, H., Xu, Y., Eccles, R., Goodall, I., & Goodacre, R. (2019). Rapid through-container detection of fake spirits and methanol quantification with handheld Raman spectroscopy. *The Analyst, 144*(1), 324–330. https://doi.org/10.1039/C8AN01702F

Gromski, P. S., Muhamadali, H., Ellis, D. I., Xu, Y., Correa, E., Turner, M. L., et al. (2015). A tutorial review: Metabolomics and partial least squares-discriminant analysis–a marriage of convenience or a shotgun wedding. *Analytica Chimica Acta, 879*, 10–23. https://doi.org/10.1016/j.aca.2015.02.012

IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. Alcohol Drinking. Lyon (FR): International Agency for Research on Cancer; 1988. (IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, No. 44.) 3, Chemical Composition of Alcoholic Beverages, Additives and Contaminants. Internet: https://www.ncbi.nlm.nih.gov/books/NBK531662.

Ledauphin, J., Saint-Clair, J. F., Lablanquie, O., Guichard, H., Founier, N., Guichard, E., et al. (2004). Identification of trace volatile compounds in freshly distilled calvados and cognac using preparative separations coupled with gas chromatography−mass spectrometry. *Journal of Agricultural and Food Chemistry, 52*(16), 5124–5134. https://doi.org/10.1021/jf040052y

Magdas, D. A., Cozar, B. I., Feher, I., Guyon, F., Dehelean, A., & Cinta Pinzaru, S. (2019). Testing the limits of FT-Raman spectroscopy for wine authentication: Cultivar, geographical origin, vintage and terroir effect influence. *Scientific Reports, 9*(1), 19954. https://doi.org/10.1038/s41598-019-56467-y

Magdas, D. A., Cristea, G., Pîrnau, A., Feher, I., Hategan, A. R., & Dehelean, A. (2021). Authentication of Transylvanian spirits based on isotope and elemental signatures in conjunction with statistical methods. *Foods, 10*(12), 3000. https://doi.org/10.3390/foods10123000

Magdas, D. A., David, M., & Berghian-Grosan, C. (2020). Fruit spirits fingerprint pointed out through artificial intelligence and FT-Raman spectroscopy. *Food Control, 133*, Article 108630. https://doi.org/10.1016/j.foodcont.2021.108630

Magdas, D. A., Guyon, F., Berghian-Grosan, C., & Muller Molnar, C. (2021). Challenges and a step forward in honey classification based on Raman spectroscopy. *Food Control, 123*, Article 107769. https://doi.org/10.1016/j.foodcont.2020.107769

Magdas, D. A., Guyon, F., Feher, I., & Cinta Pinzaru, S. (2018). Wine discrimination based on chemometric analysis of untargeted markers using FT-Raman spectroscopy. *Food Control, 85*, 385–391. https://doi.org/10.1016/j.foodcont.2017.10.024

Mangas, J., Rodríguez, R., Moreno, J., Suárez, B., & Blanco, D. (1996). Evolution of aromatic and furanic congeners in the maturation of cider brandy: A contribution to its characterization. *Journal of Agricultural and Food Chemistry, 44*(10), 3303–3307. https://doi.org/10.1021/jf950782t

Mendes, L. S., Oliveira, F. C., Suarez, P. A., & Rubim, J. C. (2003). Determination of ethanol in fuel ethanol and beverages by Fourier transform (FT)-near infrared and FT-Raman spectrometries. *Analytica chimica acta, 493*(2), 219–231. https://doi.org/10.1016/S0003-2670(03)00870-5

Parastar, H., van Kollenburg, G., Weesepoel, Y., van den Doel, A., Buydens, L., & Jansen, J. (2020). Integration of handheld NIR and machine learning to "Measure & Monitor" chicken meat authenticity. *Food Control, 112*, Article 107149. https://doi.org/10.1016/j.foodcont.2020.107149

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.

Rashad, M., Rüsing, M., Berth, G., Lischka, K., & Pawlis, A. (2013). CuO and Co3O4 nanoparticles: Synthesis, characterizations, and Raman spectroscopy. *Journal of Nanomaterials, 2013*, Article 714853. https://doi.org/10.1155/2013/714853

Santillan, J. M. J., Arboleda, D. M., Coral, D. F., Fernandez van Raap, M. B., Muraca, D., Schinca, D. C., et al. (2017). Optical and magnetic properties of Fe nanoparticles fabricated by femtosecond laser ablation in organic and inorganic solvents. *ChemPhysChem, 18*, 1192–1209. https://doi.org/10.1002/cphc.201601279

da Silva, A. G., Santana, H. S., Bagarolo, R., Rodrigues, A. C., Castilho, G. J., Cremasco, M. A., et al. (2019). Application of Raman spectroscopy in microfluidic devices for on-line determination of ethanol concentration in water and vegetable oil. *Chemical Engineering Transactions, 74*, 739–744. https://doi.org/10.3303/CET1974124

Socrates, G. (2001). Infrared and Raman Characteristic Group Frequencies. *Tables and Charts (3rd ed.)*. Chichester, England: John Wiley & Sons Ltd.

Spaho, N. (2017). Distillation techniques in the fruit spirits production. In Mendes, M. F. (Ed.), *Distillation - Innovative applications and modeling* (pp. 129-152). InTech. Doi: 10.5772/66774.

Šrámek, J., Švancara, I., & Sýs, M. (2019). Determination of ethanol in alcoholic drinks using Raman spectrometry. *Scientific papers of the University of Pardubice. Series A, Faculty of Chemical Technology*. 25/2019.

Tsakanikas, P., Karnavas, A., Panagou, E. Z., & Nychas, G. J. (2020). A machine learning workflow for raw food spectroscopic classification in a future industry. *Scientific Reports, 10*(1), 1–11. https://doi.org/10.1038/s41598-020-68156-2

Vaskova, H. (2014). Spectroscopic determination of methanol content in alcoholic drinks. *International Journal of Biology and Biomedical Engineering, 8*, 27–34.

Winterova, R., Mikulikova, R., Mazáč, J., & Havelec, P. (2008). Assessment of the authenticity of fruit spirits by gas chromatography and stable isotope ratio analyses. *Czech Journal of Food Science, 26*(5), 368–375.