Journal of Food Measurement and Characterization (2024) 18:437–450
https://doi.org/10.1007/s11694-023-02188-w

ORIGINAL PAPER

Chemometrics using near-infrared spectra for the quantification of robusta coffee and chicory added as adulterants in roasted arabica coffee

Leah Munyendo[1] · Majharulislam Babor[3] · Yanyan Zhang[2] · Bernd Hitzmann[1]

Received: 19 June 2023 / Accepted: 23 September 2023 / Published online: 12 October 2023

Abstract

Roasted ground coffees are targets of concern regarding intentional adulteration with cheaper foreign materials because, in this form, it may be difficult to detect due to the small particle size and the dark color. Therefore, a significant interest is developing fast, sensitive, and accurate methodologies to quantify adulterants in roasted coffees. This study investigated the potential of using near-infrared spectroscopy (NIR) to quantity robusta coffee and chicory in roasted arabica coffee. The adulterated arabica coffee samples were composed of robusta coffee or chicory ranging from 2.5 to 30% in increments of 2.5%. Four regression approaches were applied: gradient boosting regression (GBR), multiple linear regression (MLR), k-nearest neighbor regression (KNNR), and partial least squares regression (PLSR). The first three regression models were performed on the features extracted from linear discriminant analysis (LDA) or principal component analysis (PCA). Additionally, two classification methods were applied (LDA and KNN). The regression models derived based on LDA-extracted features presented better performances than PCA ones. The best regression models for the quantification of robusta coffee were GBR (pRMSEP of 13.70% and R^2 of 0.839) derived based on PCA-extracted features and MLR (pRMSEP of 1.11% and R^2 of 0.998) derived based on LDA-extracted features. For the chicory quantification, the same models derived under the same settings as mentioned above also presented the best performances (GBR, pRMSEP = 9.37%, R^2 = 0.924; MLR, pRMSEP = 1.54%, R^2 = 0.997). The PLSR prediction errors for the quantification of arabica coffee and chicory were 9.90% and 8.09%, respectively. For the classification methods, the LDA model performed well compared to KNN. Generally, some models proved to be effective in quantifying robusta and chicory in roasted arabica coffee. The results of this study indicate that NIR spectroscopy could be a promising method in the coffee industry and other legal sectors for routine applications involving quality control of coffee.

Keywords Adulteration · NIR spectroscopy · Coffee · Chicory · Regression

Introduction

Desirable sensory properties, stimulant effects of caffeine, and several health benefits are among the factors contributing to the popularity of coffee beverages globally [1]. Global coffee consumption amounted to approximately 9.98 million tonnes during the 2020/2021 period, a 1.29% increase compared to 2019/2020 [2]. Because of the continuous increase in coffee's demand and ultimately its price on the global market, industrial coffee producers may interfere with the quality of the product for economic gains. Adulteration of coffee may involve interfering with the quality of beans (geographical origin, defective beans, and species) or the addition of lower-value products such as cereals, coffee husks, and chicory among others [3]. Roasted and ground coffee is

✉ Leah Munyendo
leah.munyendo@uni-hohenheim.de

[1] Department of Process Analytics and Cereal Science, University of Hohenheim, Garbenstr. 23, 70599 Stuttgart, Germany

[2] Department of Flavor Chemistry, University of Hohenheim, Fruwirthstraße 12, 70599 Stuttgart, Germany

[3] Department of Data Science in Bioeconomy, Leibniz Institute for Agricultural Engineering and Bioeconomy, Max-Eyth-Allee 100, 14469 Potsdam, Germany

particularly vulnerable to such malpractices since adulterants with similar physical characteristics as those of coffee can be mixed and may not be easily recognized. Reports on different forms of adulteration in commercial coffees as part of economically motivated adulteration exist [4, 5].

Even though this practice generates higher profits for traders, it results in products' loss in quality (taste, aroma, and nutritional value) and may imply danger to consumers' health. Due to these factors, quality control of commercial coffee is essential for ensuring the authenticity of the product on the market and the safety of consumers. Since it is impossible to detect adulterated coffees with a simple visual inspection, different analytical methods have been developed. Núñez et al. [6] combined the HPLC method with chemometrics to quantify common adulterants added to coffee. Song et al. [7] also employed a similar technique to identify adulterated coffee samples based on different chemical indices (monosaccharides, nicotinic acid, and trigonelline). The feasibility of DNA-based approaches to detect and quantify adulterants in coffee has also been investigated [8–10]. In these techniques, the identification of adulterants is based on polymerase chain reaction (PCR) amplification of DNA regions exhibiting genetic variations that exist between the coffee and the adulterants. Souto et al. [11] and Rahman et al. [12] proposed a technique based on ultraviolet–visible spectroscopy in combination with chemometrics to identify adulterants in ground-roasted coffees. Other methods based on electrospray ionization mass spectrometry, the use of digital images and capillary electrophoresis-tandem mass spectrometry have also been proposed [13–15].

Although these methods are effective, they have been criticized for their requirement of high technical expertise, lengthy analytical time, being expensive, and being environmentally unfriendly considering the chemicals needed. Spectroscopic techniques such as NIR coupled with chemometrics offer a great replacement for the aforementioned methods for detecting and quantifying adulterants in foods [16, 17]. This is because of their rapidity, ease of use, and reliability. They allow for direct analysis of the solid samples with no or minimal sample preparation. Chakravartula et al. [18] employed NIR spectroscopy and a convolutional neural network to quantify chicory, barley, and maize in arabica coffee. Correia et al. [19] studied the feasibility of NIR spectroscopy combined with PLSR to quantify robusta coffee, corn, peels, and sticks in arabica coffee. Harohally and Thomas, [20] and Boadu et al. [21] employed a similar approach to quantify chicory and coffee husks in coffee, respectively. Forchetti and Poppi [22] proposed a methodology for the quantification of adulterants in coffee based on the combination of near-infrared hyperspectral imaging and multivariate curve resolution with errors lower than 4%. In most published work on the quantification of adulterants in coffee, one or two chemometric methods, mainly PLSR have been used to develop the models. Considering coffee fraud is one of the emerging concerns in the global coffee market, more investigations on methods for the quantification of adulterants in coffee are necessary. To the best of our knowledge, the integration of feature extraction (using PCA and LDA methods) and regression techniques into NIR spectroscopy for the quantification of adulterants in roasted arabica coffee has not been exploited.

This work aims to use NIR spectroscopy to quantify robusta coffee and chicory added as adulterants in roasted arabica coffee using different regression and classification methods. This is a further step of our previous study, which demonstrated the feasibility of NIR complemented by an autoencoder in detecting adulterants in roasted coffee [23]. Arabica and robusta coffee are among the coffee species that have major commercial and economic importance. However, arabica coffees are considered to be of higher quality due to their sensorial properties thus making them expensive [24]. Because of these price differences, mislabeling or fraudulence for economic gain is possible. Chicory, on the other hand, is used as a coffee substitute due to its flavor attributes. Nonetheless, on some occasions, it is also used as a non-declared adulterant due to its low-cost compared to coffee [25].

In the present study, we applied four different regression approaches: GBR, MLR, KNNR, and PLSR to select the best-performing method for practical applications. Two classification techniques were also applied (LDA and KNN). Specific objectives were to (1) calibrate predictive models by using different combinations of dimensionality reduction approaches (PCA and LDA) and regression and classification methods and (2) compare and select the most effective methods for the quantification of robusta coffee and chicory in roasted arabica coffee.

Materials and methods

Preparation of samples

Green coffee beans of arabica and robusta species were acquired from Buxtrade GmbH, An den Geestbergen 1, 21614 Buxtehude, Germany, and Hochland Kaffee Hunzelmann GmbH und Co. KG, Germany, respectively. Raw chicory root was sourced from Detrade UG, Bruchstrasse 14d, 28816 Stuhr, Germany. The coffee samples were roasted in a Gene Café CBR-101 coffee roaster (Gene Café, Korea) at 240 °C with varying duration to three roast levels: light (10 min), medium (15 min), and dark (20 min). Chicory root was roasted using the same roaster at the same temperatures but for 4, 5, and 6 min. Shorter times were necessary to achieve a similar color to that of arabica coffee because of its small size and low

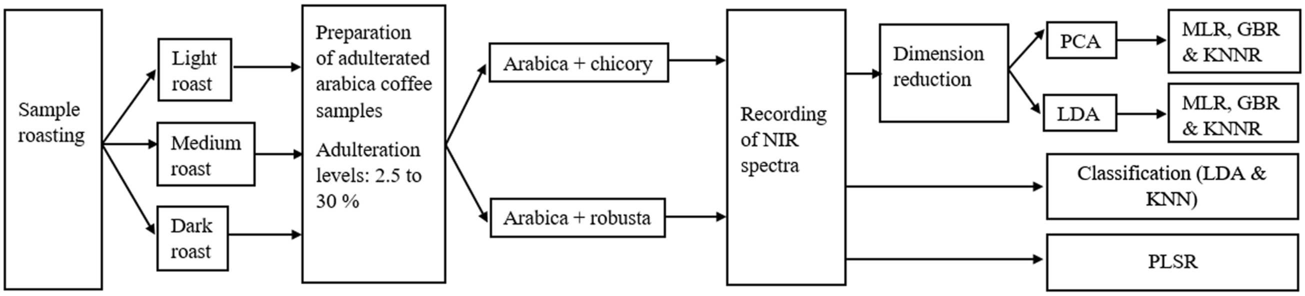


Fig. 1 Sample preparation, NIR spectra acquisition, and modelling for the quantification of robusta coffee and chicory in roasted arabica coffee

moisture content. Subsequently, samples were ground with an electric grinder (Melitta Calibra EU 1027-01 Mill 160 W, Germany) on a fine grind setting. Adulterated samples were then prepared by mixing arabica coffee with robusta coffee or chicory in different mass percentages. Specifically, appropriate amounts of arabica coffee grinds were weighed and chicory or robusta coffee added at different concentrations. Mechanical mixing using a 3D mixer (Turbula Willy A. Bachofen, Switzerland) for 5 min followed this. Each adulterant was added at 0, 2.5, 5, 7.5, 10, 12.5, 15, 17.5, 20, 22.5, 25, 27.5, and 30% (w/w) concentration levels. The light-roasted arabica coffee was mixed with adulterants (chicory or robusta coffee) roasted at the equivalent roast level. The same was applied to medium and dark-roasted samples. All adulteration levels were prepared in triplicate resulting in 117 samples for each adulterant (3 roast levels × 13 adulteration levels × 3 replicates). The prepared samples were stored in opaque zip lock bags at − 21 °C awaiting analysis. NIR spectra of the samples were acquired, followed by model calibration and testing. Figure 1 presents a summary of sample preparation, NIR spectra acquisition, and modeling for the quantification of robusta coffee and chicory in roasted arabica coffee.

NIR measurements

NIR analysis of adulterated and pure arabica coffee was performed using a Fourier Transform NIR spectrometer (Bruker Optics, Ettlingen, Germany) associated with OPUS software (Version 7, Bruker Optics, Ettlingen, Germany) for instrumental control and spectra acquisition. The spectra were collected in diffuse reflectance mode. Each sample was thoroughly mixed before transferring enough amounts to the sample holder. During measurement, the sample holder was kept in rotation to collect representative spectra of the sample. Each spectrum was recorded as an average of 64 scans. Spectral data were collected over the wavenumber range of 12,500–3600 cm^{-1} and a resolution of 4 cm^{-1}. Twenty spectra per sample were acquired for each sample triplicate.

Features extraction

This study investigated different regression methods to quantify adulterants (robusta coffee and chicory) in roasted arabica coffee using Python (version 3.9, scikit-learn package). Before applying regression models, raw spectra were subjected to PCA or LDA analysis, to reduce the number of dimensions in the spectra (4615 wavelengths) by transforming highly correlated wavelengths into a manageable set of features that could sufficiently explain the variance in the original data set. LDA finds a feature space that maximizes separability between the adulteration levels in target values while PCA focuses on finding the direction of maximum variance in the spectra [26].

LDA can be explained in the two following steps. In the first step, between-adulteration levels (S_b) and within-adulteration level variance (S_w) are computed from the spectra by using Eqs. (1) and (2), respectively. The following step involves increasing the between-adulteration levels variance and lowering the within-adulteration level variance that forms a lower-dimensional space. For LDA analysis, the singular value decomposition technique was employed to find the most discriminative transformation of independent variables (wavelengths) in lower dimensional features [27–29].

$$S_b = \sum_{i=1}^{c} n_i(\overline{x}_i - \overline{x})(\overline{x}_i - \overline{x})^T \tag{1}$$

$$S_w = \sum_{i=1}^{c} (n_i - 1)S_i = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (x_{i,j} - \overline{x}_i)(\overline{x}_{i,j} - \overline{x}_i)^T \tag{2}$$

where c is the number of total adulteration levels, n_i is the sample size of adulteration level i, $\overline{x}_i$ is the sample mean of adulteration level i, $\overline{x}$ is the overall mean, S_i is the scatter matrix for adulteration level i, $x_{i,j}$ is the jth sample of adulteration level i, and T indicates the transpose of the matrices.

LDA is ideally suited to handling object classification problems, and categorical target values are expected [28]. In this investigation, each adulteration level in the calibration set was regarded as a separate category—sorted by ascending order of adulteration levels, such as category 1 for 0% adulteration level, and category 2 for 2.5% adulteration level. By using these categories, the spectra from the calibration set were transformed into LDA features space. In the following stage, the LDA features were used to train a regression model to learn the actual adulteration levels. The following two steps were considered to test the performance of the model. First, an unknown adulteration level spectrum was transformed into LDA features space using the transformation matrix derived from the LDA of calibration spectra. The LDA features of the unknown adulteration level are combined with regression parameters in the second stage to calculate the predicted adulteration level.

The optimal number of principal components/linear discriminants was determined by evaluating the prediction performance of the models and by considering the total explained variance. Different numbers of components/discriminants were tested and those that gave the best prediction results were chosen. For the quantification of chicory adulterated samples, two linear discriminants and six principal components were considered for developing the regression models. Conversely, two linear discriminants and ten principal components were considered to derive models for the quantification of robusta coffee adulterated samples. The term LDA and PCA followed by a regression model will be used to refer to models calibrated using LDA or PCA features, respectively. For example, the LDA-multiple linear regression model means that a multiple linear regression model was calibrated using LDA features.

Calibration and test sets

Twenty spectra per sample were acquired for each sample triplicate (meaning sixty spectra for each pure and adulterated arabica coffee sample). Thus, at every roast level for each adulterant, 780 spectra were obtained (13 adulteration levels × 60 spectra). Subsequently, spectra for the three roast levels (light, medium, and dark) for each adulterant were mixed resulting in two types of adulterated samples: one containing arabica coffee adulterated with chicory and another containing arabica coffee adulterated with robusta coffee. After feature extraction using LDA and PCA methods, the total data set was divided into calibration and test sets for regression modeling. The spectra of adulterated arabica coffee samples with 0%, 5%, 10%, 15%, 20%, 25%, and 30% of robusta coffee or chicory were used as calibration set and the rest as test set (2.5%, 7.5%, 12.5%, 17.5%, 22.5%, and 27.5%). In this orientation, the objective is to validate the performance of models with unseen/unknown spectra of particular adulteration levels that are absent in the calibration set. The calibration set was used to calibrate predictive models while the test set was used to assess the performance of the models. Modeling using extracted features did not apply to the PLSR method (Fig. 1), however, the separation of data into calibration and test sets was done in the same way. For the classification modeling, two replicates were used as a calibration set, and the third replicate as a test set. The modeling was performed using raw spectra data. For the spectra presented in this study (Fig. 2), NIR data were pre-processed using the standard normal variate (SNV) method to remove scattering effects for better visualization and comparison [30].

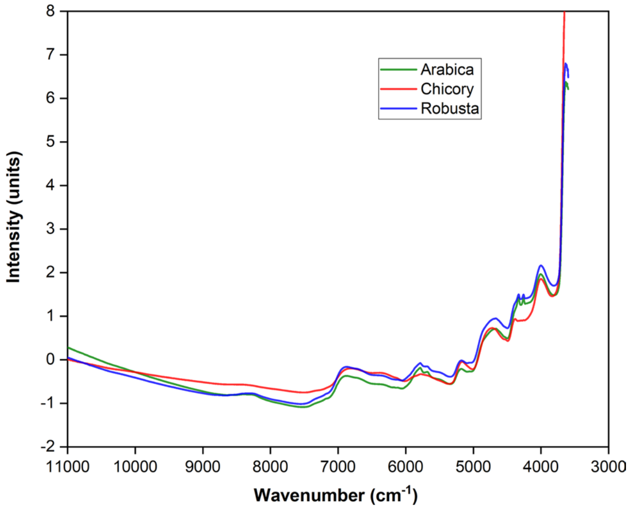


Fig. 2 The average SNV transformed spectra of chicory, arabica, and robusta coffee

Modeling

Partial least squares regression

PLSR models were calibrated by using NIR spectra as X -matrix (independent variables) and adulteration concentrations as y-vectors (dependent variables). The optimal number of latent variables in the PLSR models was estimated by a cross-validation procedure with 10 data splits. Precisely, the number of latent variables, which provided the lowest root mean square error of cross-validation, were selected to ensure good generalizability of the models [31]. Eight and five latent variables were considered for developing the regression models for the quantification of robusta coffee and chicory in arabica coffee, respectively.

Multiple linear regression

Multiple linear regression is a mathematical technique used to model the relationship between multiple independent predictor variables (X) and a single dependent outcome variable (y). Approximation of y was done by a linear combination of extracted features using PCA or LDA. The regression coefficients were estimated by minimizing the error between predicted and observed response values by using the least square method [32]. The mathematical expression for the LDA-multiple linear regression model is shown in the following equation.

$$y_{pred} = p_0 + (p_1 \times LD_1) + (p_2 \times LD_2) \tag{3}$$

where p_0, p_1, and p_2 are the regression coefficients, LD_1 and LD_2 are the first linear discriminant and second linear discriminant, respectively, obtained from LDA of spectra, and y_{pred} is the predicted adulteration level (%) of an adulterant in arabica coffee.

Gradient boosting regression

Gradient boosting regression is an ensemble method with the advantage of capturing the non-linear relationship between the target variable and features, which minimizes the loss function by iteratively adding a new weak model at each step. The algorithm starts by initializing the model with a first guess, which is a decision tree that greatly decreases the loss function (mean square error). Then at each subsequent step, a new model is fitted to the existing residual and added to the previous model to update the residual. Fitting consecutive models to the residuals improve the performance of the model [33]. An optimization process is employed to fit a model against the residuals. The addition of a new model is continued until no further improvement in calibration error is observed. Better results are achieved if, at each iterative step, the contribution of the added model is shrunk using a shrinkage parameter α, called the learning rate. This parameter can take a value between 0 and 1 and the smaller it is, the more accurate the model [33]. In this study, α was set at 0.4, which was tuned based on the lowest calibration error.

K-nearest neighbor regression

The k-nearest neighbor method is a non-parametric classical algorithm that predicts a class of an object based on a similarity metric i.e. distance. It uses different distance metrics to calculate the distance between the target values in the multi-dimensional feature space. Based on the selected metric, the k-nearest neighbor algorithm searches the training data set for the k samples, which are nearest to the sample to be classified. The minimum distance is calculated and the new sample is assigned to the corresponding group [26]. In the case of KNNR, the target variable is numerical. From unknown independent variables, the KNNR model tries to find the k-nearest neighbors who are close to each other in the feature space and predict the target value based on the known target values of the neighbors provided in the calibration set [34]. In this study, the KNNR model was calibrated using features derived by PCA or LDA, with $k = 5$ serving as the number of neighbors and the Euclidean distance metric used to quantify distances.

Classification methods

In order to classify arabica coffee samples adulterated with different concentrations of adulterants, classification strategies based on LDA and KNN were adopted. The aim of the classification methods is to build predictive models for qualitative responses; each of the possible values of the responses is considered a class or category [35]. For the KNN, the parameter k was equal to three. The LDA models were developed using two linear discriminants.

Evaluation of model performance

The performance of the calibrated models was evaluated based on their accuracy in quantifying the adulterants in a test set. Statistical parameters used to compare the models' performance included: the root mean square error of prediction (RMSEP), percentage root mean square error of prediction (pRMSEP), the range error ratio of prediction (RER), the ratio of prediction to deviation (RPD), and the coefficient of determination (R^2). Low RMSEP and high R^2 indicate better performance of the model. Values of RER and RPD greater than 10 and 3, respectively, indicate good predictive capability of the model [31, 34]. For the classification models, their performance were evaluated based on the sensitivity and specificity as calculated by Basri et al. [36]. The other parameters were calculated using the following equations [34, 35].

$$\mathrm{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n}} \tag{4}$$

$$pRMSEP[\%] = \frac{RMSEP}{m} \times 100 \tag{5}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \tag{6}$$

$$RPD = \frac{SD}{RMSEP} \tag{7}$$

$$RER = \frac{m}{RMSEP}, \tag{8}$$

where y_i and $\hat{y}_i$ are the actual and predicted adulteration concentrations [%] in sample i, respectively, *n* is the total number of samples in the test set, *m* is the range of adulteration concentration [%] calculated from the difference between maximum and minimum concentration in the test set, *SD* is the standard deviation, $\bar{y}$ is the mean of all actual adulteration concentrations in the test set.

Results and discussion

Overview of the spectra

The average SNV transformed spectra for arabica coffee, chicory, and robusta coffee are shown in Fig. 2. The main difference in the absorption band among the three samples is in the region 5600–6000 cm^{-1}. This spectral band is related to the first overtones of C–H stretching attributed to long-chain fatty acids [37]. For chicory, a curve trend in this region is different from that of arabica and robusta coffee clearly indicating differences in chemical composition between coffee and chicory. The same band is also characterized by the first overtone and combinations of C–H vibrations in CH_3 groups which is related to caffeine [38]. This could explain the observed high spectral intensities in robusta coffee since it has considerably higher caffeine content than arabica coffee [39]. Additionally, chicory is caffeine free and this could explain the difference in the curve trend in this region [40].

The same trend can be seen at the absorption band between 4000 and 4300 cm^{-1} mainly associated with lipids, caffeine, and chlorogenic acids [38, 39] (Fig. 2). The observed absorption differences could be explained by the variances in the content of the above-mentioned compounds in the samples. Robusta coffee has higher caffeine and chlorogenic acid contents compared to arabica coffee and thus, higher absorption intensities in this region [39]. Absorption bands around 7000, 5000, 4800, and 3900 cm^{-1} are mainly associated with water, proteins, carbohydrates, chlorogenic acids, and lipids [39–41] (Fig. 2). In general, the spectra of chicory, arabica, and robusta coffee show differences in their absorption band intensities.

Features extraction

Before model calibration, two well-known dimensionality reduction techniques i.e. LDA and PCA were applied to the spectra. The scatter plots showing scores of the first two components/discriminants of PCA and LDA for the arabica coffee adulterated with robusta coffee and chicory are presented in Fig. 3. PCA (Fig. 3a, c) showed less separability between the levels of adulteration in projected scores compared to LDA (Fig. 3b, d).

The two LDA discriminants discriminated adulteration levels in the feature space more accurately giving thirteen groups (Fig. 3b, d) compared to PCA where intermixing of adulteration levels occurred (Fig. 3a, c). This could be explained by the fact that LDA identifies the feature subspace that maximizes the separability of adulteration levels and minimizes the variance within the adulteration levels. While PCA concentrates on identifying the direction of maximum variance regardless of the adulteration levels [26]. This may imply that a model trained with LDA-extracted features may perform better than PCA-extracted features.

The loading plots in Fig. 4 provide information on the wavenumbers that contribute to the differentiation between the samples. Examining PCA loadings (Fig. 4a, c), the main wavenumber region responsible for the separation of arabica coffee with different adulterant concentrations (Fig. 3a, c) are between 7500 and 11,000 cm^{-1} as exhibited by high loading values. This region is related to the second and third overtone of C–H vibrations in CH_3 groups attributed to carbohydrates, lipids, caffeine, fatty acids, amino acids, and phenolic acids [19, 38, 42]. For LDA loadings (Fig. 4b, d), the NIR region of high influence as shown with high loading values is between 3500 and 7000 cm^{-1}. The first overtone and combinations of C–H vibrations in CH_3 groups, first overtones of O–H and N–H characterize this region corresponding to proteins, water, chlorogenic acids, caffeine, carbohydrates, and trigonelline [36, 43]. The NIR region of important influence is different for PCA and LDA and this could imply differences in the performance of the models calibrated using the extracted features of the two techniques.

Calibration models

Different regression models were constructed to quantify the contents of robusta coffee in roasted arabica coffee (Table 1). The chemical composition of the two species exhibits differences mainly in caffeine, phenolic acids, lipids, and sugars [44–46]. However, during the roasting process, changes in the content of these compounds due to degradation could lead to difficulties in quantifying a particular species in a blend. Regression models constructed

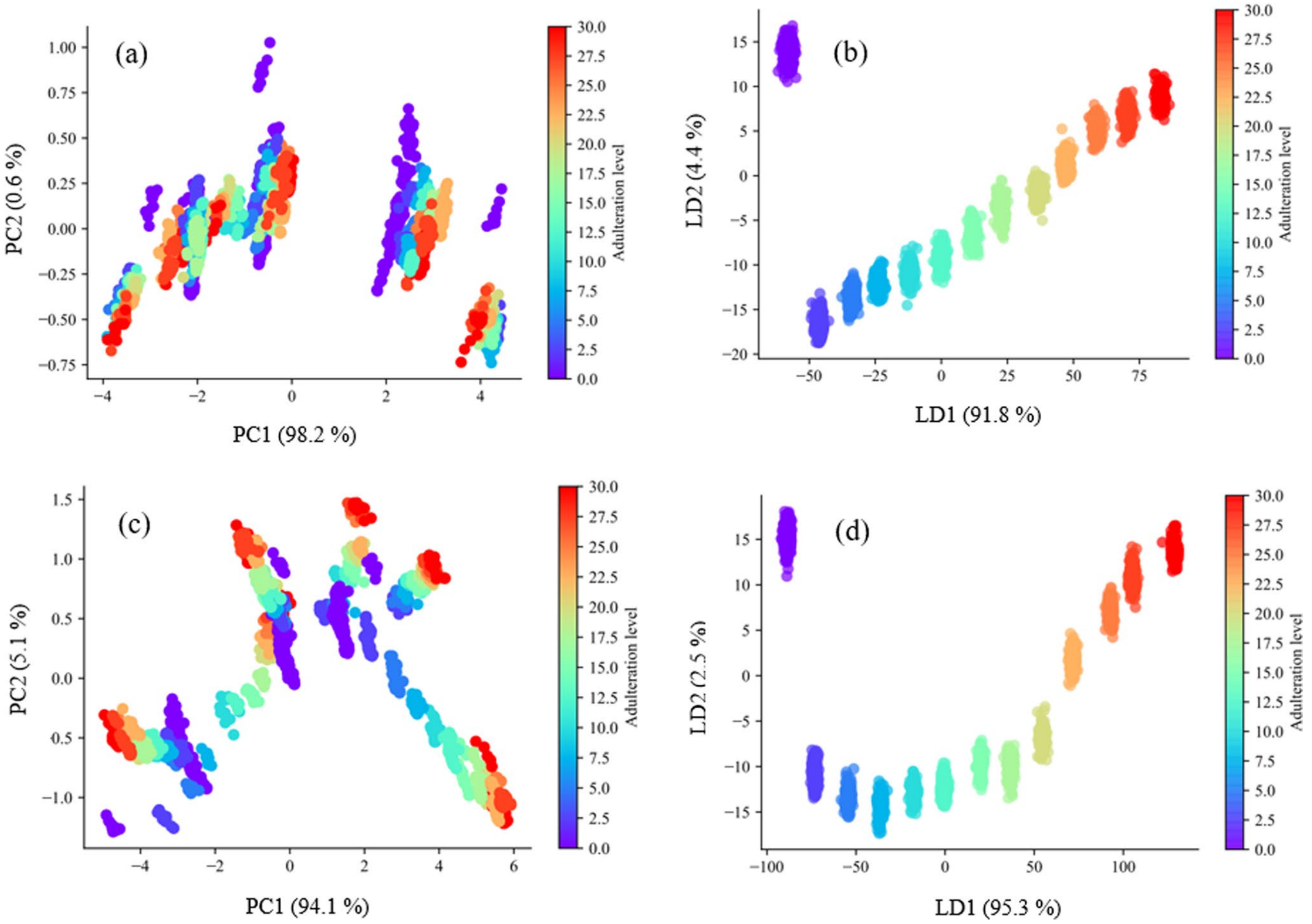


Fig. 3 PCA and LDA score plots of the samples' raw spectra. Plots 3a (PCA) and 3b (LDA) are for the arabica coffee adulterated with robusta coffee while plots 3c (PCA) and 3d (LDA) are for the arabica coffee adulterated with chicory

from LDA-extracted features presented the best results with pRMSEP and R^2 of between 1.11–9.33% and 0.925–0.998, respectively (Table 1). Among these models, the multiple linear regression model recorded the lowest prediction error (pRMSEP = 1.11%) and the highest R^2 (0.998) indicating its efficiency in quantifying robusta coffee in arabica coffee (Fig. 5). The model also presented satisfactory RER (89.84) and RPD (30.70) values (Table 1).

Although regression models derived based on the PCA-extracted features required a greater number of components (ten) than those based on the LDA-extracted features (two), they did not present good quantification capabilities as highlighted by the high pRMSEP (above 13%) and the low R^2 (below 0.9) (Table 1). Moreover, the RER and RPD values are below 10 and 3, respectively. According to the literature, the values of R^2 greater than 0.9, RER greater than 10 and the RPD greater than 3 indicate a calibration model with good performance [47]. As already mentioned, PCA did not clearly separate all the adulteration levels in the features space (Fig. 3a, c) as observed with LDA and thus, this could explain the poor performance of the models calibrated based on these features.

The accuracies of the models in quantifying specific adulteration levels were also evaluated to provide information on which adulteration level was predicted with low or high errors by the models. The multiple linear and the k-nearest neighbor regression models calibrated with PCA features presented the lowest prediction accuracies (pRMSEP > 5%) for 2.5% and 7.5% adulteration levels, respectively (Table 2). The models calibrated with LDA-extracted features exhibited the highest prediction accuracies for the specific adulteration levels compared to PCA-extracted features; with the LDA-multiple linear regression model recording the lowest RMSEP of below 0.5% for all the adulteration levels (Table 2).

The PLSR model did not show good reliability in predicting the quantity of adulterant robusta coffee in arabica coffee samples as indicated by high pRMSEP (9.90%) and

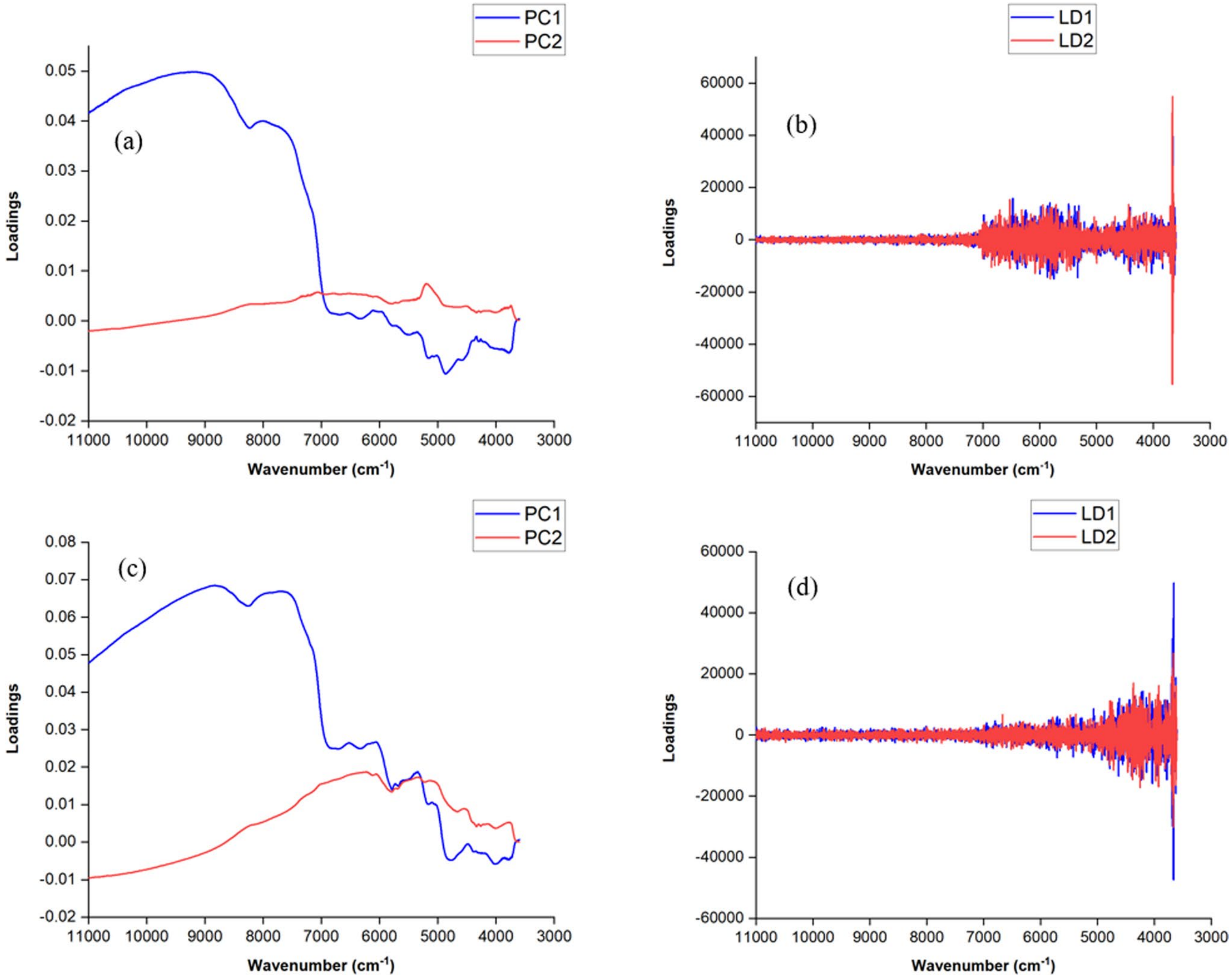


Fig. 4 PCA and LDA loading plots of the samples' raw spectra. Plots 4a (PCA) and 4b (LDA) are for the arabica coffee adulterated with robusta coffee while plots 4c (PCA) and 4d (LDA) are for the arabica coffee adulterated with chicory

relatively low R^2 (0.916) (Table 1). On the contrary, Correia et al. [19] developed robust PLSR models that could quantify robusta coffee in arabica coffee with pRMSEP values between 2.8–6.6% and R^2 values of between 0.957–0.993. Better accuracies of the developed models could have been because of spectral data preprocessing (Savitzky–Golay), which, was not the case in this study where modeling was performed using raw spectra data. Regarding the quantification of individual adulteration levels, the pRMSEP ranged between 2.08 and 2.84% (Table 2). The best prediction was a 7.5% adulteration level with the lowest error of 2.08% (Table 2).

Predictive models for quantifying chicory in the arabica coffee presented considerably better results compared to those for robusta coffee quantification (Table 1). The results could be attributed to the greater differences between the samples in terms of their chemical composition, facilitating their efficient quantification by the models. The spectral feature differences between arabica and chicory can be seen in Fig. 2. Chicory does not contain caffeine, one of coffee beans' main compounds [40]. Considering PCA and LDA features, models calibrated based on LDA-extracted features exhibited good quantification ability of chicory in arabica coffee with low pRMSEP (Table 1). The multiple linear regression model (based on LDA-extracted features), showed the lowest pRMSEP (1.54%) and the highest R^2 (0.997) (Table 1), confirming its reliability in quantifying chicory in unidentified arabica coffee samples. The achieved RER and RPD values of 64.98 and 22.20, respectively, were also satisfactory. Figure 6 shows high accuracy in the prediction of chicory

Table 1 Performance comparison of regression models on the test set for the quantification of robusta coffee and chicory used to adulterate arabica coffee

Adulterant	Model	Components/factors	RMSEP	pRMSEP (%)	R^2	RPD	RER
Robusta coffee	PCA-multiple linear	10	4.19	16.74	0.759	2.04	5.97
	PCA-gradient boosting	10	3.42	13.70	0.839	2.49	7.30
	PCA-k-nearest neighbor	10	4.07	16.29	0.772	2.10	6.14
	LDA-multiple linear	2	0.28	1.11	0.998	30.70	89.84
	LDA-gradient boosting	2	2.33	9.33	0.925	3.66	10.72
	LDA-k-nearest neighbor	2	2.31	9.22	0.927	3.71	10.84
	PLSR	8	2.47	9.90	0.916	3.45	10.10
Chicory	PCA-multiple linear	6	2.45	9.80	0.917	3.49	10.20
	PCA-gradient boosting	6	2.34	9.37	0.924	3.63	10.63
	PCA-k-nearest neighbor	6	2.89	11.55	0.885	2.96	8.66
	LDA-multiple linear	2	0.38	1.54	0.997	22.20	64.98
	LDA-gradient boosting	2	1.98	7.91	0.946	4.26	12.47
	LDA-k-nearest neighbor	2	2.32	9.28	0.926	3.68	10.77
	PLSR	5	2.02	8.09	0.943	4.22	12.35

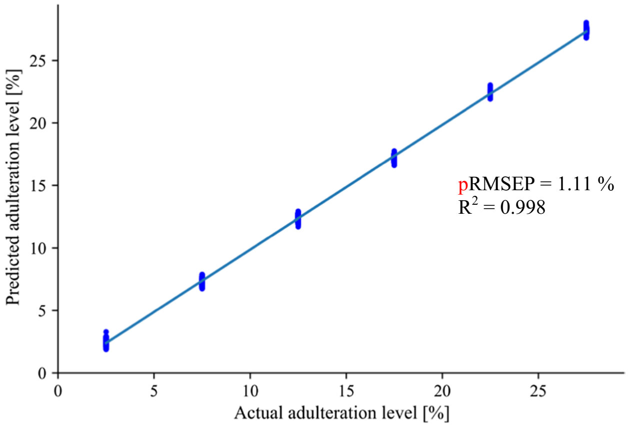


Fig. 5 The actual and predicted concentrations of robusta coffee (%) used to adulterate arabica coffee as determined by the LDA—multiple linear regression model

content mixed in arabica coffee as determined by the LDA-multiple linear regression model. The models based on PCA-extracted features included six components as opposed to those based on LDA-extracted features where two discriminants were used. Among all the constructed models, the PCA-gradient boosting regression model had the lowest pRMSEP (9.37%) and the highest R^2 (0.924) values (Table 1).

The PLSR model for quantifying chicory in arabica coffee exhibited better performance as compared to that developed for quantifying robusta coffee although fewer latent variables (five) were used. The model presented a pRMSEP of 8.09%, an R^2 of 0.943, RER values of 12.35, and RPD values of 4.22 (Table 1). According to Leoni et al. [48], models with RER values between seven and twenty are adequate for quality screening while those with RER values above twenty could be applied for quality control. On the other hand, models with RPD values greater than three could be applied for screening applications and greater than five for quality control applications. Chakravartula et al. [18] proposed PLSR models for the quantification of chicory in arabica coffee with excellent performances as highlighted by R^2 values of above 0.98. It is important to note that in this work, the PLSR modeling was performed using raw spectra data as opposed to the above-mentioned study where the spectra were subjected to different pre-processing methods prior to modeling.

Models were also examined for their ability to quantify individual adulteration levels of the chicory in roasted arabica coffee. For models calibrated using PCA-extracted features, the k-nearest neighbor regression model exhibited the highest pRMSEP for most adulteration levels (Table 2). Compared to models based on PCA-extracted features, those calibrated using LDA-extracted features recorded the best capabilities to quantify individual adulteration levels with the multiple linear regression model presenting the lowest pRMSEP ranging between 0.14 and 0.79% (Table 2).

The presented results can be compared with other studies that have used NIR spectroscopy to quantify adulterants in roasted coffee. For example, Chakravartula et al. [18] described the application of PLSR and convolutional neural network for the quantification of chicory, barley, and maize in roasted arabica coffee and obtained R^2 values of

Table 2 Performance comparison of regression models on the quantification of individual adulteration levels in roasted arabica coffee

Adulterant	Model	Components/factors	RMSEP for individual adulteration levels					
			2.5%	7.5%	12.5%	17.5%	22.5%	27.5%
Robusta coffee	PCA-multiple linear	10	5.24	4.23	4.17	3.51	4.21	3.52
	PCA-gradient boosting	10	3.04	2.65	3.51	3.83	4.27	2.98
	PCA-k-nearest neighbor	10	3.53	5.06	3.83	3.20	3.92	4.61
	LDA-multiple linear	2	0.22	0.30	0.25	0.38	0.24	0.25
	LDA-gradient boosting	2	1.93	2.50	2.50	2.50	1.98	2.50
	LDA-k-nearest neighbor	2	2.50	2.40	2.12	2.33	2.34	2.12
	PLSR	8	2.29	2.08	2.60	2.57	2.84	2.39
Chicory	PCA-multiple linear	6	2.98	1.62	1.47	1.84	1.73	4.00
	PCA-gradient boosting	6	1.69	1.92	2.28	2.10	2.77	3.01
	PCA-k-nearest neighbor	6	2.92	3.03	2.48	2.17	3.53	3.01
	LDA-multiple linear	2	0.21	0.14	0.21	0.16	0.36	0.79
	LDA-gradient boosting	2	1.09	2.50	2.50	2.50	0.75	1.72
	LDA-k-nearest neighbor	2	2.50	2.07	2.39	2.14	2.29	2.50
	PLSR	5	2.31	1.64	1.33	1.61	1.23	3.27

above 0.98. Correia et al. [19] also described the application of PLSR modeling for the quantification of robusta coffee, corn, peels, and sticks in roasted arabica coffee attaining R^2 values of between 0.859 and 0.993. Harohally and Thomas, [20] and Boadu et al. [21] employed a similar approach to quantify chicory and coffee husks in coffee, respectively, presenting models with good performance abilities ($R^2 > 0.97$). In all the studies, PLSR models presented better accuracies in the quantification of adulterants in coffee in terms of R^2 values compared to the ones reported in this study. The reason could be that the spectra were subjected to different pre-processing methods to reduce non-linearity and scattering effects prior to modeling which was not the case in this study. However, other models i.e. GBR, MLR, and KNNR performed fairly well with the raw spectra data depending on the extracted features used to develop them. For instance, multiple linear regression models derived based on LDA-extracted features presented R^2 values of 0.99. In literature, these models have not been explored concerning the quantification of adulterants in coffee and yet they have great potential as exhibited by their reasonably good performance even on the raw spectra.

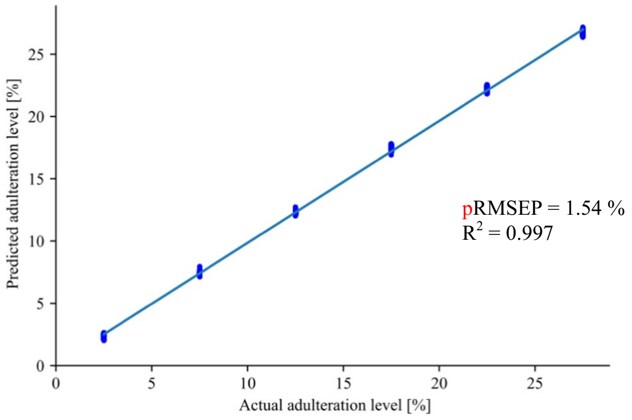
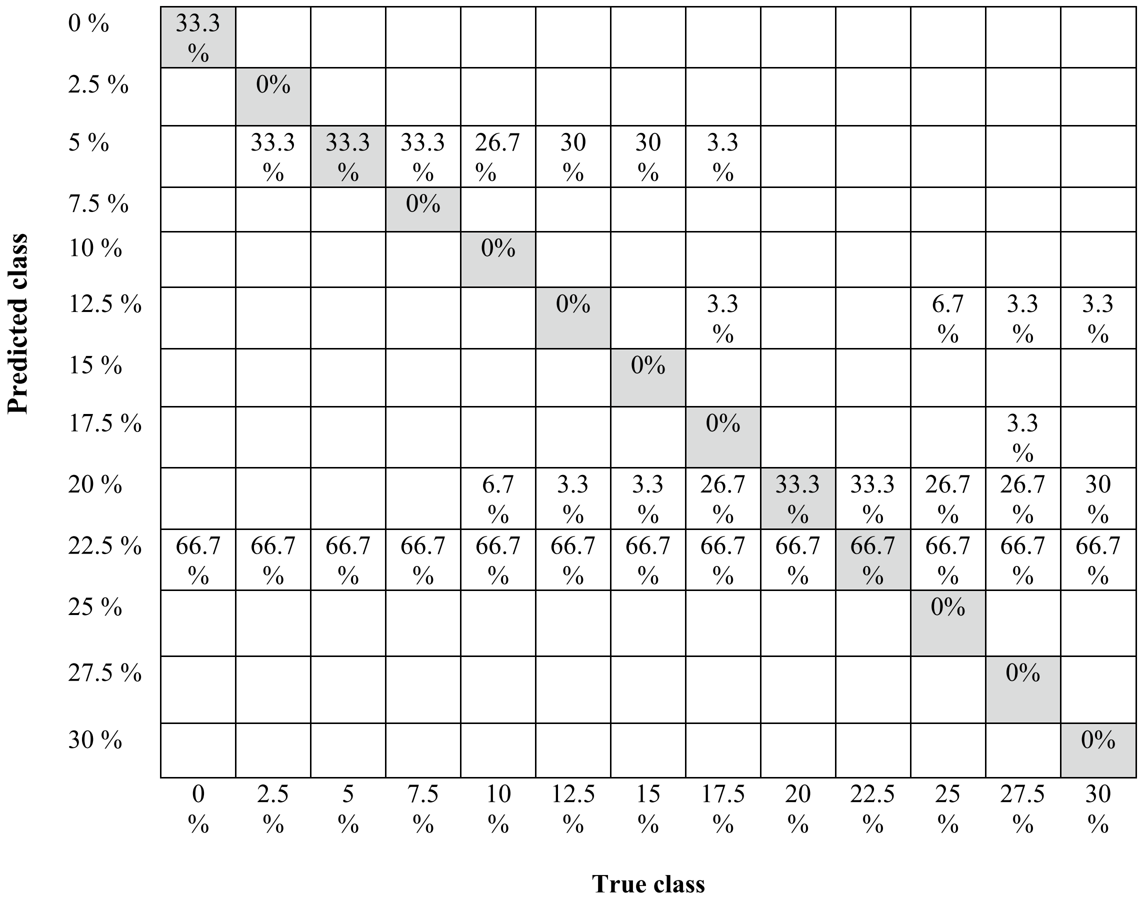


Fig. 6 The actual and predicted concentrations of chicory (%) used to adulterate arabica coffee as determined by the LDA-multiple linear regression model

The confusion matrices shown in Figs. 7 and 8 demonstrate the effect of different adulterant concentrations (robusta coffee) on the classification performance by the KNN and LDA models. Only the results for the arabica coffee adulterated with the robusta coffee are presented in this paper since almost similar findings were observed with samples adulterated with chicory. In the case of the KNN model (Fig. 7), a high percentage of all the classes were misclassified. According to the figure, the 22.5% class achieved the best classification where 66.7% of its samples were correctly classified. The rest were misclassified as those containing 20% robusta coffee. For the remaining classes, 66.7% of the samples of each class were misclassified as 22.5% (Fig. 7). The poor performance of the model can also be observed by its low sensitivity in correctly classifying the classes (Table 3). Conversely, the LDA model made a perfect classification for all the classes with high sensitivity (Fig. 8 and Table 3). The LDA algorithm considers class labels when finding the direction of maximum variance with the goal of maximizing the separation between classes [28]. Thus, this could explain the correct classification of all the classes by the model in the test set.

In summary, considerable differences in the models' performances can be noted from the results. Regarding the dimensionality reduction methods used, LDA transformed



Predicted class	0 %	2.5 %	5 %	7.5 %	10 %	12.5 %	15 %	17.5 %	20 %	22.5 %	25 %	27.5 %	30 %
0 %	33.3 %												
2.5 %		0%											
5 %		33.3 %	33.3 %	33.3 %	26.7 %	30 %	30 %	3.3 %					
7.5 %				0%									
10 %					0%								
12.5 %						0%		3.3 %			6.7 %	3.3 %	3.3 %
15 %							0%						
17.5 %								0%				3.3 %	
20 %					6.7 %	3.3 %	3.3 %	26.7 %	33.3 %	33.3 %	26.7 %	26.7 %	30 %
22.5 %	66.7 %	66.7 %	66.7 %	66.7 %	66.7 %	66.7 %	66.7 %	66.7 %	66.7 %	66.7 %	66.7 %	66.7 %	66.7 %
25 %											0%		
27.5 %												0%	
30 %													0%

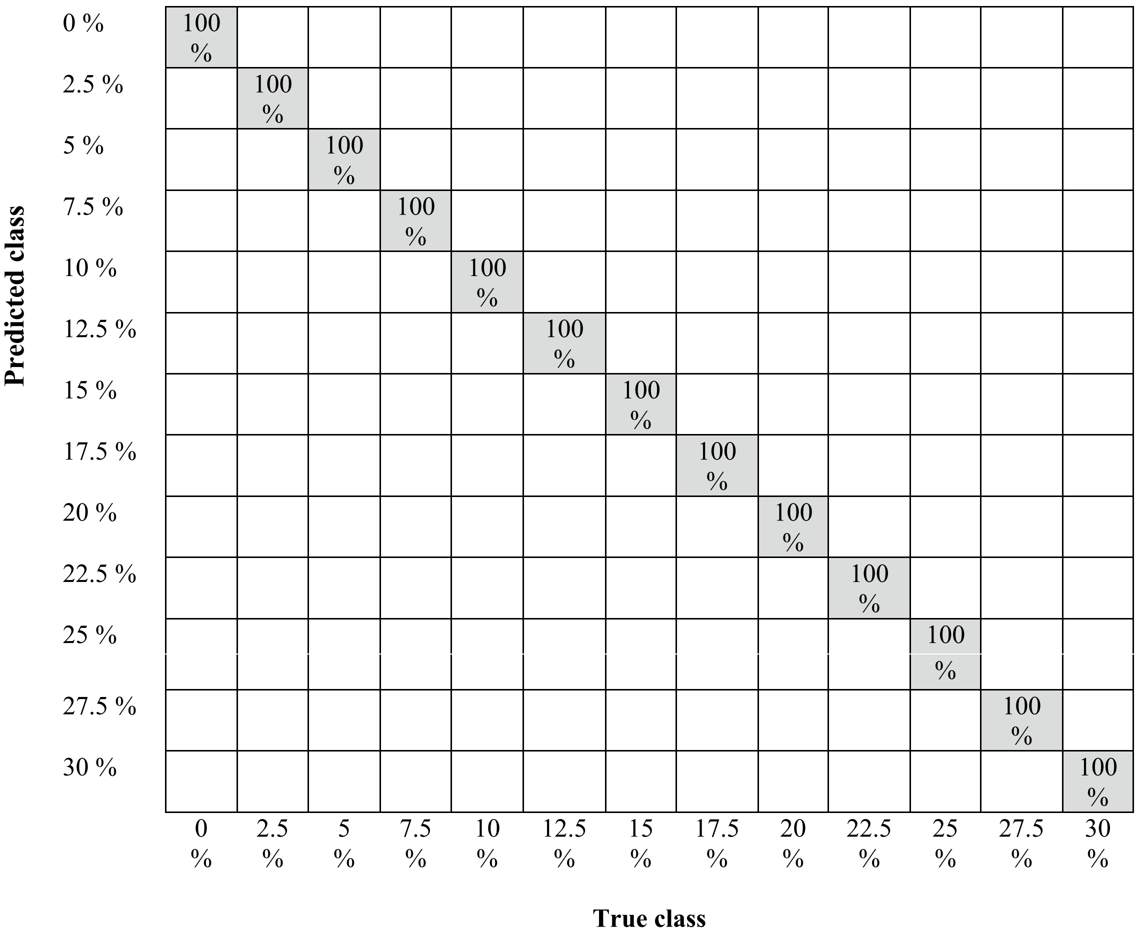
True class

Fig. 7 The confusion matrix for the KNN classification model on the test set of arabica coffee adulterated with robusta coffee (correct classifications are shown in gray)

the high dimensional spectra into lower dimensional features effectively, which could explain the better performance of the models from this approach compared to PCA. It seems to reason that while LDA identifies the feature subspace that maximizes adulteration levels separability and minimizes variance within the adulteration levels, PCA concentrates on identifying the direction of maximum variance regardless of the adulteration levels [26]. Thus with LDA, different adulteration levels were separated into feature space producing regression models suitable to quantify adulterants in unknown arabica coffee samples from the unseen spectra (spectra unknown to the model). Overall, LDA significantly improved the performance of all regression models with only two components as compared to PCA, which included six and ten components for the quantification of chicory and robusta coffee, respectively. For the classification methods, the LDA model performed better compared to KNN. Comparing regression and classification methods presented in this study, the regression technique has the advantage that one can predict adulteration concentrations, which are in between the trained values (2.5%, 7.5%, 12.5%, 17.5%, 22.5%, and 27.5%).

Conclusion

This study evaluated the potential of NIR spectroscopy coupled with features extraction (PCA and LDA) and regression techniques for the quantification of robusta coffee and chicory in roasted arabica coffee. Two classification methods were also studied (LDA and KNN). Regression models derived based on the LDA-extracted features presented better performances than those derived based on the PCA-extracted features. The LDA-multiple linear regression model showed the best prediction performances with pRMSEP of below 1.6% and R^2 values of 0.99 for all the adulterants. Although other studied models presented high pRMSEP (7.91–16.74) and low R^2 values (0.759–0.946), their accuracies could be improved by spectra pre-processing prior to modeling. For the classification methods, the LDA model performed well


Predicted class
True class

	0 %	2.5 %	5 %	7.5 %	10 %	12.5 %	15 %	17.5 %	20 %	22.5 %	25 %	27.5 %	30 %
0 %	100 %												
2.5 %		100 %											
5 %			100 %										
7.5 %				100 %									
10 %					100 %								
12.5 %						100 %							
15 %							100 %						
17.5 %								100 %					
20 %									100 %				
22.5 %										100 %			
25 %											100 %		
27.5 %												100 %	
30 %													100 %

Fig. 8 The confusion matrix for the LDA classification model on the test set of arabica coffee adulterated with robusta coffee (correct classifications are shown in gray)

Table 3 The sensitivity and specificity of the test set (arabica coffee adulterated with robusta coffee) for each class using the k-nearest neighbors and linear discriminant analysis classification models

Class (%)	K-nearest neighbors		Linear discriminant analysis	
	Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)
0	33.33	100	100	100
2.5	0	100	100	100
5	33.33	86.94	100	100
7.5	0	100	100	100
10	0	100	100	100
12.5	0	98.61	100	100
15	0	100	100	100
17.5	0	99.72	100	100
20	33.33	86.94	100	100
22.5	66.67	33.33	100	100
25	0	100	100	100
27.5	0	100	100	100
30	0	100	100	100

compared to KNN. Further studies could be carried out to explore the performance of the classification and regression models explored in this study for the quantification of other possible coffee adulterants and their mixtures. Overall, some models proved to be effective in quantifying robusta and chicory in roasted arabica coffee. The results of this study indicate that NIR spectroscopy could be a promising method in the coffee industry and other legal sectors for routine applications involving quality control of coffee.

Author contributions LM: investigation, methodology, writing—original draft, writing—reviewing and editing, funding acquisition. MB: writing—reviewing and editing, methodology, and formal analysis. YZ: Supervision, writing—reviewing, and editing. BH: conceptualization, methodology, writing—reviewing and editing, validation, and supervision.

Funding Open Access funding enabled and organized by Projekt DEAL. The National Research Fund of Kenya and the German Academic Exchange Service (DAAD), funding ID 57524989, funded this study.

Data availability The datasets generated during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

1. A. Samoggia, B. Riedel, Nutrients **11**, 653 (2019)
2. International Coffee Organization, World Coffee Consumption. (ICO, 2021). http://www.ico.org/trade_statistics.asp?section=Statistic. Accessed 10 January 2023
3. A.T. Toci, A. Farah, H.R. Pezza, L. Pezza, Crit. Rev. Anal. Chem. **46**, 83 (2016)
4. European Commission, Food Fraud Summary, May 2018. (EC, 2018). https://knowledge4policy.ec.europa.eu/publication/food-fraud-summary-may-2018_en. Accessed 20 April 2023
5. European Commission, Food Fraud Summary, January 2022. (EC, 2022). https://knowledge4policy.ec.europa.eu/publication/food-fraud-summary-january-2022_en. Accessed 20 April 2023
6. N. Núñez, J. Saurina, O. Núñez, Foods **10**, 840 (2021)
7. H.Y. Song, H.W. Jang, T. Debnath, K.-G. Lee, Int. J. Food Sci. Technol. **54**, 256 (2019)
8. T. Ferreira, A. Farah, T.C. Oliveira, I.S. Lima, F. Vitório, E.M.M. Oliveira, Food Chem. **199**, 433 (2016)
9. M.C. Combes, T. Joët, P. Lashermes, Food Control **88**, 198 (2018)
10. A.T. Uncu, A. Uncu, Food Control **91**, 32 (2018)
11. U.T.D.C.P. Souto, M.F. Barbosa, H.V. Dantas, A.S. de Pontes, W. da Silva Lyra, P.H.G.D. Diniz, M.C.U. de Araújo, E. da Silva, LWT Food Sci. Technol. **63**, 1037 (2015)
12. M.Z.F.B.A. Rahman, H.W. Chong, V. Lim, Malays. J. Med. Health Sci. **14**, 147 (2018)
13. F.J. Aquino, R. Augusti, J.D.O. Alves, M.E. Diniz, S.A. Morais, B.H. Alves, E.A. Nascimento, A.A. Sabino, Microchem. J. **117**, 127 (2014)
14. U.T.D.C.P. Souto, M.F. Barbosa, H.V. Dantas, A.S. de Pontes, W. da Silva Lyra, P.H.G.D. Diniz, M.C.U. de Araújo, E. da Silva, Food Anal. Methods **8**, 1515 (2014)
15. D. Daniel, F.S. Lopes, V.B. Dos Santos, C.L. Lago, Food Chem. **243**, 305 (2018)
16. J. Yin, S. Hameed, L. Xie, Y. Ying, J. Food Meas. Charact. **15**, 189 (2021)
17. R. Aslam, S.R. Sharma, J. Kaur, A.S. Panayampadan, O.I. Dar, J. Food Meas. Charact. **17**, 3094 (2023)
18. S.S.N. Chakravartula, R. Moscetti, G. Bedini, M. Nardella, R. Massantini, Food Control **135**, 108816 (2022)
19. R.M. Correia, F. Tosato, E. Domingos, R.R.T. Rodrigues, L.F.M. Aquino, P.R. Filgueiras, V. Lacerda, W. Romão, Talanta **176**, 59 (2018)
20. N.V. Harohally, C. Thomas, ACS Food Sci. Technol. **1**, 524 (2021)
21. V.G. Boadu, E. Teye, C.L.Y. Amuah, F.P. Lamptey, L.K. Sam-Amoah, Processes **11**, 1140 (2023)
22. D.A.P. Forchetti, R.J. Poppi, Food Anal. Methods **13**, 44 (2020)
23. L. Munyendo, D. Njoroge, Y. Zhang, B. Hitzmann, Int. J. Food Sci. Technol. **58**, 1284 (2023)
24. M.S.G. Barbosa, M.B.S. Scholz, C.S.G. Kitzberger, M.T. Benassi, Food Chem. **292**, 275 (2019)
25. V.C. Martins, R.L.O. Godoy, A.C.M.S. Gouvêa, M.C.P.A. Santiago, R.G. Borguini, E.C.O. Braga, S. Pacheco, L.S.M. Nascimento, Food Qual. Saf. **2**, 121 (2018)
26. D. Saha, A. Manickavasagan, Res. Food Sci. **4**, 28 (2021)
27. R. Fisher, Ann. Eugen. **7**, 179 (1936)
28. G. Shao, N. Sang, Pattern Recognit. **66**, 353 (2017)
29. M.R. Rana, M. Babor, A.A. Sabuz, J. Agric. Food Res. **5**, 100155 (2021)
30. A. Rinnan, F. Berg, S.B. Engelsen, Trends Anal. Chem. **28**, 1201 (2009)
31. T. Mehmood, S. Sæbø, K.H. Liland, J. Chemom. **34**, e3226 (2020)
32. B.M. Nicolaï, K. Beullens, E. Bobelyn, A. Peirs, W. Saeys, K.I. Theron, J. Lammertyn, Postharvest Biol. Technol. **46**, 99 (2007)
33. S. Touzani, J. Granderson, S. Fernandes, Energy Build. **158**, 1533 (2018)
34. S. Tajmouati, B. Wahbi, A.Bedoui, A. Abarda, M. Dakkon, arXiv preprint (2021). arXiv:2103.14200
35. C. Ingallina, A. Cerreto, L. Mannina, S. Circi, S. Vista, D. Capitani, M. Spano, A.P. Sobolev, F. Marini, Metabolites **9**, 65 (2019)
36. K.N. Basri, M.N. Hussain, J. Bakar, Z. Sharif, M.F. Khir, A.S. Zoolfakar, Spectrochim. Acta Part A Mol. Biomol. Spectrosc. **173**, 335 (2017)

37. I. Esteban-Díez, J.M. González-Sáiz, C. Pizarro, Anal. Chim. Acta **525**, 171 (2004)
38. S. Mayr, J. Schmelzer, C.G. Kirchler, C.K. Pezzei, K.B. Beć, J. Grabska, C.W. Huck, Talanta **221**, 121165 (2021)
39. L.M. Caracostea, R. Sîrbu, F. Busuricu, Eur. J. Nat. Sci. Med. **4**, 69 (2021)
40. I.C. Nwafor, K. Shale, M.C. Achilonu, Sci. World J. **3**, 1 (2017)
41. A. Giraudo, S. Grassi, F. Savorani, G. Gavoci, E. Casiraghi, F. Geobaldo, Food Control **99**, 137 (2019)
42. J.S. Ribeiro, M.M.C. Ferreira, T.J.G. Salva, Talanta **83**, 1352 (2011)
43. J.R. Santos, M. Lopo, A.O.S.S. Rangel, J.A. Lopes, Food Control **60**, 408 (2016)
44. S.T. Alemneh, S.A. Emire, M. Jekle, O. Paquet-Durand, A. Von-Wrochem, B. Hitzmann, Foods **11**, 1171 (2022)
45. P. Khuwijitjaru, K. Boonyapisomparn, C.W. Huck, Int. Food Res. J. **27**, 287 (2020)
46. A. Tugnolo, V. Giovenzana, C. Malegori, P. Oliveri, A. Casson, M. Curatitoli, R. Guidetti, R. Beghi, Food Control **130**, 108312 (2021)
47. T. Fearn, NIR News **13**, 12 (2002)
48. E. Leoni, M. Mancini, D. Duca, G. Toscano, Processes **8**, 1413 (2020)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.