agronomy

MDPI

Article

Detection of Adulterations in Fruit Juices Using Machine Learning Methods over FT-IR Spectroscopic Data

José Luis P. Calle [1], Marta Ferreiro-González [1,*], Ana Ruiz-Rodríguez [1,*], Daniel Fernández [2,3,4] and Miguel Palma [1]

[1] Department of Analytical Chemistry, Faculty of Sciences Agrifood Campus of International Excellence (ceiA3), IVAGRO, University of Cadiz, 11510 Puerto Real, Spain; joseluis.perezcalle@uca.es (J.L.P.C.); miguel.palma@uca.es (M.P.)
[2] Department of Statistics and Operations Research (DEIO), Universitat Politècnica de Catalunya—BarcelonaTech (UPC), 08028 Barcelona, Spain; daniel.fernandez.martinez@upc.edu
[3] Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM), Instituto de Salud Carlos III, 28029 Madrid, Spain
[4] Institute of Mathematics of UPC (IMTech), Universitat Politècnica de Catalunya—Barcelona Tech, 08028 Barcelona, Spain
* Correspondence: marta.ferreiro@uca.es (M.F.-G.); ana.ruiz@uca.es (A.R.-R.); Tel.: +34-956-016-359 (M.F.-G.); +34-956-016-363 (A.R.-R.)

Abstract: Fruit juices are one of the most adulterated beverages, usually because of the addition of water, sugars, or less expensive fruit juices. This study presents a method based on Fourier transform infrared spectroscopy (FT-IR), in combination with machine learning methods, for the correct identification and quantification of adulterants in juices. Thus, three types of 100% squeezed juices (pineapple, orange, and apple) were evaluated and adulterated with grape juice at different percentages (5%, 10%, 15%, 20%, 30%, 40%, and 50%). The results of the exploratory data analysis revealed a clear clustering trend of the samples according to the type of juice analyzed. The supervised learning analysis, based on the development of models for the detection of adulteration, obtained significant results for all tested methods (i.e., support-vector machines or SVM), random forest or RF, and linear discriminant analysis or LDA) with an accuracy above 97% on the test set. Regarding quantification, the best results are obtained with the support vector regression and with partial least square regression showing an R^2 greater than 0.99 and a root mean square error (RMSE) less than 1.4 for the test set.

Keywords: FT-IR; fruit juices; food control; machine learning; spectroscopy; regression; authentication; classification

Citation: Calle, J.L.P.; Ferreiro-González, M.; Ruiz-Rodríguez, A.; Fernández, D.; Palma, M. Detection of Adulterations in Fruit Juices Using Machine Learning Methods over FT-IR Spectroscopic Data. *Agronomy* **2022**, *12*, 683. https://doi.org/10.3390/agronomy12030683

Academic Editor: Pedro Javier Zapata

Received: 20 February 2022
Accepted: 8 March 2022
Published: 11 March 2022

1. Introduction

One of the largest sectors in the beverage industry is the production of fruit juices. In fact, 9067 million liters were consumed in Europe during 2018 according to the Association of the Juice and Nectar Industry of the European Union (AIJN) [1]. Additionally, the interest in fruit juices is increasing, as these drinks provide nutritional and dietary benefits, being an excellent source of vitamin and key nutrients such as potassium, folate, magnesium, among others [2,3].

In Europe, there is strict regulation to guarantee the quality and origin in the manufacture of juices established by the 2012/12 EU directive [4]. Moreover, this directive describes that fruit juices must be 100% squeezed from healthy and ripe fruits where the addition of sugars is not allowed. Despite the regulations, these products have always been subjected to adulteration in the market, being reported as one of the seven most common foods for adulteration between 1980 and 2010 [5]. Among the most frequent adulterations is the dilution with water, the addition of artificial sweeteners, or less expensive fruit juices [6]. The latter is very popular due to its greater difficulty to be detected [7]. It is important to be

aware that adding other fruit juices and not informing the consumer entails an additional risk of allergic reactions [8]. Therefore, it is food fraud for economic benefits, which could also affect the health of the consumer [9].

Currently, analytical separation techniques, including both liquid and gas chromatography, are the most widely used methodologies for the detection of adulterants in juices [6,8,10–13]. However, other techniques such as isotope-based techniques and elemental techniques [14–16], physicochemical analysis [17,18], DNA-based techniques [19,20], and spectroscopic techniques [6,8,12,13,21] have been employed. Among them, spectroscopic techniques are becoming more and more important due to their numerous advantages such as low analysis time and costs, ease of use, high reproducibility, and greater portability, allowing in situ analysis. In this context, Fourier transform infrared spectroscopy (FT-IR) has been one of the most extensive techniques to successfully detect adulterations in juices. It has been employed for the detection of sugar addition in apple [22–24], mango [25], and orange [12] juices, the authentication of different commercial juices [26], the detection of grape juice in pomegranate juice [27], the discrimination of lime juice adulterated with isocitric and/or citric acid [21], as well as the authentication of Concord grape juice in grape blends [10].

Most of these FT-IR methods are based on the identification of one or a few individual compounds that are used as biomarkers to detect the adulterant or the authenticity of the juice. Therefore, these methods have some limitations, as they are easy to elude and, consequently, less and less successful. An alternative to developing more robust methods could be focusing on the whole or a wide spectral range that can be used as a pattern, or using *spectralprint* techniques characteristic of each type of sample [28]. In that manner, FT-IR spectroscopy, used as a global profiling or screening technique, generates a large amount of information in a few seconds. With the aim of handling and extracting useful information from these data matrices, the application of data analysis methods (including machine learning methods) is crucial. Thus, both the transformation of data into interpretable information and the fitting of predictive models with interactive applications to automate quality control processes are possible by employing languages of coding, so these global profiling methods are becoming increasingly popular [9] and have been used in various fields such as forensic chemistry [29,30], agri-food [31,32], pharmacological industry [33,34], among others. It is worth it since juices are complex matrices and the differences between spectra are sometimes very substantial, thus identifying a small number of markers can be difficult and unsuccessful. However, holistic methods consider small differences that may be important in many cases and consequently offer better results than individual identification. In addition, the process can be automated which requires less time for the correct characterization of the sample.

Multivariate parametric analysis methods such as principal component analysis (PCA) for dimensionality reduction, linear discriminant analysis (LDA) for classification, and partial least squares regression (PLS) for adulterant quantification stand out among the most commonly employed machine learning methods in juice characterization studies. However, non-parametric methods such as random forests (RF) or support-vector machines (SVM) have been less widely used, even though they have reported better results for similar purposes [35,36]. Both SVM and RF have been used in combination with NIRS spectroscopy data for the purity assessment of lime juice, with optimal results [37]. Additionally, another study assessed the quality change of tomato juice using near-infrared spectroscopy coupled to global profiling methods, obtaining better results with support vector machines than with partial least squares [38]. This highlights the superiority of non-parametric techniques in some situations, especially when the matrices analyzed are complex and the relationship between the spectrum and the response variable is not linear.

This research study focuses on the development of a novel method based on machine learning models (parametric and non-parametric) for the detection and quantification of juices-to-juices adulterations in several fruit juices through the spectral information

generated by FT-IR spectroscopy as an objective, robust, fast, and automatic method for the detection of this illegal practice.

2. Materials and Methods

2.1. Samples

Four types of 100% fruit juices (apple, pineapple, orange, and grape) from different suppliers in Spain were selected. A minimum of three different brands were chosen and, from each of them, several batches to increase heterogeneity within each type of juice. In that manner, the samples were labeled according to the following character code: "A_BB_C" where "A" stands for the type of juice, i.e., A (apple), P (pineapple), O (orange), and G (grape). "BB" indicates the brand and "C" the lot number (1, 2, or 3). Additionally, each of them was analyzed twice, being labeled "R1" or "R2". Thus, the label corresponding to the second batch of an "HC" brand apple juice would be "A_HC_2_R1" for the first replica and "A_HC_2_R2" for the second one. The overall number of unadulterated juice samples reaches up to 76 with a minimum of 18 samples for each type of juice, which were subsequently analyzed by FT-IR.

2.2. Adulteration

Grape juice was selected as an adulterant since this juice is commonly used for its low cost compared to other types of juices [15]. Thus, two samples of each type of juice were created by mixing the different brands of juice, i.e., the 4 brands for orange juice and 3 brands for all other juices from a randomly selected batch were proportionally mixed in order to cover the widest heterogeneity. Those two samples of each type of juice (orange, pineapple, and apple) were those adulterated with grape juice at the following ratios: 5%, 10%, 15%, 20%, 30%, 40%, and 50%. Non-adulterated fruit juices samples (0%) and grape juices (100%) were also analyzed. In addition, the grape juice was diluted with the sugar concentration of the juice (orange, pineapple, or apple) for the adulteration process. This step was necessary because it is well known that FT-IR is sensitive to the number of sugars, allowing even the quantification of sugars in juices [39–41]. Therefore, the dilution of the juices was necessary to prevent the fraud from being easily eluted and to get a more robust model, independent of the sugar content. A final set of 108 adulterated samples was obtained: 3 types of juices $\times$ 2 different samples $\times$ 9 adulteration ratios (from 0 to 50% and 100%) $\times$ 2 replicates. The samples were labeled according to the following character code: "AB_C" where "A" stands for the type of juice, "B" indicates the sample used (1 or 2), and "C" the ratio of adulteration (0, 5, 10, 15, 20, 30, 40, 50, and 100). In addition, each of them was analyzed twice, being labeled "R1" or "R2".

2.3. Fourier Transform Infrared Spectroscopy (FT-IR)

Infrared spectra were measured by Fourier transform for all samples using a MultiSpec (TDI, Barcelona, Spain). Previously, the samples had to be centrifuged and filtered with 0.45 µm filters to reduce turbidity and eliminate impurities. The sample volume analyzed was 7 mL (standard setting) and was pumped through the system. Spectra were recorded in the range of 952–3070 cm^{-1} with a resolution of 3.86 cm^{-1} and an optical path length of 20 µm. The region from 1610 to 1670 cm^{-1} was eliminated for presenting a high variability since it is the characteristic signal of water. The working temperature was set at 25 °C and the total analysis time per sample was 1 min.

2.4. Data Analysis

The raw spectral data were acquired for the region previously described and were placed into $D_{n\times p}$ matrix where *n* denotes the number of samples and *p* denotes the number of variables. Thus, the complete total matrix ($D_{184\times 540}$) consisted of 540 variables (wavenumbers) and 184 samples. The entire computer analysis necessary to carry out this study was performed with the statistical software RStudio v.4.0.2 (Rstudio Team 2021, Boston, MA, USA). All visualizations were performed with the *ggplot2* package [42]. The

unsupervised learning analysis, including principal component analysis (PCA) and hierarchical cluster analysis (HCA), was performed with the *stats* package [43]. In the HCA analysis, the Manhattan distance was used and the linkage method was chosen based on the highest value of the correlation coefficient between the cophenetic distance of the dendrogram and the original distance matrix. The methods tested were: single, average, complete, ward, and centroid, obtaining the best result for the average method.

The supervised learning analysis, which includes regression models: partial least squares and shrinkage methods (lasso, ridge, and elastic net), and classification models: linear discriminant analysis, support-vector machine, and random forest, was performed with the *caret* package [44]. The performance metrics for the classification models measured the accuracy and for the regression models, the root mean square error (RMSE) and the coefficient of determination (R^2). Finally, the development of the application was carried out using the *shiny* package [45].

3. Results and Discussion

3.1. Exploratory Data Analysis (EDA)

Initially, the goal was to observe whether there is a spectral difference between the different types of fruit juices. For this purpose, we used the 76 unadulterated juice samples and the resulting 540 spectral variables (wavenumbers). Therefore, a data matrix $D_{76\times540}$ was subjected to HCA, which was performed using both the Manhattan distance and the average method. It should be noted that the choice of method was based on the correlation coefficient between the cophenetic distance of the dendrogram (height of the nodes) and the original distance matrix. Therefore, different common methods were assessed: single, complete, average, ward, and centroid, which values are shown in Table 1. In general, high values are obtained for all of them, but the best result is obtained through the average method (0.9848), which indicates that the resulting dendrogram (depicted in Figure 1) reflects very well the true similarity between the observations. This dendrogram is represented circularly and the samples are colored according to the type of juice to facilitate the interpretation of the dendrogram.

Table 1. Correlation between the cophenetic distance and the distance matrix for the different clustering methods using the FT-IR (Fourier transform infrared spectroscopy) spectrum of all the pure juice samples ($D_{76\times540}$).

Method	Cophenetic Distance
Single	0.9835
Complete	0.9835
Average	0.9848
Ward	0.9836
Centroid	0.9791

The dendrogram shows that there are four perfectly differentiated groups and each of them corresponds to a type of juice. Furthermore, the grape juice is the one that seems to differentiate itself the most from the rest. Moreover, all the pineapple samples (colored yellow) fall into an independent cluster, which is joined to the cluster containing all the orange samples (colored orange). Finally, this cluster merges with the one containing all the apple samples (colored red). In this case, it seems that there is a perfect clustering of the samples according to the brand within each type of juice, i.e., all the samples fall into the same subcluster. Therefore, the FT-IR spectra are clearly influenced firstly by the type of juice and secondly by the brand used. In general, HCA allowed perfect distinction of the type of juice analyzed regardless of the brand or sample used. However, it is important to ensure that this differentiation is not exclusively due to the sugar content present in the

sample (Average values of the sugar content of each type of fruit juice can be found in Table S1 of Supplementary Materials).

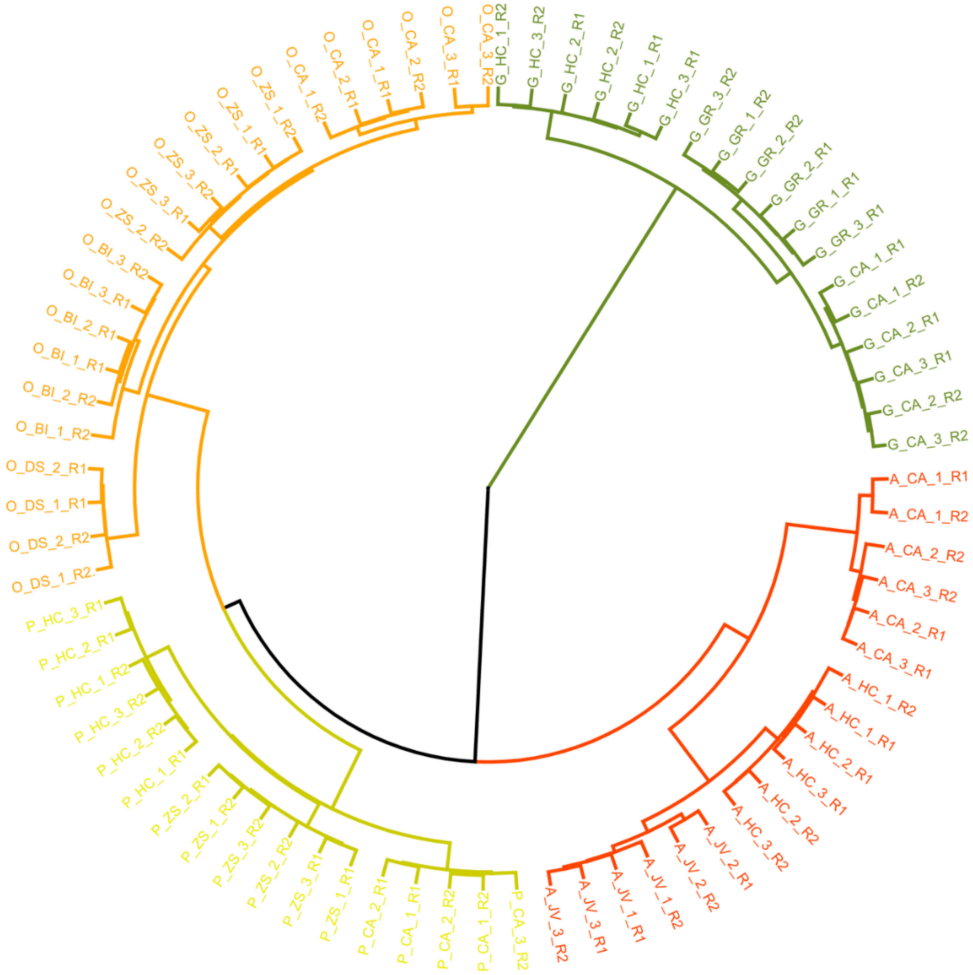

Figure 1. Dendrogram from HCA (hierarchical cluster analysis) using the Manhattan distance and average method for all FT-IR spectra of pure juices ($D_{76\times540}$). Samples are colored according to juice type: P for pineapple (yellow), G for grape (dark green), A for apple (red), and O for orange (orange).

We perform a PCA to identify regions and, consequently, to determine the most influential functional groups in the juice classification. Figure 2A shows the scores obtained by the samples for the first two principal components (PC1 & PC2) and Figure 2B shows the loadings of these components. As can be seen in Figure 2A, PC1 explains 87.8% of the variability of the data, which allows to distinguish between the grape juice samples and the other juices. Thus, positive loadings of PC1 are associated with grape juices while negative loadings correspond to apple, pineapple, and orange juices. This is quite relevant because grape juice is the one used as an adulterant and, therefore, the greater the separation, the easier it will be to detect it. Besides, PC2 explains 6.1% of the total variability and allows to distinguish mainly pineapple from apple. In this case, positive loadings are associated with apple, more negative loadings represent pineapple, and loadings around 0 would be most associated with orange. It can also be observed that juices of the same brand and type tend to appear closer together, indicating a tendency for grouping according to brand.

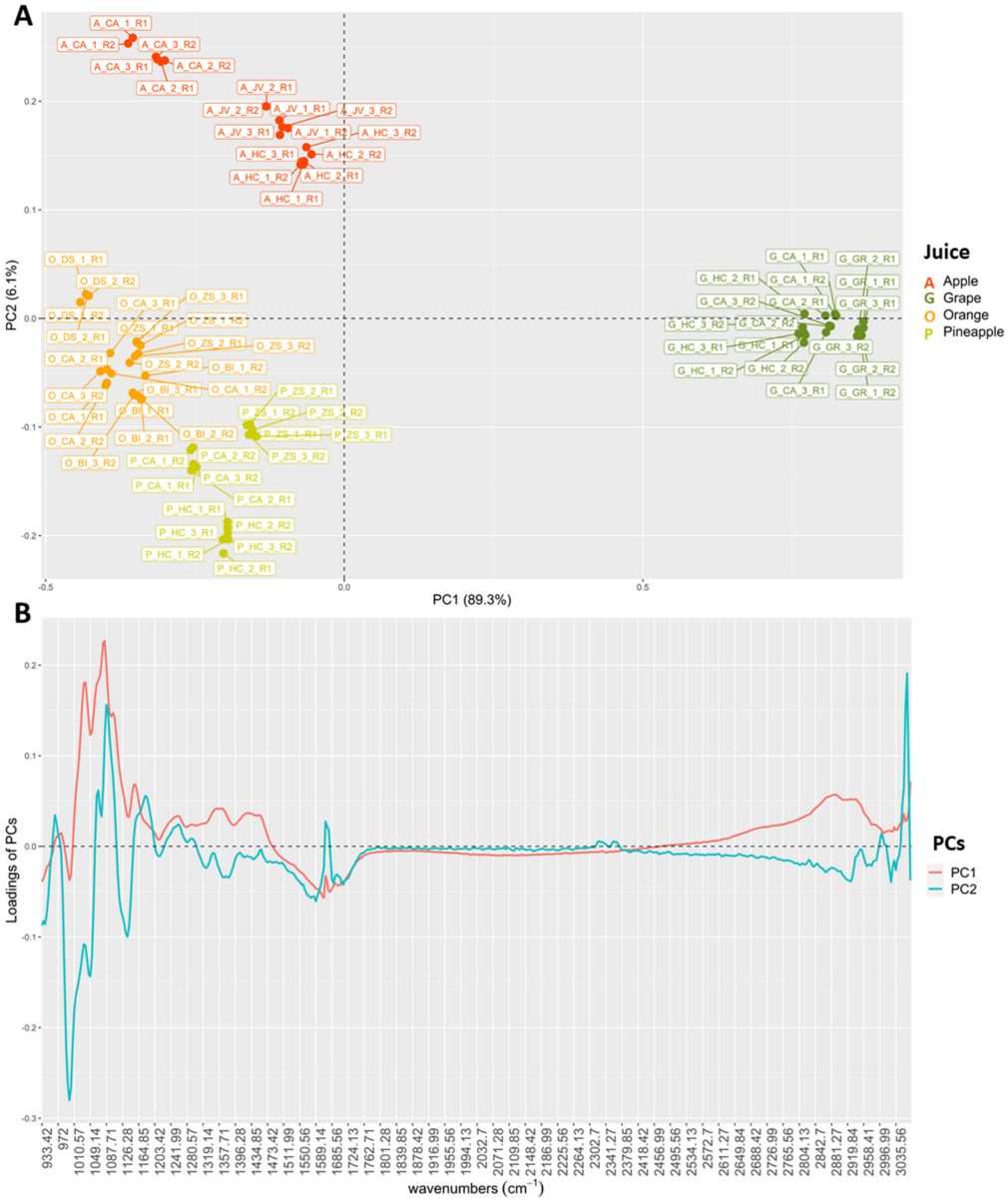

Figure 2. (**A**) Representation of the scores obtained using the FT-IR spectrum of all the pure juice samples ($D_{76\times540}$) as a function of the first two principal components (PC1 and PC2). (**B**) Representation of the loadings of the first two PCs.

As can be seen in Figure 2B, some regions are influential in the separation of the samples. In general, weights greater than |0.07| can be seen approximately in the region 930–1170 cm^{-1} and around |0.03| in the region 1199–1400 cm^{-1}. According to previous studies in the analysis of fruit juices by infrared spectroscopy [39], the 900–1400 cm^{-1} region is related to the sugars present in these beverages. These include glucose, sucrose, and fructose, which present characteristic and intense bands in this region. Thus, the region of 900–1153 cm^{-1} is related to the stretching vibrations produced by the C–O and CC bonds, which is exactly where the highest weights are acquired. The 1199–1400 cm^{-1} region, which acquires moderately high weights, is related to the bending vibrations of the O–C–H, C–CH, and C–O–H bonds. In addition, previous studies have identified these regions as important for the detection of adulterations in different types of fruit juices [12,24]. The last region of the spectrum, approximately from 2700 to 3030 cm^{-1}, also acquired high values in the loadings. According to the existing literature, the region from 2500 to 3300 cm^{-1} is related to the stretching vibrational motions produced by the O–H bonds of carboxylic

acids [46]. These higher weights may be due to the multitude of acids naturally present in fruit such as malic acid, which is found in high concentration in apples, citric acid, which is scarce in grapes and abundant in pineapple and orange, and tartaric acid, which is very characteristic of grapes, among others [47–49].

To sum up, this classification trend could be due to the difference in the number of initial sugars since, according to the data provided by the manufacturer and checked by densimeter, grape juice is the one with the highest amount (≈15.8 g/100 mL). For this reason, to prevent the applied supervised learning models from discriminating between the different types of juice based on sugars, the samples of the adulteration process were diluted according to the sugar concentration. Additionally, PCA allowed a good separation of the juices according to their type as well as the selection of the wavenumbers responsible for this separation.

3.2. Classification Methods for Adulterant Detection

Once the ability of FT-IR to group juices according to the type of fruit has been checked, an evaluation of the feasibility of the technique to identify and quantify adulteration in the different fruit juices was carried out. For this purpose, the whole data matrix was used, which is composed by the 184 samples from the different types of juices (orange, pineapple, apple, and grape) and the unadulterated and adulterated samples at different ratios (5%, 10%, 15%, 20%, 30%, 40%, and 50%). Therefore, the resulting matrix has 540 wavenumbers and 184 samples ($D_{184\times540}$).

For the supervised learning methods, a total of four groups were established a priori according to the type of juice used ("Pineapple", "Apple", and "Orange") and the presence of adulteration ("Adulterated"). The latter includes adulterated samples of all juices at different percentages with the grape juice, and also samples of pure grape juice. The complete data set was randomly split up on 75% of the samples for the training set and the remaining 25% for the test set. Additionally, it was ensured that they were balanced, and the test contained at least one sample of each type of juice at each of its percentages of adulteration. Thus, the test set contains 46 independent samples that are never part of the model and is used as external validation of all trained models, leading to an unbiased error. It is important to remark that both the undiluted grape samples used in the exploratory analysis and the diluted ones used in the adulterations are adjusted on the models. Thus, the models do not depend on sugar content as seen in the applied PCA and, therefore, a greater robustness of the models is achieved. The classification models evaluated were SVM with Gaussian kernel function, RF, and, LDA. A summary of the accuracy obtained by the fitting of the different models is shown in Table 2.

Table 2. Summary of the accuracy obtained by the classification models tested.

Model	Hyperparameter	Training Set Accuracy	Test Set Accuracy
lda	-	100%	100%
SVm	C = 2.83 Y = 0.022	100%	100%
RF	*mtry* = 23 *ntree* = 500	100%	97.67%

3.2.1. Support Vector Machines (SVM) with Gaussian Kernel Function

Gaussian kernel SVM models contain two hyperparameters (γ and C) that must be selected by the analyst. Thus, γ controls the behavior of the kernel and therefore, increasing its value increases the flexibility of the model. C controls the penalty, i.e., the bias-variance trade-off [50]. Optimization was performed by five-fold cross-validation with a grid search method with exponentially growing C and γ sequences [51]. Specifically, the values of γ and C ranged from $\log_2\gamma$, $\log_2$C in the range of $[-10, 10]$ taking values every 0.5. The result of this optimization is represented in Figure S1 of the Supplementary Materials, showing

the best accuracy values for γ = 0.022 and C = 2.83. Note that the five-fold validation process has been performed on the training set to avoid overfitting. In this way, the test set does not participate in the optimization process and gives rise to an unbiased error. The model fitted with the previous values resulted in an accuracy of 100% in both the training and test sets.

3.2.2. Random Forest (RF)

There are two hyperparameters to be determined in random forest (RF). One of them is the value of *mtry* which, for classification problems, is recommended to use as the root of the number of predictors [52]. Therefore, this was set to 23 (540 variables). Such value represents the number of predictors evaluated before setting the cutoff for each individual decision tree. The other hyperparameter is the number of trees, which was set at 500 since it is a large size to achieve the stabilization of the error. The result led to 100% accuracy in the training set and 97.67% accuracy in the test set, in which a 5% adulterated orange sample was incorrectly classified as pure orange juice. Finally, the resulting kappa was 0.9669.

3.2.3. Linear Discriminant Analysis (LDA)

The fit of LDA provided 100% accuracy in both the training and test sets. In addition, the probabilities of group belonging for each sample were very high, with all the probabilities being above 0.99 except for one sample of grape juice.

To sum up, the best performing models in our framework are from LDA and SVM with Gaussian kernel. None of them reach any errors neither in the training set nor in the test. Previous studies in juice analysis report better results in detecting adulteration when using SVM models, rather than LDA [53], while others report similar results [54]. In this case, either of the two may be applicable for the detection of the adulterant (grape juice) in the rest of the juices studied. In this case, it seems that there are no differences between the performances of the non-parametric and parametric multivariate methods. However, the SVM with a Gaussian kernel could also be used since the complexity of the model is not a prerequisite for the purpose of this research.

3.3. Regression Methods for Global Adulterant Quantification

After the algorithms for the identification of adulteration were trained, the next step was to identify the percentage of adulteration based on the FT-IR data. For this purpose, a global regression was performed using all the samples generated in the adulteration process. The sample was 96 (3 types of juices $\times$ 8 percentages $\times$ 4 points) and this was randomly split into a training set of 72 samples, i.e., three points for each percentage of adulteration and type of juice. In that manner, the test set consists of 24 independent samples, selected in a balanced manner and represents the entire data set. In this way, the test set was used as an external validation, since these samples were never used for the development and optimization of the models. The regression models evaluated were both parametric, such as partial least square (PLS) and shrinkage methods (lasso, ridge, and elastic net), and non-parametric, such as support vector regression (SVR) with Gaussian kernel function and RF regression. Additionally, a summary table of the results obtained for each model can be found in Table 3.

3.3.1. Partial Least Square Regression (PLS)

The optimal number of components for PLS was determined by leave-one-out cross-validation (LOOCV) on the training set data. Following the criterion of lower root-mean-square error (RMSE), the final model is formed by 10 components, with an RMSE of 1.366 and an R^2 of 0.9927 for the LOOCV. Figure S2 in the Supplementary Materials depicts its evolution graphically. Regarding the training set, the RMSE was 0.814 and the R^2 was 0.998, while the RMSE was 1.357 and the R^2 was 0.993 in the test. Therefore, a high correlation between the real values and the estimated ones exists.

Table 3. Results obtained for each regression method applied in the quantification of the global adulterant by using the FT-IR spectra of all adulterated juice samples ($D_{96\times540}$).

Model	Hyperparameter	LOOCV Performance	Training Set Perfomance	Test Set Performance
PLS	10 principal components	RMSE = 1.366 R^2 = 0.993	RMSE = 0.814 R^2 = 0.998	RMSE = 1.357 R^2 = 0.993
SVR	C = 11.31 $\gamma = 9.77\times10^{-4}$	RMSE = 2.775 R^2 = 0.977	RMSE = 1.446 R^2 = 0.994	RMSE = 1.243 R^2 = 0.995
RF	*mtry* = 139	RMSE = 5.742 R^2 = 0.926	RMSE = 2.298 R^2 = 0.991	RMSE = 3.873 R^2 = 0.973
Lasso	λ = 0.0756	RMSE = 1.732 R^2 =0.989	RMSE = 1.009 R^2 = 0.996	RMSE = 1.707 R^2 = 0.991
Ridge	λ = 10	RMSE = 6.416 R^2 = 0.941	RMSE = 5.579 R^2 = 0.960	RMSE = 5.189 R^2 = 0.942
Elastic net	λ = 0.196 α = 0.436	RMSE = 1.644 R^2 = 0.991	RMSE = 1.009 R^2 = 0.997	RMSE = 1.808 R^2 = 0.989

3.3.2. Support Vector Regression (SVR)

Analogous to SVM classification, the SVR requires the optimization of the previously discussed hyperparameters (C and γ). Those hyperparameters were again optimized with a grid search method with exponentially growing C and γ sequences, taking values from $\log_2\gamma$, $\log_2$C in the range of [−10, 10] every 0.5. The learning rate controlled by the epsilon (ε) hyperparameter was kept constant at 0.1. Thus, the best result was obtained for a γ of 9.77×10^{-4} and a C of 11.31, which provided an RMSE of 2.775 and an R^2 of 0.977. For the training set, the RMSE was 1.446 and the R^2 of 0.994, while the RMSE was 1.243 and the R^2 0.995 for the test set. The result obtained is similar to the PLS, however, the interpretability of this type of algorithm is lower.

3.3.3. Random Forest Regression

The number of trees was kept at 500 and the *mtry* value was optimized by a random search of values testing 30 of them. The model with the lowest RMSE presented a *mtry* of 139, which resulted in an R^2 of 0.926 and an RMSE of 5.742 for LOOCV. In the training and test set, an RMSE of 2.298 and 3.873, as well as an R^2 of 0.991 and 0.973, respectively, were obtained.

3.3.4. Lasso Regression

In this type of linear regression that uses shrinkage, the coefficients of the predictors that do not contribute to the model are penalized, forcing them to be 0, which excludes them from the analysis and makes the model more parsimonious. In this way, a new hyperparameter to optimize called lambda (λ) appears, which controls the degree of penalty. The λ value obtained as optimal was 0.0756, with an RMSE of 1.732 and an R^2 of 0.989. In the training and test sets, the RMSE obtained was 1.009 and 1.707 with an R^2 of 0.996 and 0.991, respectively. The fitting of lasso regression is very useful as it allows the selection of the important predictors, facilitating the understanding of the result. This method selected only 34 variables out of 540, depicted in Figure 3, where most of them, and particularly the most important ones, were in the vibrational region of sugars and carboxylic acids. These most important variables are, on the one hand, the wavenumber 1276.71 cm^{-1} which acquires a coefficient lower than −1250 and, on the other hand, the wavenumber 2240.99 cm^{-1}, which acquires a coefficient greater than 1250. The first one is related to the bending vibrations of the O–C–H, C–C–H, and C–O–H bonds presented in sugars and differences in polysaccharides in fruit juices already reported in this region [12]. The second one is related to C = C stretching vibrations groups and a previous study on the detection of adulterations in apple juice observed absorption bands important for discrimination in this region [55].

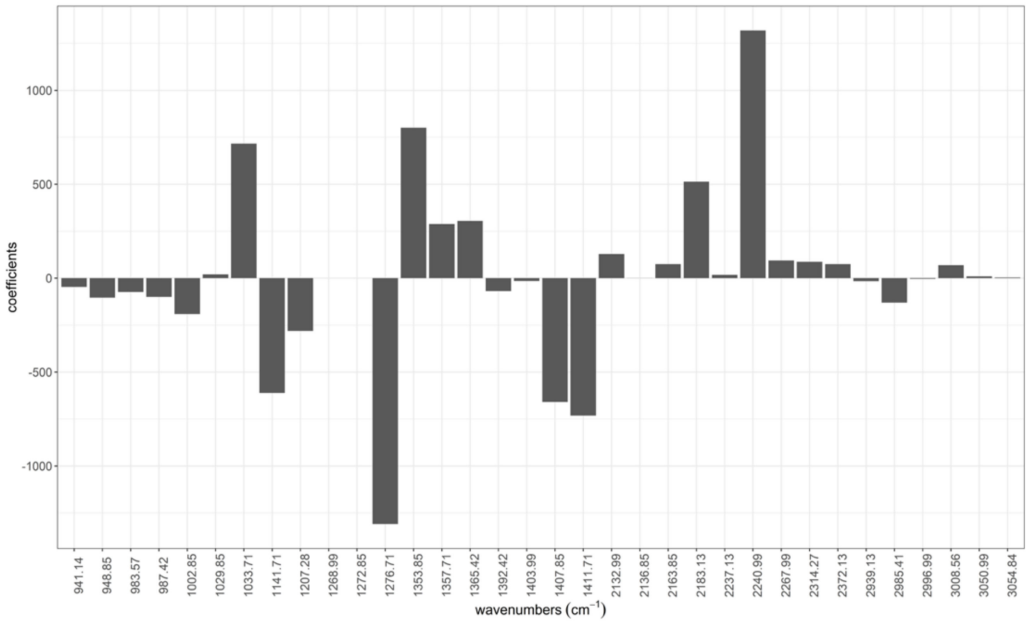

Figure 3. Coefficients selected by the lasso method for the quantification of the adulterant using the FT-IR spectrum of the adulterated juice samples.

3.3.5. Ridge Regression

Similarly to the previous analysis, the coefficients of variables that are not important are penalized so their values are reduced but without reaching 0 and, therefore, predictors are not excluded. In this shrinkage regression, the value of the lambda hyperparameter had to be specified using the same search method as the one employed in lasso regression. The value of λ selected as optimal was 10, achieving an RMSE of 6.416 and an R^2 of 0.941 for LOOCV. For the training and test sets, the RMSE values were 5.579 and 5.189 with an R^2 of 0.960 and 0.972, respectively. These results are considerably lower than the equivalents obtained with the other methods.

3.3.6. Elastic Net

In elastic net, a balance is sought between the exclusion of predictors (lasso) and the reduction of coefficients (ridge). Therefore, there are two hyperparameters to optimize: lambda, which controls the degree of penalty, and a new hyperparameter called alpha (α), which controls the degree of influence of each of the penalties (ridge and lasso). Their optimization was performed by testing 60 random combinations of values, and the best model will be decided based on the one that achieves a lower RMSE. In this case, the combination taken as optimal is a value of λ of 0.196 and an α of 0.436, which indicates that it is more similar to the ridge regression. For that combination, an RMSE of 1.644 and an R^2 of 0.991 were achieved for LOOCV while in the training set an RMSE of 1.009 and R^2 of 0.997 were obtained. Finally, for the test set, an R^2 of 0.989 and an RMSE of 1.808 were obtained.

For the global quantification of the adulterant, the results were more than satisfactory for all the methods applied (Table 3). The result obtained indicates a slightly higher potential for the use of the non-parametric method (SVR) in the test set (RMSE = 1.243), although, the second-best result (RMSE = 1.357) is obtained with a parametric approach (PLS). These results suggest that in the FT-IR spectroscopic data of juice samples none of the strategies is superior to the other. It should be noted that there are studies based on infrared spectroscopy data where better results are obtained with PLS than with SVR [56,57], while in other cases practically identical performance is reported [58] or even better with SVR [59]. Based on the previous literature and the results obtained in this study, the use of one or other methods may be of interest depending on the objective.

As a last remark, due to the high performance obtained and with the aim of sharing the models created and facilitating the detection and quantification of the adulterant for other users, a web application has been created (link and a short description of use available in the Supplementary Materials).

4. Conclusions

FT-IR spectroscopy, combined with suitable machine learning methods, has been empirically proven to be a reliable analytical technique for the detection and quantification of grape juice used as an adulterant in other juices. It has been observed that the FT-IR spectra of the juices are mainly influenced by the type of fruit and, to a lesser extent, by the brand used. Additionally, both the regression and classification models obtained perform more than satisfactorily. In the case of the classification problem, the best results were obtained with both LDA and with the non-parametric SVM (100% accuracy in the test and training set). Regarding the regression problem, the best results were obtained with the non-parametric method SVR and with the parametric PLS (both with R^2 greater than 0.99 and RMSE less than 1.4 in the test set). In that manner, the use of global profile methods, compared to individual identification, allows to eliminate subjectivity and automate the process. Thus, an application has been developed to share the models created with researchers and practitioners, and to ease the detection of adulterations in juices. Those models can learn as more samples are analyzed and a common, open database can be created in order to increasingly cover the needs of the beverage industry. Additionally, the proposed methodology is faster, cleaner, more objective, easy to use, and cheaper than traditional chromatographic-based methods.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/agronomy12030683/s1, Table S1: Theoretical (provided by the manufacturer) and experimental (measured by densimeter) average sugar values of fruit juices; Figure S1: Search for the best combination of hyperparameters (C and γ) for the Gaussian SVM model obtained by CV of 5 folds using the FT-IR spectrum of all training set samples ($D_{540\times138}$); Figure S2: Evolution of the root mean square error (RMSE), as a function of the number of components used in PLS analysis. The LOOCV error has been used for the FT-IR spectrum of the adulterated and unadulterated juice samples from the training set.

Author Contributions: Conceptualization, M.F.-G. and M.P.; data curation, J.L.P.C. and A.R.-R.; formal analysis, J.L.P.C. and D.F.; investigation, J.L.P.C. and A.R.-R.; methodology, J.L.P.C., M.F.-G. and A.R.-R.; resources, D.F. and M.P.; software, J.L.P.C.; supervision, M.F.-G., D.F. and M.P.; validation, A.R.-R.; writing—original draft, J.L.P.C.; writing—review and editing, M.F.-G., D.F. and M.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by: Marsden grant E2987-3648 administrated by the Royal Society of New Zealand, Grant 2017 SGR 622 (GRBIO) administrated by the Departament d'Economia i Coneixement de la Generalitat de Catalunya (Spain), Ministerio de Ciencia e Innovación (Spain) (PID2019-104830RB-I00/ DOI (AEI): 10.13039/501100011033) and by Proyecto Singular AgroMIS. ceiA3 Instrumentos Estratégico hacia un tejido productivo Agroalimentario Moderno, Innovador y Sostenible: motor del territorio rural andaluz. Programa Operativo FEDER 2014-2020 de Andalucía—PAI-TAN-AT2019-AGROMIS-EC.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: José Luis Pérez Calle gratefully thanks the Ministry of Science and Innovation of Spain for a Ph.D. contract under the program FPU (FPU20/03377). The authors are grateful to the Instituto de Investigación Vitivinícola y Agroalimentario (IVAGRO) for providing the necessary facilities to carry out this research. Daniel Fernández is a Serra Húnter Fellow.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. AIJN. Liquid Fruit Market Report | AIJN—PlEuropean Fruit Juice Association. Available online: https://aijn.eu/en/publications/market-reports-1 (accessed on 19 February 2021).
2. Rampersaud, G.C.; Valim, M.F. 100% citrus juice: Nutritional contribution, dietary benefits, and association with anthropometric measures. *Crit. Rev. Food Sci. Nutr.* **2017**, *57*, 129–140. [CrossRef] [PubMed]
3. Rajauria, G.; Tiwari, B.K. Fruit Juices: An Overview. In *Fruit Juices: Extraction, Composition, Quality and Analysis*; Elsevier Inc.: Amsterdam, The Netherlands, 2018; pp. 3–13, ISBN 9780128024911.
4. European Parliament, *European Parliament Directive 2012/12/EU of the European Parliament and of the Council of 19 April 2012 amending Council Directive 2001/112/EC relating to fruit juices and certain similar products intended for human consumption*; European Parliament: Luxembourg, 2012; pp. 1–11.
5. Moore, J.C.; Spink, J.; Lipp, M. Development and Application of a Database of Food Ingredient Fraud and Economically Motivated Adulteration from 1980 to 2010. *J. Food Sci.* **2012**, *77*, R118–R126. [CrossRef] [PubMed]
6. Ammari, F.; Redjdal, L.; Rutledge, D.N. Detection of orange juice frauds using front-face fluorescence spectroscopy and Independent Components Analysis. *Food Chem.* **2015**, *168*, 211–217. [CrossRef] [PubMed]
7. Różańska, A.; Dymerski, T.; Namieśnik, J. Novel analytical method for detection of orange juice adulteration based on ultra-fast gas chromatography. *Mon. Fur Chem.* **2018**, *149*, 1615. [CrossRef] [PubMed]
8. Boggia, R.; Casolino, M.C.; Hysenaj, V.; Oliveri, P.; Zunin, P. A screening method based on UV-Visible spectroscopy and multivariate analysis to assess addition of filler juices and water to pomegranate juices. *Food Chem.* **2013**, *140*, 735–741. [CrossRef]
9. Dasenaki, M.E.; Thomaidis, N.S. Quality and authenticity control of fruit juices-a review. *Molecules* **2019**, *24*, 1014. [CrossRef]
10. Snyder, A.B.; Sweeney, C.F.; Rodriguez-Saona, L.E.; Giusti, M.M. Rapid authentication of concord juice concentration in a grape juice blend using Fourier-Transform infrared spectroscopy and chemometric analysis. *Food Chem.* **2014**, *147*, 295–301. [CrossRef]
11. Shojaee AliAbadi, M.H.; Karami-Osboo, R.; Kobarfard, F.; Jahani, R.; Nabi, M.; Yazdanpanah, H.; Mahboubi, A.; Nasiri, A.; Faizi, M. Detection of lime juice adulteration by simultaneous determination of main organic acids using liquid chromatography-tandem mass spectrometry. *J. Food Compos. Anal.* **2022**, *105*, 104223. [CrossRef]
12. Ellis, D.I.; Ellis, J.; Muhamadali, H.; Xu, Y.; Horn, A.B.; Goodacre, R. Rapid, high-throughput, and quantitative determination of orange juice adulteration by Fourier-transform infrared spectroscopy. *Anal. Methods* **2016**, *8*, 5581–5586. [CrossRef]
13. Chang, J.-D.; Zheng, H.; Mantri, N.; Xu, L.; Jiang, Z.; Zhang, J.; Song, Z.; Lu, H. Chemometrics coupled with ultraviolet spectroscopy: A tool for the analysis of variety, adulteration, quality and ageing of apple juices. *Int. J. Food Sci. Technol.* **2016**, *51*, 2474–2484. [CrossRef]
14. Bononi, M.; Quaglia, G.; Tateo, F. Preliminary LC-IRMS Characterization of Italian Pure Lemon Juices and Evaluation of Commercial Juices Distributed in the Italian Market. *Food Anal. Methods* **2016**, *9*, 2824–2831. [CrossRef]
15. Nuncio-Jáuregui, N.; Calín-Sánchez, Á.; Hernández, F.; Carbonell-Barrachina, Á.A. Pomegranate juice adulteration by addition of grape or peach juices. *J. Sci. Food Agric.* **2014**, *94*, 646–655. [CrossRef] [PubMed]
16. Cristea, G.; Dehelean, A.; Voica, C.; Feher, I.; Puscas, R.; Magdas, D.A. Isotopic and Elemental Analysis of Apple and Orange Juice by Isotope Ratio Mass Spectrometry (IRMS) and Inductively Coupled Plasma–Mass Spectrometry (ICP-MS). *Anal. Lett.* **2021**, *54*, 212–226. [CrossRef]
17. Lorente, J.; Vegara, S.; Martí, N.; Ibarz, A.; Coll, L.; Hernández, J.; Valero, M.; Saura, D. Chemical guide parameters for Spanish lemon (*Citrus limon* (L.) Burm.) juices. *Food Chem.* **2014**, *162*, 186–191. [CrossRef]
18. Dzugan, M.; Wesołowska, M.; Zaguła, G.; Puchalski, C. The comparison of the physicochemical parameters and antioxidant activity of homemade and commercial pomegranate juices. *Acta Sci. Pol. Technol. Aliment.* **2018**, *17*, 59–68. [CrossRef]
19. Liang, Y.L.; Ding, Y.J.; Liu, X.; Zhou, P.F.; Ding, M.X.; Yin, J.J.; Song, Q. hou A duplex PCR–RFLP–CE for simultaneous detection of mandarin and grapefruit in orange juice. *Eur. Food Res. Technol.* **2021**, *247*, 1–7. [CrossRef]
20. Pardo, M.A. Evaluation of a dual-probe real time PCR system for detection of mandarin in commercial orange juice. *Food Chem.* **2015**, *172*, 377–384. [CrossRef]
21. Jahani, R.; Yazdanpanah, H.; van Ruth, S.M.; Kobarfard, F.; Alewijn, M.; Mahboubi, A.; Faizi, M.; Aliabadi, M.H.S.; Salamzadeh, J. Novel application of near-infrared spectroscopy and chemometrics approach for detection of lime juice adulteration. *Iran. J. Pharm. Res.* **2020**, *19*, 34–44. [CrossRef]
22. Dhaulaniya, A.S.; Balan, B.; Yadav, A.; Jamwal, R.; Kelly, S.; Cannavan, A.; Singh, D.K. Development of an FTIR based chemometric model for the qualitative and quantitative evaluation of cane sugar as an added sugar adulterant in apple fruit juices. *Food Addit. Contam. Part A Chem. Anal. Control. Expo. Risk Assess.* **2020**, *37*, 539–551. [CrossRef]
23. Sivakesava, S.; Irudayaraj, J.M.K.; Korach, R.L. Detection of Adulteration in Apple Juice Using Mid Infrared Spectroscopy. *Appl. Eng. Agric.* **2001**, *17*, 815–820. [CrossRef]
24. Kelly, J.F.D.; Downey, G. Detection of sugar adulterants in apple juice using fourier transform infrared spectroscopy and chemometrics. *J. Agric. Food Chem.* **2005**, *53*, 3281–3286. [CrossRef] [PubMed]
25. Jha, S.N.; Gunasekaran, S. Authentication of sweetness of mango juice using Fourier transform infrared-attenuated total reflection spectroscopy. *J. Food Eng.* **2010**, *101*, 337–342. [CrossRef]
26. He, J.; Rodriguez-Saona, L.E.; Giusti, M.M. Midinfrared spectroscopy for juice authentication-rapid differentiation of commercial juices. *J. Agric. Food Chem.* **2007**, *55*, 4443–4452. [CrossRef] [PubMed]

27. Vardin, H.; Tay, A.; Ozen, B.; Mauer, L. Authentication of pomegranate juice concentrate using FTIR spectroscopy and chemometrics. *Food Chem.* **2008**, *108*, 742–748. [CrossRef] [PubMed]
28. Ríos-Reina, R.; Camiña, J.M.; Callejón, R.M.; Azcarate, S.M. Spectralprint techniques for wine and vinegar characterization, authentication and quality control: Advances and projections. *TrAC-Trends Anal. Chem.* **2021**, *134*, 116121. [CrossRef]
29. Martín-Alberca, C.; Ortega-Ojeda, F.E.; García-Ruiz, C. Analytical tools for the analysis of fire debris. A review: 2008–2015. *Anal. Chim. Acta* **2016**, *928*, 1–19. [CrossRef]
30. Falatová, B.; Ferreiro-González, M.; Luis, J.; Calle, P.; Ángel Álvarez, J.; Palma, M. Discrimination of Ignitable Liquid Residues in Burned Petroleum-Derived Substrates by Using HS-MS eNose and Chemometrics. *Sensors* **2021**, *21*, 801. [CrossRef]
31. Lachenmeier, D.W. Rapid quality control of spirit drinks and beer using multivariate data analysis of Fourier transform infrared spectra. *Food Chem.* **2007**, *101*, 825–832. [CrossRef]
32. Pérez Calle, J.L.; Ferreiro González, M.; Ruiz Rodríguez, A.; Fernández Barbero, G.; Álvarez Saura, J.Á.; Palma Lovillo, M.; Ayuso Vilacides, J. A Methodology Based on FT-IR Data Combined with Random Forest Model to Generate Spectralprints for the Characterization of High-Quality Vinegars. *Foods* **2021**, *10*, 1411. [CrossRef]
33. Tôrres, A.R.; de Oliveira, A.D.P.; Grangeiro, S.; Fragoso, W.D. Multivariate statistical process control in annual pharmaceutical product review. *J. Process Control* **2018**, *69*, 97–102. [CrossRef]
34. Tiwari, P.K.; Awasthi, S.; Kumar, R.; Anand, R.K.; Rai, P.K.; Rai, A.K. Rapid analysis of pharmaceutical drugs using LIBS coupled with multivariate analysis. *Lasers Med. Sci.* **2017**, *33*, 263–270. [CrossRef] [PubMed]
35. Dankowska, A.; Kowalewski, W. Tea types classification with data fusion of UV–Vis, synchronous fluorescence and NIR spectroscopies and chemometric analysis. *Spectrochim. Acta-Part A Mol. Biomol. Spectrosc.* **2019**, *211*, 195–202. [CrossRef] [PubMed]
36. Jia, W.; Liang, G.; Tian, H.; Sun, J.; Wan, C. Electronic Nose-Based Technique for Rapid Detection and Recognition of Moldy Apples. *Sensors* **2019**, *19*, 1526. [CrossRef] [PubMed]
37. Shafiee, S.; Minaei, S. Combined data mining/NIR spectroscopy for purity assessment of lime juice. *Infrared Phys. Technol.* **2018**, *91*, 193–199. [CrossRef]
38. Xie, L.J.; Ying, Y. Bin Use of near-infrared spectroscopy and least-squares support vector machine to determine quality change of tomato juice. *J. Zhejiang Univ. Sci. B* **2009**, *10*, 465–471. [CrossRef] [PubMed]
39. Leopold, L.F.; Leopold, N.; Diehl, H.A.; Socaciu, C. Quantification of carbohydrates in fruit juices using FTIR spectroscopy and multivariate analysis. *Spectroscopy* **2011**, *26*, 93–104. [CrossRef]
40. Duarte, I.F.; Barros, A.; Delgadillo, I.; Almeida, C.; Gil, A.M. Application of FTIR spectroscopy for the quantification of sugars in mango juice as a function of ripening. *J. Agric. Food Chem.* **2002**, *50*, 3104–3111. [CrossRef]
41. Bureau, S.; Ruiz, D.; Reich, M.; Gouble, B.; Bertrand, D.; Audergon, J.M.; Renard, C.M.G.C. Application of ATR-FTIR for a rapid and simultaneous determination of sugars and organic acids in apricot fruit. *Food Chem.* **2009**, *115*, 1133–1140. [CrossRef]
42. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*; Springer-Verlag: Berlin, Germany, 2016; ISBN 978-3-319-24277-4.
43. R Core Team. *R: A Language and Environment for Statistical Computing*; R Core Team: Vienna, Austria, 2020.
44. Kuhn, M. Caret: Classification and Regression Training. 2020. Available online: https://cran.r-project.org/web/packages/caret/caret.pdf (accessed on 18 December 2021).
45. Chang, W.; Cheng, J.; Allaire, J.J.; Xie, Y.; McPherson, J. Shiny: Web Application Framework for R. 2020; Available online: https://cran.r-project.org/web/packages/shiny/index.html (accessed on 19 February 2021).
46. IR Spectrum Table & Chart | Sigma-Aldrich. Available online: https://www.sigmaaldrich.com/technical-documents/articles/biology/ir-spectrum-table.html (accessed on 16 May 2021).
47. Li, J.; Zhang, C.; Liu, H.; Liu, J.; Jiao, Z. Profiles of Sugar and Organic Acid of Fruit Juices: A Comparative Study and Implication for Authentication. *J. Food Qual.* **2020**, *2020*, 7236534. [CrossRef]
48. Zhang, H.; Xie, Y.; Liu, C.; Chen, S.; Hu, S.; Xie, Z.; Deng, X.; Xu, J. Comprehensive comparative analysis of volatile compounds in citrus fruits of different species. *Food Chem.* **2017**, *230*, 316–326. [CrossRef] [PubMed]
49. Amakura, Y.; Okada, M.; Tsuji, S.; Tonogai, Y. Determination of phenolic acids in fruit juices by isocratic column liquid chromatography. *J. Chromatogr. A* **2000**, *891*, 183–188. [CrossRef]
50. Géron, A. *Hands-On Machine Learning with Scikit-Learn and TensorFlow*, 2nd ed.; Roumeliotis, R.N.T., Ed.; O'Reilly Media, Inc.: Newton, MA, USA, 2019.
51. Hastie, T.; Tibshirani, R.; Friedman, J.H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: New York, NY, USA, 2009; ISBN 0387848576.
52. Rodriguez-Galiano, V.F.; Ghimire, B.; Rogan, J.; Chica-Olmo, M.; Rigol-Sanchez, J.P. An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS J. Photogramm. Remote Sens.* **2012**, *67*, 93–104. [CrossRef]
53. Rasekh, M.; Karami, H. Application of electronic nose with chemometrics methods to the detection of juices fraud. *J. Food Process. Preserv.* **2021**, *45*, e15432. [CrossRef]
54. Qiu, S.; Wang, J.; Gao, L. Discrimination and characterization of strawberry juice based on electronic nose and tongue: Comparison of different juice processing approaches by LDA, PLSR, RF, and SVM. *J. Agric. Food Chem.* **2014**, *62*, 6426–6434. [CrossRef] [PubMed]
55. Downey, G.; Kelly, J.D.; León, L. Detection of Apple Juice Adulteration Using Near-Infrared Transflectance Spectroscopy. *Appl. Spectrosc.* **2005**, *59*, 593–599.

56. Moura, H.O.M.A.; Câmara, A.B.F.; Santos, M.C.D.; Morais, C.L.M.; de Lima, L.A.S.; Lima, K.M.G.; de Carvalho, L.S. Advances in chemometric control of commercial diesel adulteration by kerosene using IR spectroscopy. *Anal. Bioanal. Chem.* **2019**, *411*, 2301–2315. [CrossRef]
57. Leng, T.; Li, F.; Chen, Y.; Tang, L.; Xie, J.; Yu, Q. Fast quantification of total volatile basic nitrogen (TVB-N) content in beef and pork by near-infrared spectroscopy: Comparison of SVR and PLS model. *Meat Sci.* **2021**, *180*, 108559. [CrossRef]
58. Sugami, Y.; Minami, E.; Saka, S. Renewable diesel production from rapeseed oil with hydrothermal hydrogenation and subsequent decarboxylation. *Fuel* **2016**, *166*, 376–381. [CrossRef]
59. Borin, A.; Ferrão, M.F.; Mello, C.; Maretto, D.A.; Poppi, R.J. Least-squares support vector machines and near infrared spectroscopy for quantification of common adulterants in powdered milk. *Anal. Chim. Acta* **2006**, *579*, 25–32. [CrossRef] [PubMed]