Journal of Food Measurement and Characterization (2025) 19:2595–2610
https://doi.org/10.1007/s11694-025-03133-9

ORIGINAL PAPER

Potential use of the ATR-FTIR spectroscopy as an almond cultivar recognition tool: impact of sample and spectral pre-treatments

Sandra Lamas[1] · Nuno Rodrigues[1] · Arantzazu Santamaria-Echart[1] · Igor Palu[1] · Jocyla R. Manhique[1] · Baudilio Herrero[3] · Isabel López-Cortés[2] · José Alberto Pereira[1] · António M. Peres[1]

Received: 18 March 2024 / Accepted: 7 February 2025 / Published online: 28 February 2025

Abstract

Three almond cultivars (Lauranne, Marinada, and Vairo) were studied, considering morphological parameters that showed statistical cultivar-dependence but not enabled accurate cultivar recognition. Alternatively, attenuated total reflectance Fourier transform infrared spectroscopy (ATR-FTIR) was applied to whole endocarp, kernel, and ground almonds. Both transmittance spectra and respective derivatives (3400–2800 and 1900–600 cm^{-1}) of the three matrices were used to establish multivariate linear discriminant models, based on subsets of selected wavenumbers (6–33), allowing 90.5–99.7% of correct cultivar classification for repeated K-fold cross-validation. Ground almonds yielded the best results regardless of spectrum pre-treatment. While epicarp analysis offers less invasiveness, the use of raw transmittance spectra of ground almonds resulted into the most practical approach due to the need of fewer independent variables (less complex models), proving effective for cultivar identification via ATR-FTIR -chemometric tools. Overall, the findings point out that ATR-FTIR is a reliable tool for almond cultivar traceability.

Keywords Dry fruits · Cultivar traceability · Principal Component Analysis · Linear Discriminant Analysis

Introduction

The almond, *Prunus dulcis* (Mill.), is one of the oldest and most relevant crops, and in recent years, there has been a significant global increase in its production, due to the heightened consumer demand, with the United States of America leading as the top producer (1,858,010 tons), followed by Australia (360,328.03 tons), and Spain with a production of 245,990 tons [1]. Almonds, a typical dry fruit, are highly prized for their nutritional, health-promoting, and sensory attributes. They showcase remarkable anti-inflammatory and antioxidant properties, as well as neuroprotective and cholesterol-lowering benefits, attributed to their abundant content of phenolic and polyphenolic compounds, high levels of vitamin E, fiber, and essential minerals such as magnesium, manganese, and copper [2]. The quality of almonds hinges on various factors, including the physical appearance of the kernel and shell, sensory characteristics, and chemical composition (e.g., protein content, lipids, sugars, among others) [3]. These characteristics can be influenced by climatic conditions, geographic location, and cultural practices/techniques [4]. Moreover, the beneficial features of almonds may vary among cultivars due to differences in nutrient levels and fatty acid composition [5, 6]. Given the rich fatty acid content, almonds are susceptible to oxidation reactions and thermal degradation processes, which can be induced by environmental and storage factors such as moisture, temperature, light, and exposure to oxygen [7].

As for other food products, it has become imperative to develop methods for controlling the traceability of almonds. Such methods would equip producers and authorities with rapid and straightforward analytical tools to ensure the authenticity of almonds concerning their origin and/or cultivar.

✉ Nuno Rodrigues
nunorodrigues@ipb.pt

✉ António M. Peres
peres@ipb.pt

[1] CIMO, LA SusTEC, Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal

[2] Department of Plant Production, COMAV, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain

[3] Dpto. de Ciencias Agroforestales, Universidad de Valladolid, Avda. de Madrid, 57, 34004 Palencia, Spain

For instance, various strategies have been explored to establish distinctive almond profiles for accurate geographical origin recognition. Many of these efforts have concentrated on scrutinizing lipid composition, particularly emphasizing fatty acids, triacylglycerols, phospholipids, and tocopherols [4, 8–12]. However, a key limitation arising from these studies is the insufficient validation of the developed methodologies and the potential inadequacy of selected compounds as reliable biomarkers, mainly due to a targeted approach [8]. This drawback underscores uncertainties about the accuracy and routine applicability of lipid-based classification methodologies. Hence, there is a demand for comprehensive validation of methodologies and alternative discriminators better suited for distinguishing almonds. Presently, the utilization of non-destructive analytical techniques, particularly spectroscopy-based methods combined with chemometric approaches, has proven advantageous for tracing and authenticating various food products. These techniques facilitate rapid non-invasive analysis with high accuracy and sensitivity, eliminating the need for expensive, complex, and time-consuming sample pre-treatments [3, 13, 14]. Furthermore, these techniques yield spectral data that can be correlated with the chemical composition of food products. The analytical techniques encompass near-infrared (NIR), infrared (IR), Raman spectroscopy, and Fourier-transform infrared spectroscopy (FTIR), which, when coupled with diverse multivariate statistical techniques, have proven successful in food analysis [3, 13].

FTIR spectroscopy, in particular, has been employed for detecting adulterations, discriminating almonds based on geographical origin/cultivar, and playing a pivotal role in establishing the authenticity and quality of the product. This technique allows the correlation of bands detected in the IR spectra with specific vibrations of functional groups, thereby linking them to the chemical composition of almonds [3, 5, 13, 15], demonstrating the potential use of ATR-FTIR as a tool for discriminating between different almond cultivars [3, 5, 16], distinguishing bitter from sweet almonds [15], and evaluating nutritionally relevant functional groups in almonds. This provides specific chemical information for each cultivar, contributing to a comprehensive nutritional characterization of various almond types [17].

The primary objective of the present study is to assess the potential application of ATR-FTIR spectroscopy as a tool for identifying almond cultivars, specifically cvs. Lauranne, Marinada, and Vairo. Additionally, it is aimed to explore the comparative advantages of utilizing three distinct almond matrices: epicarp, kernel, and ground almonds. The goal is to determine the most promising matrix for accurate cultivar recognition. Furthermore, the study seeks to address the necessity of spectra pre-processing. To achieve this, a comparison will be carried out between the predictive classification performances of different ATR-FTIR-chemometric models. These models will be based on both raw transmittance spectra and transformed spectra forms, including the 1st and 2nd derivatives of the raw spectra, with the objective of identifying the most effective and straightforward approach. To achieve this, ATR-FTIR spectra obtained in the wavenumber range from 4000 to 500 cm^{-1} will be employed. Both unsupervised and supervised pattern recognition techniques will be applied to analyse and interpret the data. The overarching aim is to contribute valuable insights into the optimal use of ATR-FTIR spectroscopy for almond cultivar identification and to determine the most effective spectral pre-processing technique as well as the most adequate almond matrix to fulfil the established objectives.

Materials and methods

Sampling

In the present study, almond independent samples from three cultivars were collected from different almond groves located in the region of Valencia (Spain), during the crop year of 2022/2023. The sampling was focused on cv. Lauranne, cv. Marinada, and cv. Vairo due to its importance and representativity on the market.

Morphologic parameters

For each almond cultivar, 40 seeds were randomly selected. The characterization of the seeds (with some modifications from UPOV [18]) was carried out through measurements using a digital caliper (Powerfix Profi Electronic Digital Caliper, Owim GmbH &Co. KG, Neckarsulm, Germany), considering the parameters of length (mm), maximum diameter (mm) and transverse diameter (mm). The parameters of symmetry A, symmetry B, apex, base, and rugosity of surface of the seed were evaluated by observing the respective seeds. The shape of the seed was calculated based on length and width, and the corresponding weight (g) of each seed was registered, using the analytical balance (Radwag AS 220.R2 Plus, Poland).

Sample preparation

To obtain the kernels, the outer shell had to be removed using a hammer. The samples were prepared immediately before analysis to avoid oxidation reactions. To obtain the ground almonds, two similar almonds were placed in the ZM200 ultra centrifugal mill for 15 s. The flour generated by the milling was carefully removed from the mill plate with a spatula and placed in a goblet. Five milling operations were carried out, totaling ten readings for each variety, i.e. a new milling operation was carried out every two ATR-FTIR

readings. The ring sieve was cleaned after each milling, as it clogged up during the process. To take the kernel reading, a small sample was taken from the inside of the kernel using a scalpel so that the piece completely covered the crystal. The same was done for the epicarp, a piece of which was carefully removed with the help of a scalpel so that it covered the entire crystal.

ATR-FTIR spectroscopy: apparatus and analysis of almonds

The ATR-FTIR analysis was performed using an MB300 FTIR by ABB (Zurich, Switzerland), equipped with a horizontal cell and a diamond crystal for attenuated total reflectance (ATR) mode. Spectral data were acquired with a scanning rate of 32 scans per minute and a resolution of 4 cm^{-1}, covering the wavenumber range from 4000 to 500 cm^{-1}. The acquired spectra were processed using the software Horizon MB version 3.4. For each almond cultivar, 30 independent almond samples were evaluated being the ATR-FTIR spectra acquired in duplicate for 10 different almonds from each independent sample and for each sample matrix studied (epicarp/exocarp, kernel and ground almonds as shown in Fig. 1).

For statistical analysis, the raw ATR-FTIR spectra (transmittance in percentage) and two transformed spectra forms (the 1st and 2nd derivatives of the transmittance) were used. To avoid using replicas of the same independent almond sample, for each one, an average representative spectrum was calculated from the 20 ATR-FTIR raw spectra and respective transformed forms (10 almonds × 2 replicas per matrix and per cultivar studied). Three final datasets were constructed (one per type of matrix) comprising 30 independent samples × 3 almond cultivar (cvs. Lauranne, Marianda and Vairo) × 3 pre-processed average ATR-FTIR spectra (raw transmittance, 1st derivative spectra and 2nd derivative spectra). All almonds were pre-processed prior to analysis to prevent/minimize oxidation reactions. The almond external shells were removed using a hammer, followed by a longitudinal cut made with a scalpel to obtain the epicarp and kernel almond. Additionally, almonds without the external shells were ground individually, for 15 s, using a mill device (Retsch ZM 200) to obtain the ground samples.

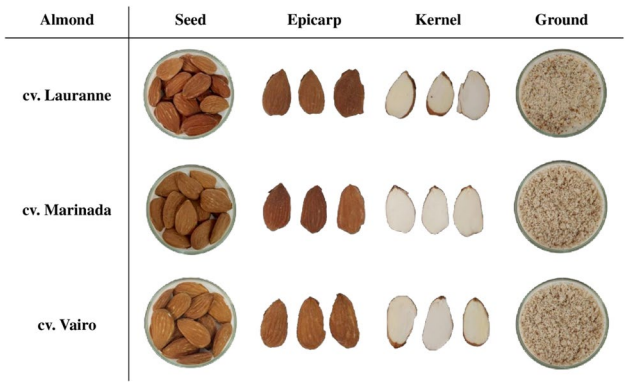

Fig. 1 Examples of the three types of almond evaluated matrices (epicarp, kernel and ground almonds)

Statistical analysis

Morphologic data

The morphologic parameters (weight, length, shape, maximum diameter and transversal diameter) of the almonds of the three cultivars under study were subjected to a one-way ANOVA followed by Tukey's post-hoc multiple comparison test. Each parameter was assessed in 40 independent almonds for each of the three cultivars under study. The aim was to infer about the existence or not of a statistically significant effect of the cultivar on each parameter. Principal Component Analysis (PCA) was applied as an unsupervised multivariate pattern recognition tool to verify if the morphologic data could be used to differentiate the three almond cultivars, using a dataset of 40 almonds × 3 cultivars × 5 morphological parameters. Additionally, a Linear Discriminant Analysis (LDA) coupled with the Simulated Annealing (SA) variable selection algorithm was implemented aiming to identify the most discriminant subset of morphological parameters and the related classification performance for training and leave-one-out cross-validation internal variant (LOO-CV). The statistical analysis was performed using the open-source statistical program R (version 3.6.2) and its related software packages.

ATR-FTIR spectroscopy data

Principal component analysis (PCA) was applied to analyse the ATR-FTIR spectra of epicarp almond, kernel almond, and ground almond samples. PCA was used as a multivariate recognition technique to determine which spectra pretreatment (i.e., raw spectra and the 1st and 2nd derivatives spectra, comprising each type the outputs of 1790 wavenumbers) would allow the best differentiation between (i) the type of studied matrix (i.e., epicarp, kernel and ground almonds) for each of the thee almond cultivars; or (ii) the almond cultivar for each of the three matrices considered. The average values of the replicates conducted for each independent sample/cultivar/matrix were used for PCA. Considering the three almond cultivars, the ten independent samples from each cultivar, the 10 almonds analysed per sample in duplicate the initial dataset had a total of 30 independent samples × 10 almonds × 3 cultivars × 3 matrices × 2 ATR-FTIR assays. To assess the discrimination performance based on the ATR-FTIR data, LDA was implemented using

an average spectrum per independent sample/cultivar/matrix from those spectra acquired during the analysis of the 10 almonds in duplicate. The final dataset, per type of matrix, comprised 30 independent samples × 3 almond cultivars × 3 independent average spectra (raw, 1st and 2nd derivatives, each one comprising responses at 1790 wavenumbers). To select the subsets of wavenumbers that would allow the best discrimination performance, the final spectra database was initially reduced from 1790 to 1095 wavenumbers, by only considering the regions of the spectrum where bands were detected and transmittance values were different from 100% (i.e., non-zero absorbance). The meta-heuristic SA variable selection technique was used to determine which wavenumber subsets (and related band intensities) allowed the best discrimination performance. The LOO-CV and the repeated K-fold-CV internal validation variants were used to discuss the LDA classification performance. The latter CV variant (repeated K-fold-CV with K = 4 and 10 repeats) is more robust than the LOO-CV, being a better approximation to an external validation, since in each validation run it keeps 25% of the data for validation using the remaining 75% for training. Thus, contrary to LOO-CV, where at each stage only one sample is used at each run for validation purposes, in K-fold-CV between 22 and 23 (randomly selected at each stage) of the 90 samples for each of the three almond matrices evaluated, were used for validation purposes (samples not used in training), of which 7–8 samples belong to each of the three almond cultivars under study. For the original (training) data, the quality of the models was also assessed by visual inspection of the two-dimensional graphs (2D), obtained for the first two primary discriminant functions (DF), in which the 90 samples under study for each matrix studied are represented as well as the confidence ellipse of each supervised group, calculated according to the Bayes' theorem [19]. Statistical analysis was carried out at a 5% confidence level, using the open-source statistical program R (version 3.6.2) and respective libraries.

Results and discussion

Morphologic parameters

The mean values with standard errors (± SE) of the five morphological parameters are shown in Table 1 The results of the one-way ANOVA indicated a significant statistical effect of the cultivar (P-value < 0.05) on all five evaluated parameters. In terms of weight, almonds from the cv. Vairo exhibited the highest weight at 1.21 g, followed by those from the cv. Lauranne at 1.17 g and the cv. Marinada at 0.85 g. Conversely, almonds from the cv. Lauranne were the longest, measuring 25.13 mm, with almonds from the cv. Vairo following closely at 24.20 mm and those from the cv. Marinada at 23.11 mm. As for diameter, almonds from the cv. Vairo displayed the largest maximum diameter at 14.95 mm, succeeded by almonds from the cv. Lauranne at 13.96 mm and the cv. Marinada at 13.15 mm. A similar pattern was observed for the transverse diameter, with measurements of 7.33 mm, 6.81 mm, and 5.75 mm for almonds of the cvs. Vairo, Lauranne, and Marinada, respectively. The calculated shape, determined by the length/width ratio, ranged from 1.80 for the cv. Lauranne to 1.62 for the cv. Marinada. Based on these values, it can be concluded that almonds from all three cultivars possess an elongated shape, as the shape ratio exceeds 1.45.

The recorded values of the examined parameters align with those previously documented in the literature for the same three almond cultivars. For instance, Mougiou et al. [20] studied almonds of the cv. Lauranne, noting weights of 1.30 g, a length of 26.30 mm, and a width of 13.70 mm (shape = 1.92). Similarly, Graeff et al. [21] reported results for cv. Marinada, indicating a weight of 1.70 g, a length of 26.90 mm, and a width of 14.60 mm (shape = 1.84). Regarding cv. Vairo, Rabadán et al. [22] provided data showing weights of 1.17 g, a length of 24.60 mm, and a width of 15.00 mm (shape = 1.64).

The evaluation also involved the visual assessment of the parameters of symmetry A, symmetry B, apex, base, and surface rugosity for all the studied almonds. Symmetry A was determined by examining the front of the almond, while symmetry B was analysed on the almond's side. In the case of cv. Lauranne, 88% of the almonds were categorized as slightly asymmetrical, 8% as symmetrical, and 5% as asymmetric. For cv. Marinada, the majority of the almonds were slightly asymmetrical (70%), with symmetrical and asymmetric almonds accounting for 23% and 8%, respectively. In the case of cv. Vairo, a significant percentage of almonds (98%) exhibited slight asymmetry, while 2% were classified as asymmetric. Distinct patterns in symmetry B were observed among the cultivars. In cv. Lauranne, there was an almost equal distribution between symmetrical (48%) and slightly asymmetrical (53%) almonds. For cv. Marinada, the distribution was evenly split between symmetrical (50%) and slightly asymmetrical (50%) almonds. In contrast, cv. Vairo predominantly had slightly asymmetrical almonds (88%), with only 13% classified as symmetrical. Additionally, all almonds from cvs. Lauranne and Marinada displayed an acute apex, while in cv. Vairo, 65% of the almonds had an acute apex, and 35% had an obtuse apex. All evaluated almonds from cv. Lauranne had a rounded base (100%). For cv. Marinada, 95% of the almonds had a rounded base, with 5% showing a truncated base- On the other hand, for cv. Vairo, all almonds had a truncated base (100%). In which concerns the surface rugosity, 98% of the almonds of cv. Lauranne showed a medium rugosity, and the other 2% had a weak rugosity. For cv. Marinada, all almonds

Table 1 Mean values and standard deviation (minimum – maximum) of the morphologic parameters of the almonds studied (cvs. Lauranne, Marinada and Vairo)

Morphologic parameters	Studied almonds			*P*-value
	cv. Lauranne	cv. Marinada	cv. Vairo	
Weight (g)	1.17 ± 0.16^{a} (0.94–1.50)	0.85 ± 0.18^{b} (0.44–1.23)	1.21 ± 0.19^{a} (0.63–1.69)	< 0.0001
Length (mm)	25.13 ± 1.36^{a} (22.68–27.57)	23.11 ± 1.59^{c} (19.16–26.00)	24.20 ± 1.84^{b} (19.08–28.49)	< 0.0001
Maximum diameter (mm)	13.96 ± 0.75^{b} (12.39–15.29)	13.15 ± 1.10^{c} (10.35–15.04)	14.95 ± 1.20^{a} (11.86–18.41)	< 0.0001
Transverse diameter (mm)	6.81 ± 0.60^{b} (5.66–7.95)	5.75 ± 0.60^{c} (4.22–6.85)	7.33 ± 0.60^{a} (5.88–8.24)	< 0.0001
Shape (length/width)	1.80 ± 0.09^{a} (1.62–1.97)	1.76 ± 0.12^{a} (1.56–2.03)	1.62 ± 0.10^{b} (1.43–1.91)	< 0.0001
Symmetry A (%)				
Symmetrical	8	23	0	–
Slightly asymmetrical	88	70	98	–
Asymmetric	5	8	2	–
Symmetry B (%)				
Symmetrical	48	50	13	–
Slightly asymmetrical	53	50	88	–
Asymmetric	0	0	0	–
Apex (%)				
Acute	100	100	65	–
Obtuse	0	0	35	–
Rounded	0	0	0	–
Base (%)				
Truncated	0	5	100	–
Rounded	100	95	0	–
Rugosity of surface (%)				
Weak	3	0	8	–
Medium	98	100	80	–
Strong	0	0	13	–

Different letters on values of the same parameter indicate significant differences ($p \leq 0.01$)

The test applied was one-way ANOVA followed by the application of the Tukey's test

showed medium rugosity (100%). In cv. Vairo, 80% showed a medium rugosity, 13% had a strong rugosity and 8% had a weak rugosity.

Different letters on values of the same parameter indicate significant differences ($p \leq 0.01$). The test applied was one-way ANOVA followed by the application of the Tukey's test.

The preceding analysis demonstrated that the morphological data of the studied almonds were influenced by the specific almond cultivar. Consequently, a PCA was conducted using five quantitative morphological parameters (weight, length, shape, maximum and transverse diameters). The outcomes revealed a partial unsupervised differentiation (Fig. 2) of the examined almonds based on their cultivar (cvs. Lauranne, Marinada, and Vairo). The first three principal components (PCs) explained 62.1%, 26.8%, and 10.2% of the data variability, respectively. Despite the substantial percentage of explained variance, there was a noticeable overlap of the confidence ellipses, suggesting that relying solely on morphological data as almond cultivar biomarkers could result in a high rate of misclassifications.

Employing a LDA-SA approach, an optimal classification multivariate model was established based on the information from four out of the five morphological parameters (selected by the SA algorithm: weight, length, maximum, and transverse diameters) with a sensitivity of only 80% for both training and LOO-CV procedures. The high misclassification rate (20%) strengthened the previous finding and highlighted the need for a different classification strategy to ensure a more accurate identification of the almond cultivars. In this context, a FTIR-chemometric approach was subsequently evaluated to assess its feasibility and versatility in discriminating almond cultivars.

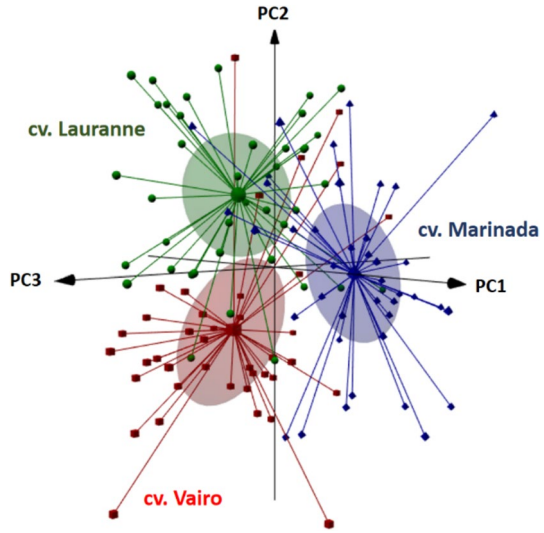

Fig. 2 3D PCA plot: unsupervised differentiation of the three almond cultivars studied (cvs. Lauranne, Marinada and Vairo) based on the data of five morphological parameters evaluated (weight, length, maximum diameter, transverse diameter and shape)

ATR-FTIR spectroscopy

ATR-FTIR analysis was conducted on distinct independent samples of three almond cultivars, namely cv. Lauranne, cv. Marinada, and cv. Vairo. This allowed the recording of transmittance spectra across the wavenumber range from 4000 to 500 cm^{-1}. Figures 3, 4, 5 display the raw transmittance spectra, along with the transformed forms (1st and 2nd derivatives) for each of the three cultivars and three types of matrices (ground, kernel, and epicarp). In these spectra, common bands were identified for all studied cultivars, corresponding to specific vibrations of functional groups present in chemical compounds within almonds. These compounds include proteins, carbohydrates, water, alcohols, esters, and peptides, as documented by García et al. [5] and Sharma & Neeraj [17]. The ATR-FTIR analysis unveiled distinct transmittance band spectral profiles within the 4000 to 500 cm^{-1} region, particularly emphasizing bands within the wavenumber ranges of 4000–2825 cm^{-1} and 1766–549 cm^{-1}. Notably, no discernible peaks were observed in the wavenumber range

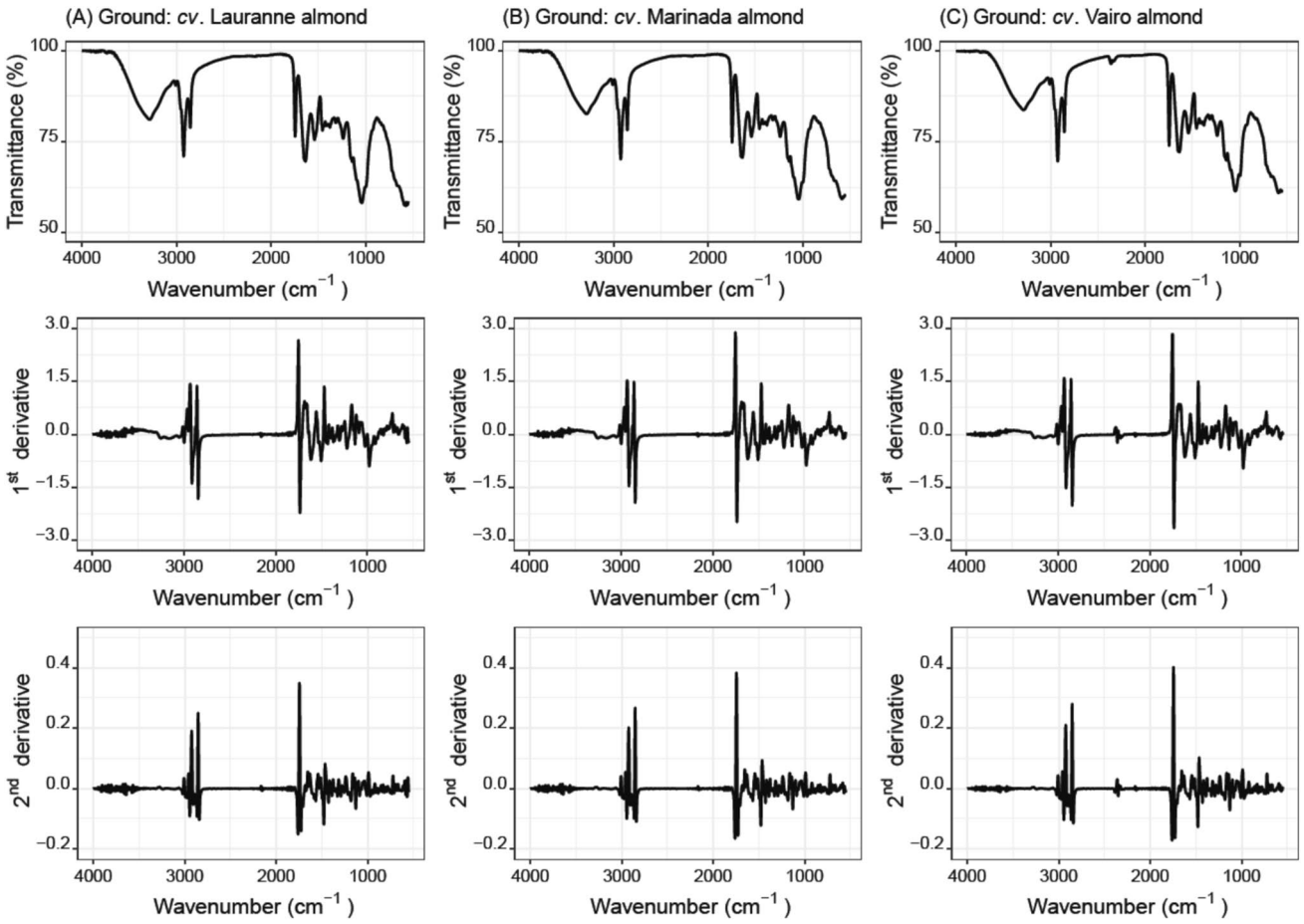

Fig. 3 ATR-FTIR spectra (wavenumbers ranging from 4000 to 500 cm^{-1}) and respective 1st and 2nd derivatives regarding the almond's ground analysis for the three cultivars studied: **A** cv. Lauranne, **B** cv. Marinada, and **C** cv. Vairo

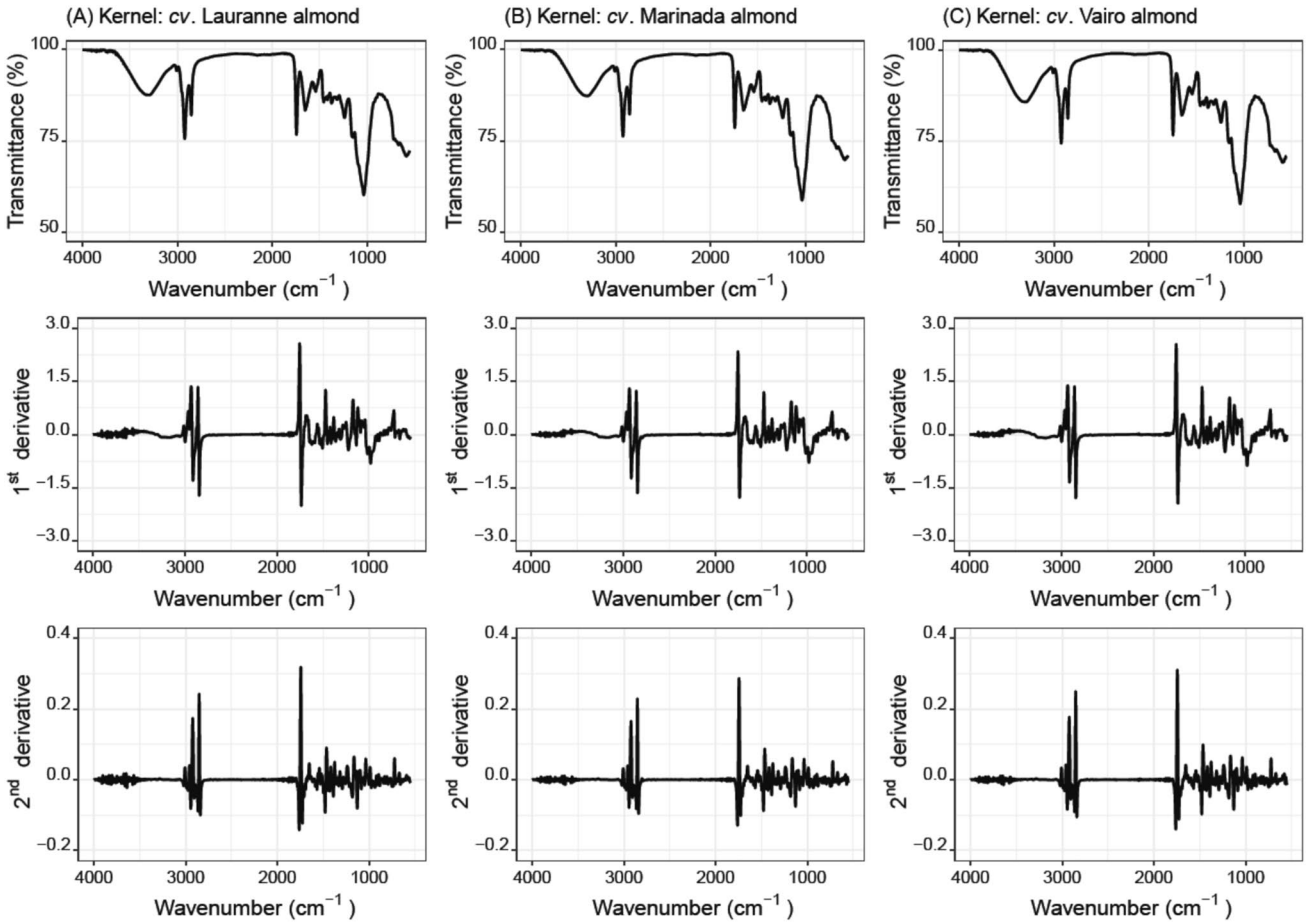

Fig. 4 ATR-FTIR spectra (wavenumbers ranging from 4000 to 500 cm^{-1}) and respective 1st and 2nd derivatives regarding the almond's kernel analysis for the three cultivars studied: **A** cv. Lauranne, **B** cv. Marinada, and **C** cv. Vairo

of 2825–1791 cm^{-1}, which is considered a neutral spectral region.

Table 2 presents the wavenumber regions encompassing the overall range of 4000–500 cm−1, where ATR-FTIR bands were identified during the analysis of ground, kernel, or epicarp almonds from the three assessed cultivars. While the overall spectra obtained (transmittance, 1st and 2nd derivatives) for the three almond cultivars and matrices exhibit visual similarities (Figs. 3, 4, 5), variations exist in the number and intensity of the detected bands within each cultivar-matrix pairing. Distinct ATR-FTIR bands were discerned in the spectra of the different almond matrices under investigation, although certain bands were common across two or three matrices. In general, each matrix exhibited characteristic bands. Ground almonds, for instance, displayed bands at 3291, 2975, 1636, 1263, 1045, and 993 cm^{-1}. Kernel almonds showcased bands at 3324, 3006, 1654, 1314, 1159, 1040, and 721 cm^{-1}. Epicarp almond exhibited bands at 3325, 1735, 1627, 1502, 1446, 1324, 1019, 819, and 761 cm^{-1}. Conversely, bands at 2923, 2854, 1377, 1237, 1097, and 906 cm^{-1} were consistently present in the spectra of all three different matrices, regardless of the almond cultivar.

The complete spectra of ATR-FTIR transmittance, including the 1st and 2nd derivatives at 1790 wavenumbers, were subjected to a PCA. This procedure enabled the unsupervised discrimination of the three matrices under investigation for each almond cultivar, as depicted in Fig. 6 (a 3D plot of the first three principal components, PCs). Subsequently, a parallel analysis was conducted to assess whether PCA applied to the full ATR-FTIR spectra data could effectively distinguish almond cultivars within each analysed matrix, considering the three spectrum pre-treatments (transmittance, 1st and 2nd derivatives). The findings, illustrated

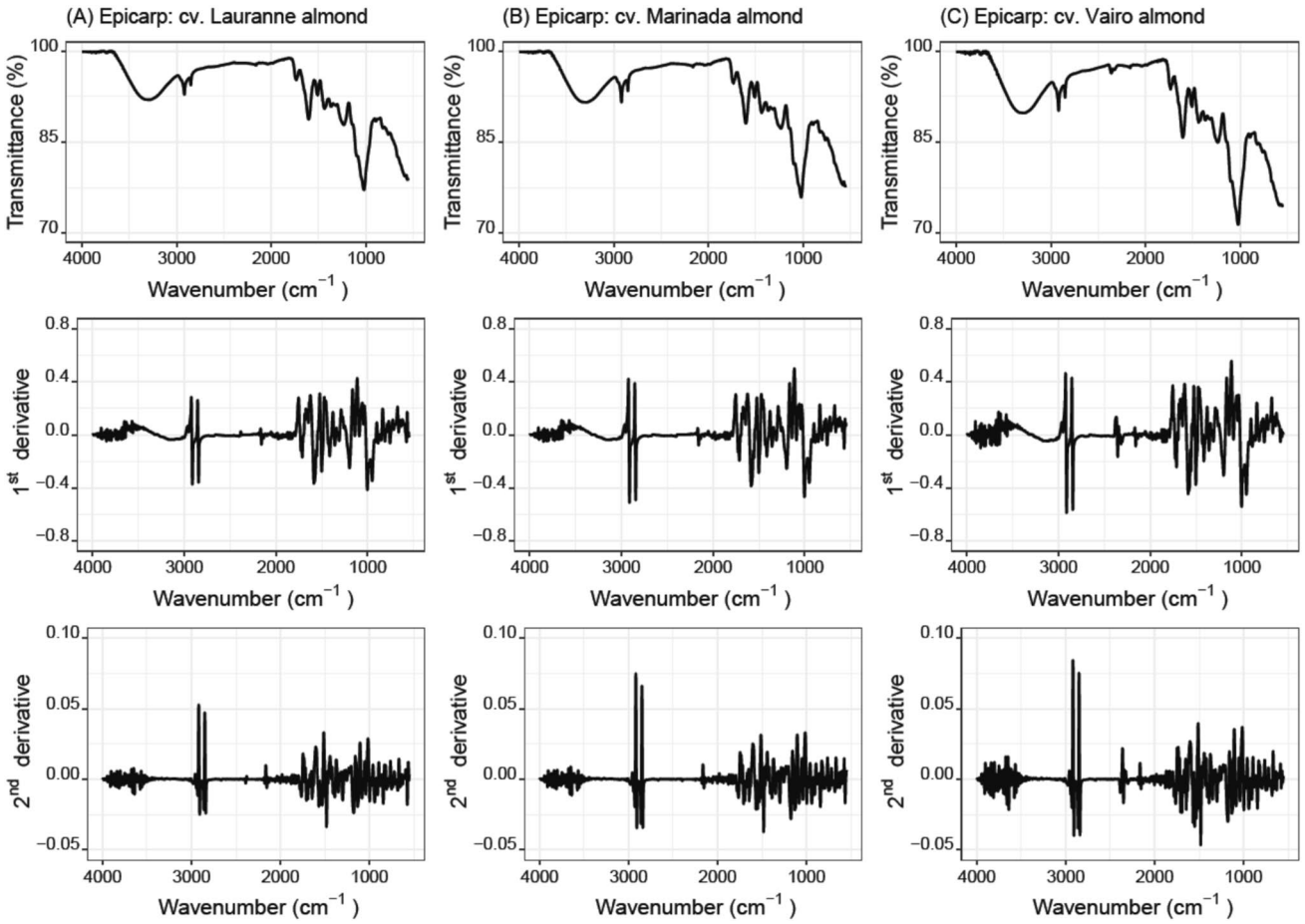

Fig. 5 ATR-FTIR spectra (wavenumbers ranging from 4000 to 500 cm^{-1}) and respective 1st and 2nd derivatives regarding the almond's epicarp analysis for the three cultivars studied: **A** cv. Lauranne, **B** cv. Marinada, and **C** cv. Vairo

in Fig. 7, clearly demonstrate that utilizing the complete spectrum did not yield a distinct separation among almond cultivars. Notably, there was significant overlap, particularly evident in two matrices, i.e., epicarp and kernel almonds. Interestingly, ground almonds exhibited a partial separation of the three cultivars, suggesting that this matrix might be the most suitable for achieving the primary objective of the study: cultivar recognition. Additionally, the results underscore that almonds of the cv. Vairo are the most readily distinguishable among the studied cultivars.

In this study, LDA models were established using ATR-FTIR spectra, as well as the 1st and 2nd derivatives. These models were built based on transmittances or their respective derivatives recorded at wavenumbers chosen by the SA algorithm. The used datasets comprised 1095 out of a total of 1790 wavenumbers within the range of 4000–550 cm^{-1}. The selected intervals aligned with regions exhibiting distinctive bands in the ATR-FTIR spectra and where transmittance deviated from 100%. Table 3 provides a comprehensive overview of the ATR-FTIR -LDA-SA models developed for the classification of almonds according to their specific cultivars (cvs. Lauranne, Marinada, and Vairo). The classification was performed for each type of matrix (epicarp, kernel, and ground almonds) and three pre-treated ATR-FTIR spectra (raw transmittance, 1st and 2nd derivatives). Table 3 includes information such as the number and identification of independent parameters incorporated into each classification model, selected through the SA algorithm. Additionally, the table presents the percentages of correct classifications (sensitivities) obtained during both the training and cross-validation procedures (LOO-CV and K-fold-CV). It's noteworthy that, in the case of the repeated K-fold-CV variant, 25% of the original data (i.e., 22–23 independent samples out of a total of 90, including 7–8 samples from each cultivar) are reserved for evaluating the predictive performance in each validation run. This approach enhances the robustness of the technique compared to LOO-CV, where only 1 out of

Table 2 Bands observed in the 4000–500 cm^{-1} wavenumber region of the ATR-FTIR spectra recorded during the analysis of the different almond matrices studied for the three almond cultivars considered (cvs. Lauranne, Marinada and Vairo)

Wavenumber region (cm^{-1})	ATR-FTIR raw bands found at a nominal wavenumber (cm^{-1})	Studied almonds			Identification	Reference
		Ground almonds	Kernel almonds	Epicarp almonds		
3400–3000	3325	–	–	X	Stretching vibration of the surface –OH, possibly related with the cellulose and hemicellulose	[23]; [24]
	3324	–	X	–	Stretching vibration of the surface –OH	[23]
	3291	X	–	–	Hydroperoxides formation	[25]
	3006	–	X	–	Stretching vibration of the C=CH cis- and CH cis-olefinic groups	[5]; [13]; [16]; [25]; [26]
3000–2800	2975	X	–	–	Asymmetric stretching vibrations of aliphatic CH_2 functional groups	[25]
	2923	X	X	X	Symmetric stretching vibrations of aliphatic CH_2 functional groups and methylene C–H asymmetric/symmetric stretch (saturated aliphatic alkane/alkyl group frequencies)	[5]; [17]; [23]
	2854	X	X	X	Asymmetric stretching vibrations of aliphatic CH_2 functional groups, methylene C–H asymmetric/symmetric stretch (saturated aliphatic alkane/alkyl group frequencies) and asymmetric stretching of methyl CH_3 groups	[5]; [17]; [25]; [27]
1900–1400	1743	X	X	–	Extension movement of the typical ester carbonyl functional group (C=O) of the triglyceride esters	[3]; [5]; [15]; [16]; [17]; [26] [27];
	1735	–	–	X	Stretching vibration of the aliphatic acid C=O, possibly related with the lignin and hemicellulose	[23]; [24]
	1654	–	X	–	Stretching vibration of cis –C=C–	[15]
	1636	X	–	–	Stretching vibration of unsaturated group like alkene C=C and amide	[17]; [23]
	1627	–	–	X	Stretching vibration of cis C=C of unsaturated acyl groups	[5]
	1541	X	X	–	Presence of aliphatic nitro compounds, simple hetero-oxy compounds and bending vibration of N–H functional group mainly observed in amide I and amide II of protein compounds	[5]; [13]; [15]; [17]
	1502	–	–	X	Bending vibration of the aliphatic secondary amine, the N–H	[27]
	1460	X	X	–	Presence to the bending vibrations of CH in CH_2 and CH_3	[5]; [13]; [25]
	1446	–	–	X	Asymmetric bending vibration of the –C–H(CH_3)	[15]

Table 2 (continued)

Wavenumber region (cm^{-1})	ATR-FTIR raw bands found at a nominal wavenumber (cm^{-1})	Studied almonds			Identification	Reference
		Ground almonds	Kernel almonds	Epicarp almonds		
1400–1000	1377	X	X	X	Bending in of plane of the O–H	[15]
	1324	–	–	X	Bending in of plane of the O–H	[15]
	1314	–	X	–	Carboxylates (carboxylic acid salts) and carbonyl compound group frequencies	[17]
	1263	X	–	–	Stretching vibration of –C–O	[15]
	1237	X	X	X	Aromatic ethers, oxy compounds group frequencies and presence to the bending vibrations of CH in CH_2 and CH_3	[5]; [13]; [17]
	1159	–	X	–	Stretching vibrations of C–O of the ketones	[15]; [27]
	1143	X	–	X	Stretching vibrations of C–O of the ketones	[15]; [27]
	1097	X	X	X	Stretching vibrations of C–O of the ketones	[15]; [27]
	1045	X	–	–	Stretching vibration of C–O functional groups characteristic of the carbohydrate fraction	[5]; [13]
	1040	–	X	–	Combination of bending vibrations of C–H (that is, the C–H bond at position C) of carbohydrates and stretching vibration of C–O functional and primary amine	[3]; [15]; [17]
	1019	–	–	X	Deformation with ring vibration of glycoside C–H and bending vibration of the OH	[23]
1000–900	993	X	–	–	Aliphatic phosphates (P–O–C stretch) grouped under simple hetero-oxy compounds	[17]
	906	X	X	X	Bending out of plane vibration of cis –HC = CH–	[15]
900–600	819	–	–	X	Bending out of plane vibration of cis –HC = CH–	[15]
	761	–	–	X	Bending out of plane of the C–H	[15]
	721	–	X	–	Alkanes C–H rocking vibration	[28]
	663	X	X	–	Bending out of plane of the C–H	[15]

the 90 samples is reserved for validation in each iterative step of the process.

The examination of Table 3 and Fig. 8 (2D representation of the two first discriminant functions, DFs, for training) enables the deduction of the satisfactory classification efficacy exhibited by the FTIR-LDA-SA models in accurately discriminating almond samples based on their specific cultivars, among the three available options. In broad terms, considering the outcomes of cross-validation (LOO-CV and repeated K-fold-CV), along with the inherent simplicity of the established models (i.e., fewer variables included into the classification model), the matrix that enabled a more straightforward approach with either equal or superior predictive performance was the ground almonds. Remarkably, the developed models only incorporated information from 6 to 8 wavenumbers, allowing correct classifications ranging

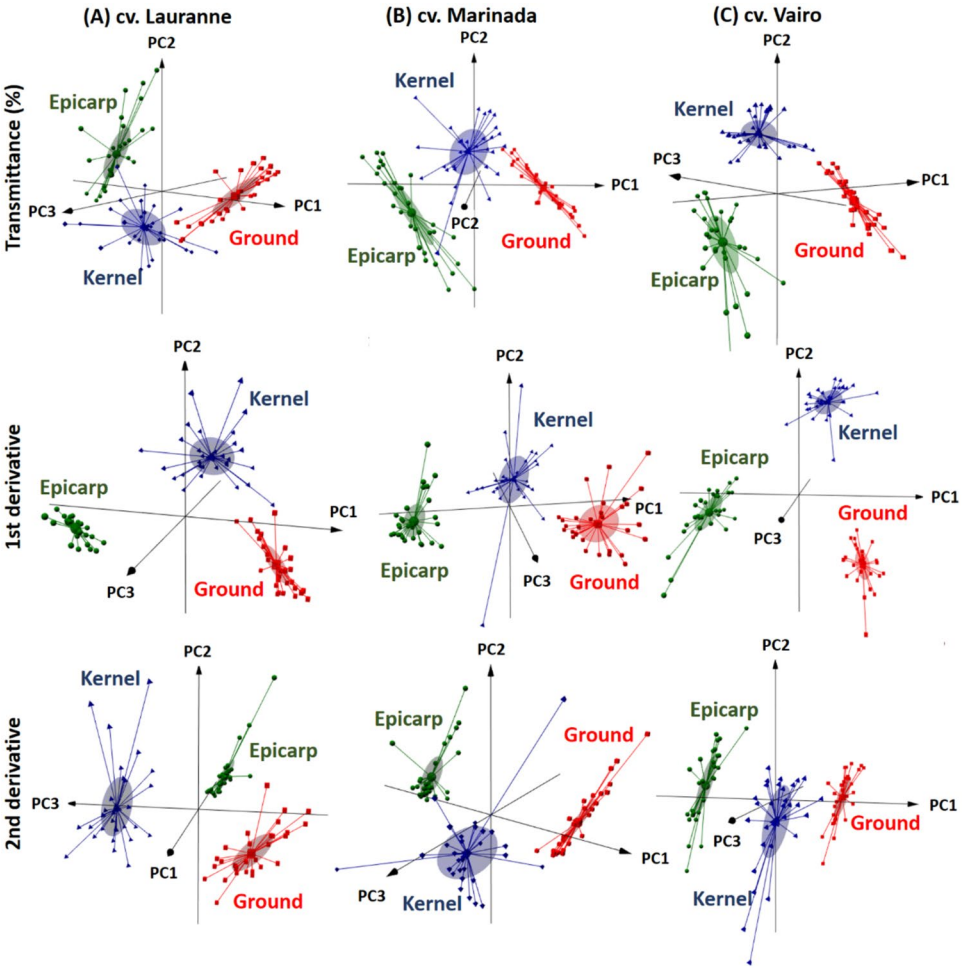

Fig. 6 Differentiation of almond samples according to the type of matrix (● epicarp almond, ▲ kernel almond and ■ ground almonds), for each of the three almond cultivars evaluated (cvs. Lauranne, Marinada and Vairo) using PCA based on information extracted from 1790 wavenumbers of ATR-FTIR spectra (transmittance, 1st derivative and 2nd derivative)

from 99.7 to 99.8%. It is noteworthy that in the model constructed using raw transmittance data, only one sample of cv. Lauranne was erroneously identified as cv. Vairo; in the model derived from 1st derivative data, a single sample of cv. Lauranne was misclassified as cv. Marinada; and in the model based on 2nd derivative data, a cv. Lauranne sample was inaccurately labelled as cv. Vairo, and reciprocally, a sample of cv. Vairo was misclassified as cv. Lauranne. The least satisfactory classification performance was observed in models developed with data from kernel almonds, yielding sensitivities between 90.5% and 97.6%. The increased uniformity of the matrix obtained post-almond milling may rationalize the superior classification results attained. Concerning the pre-processing of FTIR spectra, the results shown in Table 3 also suggested that there is no necessity to apply 1st or 2nd derivatives, as the number of required independent variables (i.e., wavenumbers) decreases in the order of transmittance spectra < 1st derivative spectra < 2nd derivative spectra.

The predictive performances reported in this study aligns with findings from previous research reported in the literature. In a study by García et al. [5], successful discrimination of three almond cultivars (one American: cv. Butte; and two Spanish: cvs. Marcona and Guara) were achieved. This discrimination was accomplished by employing a LDA coupled with a stepwise variable selection algorithm. Although the specific wavenumbers selected from the ATR-FTIR absorbance spectra were not explicitly mentioned, the LDA model achieved a 100% correct classification of the 24 samples in the training set (eight per cultivar. In a more recent study by Cortés et al. [3], the ATR-FTIR technique was applied to differentiate four Spanish almond cultivars (cvs. Guara, Rumbeta, Marcona, and Planeta) based on absorbance spectra of the whole almond. Classification models were constructed using pre-treated spectra data with two techniques: Partial Least Square Regression-Discriminant Analysis (PLS-DA) and Quadratic Discriminant Analysis (QDA). The ATR-FTIR-PLS-DA model exhibited superior classification performance, accurately predicting the almond cultivar with a success rate of 94.45% in external validation. Moreover, the same research team [15] demonstrated the successful application of the ATR-FTIR technique in

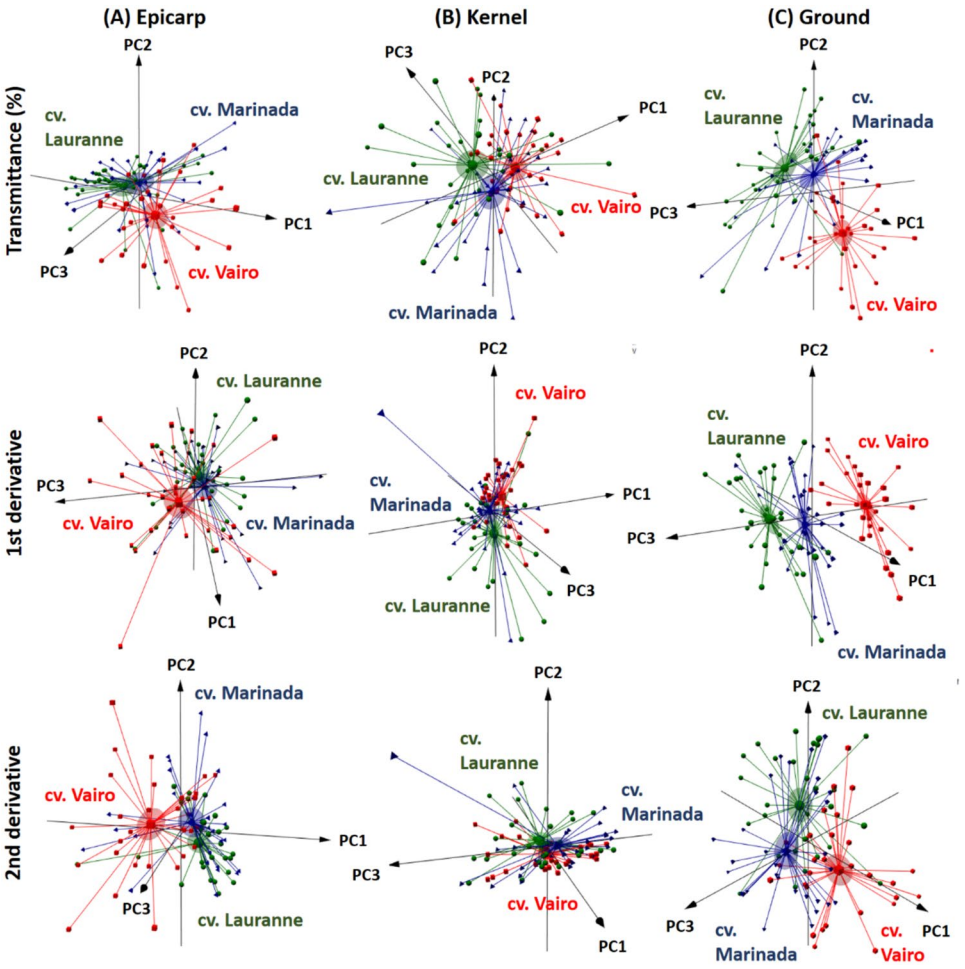

Fig. 7 Differentiation of almond samples according to cultivars (● cv. Lauranne, ▲ cv. Marinada and ■ cv. Vairo), for each of the three matrices evaluated (epicarp almond, kernel almond and ground almonds) using PCA based on information extracted from the 1790 wavenumbers of ATR-FTIR spectra (transmittance, 1st derivative and 2nd derivative)

discriminating whole almonds based on their bitterness levels (sweet almonds versus bitter almonds).

The current investigation underscores its focus on three almond cultivars (cvs. Lauranne, Marinada, and Vairo), previously unexplored through the ATR-FTIR technique. The objective was to establish LDA classification models for almond cultivar identification. Particularly, this study is the first to assess three distinct matrices (epicarp, kernel, and ground almonds) for each cultivar. The findings indicate that the ground almonds matrix emerges as the most suitable for developing simple accurate predictive cultivar classification models. However, it is important to emphasize that, despite the increased complexity of the classification models derived from the spectra generated through epicarp analysis (utilizing a higher number of independent variables, namely 18–30 wavenumbers compared to 6–8 wavenumbers for ground almonds), they offer the advantage of being based on a less invasive procedure. Additionally, the study delves into the ATR-FTIR spectra, exploring the potential benefits of pre-processing raw spectra (1st and 2nd derivatives) in comparison to using transmittance spectra. Surprisingly, it was found that the latter method yielded equally effective predictive correct classification percentages, rendering the need for such spectra pre-treatments unjustified.

Conclusions

Three almond cultivars, cvs. Lauranne, Marianda, and Vairo, were evaluated and statistically significant differences were found in morphological traits, including weight, length, shape, maximum and transverse diameters, symmetry level, apex level, base type, and surface rugosity. However, the observed ranges for these traits overlapped across the cultivars, rendering the quantitative morphological parameters ineffective as biomarkers for cultivar classification. In contrast, the study demonstrated the suitability of ATR-FTIR spectroscopy as an analytical tool for cultivar identification. Among the matrices tested, namely,

Table 3 ATR-FTIR -LDA-SA models developed for discriminating almonds according to correct cultivar (cvs. Lauranne, Marinada and Vairo) and respective classification performances (training, internal validation variants: LOO-CV and K-fold-CV)

Matrix: Epicarp almonds				
ATR-FTIR spectrum	Number of variables in the LDA model	Correct classification (sensitivity, %)		
		Training	LOO-CV	K-fold-CV
Transmittance	30[a]	100	98.9	99.2 ± 1.7
1st derivative	19[b]	100	100	99.5 ± 1.8
2nd derivative	18[c]	100	100	99.3 ± 1.6
Matrix: Kernel almonds				
ATR-FTIR spectrum	Number of variables in the LDA model	Correct classification (sensitivity, %)		
		Training	LOO-CV	K-fold-CV
Transmittance	33[d]	98.9	94.4	92.8 ± 3.8
1st derivative	30[e]	100	98.9	97.6 ± 3.5
2nd derivative	31[f]	98.9	93.3	90.5 ± 6.7
Matrix: Ground almonds				
ATR-FTIR spectrum	Number of variables in the LDA model	Correct classification (sensitivity, %)		
		Training	LOO-CV	K-fold-CV
Transmittance	8[g]	100	100	99.8 ± 1.0
1st derivative	7[h]	100	100	99.8 ± 0.9
2nd derivative	6[i]	100	100	99.7 ± 1.2

[a]Wavenumbers: 732, 734, 968, 1002, 1004, 1045, 1189, 1232, 1259, 1627, 1660, 1687, 1704, 1731, 1853, 1876, 2306, 2308, 2322, 2329, 2331, 2376, 2842, 2912, 3043, 3139, 3496, 3562 e 3596, 3645 cm^{-1}

[b]Wavenumbers: 667, 727, 829, 844, 1006, 1045, 1072, 1091, 1568, 2291, 2318, 2343, 2351, 2356, 2846, 2894, 2908, 3076 e 3475 cm^{-1}

[c]Wavenumbers: 727, 1012, 1186, 1226, 1352, 1419, 1504, 1704, 1708, 1737, 2293, 2300, 2308, 2333, 2339, 2850, 2904 e 3562 cm^{-1}

[d]Wavenumbers: 663, 850, 931, 991, 1172, 1182, 1201, 1236, 1240, 1249, 1297, 1303, 1326, 1332, 1340, 1375, 1544, 1546, 1587, 1637, 1654, 1666, 1689, 1733, 2823, 2950, 3068, 3109, 3454, 3535, 3693, 3708 e 3722 cm^{-1}

[e]Wavenumbers: 642, 694, 709, 734, 831, 858, 919, 952, 962, 1054, 1074, 1145, 1178, 1270, 1454, 1475, 1556, 1591, 1647, 1755, 1859, 1863, 1868, 2530, 2939, 3049, 3058, 3093, 3103 e 3425 cm^{-1}

[f]Wavenumbers: 638, 644, 833, 891, 948, 968, 1000, 1012, 1027, 1178, 1180, 1257, 1396, 1512, 1535, 1564, 1598, 1691, 1785, 1814, 1826, 1855, 2821, 2914, 3006, 3070, 3087, 3608, 3649, 3654 e 3656 cm^{-1}

[g]Wavenumbers: 970, 995, 1054, 1382, 1595, 1604, 1643 e 3679 cm^{-1}

[h]Wavenumbers: 1101, 1222, 1328, 1388, 1623, 2327 e 2372 cm^{-1}

[i]Wavenumbers: 850, 970, 1027, 1114, 1620 e 1652 cm^{-1}

epicarp, kernel, and ground almonds, ground almonds emerged as the optimal choice for developing a simple yet accurate linear multivariate discriminant model for cultivar discrimination. The research further suggested that while more complex models might be required, ATR-FTIR spectra obtained from the epicarp could provide a viable non-invasive alternative for almond cultivar identification. Particularly, the study found that pre-processing ATR-FTIR spectra was unnecessary, as classification models based on raw transmittance spectra performed comparably to those using transformed forms, such as first and second derivatives. In conclusion, the research proposed a classification approach using ATR-FTIR-LDA-SA based on raw transmittance spectra of ground almonds. This method emerged as a rapid, cost-effective, and minimally invasive tool for almond producers and industry stakeholders to ensure cultivar traceability throughout the almond supply chain. However, the study's validation will require a broader range of almond cultivars.

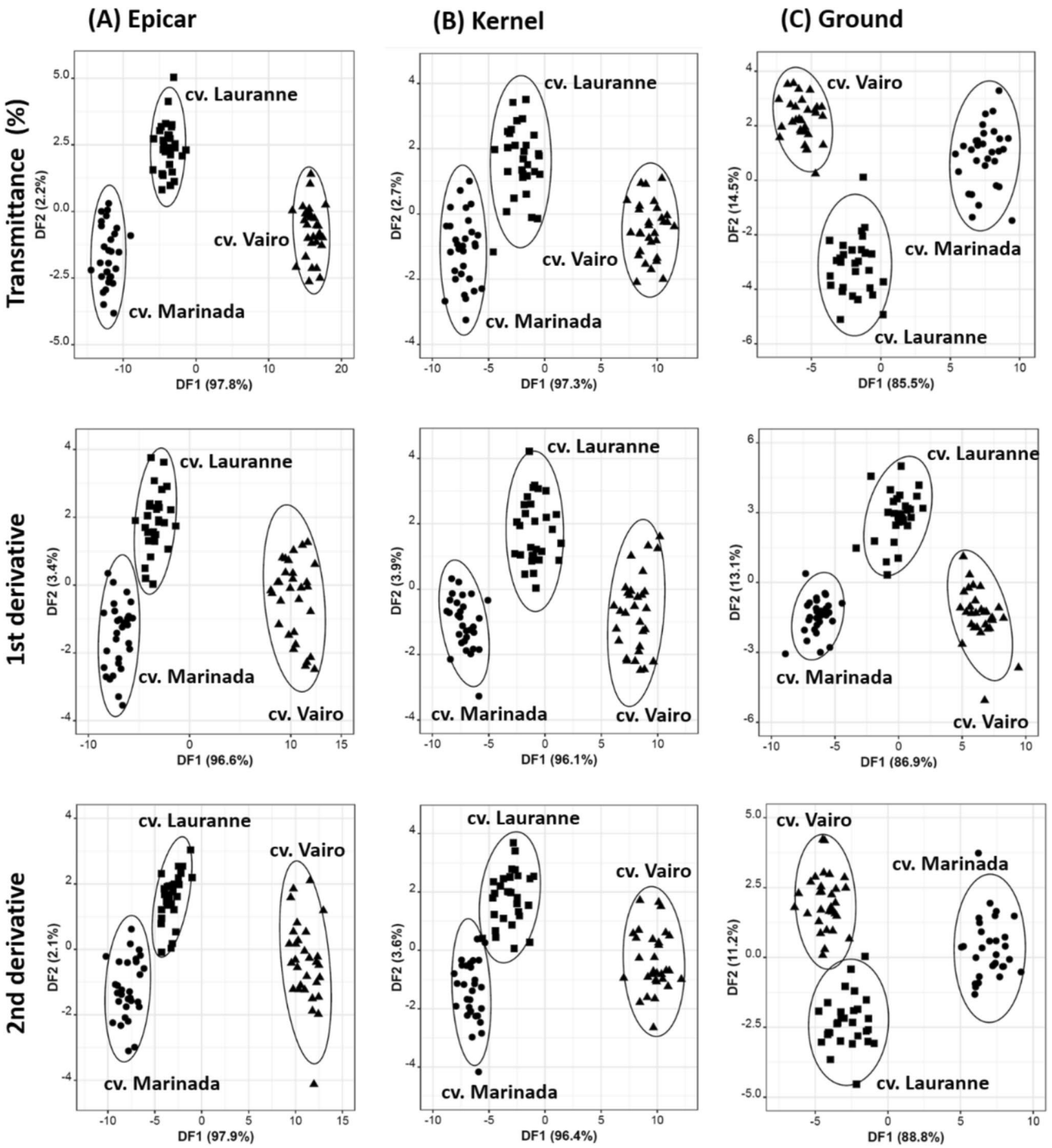

Fig. 8 Discrimination of almond samples (training: original data) according to their cultivars (■ cv. Lauranne, ● cv. Marinada, and ▲ cv. Vairo) using LDA models established using data from ATR-FTIR spectra (transmittance, 1st derivative and 2nd derivative) referring to subsets of wavenumbers selected by the SA algorithm

Acknowledgements This work was supported by national funds through FCT/MCTES (PIDDAC): CIMO, UIDB/00690/2020 (DOI: 10.54499/UIDB/00690/2020) and UIDP/00690/2020 (DOI: 10.54499/UIDP/00690/2020); and SusTEC, LA/P/0007/2020 (DOI: 10.54499/LA/P/0007/2020). National funding by FCT- Foundation for Science and Technology, through the institutional scientific employment program-contract with Nuno Rodrigues.

Author contributions Conceptualization, N.R., I.L-C., B.H., J.A.P. and A.M.P.; methodology, S.L., A.S.-E., I.P., B.H., J.M. and N.R.; software, A.M.P.; formal analysis, S.L., A.S.-E., I.P. J.M. and N.R.;

investigation, S.L., N.R., A.S.-E., I.P., B.H., J.M., I.L-C., J.A.P. and A.M.P.; resources, N.R., J.A.P. and A.M.P.; writing—original draft preparation, S.L.; writing—review and editing, N.R., I.L-C., B.H., J.A.P. and A.M.P.; supervision, N.R., A.S.-E., J.A.P. and A.M.P project administration, N.R., J.A.P. and A.M.P.; funding acquisition, N.R. and A.M.P.; All authors have read and agreed to the published version of the manuscript.

Funding Open access funding provided by FCT|FCCN (b-on).

Data availability Not applicable.

References

1. FAOSTAT, Food and Agriculture Organizations of the United Nations – FAO Database, 2022. Available online: https://www.fao.org/faostat/en/#data. (accessed on 11 October 2023).
2. C.E. Berryman, A.G. Preston, W. Karmally, R.J. Deckelbaum, P.M. Kris-Etherton, Effects of almond consumption on the reduction of LDL-cholesterol: A discussion of potential mechanisms and future research directions. Nutr. Rev. **69**, 171–185 (2011)
3. V. Cortés, J.M. Barat, P. Talens, J. Blasco, M.J. Lerma-García, A comparison between NIR and ATR-FTIR spectroscopy for varietal differentiation of Spanish intact almonds. Food Control **94**, 241–248 (2018)
4. M.A. Askin, M.F. Balta, F.E. Tekintas, A. Kazankaya, F. Balta, Fatty acid composition affected by kernel weight in almond (Prunus dulcis (Mill.) D.A. Webb.] genetic resources. J. Food Compos. Anal. **20**, 7–12 (2007)
5. A.V. García, A.B. Sanahuja, M.C.G. Selva, Characterization and classification of almond cultivars by using spectroscopic and thermal techniques. J. Food Sci. **78**, 1–7 (2013)
6. A. Piscopo, F.V. Romeo, B. Petrovicova, M. Poiana, Effect of the harvest time on kernel quality of several almond varieties (Prunus dulcis (Mill.) D.A. Webb). Sci. Hortic. **125**, 41–46 (2010)
7. C.L. Salcedo, B.A.L. Mishima, M.A. Nazareno, Walnuts and almonds as model systems of foods constituted by oxidisable, pro-oxidant and antioxidant factors. Food Res. Int. **43**, 1187–1197 (2010)
8. R.G. Solsona, C. Boix, M. Ibáñez, J.V. Sancho, The classification of almonds (Prunus dulcis) by country and variety using UHPLC-HRMS-based untargeted metabolomics. Food Addit. Contam.: Part A **35**, 395–403 (2017)
9. O. Kodad, G. Estopañán, T. Juan, J.M. Alonso, M.T. Espiau, R. Socias i Company, Oil content, fatty acid composition and tocopherol concentration in the Spanish almond genebank collection. Sci. Hortic. **177**, 99–107 (2014)
10. M.M. Özcan, A review on some properties of almond: impact of processing, fatty acids, polyphenols, nutrients, bioactive properties, and health aspects. J. Food Sci. Technol. **60**, 1493–1504 (2022)
11. A. Valdés, L. Vidal, A. Beltrán, A. Canals, M.C. Garrigós, Microwave-assisted extraction of phenolic compounds from almond skin byproducts (Prunus amygdalus): a multivariate analysis approach. J. Agric. Food Chem. **63**, 5395–5402 (2015)
12. A.J. Zamany, G.R. Samadi, D.H. Kim, Y.S. Keum, R.K. Saini, Comparative study of tocopherol contents and fatty acids composition in twenty almond cultivars of Afghanistan. J. Am. Oil Chem. Soc. **94**, 805–817 (2017)
13. M.A. Faqeerzada, S. Lohumi, R. Joshi, M.S. Kim, I. Baek, B.K. Cho, Non-targeted detection of adulterants in almond powder using spectroscopic techniques combined with chemometrics. Foods **9**, 879 (2020)
14. P. Firmani, R. Bucci, F. Marini, A. Biancolillo, Authentication of "Avola almonds" by near infrared (NIR) spectroscopy and chemometrics. J. Food Compos. Anal. **82**, 103235 (2019)
15. V. Cortés, P. Talens, J.M. Barat, M.J. Lerma-García, Discrimination of intact almonds according to their bitterness and prediction of amygdalin concentration by Fourier transform infrared spectroscopy. Postharvest Biol. Technol. **148**, 236–241 (2019)
16. A.B. Sanahuja, M.S.P. Moya, S.E.M. Pérez, N.G. Teruel, M.L.M. Carratalá, Classification of four almond cultivars using oil degradation parameters based on FTIR and GC data. J. Am. Oil Chem. Soc. **86**, 51–58 (2009)
17. S. Sharma, P. Neeraj, FTIR spectroscopic characterisation of almond varieties (Prunus dulcis) from Himachal Pradesh (India). Int. J. Curr. Microbiol. Appl. Sci. **7**, 887–898 (2018)
18. UPOV, International Union for the Protection of New Varieties of Plants. Guidelines for the Conduct of Tests for Distinctness, Uniformity and Stability. Almond (Prunus dulcis (Mill.) D. A. Webb). Geneva, Switzerland, TG/56/4, 2011. Available online: https://www.upov.int/meetings/en/doc_details.jsp?meeting_id=22092&doc_id=177773 (accessed on 04 January 2024).
19. C.M. Bishop, *Pattern recognition and machine learning*, 1st edn. (Springer, New York, 2006)
20. N. Mougiou, P. Maletsika, A. Konstantinidis, K. Grigoriadou, G. Nanos, A. Argiriou, Morphological and molecular characterization of a new self-compatible almond variety. Agriculture **13**, 1362 (2023)
21. F. Graeff, L. Fernandes, J. A. Pereira, C. Garcia, E. Ramalhosa, Caracterização físico-química de dez variedades de amêndoa (Prunus dulcis), Revista de Ciências Agrárias **43**(Especial 2): 084-090 (2020)
22. A. Rabadán, M. Álvarez-Ortí, J.E. Pardo, Morphological, mechanical and sensory properties of almond (Prunus dulcis L.) and pistachio (Pistacia vera L.) cultivars grown in Spain. Kuwait J. Sci. **46**(3), 103–113 (2019)
23. A. Das, M. Banerjee, N. Bar, S.K. Das, Adsorptive removal of Cr(VI) from aqueous solution: kinetic, isotherm, thermodynamics, toxicity, scale-up design, and GA modelling. SN Appl. Sci. **1**, 776 (2019)
24. X. Li, Y. Liu, J. Hau, W. Wang, Study of almond shell characteristics. Materials **11**, 1782 (2018)
25. A. Valdés, A. Beltrán, I. Karabagias, A. Badeka, M.G. Kontominas, M.C. Garrigós, Monitoring the oxidative stability and volatiles in blanched, roasted and fried almonds under normal and accelerated storage conditions by DSC, thermogravimetric analysis and ATR-FTIR. Eur. J. Lipid Sci. Technol. **117**, 1199–1213 (2015)
26. A.R. Sidhu, S. Naz, S.A. Mahesar, A.R. Khaskheli, H. Shoaib, H. Khan, M.Q. Samejo, F.J. Siyal, Z.H. Shar, S. Pirzada, A.A. Unar, A.A. Kandhro, Quality assessment on the oxidative stability of almond kernels during extensive storage time. Pak. J. Anal. Environ. Chem. **22**, 332–343 (2021)
27. M.F. Manzoor, X.A. Zeng, A. Rahaman, A. Siddeeg, R.M. Aadil, Z. Ahmed, J. Li, D. Niu, Combined impact of pulsed electric field

and ultrasound on bioactive compounds and FT-IR analysis of almond extract. J. Food Sci. Technol. **56**, 2355–2364 (2019)
28. C. Chun-Song, W. Can-Jian, L. Jie, L. Chi-Chou, Z. Hua, Z. Zhi-Feng, A new approach for identification of medicinal almonds by fourier transform infrared spectroscopy and systematic clustering of characteristic peaks. Chin. J. Nat. Med. **15**, 0703–0709 (2017)