Hindawi
Journal of Spectroscopy
Volume 2021, Article ID 5845422, 9 pages
https://doi.org/10.1155/2021/5845422

Research Article

Comparison of Machine Learning Classification Methods for Determining the Geographical Origin of Raw Milk Using Vibrational Spectroscopy

Aimen El Orche,[1] **Amine Mamad**,[2] **Omar Elhamdaoui**,[2] **Amine Cheikh**,[3] **Miloud El Karbane**,[2] **and Mustapha Bouatia**[2]

[1]*Team of Analytical and Computational Chemistry,Nanotechnology and Environment, Faculty of Sciences and Techniques, University of Sultan Moulay Slimane, Beni Mellal, Morocco*
[2]*Laboratory of Analytical Chemistry, Faculty of Medicine and Pharmacy, Mohammed V University, Rabat, Morocco*
[3]*Faculty of Medicine, Abulcasis University, Rabat, Morocco*

Correspondence should be addressed to Aimen El Orche; aimen.elorche@gmail.com

Received 1 October 2021; Revised 22 November 2021; Accepted 26 November 2021; Published 8 December 2021

Academic Editor: Ana Domi nguez Vidal

One of the significant challenges in the food industry is the determination of the geographical origin, since products from different regions can lead to great variance in raw milk. Therefore, monitoring the origin of raw milk has become very relevant for producers and consumers worldwide. In this exploratory study, midinfrared spectroscopy combined with machine learning classification methods was investigated as a rapid and nondestructive method for the classification of milk according to its geographical origin. The curse of dimensionality makes some classification methods struggle to train efficient models. Thus, principal component analysis (PCA) has been applied to create a smaller set of features. The application of machine learning methods such as PLS-DA, PCA-LDA, SVM, and PCA-SVM demonstrates that the best results are obtained using PLS-DA, PCA-LDA, and PCA-SVM methods which show a correct classification rate (CCR) of 100% for PLS-DA and PCA-LDA and 94.95% for PCA-SVM, whereas the application of SVM without feature extraction gives a low CCR of 66.67%. These findings demonstrate that FT-MIR spectroscopy, combined with machine learning methods, is an efficient and suitable approach to classify the geographical origins of raw milk.

1. Introduction

Consumers are increasingly demanding guarantees concerning the quality and safety of food products, especially when they are of animal origin. Food authentication consists of checking that the product is consistent with the statements made on the label [1]. Falsification or willful mislabeling is usually used to reduce production costs [2]. In the milk industry especially, mislabeling can be used to confuse the industry and consumers about the origin of the milk, since products of different origins can have different qualities [1].

The European Union (EU) promotes dairy product quality programs designed to support farmers and safeguard their product names from abuse and imitation [3]. Particularly, the EU promotes two principal quality regimes that are based on the valuation of the geographical origins of foods, called Protected Designation of Origin (PDO) and Protected Geographical Indication (PGI), which respectively identify food products that are produced or closely associated with a given geographical area [4].

Food scientists support such programs by developing analytical techniques to enhance the capacity to identify the geographic origin of food [5]. These techniques are classified into two types: those based on targeted approaches and those based on untargeted approaches [6].

Targeted approaches are usually the most suitable for regulatory purposes, as they are very specific and sensitive

[7]. These approaches are based on a screening that looks for a shortlist of predetermined chemical compounds [7]. Among these techniques, we can find the following: capillary electrophoresis (CE) [8], high-performance liquid chromatography (HPLC) [9], and gas chromatography (GC) [10, 11]. In addition to their advantages, these analytical methods have certain inconveniences as they are time-consuming and may require the use of expensive and polluting reagents. Additionally, these methods are not effective to cover the increased need for an analytical workflow that takes several hours [12]. While the nontargeted approaches allow the identification of the unknowns [13]. These nontargeted analyses are therefore increasingly gaining importance in the food sector, such as nontargeted metabolomics and spectroscopy [6]. This evolution is also explained by the increasing attention to very complex authentication issues such as geographical origin or agricultural techniques, thus increasing the need for innovative analytical approaches [6, 14]. Among these methods, we find spectroscopic methods [15], such as infrared spectroscopy, UV-Visible, Raman, and fluorescence spectroscopy [12, 16, 17]. Food analysis using spectroscopic techniques has become very common and widespread since these methods are extremely fast, inexpensive, nondestructive, and have no negative impact on the environment [18, 19]. The nontargeted analysis is a very difficult task, as it requires thorough processing of the generated data set. In order to make these data meaningful, multistep strategies using chemometric tools are needed before the eventual identification of a particular signal among a forest of interfering signals [20].

In order to authenticate the milk, several spectroscopic techniques have been used, such as near-infrared spectroscopy (NIR) [21], midinfrared spectroscopy (MIR) [22–25], Raman spectroscopy [1, 26], and fluorescence spectroscopy [27, 28]. In these studies, quantitative chemometric approaches were used for the detection of adulteration, prediction of some quality parameters, and classification of milk according to species and origin.

To the best of our knowledge, there is only one study concerning the determination of the geographical origin of milk by means of MIR spectroscopy [25]. This study consists of studying the capacity of the fatty acid composition and the spectral information obtained by MIR spectroscopy for the discrimination between sheep milk coming from different geographical areas of Italy. In this study, the spectral results and the chemical composition (fatty acid) are processed by a genetic algorithm.

In this study, we develop several methods based on midinfrared spectroscopy, combined with machine learning, to classify the geographical origins of milk from 4 regions in Morocco. In detail, PLS-DA has been studied to simplify the feature extraction process while ensuring the precision and accuracy of the prediction. In addition, LDA and SVM were also applied to develop a set of classifiers with the spectral features selected by principal component analysis (PCA). The results provide a new insight and an attempt to apply MIR spectroscopy and machine learning methods in food and agri-food applications, and also to show the importance of using PCA as a preliminary method of feature reduction for building accurate, sensitive and specific classification models.

2. Materials and Methods

2.1. Sampling. This study was conducted on the raw milk of cows coming from various regions of Morocco (region 1: Mechra Bel Ksiri, region 2: Souk Larbaa, region 3: Maaziz, and region 4: Tiflet). The samples were collected during the year 2020. Samples are frozen immediately after collection in a cooler stacked with ice packs on their way to the freezer where they are stored until the day of analysis.

150 raw milk samples were selected to conduct this study; region 1 = 36, region 2 = 36, region 3 = 42, and region 4 = 36. In order to build a classification and prediction model, 100 samples are used for model training and 50 samples are used for model validation.

2.2. Spectral Acquisition. Fourier-transformed midinfrared spectra of cow's milk samples were recorded on a JASCO FTIR 460 PLUS spectrometer (PIKE Technologies, Madison, USA) in the spectral region between 600 and 4000 cm^{-1}. The instrumental resolution was 4 cm^{-1}, and each spectrum was composed of 3400 data points. Using a micropipette, each milk sample was placed on the crystalline surface of the ATR, which was cleaned for each analysis using the acetone solution, allowing both cleaning and drying of the ATR accessory. The spectra obtained were processed with the software (Spectra manager) in order to eliminate the effect of carbon dioxide in the three corresponding bands (at 2349 cm^{-1}, 1388 cm^{-1}, and 667 cm^{-1}) as well as the effect of moisture (at 3756 cm^{-1}, 3652 cm^{-1}, and 1595 cm^{-1}).

2.3. Data Analysis. In order to adequately process the spectral data obtained by midinfrared spectroscopy, multivariate data analysis (chemometrics) was used to explore the data and build classification models. The multivariate analysis became a major component of analytical chemistry. This is related to the necessity for computational approaches able to extract relevant information from increasing amounts of data provided by modern analytical instruments. Generally, these multivariate data analysis approaches concern the exploratory analysis of a single data matrix, such as principal component analysis (PCA), or the matching of an explanatory matrix to another descriptive matrix, as in regression methods such as PLS, or discriminant methods such as linear discriminant analysis (LDA), partial least squares discriminant analysis (PLS-DA), and support vector machine (SVM).

PCA is among the most commonly used methods in chemometrics since it allows to answer many objectives, such as visualization of data, detection of outliers, investigating the similarity between individuals, and the correlation between variables [29]. PCA is an unsupervised method for feature reduction that allows high-dimensional data to be projected into a new reduced-dimensional representation of the data that describes the variance of the data as much as

possible with minimal reconstruction error [30]. This method provides a new set of variables, known as principal components. Each principal component represents a linear combination of the initial variables. All principal components are mutually orthogonal; therefore, there is no redundant information. The principal components together form an orthogonal database of datasets [31]. These databases can also be used as variables for other multivariate methods.

We can also find the supervised classification methods. These methods use the membership of the samples of different classes to build a model [32], such as PLS-DA, LDA, and SVM.

The PLS-DA uses the PLS method to explain and predict the membership of individuals to several classes, based on quantitative or qualitative explanatory variables [33]. PLS-DA is a relatively new technique in chemometrics that extends and merges the functionality of principal component analysis and multiple regression [12, 34, 35]. This technique consists of performing a decomposition of the two matrices, matrix of variables X and response (Y), under the condition that the factorial coordinates extracted from X should be correlated as much as possible with the factorial coordinates extracted from Y.

The main objective of linear discriminant analysis (LDA) is to be able to classify new individuals not belonging to the initial data. The idea is based on a method looking for a linear combination of the variables Xj that maximizes the similarity between the elements within the same group [36]. In other words, it is a question of finding a linear combination of Xj that maximizes the inertia or the intergroup variance and, therefore, the one that minimizes the intragroup variance. It consists in explaining and predicting the membership of an individual to a predefined class (group) from his/her characteristics measured by means of predictive variables. For large datasets such as in image recognition and spectral data, linear lines often do not allow a good separation of the groups because LDA is not the ideal method in cases where the explanatory variables are highly correlated. In such situations, it is necessary to regularize it in order to disrupt the correlation of the predictors to obtain better results. To overcome these limitations, there are other methods to extend the LDA in order to have a better classification, especially the combination of this method with other methods of variable reduction such PCA and genetic algorithm.

SVMs are a family of machine learning algorithms that solve classification, regression, and anomaly detection problems [37]. They are known for their strong theoretical guarantees, their great flexibility, and their simplicity of use, even without much knowledge of data mining. Its principle is simple: it aims at separating the data into classes using a boundary that is as simple as possible so that the distance between the different groups of data and the boundary that separates them is maximal [38]. This distance is also called "margin," and SVMs are thus called "wide margin separators," the "support vectors" being the data closest to the border. This notion of frontier assumes that the data are linearly separable, which is rarely the case. To overcome this, SVMs often rely on the use of "kernels" [39]. These mathematical functions allow to separate the data by projecting them in a feature space [32]. The technique of margin maximization allows us to guarantee better robustness to noise and therefore a more generalizable model.

These classification approaches struggle to build efficient models when the size of the spectral data set is very high, the so-called "curse of dimensionality," since the redundancy of spectral variables affects the classification results of conventional machine learning models [40–42]. This is particularly pertinent for algorithms that use distance calculations, such as LDA and SVM [41]. Feature extraction is the most crucial step to overcome the curse of dimensionality through the creation of a smaller set of features, so spectral features were extracted using PCA as a reduction variable method [40]. This is the case in this study, in which a preliminary treatment by PCA is used to extract synthetic variables used for the construction of classification models.

In classification issues, accuracy is commonly given as an evaluation metric. However, if there are more than two categories, accuracy on its own can be misleading and will not provide reliable information. Exploiting the confusion matrix obtained by the classification methods provides more information about their performance. The confusion matrix provides insight into the types of errors committed during estimation. As a result, it shows which points are correctly classified and which are misclassified.

In the present study, accuracy, sensitivity, and specificity parameters were used to compare the classification performance of different methods employed. Accuracy evaluates the efficiency of the algorithm by displaying the probability of the true value of the class target. For our purposes, accuracy is the number of correct predictions of geographical regions in relation to the total number of predictions. However, we denote sensitivity as the proportion of positive events that are well classified and specificity as the proportion of negative events that are well classified. These parameters are calculated according to the following formulas:

$$\begin{aligned} \text{accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TN} + \text{TP} + \text{FP} + \text{FN}}, \\ \text{specificity} &= \frac{\text{TN}}{\text{TN} + \text{FP}}, \\ \text{sensitivity} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \end{aligned} \tag{1}$$

where FN, FP, TN, and TP designate false negatives, false positives, true negatives, and true positives, respectively.

The flowchart of the main procedures applied to build the classification models is presented in Figure 1.

2.4. Software. PLS-DA models were built on the basis of the partial least squares algorithm using the NIPALS algorithm (nonlinear iterative partial least squares). PCA, PLS-DA, LDA, and SVM methods were applied using the Unscrambler software 10.4.

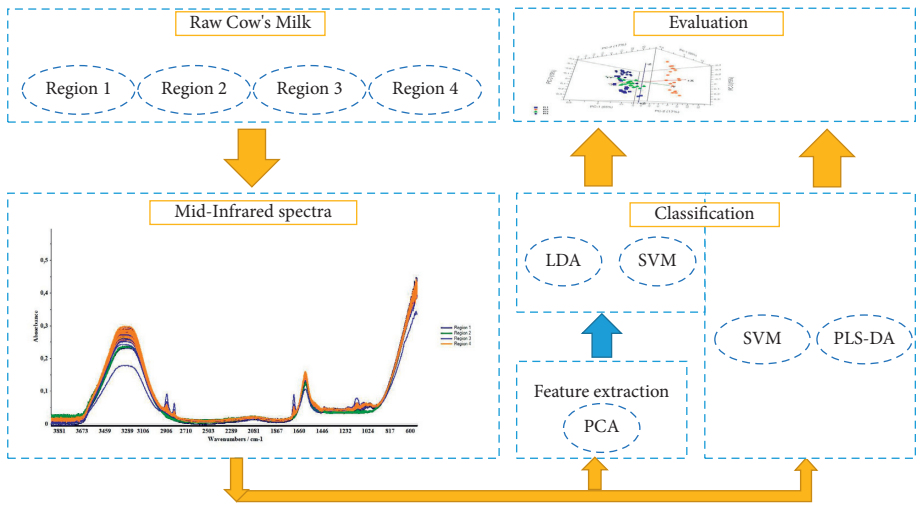

Figure 1: Principal steps employed for the classification of the geographical origin of raw milk.

3. Results and Discussion

3.1. FT-MIR Spectra of Milk Samples. The spectra of the milk obtained (Figure 2) show the presence of spectral bands of interest. A broadband between 3700 and 3100 cm^{-1} corresponds to stretching of –OH and –NH in proteins, 3000–2800 cm^{-1} coincides with C–H stretching in fatty acids, 1.750–1.650 cm^{-1} corresponds to –C=O of fatty acids and esters, and a band between 1,660 and 1,446 cm^{-1} corresponds to –C=O and –NH of the I and II amides of the proteins arising from various combinations of vibrations in the peptide linkages and secondary structure of the casein protein. Amide I vibration is mainly due to stretching of C=O bonds, and amide II vibration is due to deformation of N–H bonds and stretching of C–*N* bonds. The amide I vibration is measured in the region of 1660–1550 cm^{-1} and the amide II vibration in the region of 1550–1446 cm^{-1}. Other small bands were observed in the spectral zone 1200–800 cm^{-1}; this region corresponds to the stretching –C=O of polysaccharides and C=C of acids [43, 44].

These four groups of samples show slight differences in the band of 3000–2800 cm^{-1}; 1,750–1,650 cm^{-1}; and 1,660–1,446 cm^{-1}, which mainly represents the absorption of substances, such as proteins, fatty acids, and esters.

However, it is still difficult to directly categorize the geographical origins owing to these minor differences. Consequently, it is necessary to study the classification model in order to help to identify the geographical origins of raw milk.

To further describe the data in a very small dimensional space, a PCA was first performed on the milk spectra to exploit the data set and to obtain information about the distribution and behavior of the samples regarding the measured variables that represent the wavenumber of the MIR spectral data.

3.2. Identification of the Geographic Origin of Raw Milk

3.2.1. Principal Component Analysis. PCA was applied to raw spectral data to visualize the samples in a well-reduced space in order to get a clear view of the data distribution. PCA shows that the first three components account for 97% (PC1: 65%, PC2:17%, and PC3:15%) of the total variability contained in datasets. From Figure 3, which represents the configuration of the samples on the first three PCs, we can distinguish a grouping of the samples according to their geographical origin, indicating that the samples belonging to each group had similar FTIR properties. However, the PCA plot shows that some milk samples from different regions exhibited high overlapping due to the same compositional properties. However, the first three PCs are not sufficient to distinguish between R1, R3, and R4. For this reason, more than 3 components are selected and used as feature variables for the construction of classification models such as PLS-DA, LDA, and SVM.

3.3. Discrimination Analysis. In order to build a model able to discriminate and predict the membership of milk according to their geographical origin. In this study, three different supervised classification algorithms, including PLS-DA, LDA, and SVM, were investigated to classify different geographical origins of raw milk. For the PLS-DA and LDA, feature extraction is used directly, since these methods cannot be applied directly on spectral data without feature extraction of latent variables. While the SVM is applied without and with feature extraction of latent variables by PCA, in order to demonstrate the importance of feature extraction variables in improving classification performance, finally, the classification models were compared.

(i) Partial Least Square Discrimination Analysis. PLS-DA analysis was applied over the entire MIR spectral region (4000 cm^{-1} and 600 cm^{-1}) without spectral preprocessing and with spectral preprocessing. The choice of the optimal number of latent variables required for a good classification and prediction is made on the basis of the values obtained by the cross-validation procedure, namely, the root mean square error and R-square which are calculated on the basis of the leave one out algorithm [45]. In our case, 12 latent variables were selected.

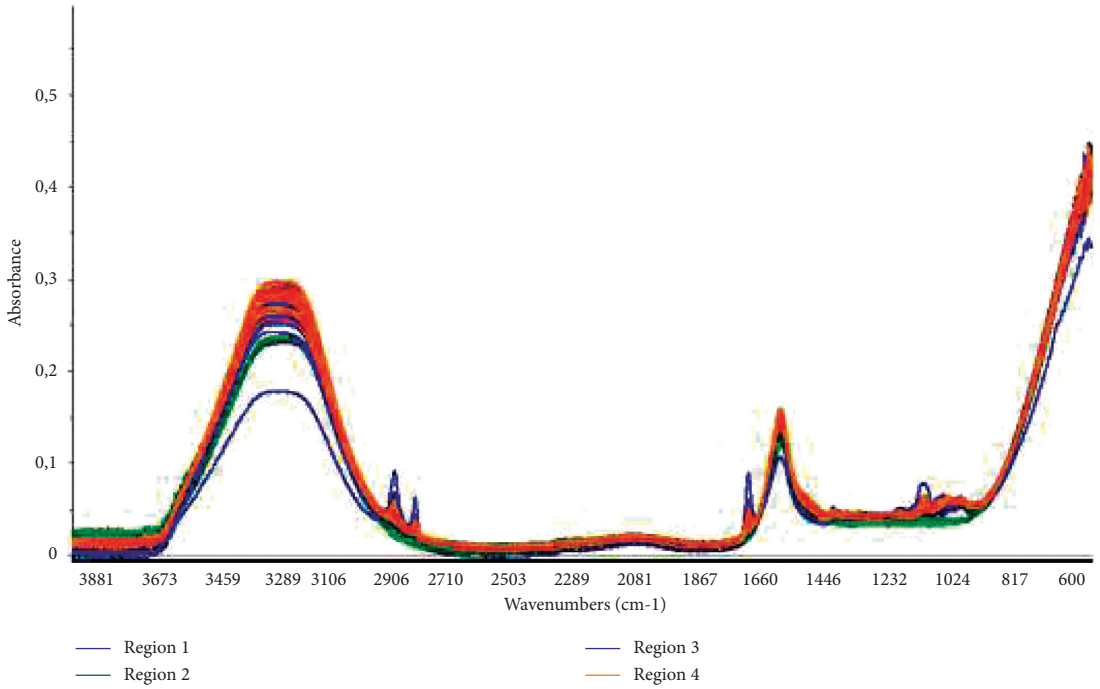

Figure 2: Midinfrared spectra of different cow's milk.

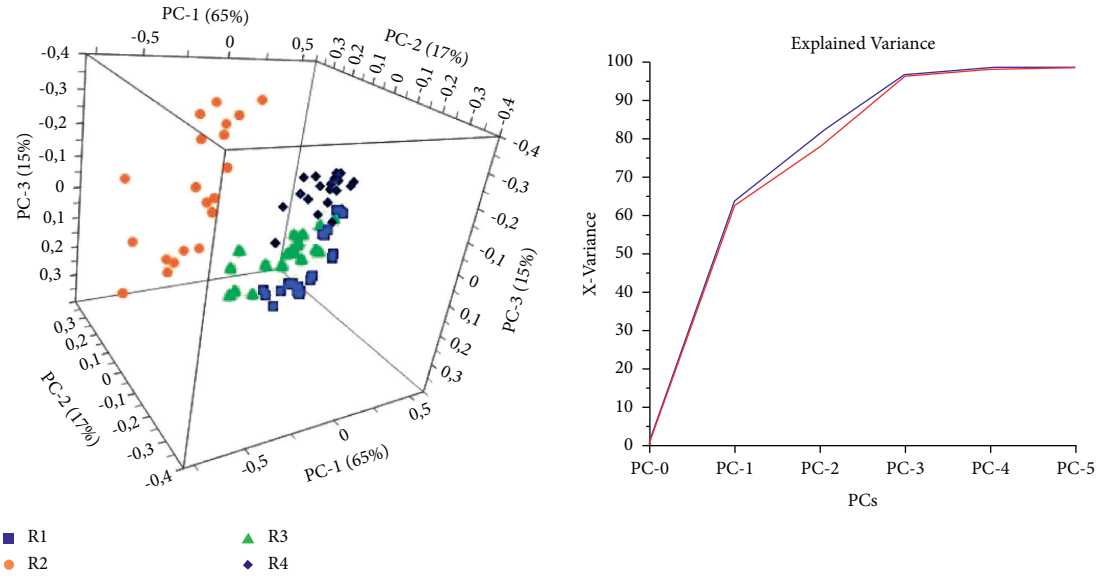

Figure 3: Scores plot of the PCA model with three principal components PC1-PC2-PC3 and a graphic of the explained variance of each principal component.

Based on the results in Table 1 obtained by the PLS-DA method, we conclude that the PLS-DA method provides good results for the classification of milk according to their geographical origin, these performances are represented by the high value of the *R*-square and the low value of the RMSE for both calibration and cross-validation results. These results also show that the spectral data preprocessed by the detrend algorithm using polynomial degree 1 gives the lowest value of RMSE and the highest value of *R*-square.

(ii) Linear Discriminant Analysis. Subsequently, the PCA-LDA model was developed to classify and predict milk according to its geographical origin. LDA was applied on the first 5 components obtained by PCA using the statistical method of Mahalanobis. In our case, the application of the LDA analysis directly on the spectral data cannot be done because the number of spectral variables is higher than the number of samples.

Based on the results obtained by PCA-LDA, we can conclude that this method gives powerful results for the classification of milk. This classification ability is represented by the high values obtained for sensitivity, specificity, and classification correct rate % (CCR), as shown in Table 2. It was clear that all the built models provide high sensitivity (otherwise

Table 1: Statistical results obtained by the PLS-DA method.

Preprocessing data	Number of LVs	Groups	$R^2_{Calibrarion}$	RMSEC	$R^2_{Cross-validation}$	RMSECV	Sensitivity (%)	Specificity (%)	% CCR
Raw	12	R 1	0.98	0.056	0.97	0.069	100	100	100
		R 2	0.99	0.036	0.99	0.041	100	100	
		R 3	0.95	0.10	0.81	0.200	100	100	
		R 4	0.93	0.11	0.79	0.200	100	100	
Detrend polynomial 1	13	R 1	0.98	0.055	0.97	0.070	100	100	100
		R 2	0.99	0.034	0.99	0.040	100	100	
		R 3	0.97	0.075	0.83	0.183	100	100	
		R 4	0.97	0.071	0.83	0.178	100	100	
Detrend polynomial 2	13	R 1	0.98	0.056	0.97	0.070	100	100	100
		R 2	0.99	0.034	0.99	0.041	100	100	
		R 3	0.97	0.075	0.84	0.185	100	100	
		R 4	0.97	0.072	0.84	0.182	100	100	

Table 2: Statistical parameters obtained by the PCA-LDA method.

Confusion matrix		Actual				Sensitivity (%)	Specificity (%)	% CCR
		R1	R2	R3	R4			
Raw	R1	24	0	0	0	100	100	99.5
	R2	0	24	0	0	100	100	
	R3	0	0	28	1	100	100	
	R4	0	0	0	24	100	98.7	
Detrend polynomial 1	R1	24	0	0	0	100	100	100
	R2	0	24	0	0	100	100	
Predicted	R3	0	0	28	0	100	100	
	R4	0	0	0	25	100	100	
Detrend polynomial 2	R1	24	0	0	0	100	100	98.5
	R2	0	24	0	0	100	100	
	R3	0	0	28	3	100	100	
	R4	0	0	0	22	100	96.2	

known as the true-positive rate), as demonstrated by the high value that reaches 100% for the three models. These models also show a high specificity, 99% for the model built on raw data, 100% for the model built on preprocessed data by detrend polynomial degree 1, and 98.5% for the model built on preprocessed data by detrend polynomial degree 2.

(iii) Support Vector Machine. The SVM method has been applied to the spectroscopic data with and without feature extraction, using the radial basis function method. The use of the SVM method without extraction of characteristics shows a CCR of 66.97, 69.68, and 67.97, respectively, for raw spectral data, spectral data corrected with detrend using a polynomial of orders 1 and 2, as shown in Table 3. According to the confusion matrix obtained by this method, we can observe that not all samples belong to their class for a certain group. This method shows an important specificity, i.e., a good capacity for the good classification of negative events. Therefore, low sensitivity has been observed for this method, i.e., low classification of positive events.

However, the application of the SVM method on the first 5 components obtained by PCA shows good results represented by the high values of CCR which reach 98.51, 98.49, and 99.00 using features extracted from spectral data without spectral preprocessing and spectral data preprocessed by detrend using polynomials of degree 1 and 2, respectively, as shown in Table 3. This method shows a sensitivity of 100%, which verifies the high capacity of this statistical approach for the prediction of correct events as well as correct events, and a high specificity ranging between 96.2% and 100%.

3.4. Validation of Classification Models. In order to evaluate the classification and predictive capacity of the constructed models, external validation of the models was carried out using different samples to those used for the construction of the models.

In the case of PLS-DA, the predicted y-value of a given sample is close to 1 (or greater than 0.5) and allocates the sample to a specific category, while a sample with a predicted y-value of less than 0.5 allocates it outside the category.

The results found by the external validation of the machine learning classification models (Table 4) show that the best results are provided by applying PLS-DA, PCA-LA, and then PCA-SVM which shows a strong classification capacity represented by the high values of CCR, specificity, and sensitivity which reaches 100% for PLS-DA and PCA-LDA and a CCR of 94.95 for PCA-SVM, while the application of the SVM method without feature extraction gives a CCR of 66.67% which is considered unsatisfactory for the corresponding classification of samples according to their

Table 3: Statistical parameters obtained by the SVM and PCA-SVM methods.

Confusion matrix (SVM)			Actual				Sensitivity (%)	Specificity (%)	% CCR
			R1	R2	R3	R4			
Raw	Predicted	R1	22	0	8	15	48.89	96.3	66,97
		R2	0	24	0	0	100	100	
		R3	2	0	20	2	46.51	87.1	
		R4	0	0	0	8	100	79,52	
Detrend polynomial 1		R1	0	0	0	0	0	69,23	69,68
		R2	0	24	0	0	100	100	
		R3	24	0	28	23	100	100	
		R4	0	0	0	2	100	69,33	
Detrend polynomial 2		R1	0	0	0	0	0	68,42	67,97
		R2	0	24	0	0	100	100	
		R3	24	0	28	25	100	100	
		R4	0	0	0	0	0	67,53	
Confusion matrix (PCA-SVM)			Actual				Sensitivity (%)	Specificity (%)	% CCR
			R1	R2	R3	R4			
Raw	Predicted	R1	24	0	0	0	100	98,68	98,51
		R2	0	24	0	0	100	100	
		R3	1	0	28	2	100	100	
		R4	0	0	0	23	100	97,44	
Detrend polynomial 1		R1	24	0	0	0	100	100	98,49
		R2	0	24	0	0	100	100	
		R3	0	0	28	3	100	100	
		R4	0	0	0	22	100	96,2	
Detrend polynomial 2		R1	24	0	0	0	100	100	99
		R2	0	24	0	0	100	100	
		R3	0	0	28	2	100	100	
		R4	0	0	0	23	100	97,44	

Table 4: Performance parameters obtained for the validation of the PLS-DA and PCA-LDA models and SVM and PCA-SVM models constructed using FT-MIR spectral data preprocessed by detrend polynomial degree 1.

Confusion matrix			Actual				Sensitivity (%)	Specificity (%)	% CCR
			R1	R2	R3	R4			
PLS-DA	Predicted	R1	12	0	0	0	100	100	100
		R2	0	12	0	0	100	100	
		R3	0	0	14	0	100	100	
		R4	0	0	0	12	100	100	
PCA-LDA		R1	12	0	0	0	100	100	100
		R2	0	12	0	0	100	100	
		R3	0	0	14	0	100	100	
		R4	0	0	0	12	100	100	
SVM		R1	0	0	0	0	0	68,42	66,67
		R2	1	12	1	0	85,71	100	
		R3	11	0	13	11	100	92,86	
		R4	0	0	0	1	100	69,44	
PCA-SVM		R1	11	0	0	0	100	97,3	94,95
		R2	0	12	0	0	100	100	
		R3	1	0	12	0	100	94,6	
		R4	0	0	2	12	85,71	100	

membership. These results show the usefulness of using PCA as a first step for classification methods. Applying this method as a primary step for SVM analysis improves considerably the classification performance as shown by the results.

4. Conclusion

The present work constitutes an exploratory investigation on the use of vibrational spectroscopy for milk samples in Morocco. In which MIR spectroscopy combined with machine learning methods has been proposed to quickly identify the geographical origin of milk. After sample collection, the MIR spectra of raw milk can be rapidly acquired. Then, the obtained MIR spectral data are modeled using four classification methods (including PLS-DA, PCA-LDA, SVM, and PCA-SVM). Then, acceptable classification results are provided by these classification approaches. Compared with the PLS-DA (CCR = 100%), the PCA-LDA and PCA-SVM methods obtained more accurate and reliable classification

results and were able to identify raw milk samples from different geographical origins with a CCR of 96% and 100% for the test set. In addition, satisfactory sensitivity and specificity were found reflecting the performance of these classification approaches. On the other hand, the application of the SVM method without feature extraction gives a low CCR of 65%. These results also prove that the use of PCA as a preliminary method for machine learning methods improves the classification performance by extracting features and reducing the data size.

The suggested exploratory approach based on MIR spectroscopy combined with machine learning methods proved to be an efficient strategy for the dairy industry to identify the geographical origin of raw milk. The above results confirm that the proposed method can be considered as a promising alternative for determining geographical origin.

Data Availability

The data are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

The authors would like to thank all the people who contributed to make this work.

References

[1] N. N. Yazgan, H. E. Genis, T. Bulat et al., "Discrimination of milk species using Raman spectroscopy coupled with partial least squares discriminant analysis in raw and pasteurized milk," *Journal of the Science of Food and Agriculture*, vol. 100, no. 13, pp. 4756–4765, 2020.

[2] R. Johnson, *Food Fraud and "Economically Motivated Adulteration" of Food and Food Ingredients*, Congressional Research Service, Washington, DC, USA, 2014.

[3] A. Tosato, "The protection of traditional foods in the eu: traditional specialities guaranteed," *European Law Journal*, vol. 19, 2021, https://onlinelibrary.wiley.com/doi/abs/10.1111/eulj.12040.

[4] M. Scampicchio, D. Eisenstecken, L. De Benedictis et al., "Multi-method approach to trace the geographical origin of alpine milk: a case study of tyrol region," *Food Analytical Methods*, vol. 9, no. 5, pp. 1262–1273, 2016.

[5] K. Katerinopoulou, A. Kontogeorgos, C. E. Salmas, A. Patakas, and A. Ladavos, "Geographical origin authentication of agri-food products: a review," *Foods*, vol. 9, no. 4, p. 489, 2020.

[6] N. Z. Ballin and K. H. Laursen, "To target or not to target? Definitions and nomenclature for targeted versus non-targeted analytical food authentication," *Trends in Food Science & Technology*, vol. 86, pp. 537–543, 2019.

[7] M. Pourchet, L. Debrauwer, J. Klanova et al., "Suspect and non-targeted screening of chemicals of emerging concern for human biomonitoring, environmental health studies and support to risk assessment: from promises to challenges and harmonisation issues," *Environment International*, vol. 139, Article ID 105545, 2020.

[8] T. de Oliveira Mendes, B. L. S. Porto, M. J. V. Bell, Í. T. Perrone, and M. A. L. de Oliveira, "Capillary zone electrophoresis for fatty acids with chemometrics for the determination of milk adulteration by whey addition," *Food Chemistry*, vol. 213, pp. 647–653, 2016.

[9] N. Bernardi, G. Benetti, N. M. Haouet et al., "A rapid high-performance liquid chromatography-tandem mass spectrometry assay for unambiguous detection of different milk species employed in cheese manufacturing," *Journal of Dairy Science*, vol. 98, no. 12, pp. 8405–8413, 2015.

[10] R. Gutiérrez, S. Vega, G. Díaz et al., "Detection of non-milk fat in milk fat by gas chromatography and linear discriminant analysis," *Journal of Dairy Science*, vol. 92, pp. 1846–1855, 2009.

[11] J. A. Park, N. K. Kim, C. Y. Yang, K. W. Moon, and J. M. Kim, "Determination of the authenticity of dairy products on the basis of fatty acids and triacylglycerols content using gc analysis," *Korean Journal of Food Science of Animal Resources*, vol. 34, 2021, https://koreascience.or.kr/article/JAKO201420249946311.

[12] A. El Orche, M. Bouatia, and M. Mbarki, "Rapid analytical method to characterize the freshness of olive oils using fluorescence spectroscopy and chemometric algorithms," *Journal of Analytical Methods in Chemistry*, vol. 2020, pp. 1–9, 2020.

[13] I. Dom, R. Biré, V. Hort, G. Lavison-Bompard, M. Nicolas, and T. Guérin, "Extended targeted and non-targeted strategies for the analysis of marine toxins in mussels and oysters by (LC-HRMS)," *Toxins*, vol. 10, no. 9, p. 375, 2018.

[14] M. M. Oliveira, J. p. Cruz-Tirado, and D. F. Barbin, "Non-targeted analytical methods as a powerful tool for the authentication of spices and herbs: a review," *Comprehensive Reviews in Food Science and Food Safety*, vol. 18, no. 3, pp. 670–689, 2019.

[15] T. F. McGrath, S. A. Haughey, J. Patterson et al., "What are the scientific challenges in moving from targeted to non-targeted methods for food fraud testing and how can they be addressed? - spectroscopy case study," *Trends in Food Science & Technology*, vol. 76, pp. 38–55, 2018.

[16] O. Elhamdaoui, A. El Orche, A. Cheikh, B. Mojemmi, R. Nejjari, and M. Bouatia, "Development of fast analytical method for the detection and quantification of honey adulteration using vibrational spectroscopy and chemometrics tools," *J. Anal. Methods Chem.*vol. 2020, Article ID e8816249, 9 pages, 2020.

[17] A. El Orche, M. Bouatia, S. Yanisse et al., "Evaluation of the capability of horizontal ATR-FTMIR and UV-visible spectroscopy in the discrimination of virgin olive oils from the Moroccan region of beni mellal-khenifra," *Journal of Spectroscopy*, vol. 2020, pp. 1–9, 2020.

[18] A. Nawrocka and J. Lamorska, "Determination of food quality by using spectroscopic methods," *IntechOpen*, 2013.

[19] Y. L. Brasil, J. P. Cruz-Tirado, and D. F. Barbin, "Fast online estimation of quail eggs freshness using portable NIR spectrometer and machine learning," *Food Control*, vol. 131, Article ID 108418, 2022.

[20] J. P. Cruz-Tirado, M. Lucimar da Silva Medeiros, and D. F. Barbin, "On-line monitoring of egg freshness using a portable NIR spectrometer in tandem with machine learning," *Journal of Food Engineering*, vol. 306, Article ID 110643, 2021.

[21] N. Liu, H. A. Parra, A. Pustjens, K. Hettinga, P. Mongondry, and S. M. van Ruth, "Evaluation of portable near-infrared

spectroscopy for organic milk authentication," *Talanta*, vol. 184, pp. 128–135, 2018.

[22] V. Federal University of Juiz De Fora, "Anjos, near and mid infrared spectroscopy to assess milk products quality: a review of recent applications," *J. Dairy Res. Technol.*vol. 3, pp. 1–10, 2020.

[23] H. Soyeurt, P. Dardenne, F. Dehareng et al., "Estimating fatty acid content in cow milk using mid-infrared spectrometry," *Journal of Dairy Science*, vol. 89, no. 9, pp. 3690–3695, 2006.

[24] S. D. Mesgaran, A. Eggert, P. Höckels, M. Derno, and B. Kuhla, "The use of milk Fourier transform mid-infrared spectra and milk yield to estimate heat production as a measure of efficiency of dairy cows," *Journal of Animal Science and Biotechnology*, vol. 11, no. 1, p. 43, 2020.

[25] M. Caredda, M. Addis, I. Ibba et al., "Building of prediction models by using Mid-Infrared spectroscopy and fatty acid profile to discriminate the geographical origin of sheep milk," *Lebensmittel-Wissenschaft und -Technologie*, vol. 75, pp. 131–136, 2017.

[26] M. K. Nieuwoudt, S. E. Holroyd, C. M. McGoverin, M. C. Simpson, and D. E. Williams, "Raman spectroscopy as an effective screening method for detecting adulteration of milk with small nitrogen-rich molecules and sucrose," *Journal of Dairy Science*, vol. 99, no. 4, pp. 2520–2536, 2016.

[27] M. Hammami, S. Dridi, F. Zaïdi et al., "Use of front-face fluorescence spectroscopy to differentiate sheep milks from different genotypes and feeding systems," *International Journal of Food Properties*, vol. 16, no. 6, pp. 1322–1338, 2013.

[28] K. S. Babu and J. K. Amamcharla, "Application of front-face fluorescence spectroscopy as a tool for monitoring changes in milk protein concentrate powders during storage," *Journal of Dairy Science*, vol. 101, no. 12, pp. 10844–10859, 2018.

[29] R. Bro and A. K. Smilde, "Principal component analysis," *Anal. Methods*, vol. 6, no. 9, pp. 2812–2831, 2014.

[30] A. N. Parveen, H. H. Inbarani, and E. N. S. Kumar, "Performance analysis of unsupervised feature selection methods," in *Proceedings of the 2012 International Conference on Computing Communication and Applications*, pp. 1–7, Dindigul, Tamilnadu, India, February 2012.

[31] I. T. Jolliffe, *Principal Component Analysis*, Springer-Verlag, New York, NY, USA, 2nd edition, 2002.

[32] H. P. Bhavsar and M. H. Panchal, "A review on support vector machine for data classification," 2012.

[33] D. Ballabio and V. Consonni, "Classification tools in chemistry. Part 1: linear models. PLS-DA," *Analytical Methods*, vol. 5, no. 16, p. 3790, 2013.

[34] A. El Orche, O. Elhamdaoui, A. Cheikh et al., "Comparative study of three fingerprint analytical approaches based on spectroscopic sensors and chemometrics for the detection and quantification of argan oil adulteration," *Journal of the Science of Food and Agriculture*, vol. 102, p. 11335, 2021.

[35] S. Wold, H. Martens, and H. Wold, "The multivariate calibration problem in chemistry solved by the PLS method," in *Matrix Pencils*, B. Kågström and A. Ruhe, Eds., Springer, Berlin, Germany, pp. 286–293, 1983.

[36] J. Peris-Vicente, M. J. Lerma-García, E. Simó-Alfonso, J. V. Gimeno-Adelantado, and M. T. Doménech-Carbó, "Use of linear discriminant analysis applied to vibrational spectroscopy data to characterize commercial varnishes employed for art purposes," *Analytica Chimica Acta*, vol. 589, no. 2, pp. 208–215, 2007.

[37] Y. Zhang, Z. Dong, S. Wang, G. Ji, and J. Yang, "Preclinical diagnosis of magnetic resonance (MR) brain images via discrete wavelet packet transform with tsallis entropy and generalized eigenvalue proximal support vector machine (GEPSVM)," *Entropy*, vol. 17, no. 4, pp. 1795–1813, 2015.

[38] S. Suthaharan, *Support Vector Machine*Springer, Boston, MA, USA, 2021, https://link.springer.com/chapter/10.1007/978-1-4899-7641-3_9.

[39] N. Cristianini and B. Schölkopf, "Support vector machines and kernel methods: the new generation of learning machines," *AI Magazine*, vol. 23, 2002.

[40] G. D. Miner, J. Elder, A. Fast, T. Hill, R. Nisbet, and D. Delen, "Feature selection and dimensionality reduction," in *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*, pp. 929–934, Elsevier, Amsterdam, Netherlands, 2012.

[41] S. Yang, C. Li, Y. Mei et al., "Determination of the geographical origin of coffee beans using terahertz spectroscopy combined with machine learning methods," *Frontiers in Nutrition*, vol. 8, p. 313, 2021.

[42] D. L. Banks and S. E. Fienberg, "Data mining, statistics," in *Encyclopedia of Physical Science and Technology*, R. A. Meyers, Ed., pp. 247–261, Academic Press, New York, NY, USA, Third edition, 2003.

[43] N. Nicolaou and R. Goodacre, "Rapid and quantitative detection of the microbial spoilage in milk using Fourier transform infrared spectroscopy and chemometrics," *The Analyst*, vol. 133, no. 10, pp. 1424–1431, 2008.

[44] A. Subramanian, V. B. Alvarez, W. J. Harper, and L. E. Rodriguez-Saona, "Monitoring amino acids, organic acids, and ripening changes in Cheddar cheese using Fourier-transform infrared spectroscopy," *International Dairy Journal*, vol. 21, no. 6, pp. 434–440, 2011.

[45] J. A. Westerhuis, H. C. J. Hoefsloot, S. Smit et al., "Assessment of PLSDA cross validation," *Metabolomics*, vol. 4, no. 1, pp. 81–89, 2008.