Journal of Pharmaceutical and Biomedical Analysis Open 2 (2023) 100015

Contents lists available at ScienceDirect

Journal of Pharmaceutical and Biomedical Analysis Open

journal homepage: www.journals.elsevier.com/journal-of-pharmaceutical-and-biomedical-analysis-open

Prediction of impurities in cocoa shell powder using NIR spectroscopy

Marciano M. Oliveira[a,c], Marcus V.S. Ferreira[b,c], Mohammed Kamruzzaman[c], Douglas F. Barbin[a,*]

[a] Department of Food Engineering and Technology, School of Food Engineering, University of Campinas, Campinas, SP, Brazil
[b] Federal Rural University of Rio de Janeiro (UFRRJ), Department of Food Technology, Seropédica, RJ, Brazil
[c] Department of Agricultural and Biological Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

ARTICLE INFO

Keywords:
Adulteration
Chemometrics
Bioactive compounds
Portable spectrometer
Authentication
Pharmaceutics

ABSTRACT

Cocoa shell is a by-product from cocoa industry which contains bioactive compounds of high and attractive value for food, pharmaceutical and cosmetics industry. However, cocoa shell can be contaminated by undesirable materials that, even in small amounts, would not change the color, aroma, and taste characteristics of the final product. Identification and prediction of this impurity are performed using expensive methods that require chemicals and produce residues. Thus, this work aims to investigate the performances of benchtop (867–2535 nm) and portable (900–1700 nm) near-infrared (NIR) spectrometer for fast prediction of cocoa shell powder impurities. Mixtures (n = 432) of cocoa shell powders with leaves, pods, stem fragments and nibs at several proportions (0–20 % w/w), were analyzed. Multivariate calibration models were developed using partial least-squares regression (PLSR) with raw spectra and various preprocessing approaches applied to the spectra. The most informative spectral variables were selected by variable importance in projection (VIP) method. Results obtained for the benchtop ($R^2_P >$ 0.99 and RMSEP < 0.71) and low-cost portable ($R^2_P >$ 0.92 and RMSEP < 1.70) devices are promising, and portable spectrometer could be used to certify cocoa shell purity.

1. Introduction

Cocoa shell is the main and most valuable by-product of the cocoa industry [1] where approximately 700 thousand tons are generated worldwide when considering world cocoa production [2]. Although cocoa shell has traditionally limited applications (mainly fuel for boilers and as animal feed or organic soil fertilizer), many studies have been conducted on its composition as well as possible industrial applications with high added value for food, pharmaceutical and cosmetics industry [3,4]. Cocoa shell is a rich source of dietary fiber and protein as well as valuable bioactive compounds (theobromine, caffeine, flavonoids, etc.). Due to its composition, it can be used as an ingredient in food processing, or in pharmaceuticals, cosmetics, or agriculture products, with new applications [4–6]. The recovery of cocoa shells has high economic value, as it is a cheap raw material for extracting many components and production of biofuel. In addition, some studies have indicated the beneficial health effects of cocoa shell compounds [7].

Therefore, it is very important to guarantee the purity and safety of this by-product [8], as some impurities found in the cocoa shell can negatively affect its application in the industry. Since it is a remaining part of the cocoa industry, control and inspection of the cocoa shells purity is not a priority. Thus, after peeling the cocoa beans, the cocoa shells may be accompanied by foreign materials, such as cocoa pods and leaves, stem fragments, and often have some nibs misplaced during peeling. On the other hand, except for the nibs, these types of impurities can be intentionally placed as bulking agents to increase companies' revenue, which can also be a potential threat to the health of consumers, since these foreign materials can be contaminated with pathogen microorganisms, and toxic substances [9–11].

However, certifying the purity of agricultural and food products is not always easy, since the techniques for identifying and quantifying the composition of these products are often arduous, expensive, and require chemical reagents, such as HPLC-ESI-QqQ-MS/MS, as reported in recent studies [11]. Therefore, there is a need for alternatives toward more sustainable practices. In this context, near-infrared spectroscopy (NIRS) is a non-destructive technique already well-established for analytical purposes, widely used in qualitative and quantitative analyses of agricultural and food products [12–14]. In the cocoa industry, many studies have shown the potential of NIRS to predict moisture, polysaccharides, fat, and protein content [15,16] as well as the phenol-related compounds, such as theobromine, catechin and organic acids in cocoa beans and products [17,18]. In addition, NIR spectroscopy has also been used to identify adulteration of cocoa powder [19], discriminate cocoa beans according to geographic origin [20], and determine the degree of fermentation [21].

* Corresponding author.
E-mail address: dfbarbin@unicamp.br (D.F. Barbin).

https://doi.org/10.1016/j.jpbao.2023.100015
Received 27 April 2023; Received in revised form 20 June 2023; Accepted 21 June 2023

The NIRS systems are found in different configurations and sensitivity levels, and recent advances have allowed these optical components to be miniaturized without excessive loss in performance [22,23]. It can be found portable devices based on Raman, mid-infrared (MIR) and near-infrared (NIR), and recently, on hyperspectral imaging (HSI) and nuclear magnetic resonance (NMR) technology [24]. While portable devices have a lower precision compared to benchtop equipment, some advantages, such as analytical capacity (online or in situ), small size, low cost, robustness, analysis simplicity, and portability, compensate for this deficiency [22]. The demand for portable devices to evaluate the quality, composition, and authentication of agricultural products has increased considerably in recent years [25], and many studies have been reported their use on cocoa beans [26,27]. There are works using NIR spectroscopy combined with multivariate analysis to authenticate cocoa shell [28], as well as to determine its bioactive compounds and antioxidant activity [29]. However, studies using portable NIR devices to detect impurities in cocoa shell have not been reported. In this sense, developing an accessible multivariate approach using portable NIR devices to identify potential impurities in cocoa shell will bring great scientific-technological contribution and, consequently, will encourage new research towards cocoa shell processing options. Therefore, the main objective of this study is to propose a methodology for predicting and screening some impurities in cocoa shells, based on NIR spectroscopy associated with multivariate analyses, to be implemented for the best use of this by-product in the industry. The specific objectives include (a) use Principal Component Analysis (PCA) to have a clear idea of the relation between samples and variables; (b) use PLSR to predict the level of impurity; (c) select some important wavelengths to develop a more effective PLSR models for impurity prediction; (d) develop a quantitative function using PLSR based on the best subset of variables for online prediction; (e) compare the predictive performance between a benchtop and portable device.

2. Materials and methods

2.1. Sample collection and preparation

Cocoa shells and the impurities (cocoa leaves, cocoa pods, cocoa stem fragments and cocoa nibs) were obtained from four suppliers from different regions in Brazil (Bahia, Espirito Santo and Para). To ensure that the distinction between the cocoa shells and impurities was due exclusively to the difference in their compositions, all samples were standardized in terms of particle size and moisture content. Initially, the samples were placed in a dryer at 74 °C for 2 h to obtain uniform sample moisture levels. After drying, the cocoa shells and impurities were individually ground using an IKA A11 Basic Analytical Mill (Königswinter, Germany). Then, a 200 mesh sieve (74 μm aperture) was used to standardize the particle size of the samples.

The binary mixtures were prepared in concentrations from 0 % to 20 % (w/w) by blending cocoa shell powders with the four different impurities at different mass proportions (0 %, 1 %, 2 %, 3 %, 4 %, 5 %, 7 %, 10 %, 12 %, 15 %, 17 % and 20 %). For each concentration, three replicate samples of 10 g (w/w) for each of the four suppliers were prepared (4 suppliers x 3 replicate x 12 concentrations = 144 samples for each impurity). Samples of all the 432 possible cocoa shell powder-impurity combinations were classified in two different levels of concentration: low concentration (0–5 %) and high concentration (7–20 %). The upper limit of 20 % was set by considering that above this concentration the presence of any impurity would become evident based on visual inspection due to changes in color. The samples were manually mixed and then transferred to a snap-cap vial. Further mixing was accomplished by placing the filled vials onto a high-speed shaker (VWR® Fixed Speed Vortex Mixers, Canada) to minimize possible dispersion effects. All mixtures were placed in hermetic plastic containers and stored at 20 ± 2 °C in a dry dark atmosphere before to analyze.

2.2. NIR data acquisition

Two different NIR spectrometers were used for data acquisition: a benchtop NIR device (Bruker Tango FT-NIR (Bruker Optik, Germany)) and a portable NIR device (NIRscan™ Nano Digital Light Processing (DLP^{R}) - Texas Instrument, USA). For benchtop device, one spectra for each sample was collected using rotating cups and each spectrum was the average of 64 scans recorded from 867 to 2535 nm with a resolution of 2 nm. For portable device, samples were scanned in the surface. Three spectra were collected for each sample to account for shorter scan times and smaller sampling surfaces, and each spectrum was also the average of 64 scans. The spectral acquisition with the portable device was made in the wavelength range of 900–1700 nm with intervals of 4 nm, using a 10 W halogen bulb as a light source and an InGaAs detector. The spectra were obtained in reflectance mode, corrected using white/dark references and transformed into absorbance units by logarithmic transformation for direct comparison between the equipment. The external white and black references are automatic in the benchtop equipment and manual for the portable device.

2.3. Data analysis

Four data sets for each equipment were used in this study. The data sets from the benchtop device consisted of 144 spectra and 949 variables (wavelengths, nm) for each impurity. On the other hand, the data sets from the portable device consisted of 144 spectra and 228 variables (wavelengths, nm) for each impurity, where an average of 3 spectra for each sample was used. The data set from each impurity was manually divided into training (84 samples) and test (60 samples) sets for external validation. The training and test sets comprised 0 %, 1 %, 3 %, 5 %, 10 %, 15 % and 20 % (w/w) and 2 %, 4 %, 7 %, 12 % and 17 % (w/w), respectively. All spectral data were preprocessed and analyzed using PLS_Toolbox (Eigenvector Research, Inc. WA, USA) under MATLAB R2022a (The Mathworks, MA, USA).

The raw spectra were preprocessed by the following methods: first and second derivatives (FD and SD, respectively), multiplicative scattering correction (MSC) and combined techniques, such as multiplicative scattering correction + first derivative (MSC + FD). Savitzky-Golay smoothing (window: 15 points; polynomial filtering: 2nd order) is recommended before applying the first (FD) and second (SD) derivatives. The application of MSC aims to remove non-uniform scattering and particle size effects from the spectrum. FD, on the other hand, eliminates baseline variations, while SD separates overlapping peaks and highlights spectral characteristics [30].

PCA was used to reduce the dimensionality of the data sets, also to identify the interrelationships between the samples and the possible clusters, to select the appropriate experimental data for the construction of the model, and finally to identify and eliminate outliers through Hotelling's T^2 statistics and F-residual values [31]. The number of principal components (PCs) was chosen based on the cumulative variance, in which the first PCs were chosen explaining together more than 70 % of the total variability.

PLS regression [32] was used to predict the amount of impurity in the cocoa shell powders. The performances of the PLSR models were estimated using random cross-validation (leave-one-out) and external validation set. For the PLSR models construction, the use of all wavelengths was considered as well as the most important ones. In this study, the variable importance in projection (VIP) method, also known as VIR scores, was used as a strategy to select the most important wavelengths for predicting impurity content in cocoa shell powders [33,34].

PLS regression models' accuracy was evaluated by the required number of latent variables (LVs), the coefficient of determination of calibration (R^2_C), the root mean squared error of calibration (RMSEC), the ratio of prediction deviation of calibration (RPD_C), the range error ratio of calibration (RER_C), the coefficient of determination of prediction (R^2_P), the root mean square error of prediction (RMSEP), the ratio of

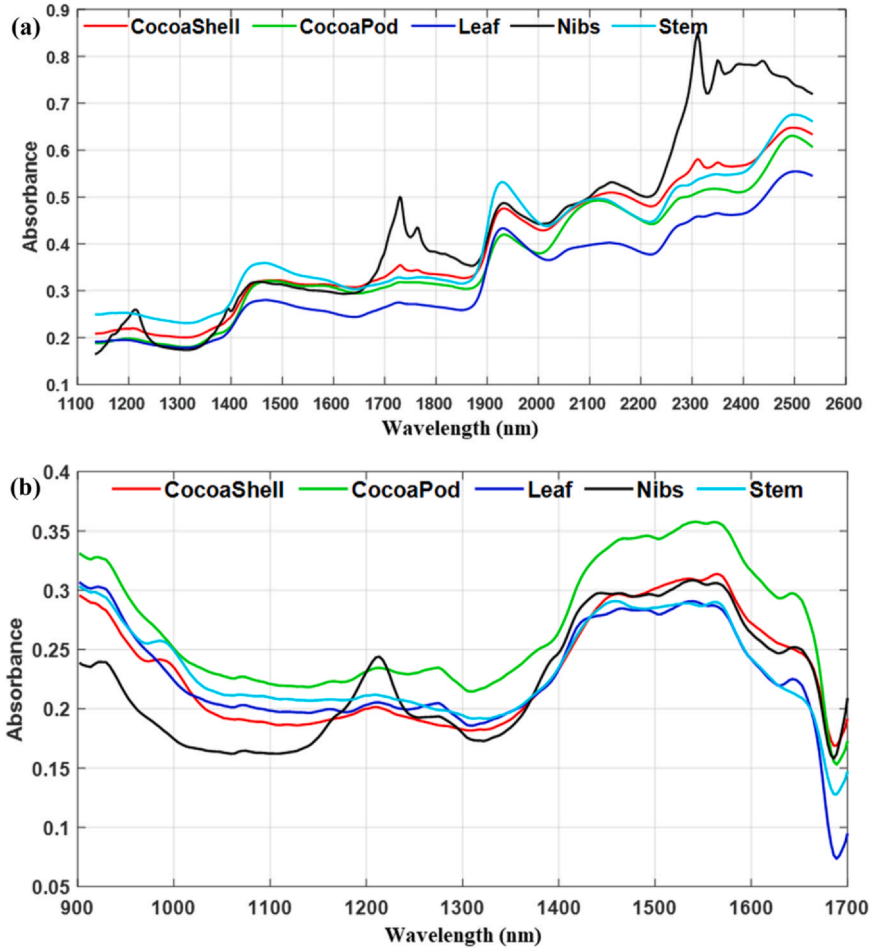

Fig. 1. Mean raw spectra of cocoa shell powder, cocoa leaf, cocoa pod, cocoa stem fragments, cocoa nibs from the (a) benchtop (1100–2535 nm) and (b) portable (900 – 1700 nm) spectrometer.

prediction deviation of prediction (RPD_P), and the range error ratio of prediction (RER_P).

3. Results and discussion

3.1. Spectral features

The spectral information in the range of 867–1100 nm of the benchtop device were not considered in the study, as this region presented some noise. The mean raw spectra of cocoa shell, cocoa pods, cocoa leaves, cocoa stem fragments and cocoa nibs from both equipment are shown in Fig. 1. The difference in the absorption bands are in the second overtone of C–H stretching (1100–1200 nm), the first overtone of the hydroxyl and amino groups (1425 nm), the first overtone of O–H and N–H stretching, which is associated with a $CONH_2$ structure (1470 nm) and first overtone of C–H (1644 nm) and the combination of C–C and C–H stretching (2146 nm) [35,36]. Furthermore, Ribeiro, Ferreira and Salva [37] showed that the band at 1730 nm could be assigned to the first overtone of C–H, while 2310 and 2350 nm are mostly related to stretching and rocking vibrations of CH_2. These absorption bands are mainly characterized as one of the various functional groups of polysaccharides, protein, fat and water [15,36]. Absorption bands around 1200, 1730 and 2350 nm are mainly associated with fat, and display higher values for the spectra of cocoa nibs due to their higher fat content in relation to cocoa shell and other impurities [19]. Absorption band in 1200 nm can also correspond to polysaccharides [38], while the three regions 1349–1386, 1661–1718 and 2161–2258 nm are related to total phenols [39]. Epicatechin absorption bands were reported at 1388, 1492, 1658, 1916, 2260 and 2324 nm, and 1764, 2092 and 2228 nm are associated with theobromine [17]. In the benchtop's spectra (Fig. 1), it can be seen that the absorption peaks in the bands 1716–1768 nm and 2280–2360 nm were more pronounced in the spectra of cocoa shell, which could be related to the migration of phenols and theobromine from the cotyledon to the cocoa shell during the fermentation process [3]. Important spectral information to differentiate cocoa shells from impurities can be extracted by comparing their NIR spectra. To a certain extent, structural and chemical information can be obtained from the NIR spectra with limitations through overlapping absorption peaks. In general, in the NIR spectra of cocoa shells there are different absorption bands intensity compared to the spectra of all impurities.

3.2. PCA analysis

The PCA was performed with the raw spectra of the four different cocoa shell suppliers to identify possible sample groupings and outliers. In the data obtained with benchtop device (Fig. 2a), the first two PCs explained 99.75 % of the total variance for the different cocoa shell suppliers, 92.86 % for PC1 and 6.69 % for PC2. In the data obtained for the portable device (Fig. 2c), PC1 explained 87.57 % and PC2 11.01 %, totaling 98.58 % of the data variance. The high explanation of the data demonstrates the quality of the analysis in transforming the original set of data and the absorbances associated with the vibrational modes of NIR spectra in principal components for both equipment.

Fig. 2a and c show the dispersion of the four different cocoa shell suppliers using the benchtop and portable device spectra, respectively.

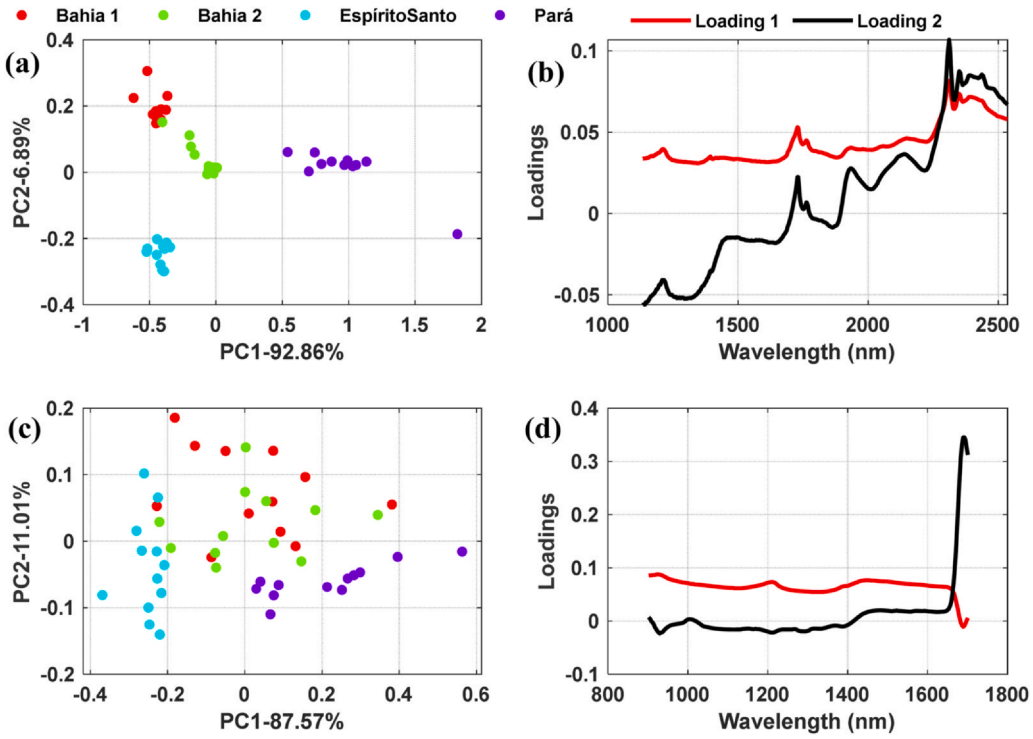

Fig. 2. PCA scores and loadings plots of mean raw spectra for the benchtop (a,b) and portable (c,d) NIR spectrometer.

PC1 was the component that best explained the data distribution, in other words, the element responsible for separating the samples considering the cocoa shell supplier. It is possible to verify that the PCA scores formed groups related to the cocoa shell suppliers for both benchtop and portable device. PC1 from both devices were the main contributors to the separation of cocoa shell samples based on their suppliers. The loadings (Fig. 2b and d) allow defining what spectral regions are involved in each relevant PC. PC1, was mainly characterized by fatty acid bands as 1730–1763 nm (1st C–H str) and 2312 nm (1st C–H str + 1st C–H def CH_2) and 2353 nm (1st C–H str + 1st C–H def). In addition, PC1 also captured some regions related to proteins such as 1438 nm, 1530 nm (1st N–H str of RNH_2), 1582 nm (1st N–H str of CONH), 1897 nm (2nd C═O of COOH) and 1926–1934 nm (2nd C═O of CONH) [37]. PC2, instead, exhibited three maxima at 2000 nm (2nd O–H def + 1st C–O def of polysaccharides), 2100 nm (2nd O–H def + 2nd C–O str of polysaccharides) and 2272 nm (1st O–H str + 1st C–C str, associated with polysaccharides) [35,36]. This indicates that both PCs mostly represents the polysaccharides content of the samples. Both loadings show similar features, which include an absorption peak at 1200 nm related to polysaccharides and fat content. The loadings also consist of a peak comprising two very weak bands related to fat at 1392 and 1414 nm, and a remarkable peak that is related to the water content of the samples (peak at 1140 nm). For the portable's loadings, a second water peak could also be observed at 970 nm.

3.3. Calibration models

3.3.1. Full spectral range

PLSR models were calculated using the full spectral range from benchtop (Table 1) and portable (Table 2) device. The best PLSR models based on the full spectral range for the portable device were constructed using the FD as a preprocessing, except the nibs' model that fits better with SD preprocessing. On the other hand, the best PLSR models for benchtop device were those built using FD as preprocessing for all impurities. As shown in Tables 1 and 2, the preprocessing applied to the raw data were effective in minimizing the influence of undesirable information, improving the prediction capability of the models. The main objective of the spectral preprocessing is to remove the scattering effects associated with the shape/structure of the samples, thereby improving subsequent multivariate regression, classification models and exploratory analysis. Preprocessing has a significant effect on spectral modeling, as a good selection of preprocessing can increase the accuracy of models, while incorrect selection can lead to inconsistencies in prediction [40]. This is mostly due to the preprocessing procedure, which can mitigate the unwanted effects on the original variables therefore, reducing the experimental error in the final models.

PLSR models based on the full spectral range achieved better results with benchtop device ($R^2_P > 0.96$, RMSEP < 1.25, RPD > 3.77 and RER > 12.02) compared with portable device ($R^2_P > 0.92$, RMSEP < 1.86, RPD > 3.41 and RER > 8.04). The PLSR models had good prediction ability, as highlighted by the high values of R^2 as well as the small difference between RMSEC and RMSEP (Tables 1 and 2). The proximity of the values of R^2_C and R^2_P along with RMSEC and RMSEP ensure that PLSR models are representative and can be applied accurately to the unknown data. Although PLSR models generated by the portable device required a greater number of LVs than the benchtop device for a better fit (except for cocoa nibs), its models presented satisfactory R^2_P, RMSEP, RPD and RER values. According to previous studies, R^2 values greater than 0.9, RPD greater than 3, and an RER greater than 10, would result in successful calibration models, indicating a greater predictability of the models to accurately predict impurities in new samples. Thus, the achieved RPD and RER index for the models for each impurity from both equipment were categorized as a good performance based on the literature [41]. Moreover, the prediction performances of the PLSR models from the two equipment developed using the full spectral range were appropriate and consistent with those presented in previous studies when predicting adulterants in cocoa products using NIR spectroscopy [19,42].

3.4. 2 Selection of informative spectral bands using VIP scores

NIR spectroscopy data can be redundant, noisy, and irrelevant, or interfering. Thus, using the full spectral range could imply the risk of overfitting, noise, and nonlinearities, which in turn can lead to

Table 1
Performance of the PLSR models at full NIR spectral range using the portable device.

Impurity	Preprocessing	#Bands	#LVs	Calibration				Prediction			
				$RMSE_C$	RPD_C	R^2_C	RER_C	$RMSE_P$	RPD_P	R^2_P	RER_P
Cocoa leaf	Raw	228	10	1.14	6.09	0.97	17.56	1.70	3.56	0.93	8.80
	FD	**228**	**7**	**1.19**	**5.54**	**0.97**	**16.80**	**1.72**	**3.66**	**0.94**	**8.71**
	SD	228	8	1.00	6.87	0.98	20.05	1.77	3.65	0.94	8.49
	MSC	228	9	1.19	5.84	0.97	16.83	1.77	3.47	0.92	8.46
	Raw	228	10	1.21	5.79	0.97	16.47	1.60	3.56	0.92	9.37
Cocoa pod	**FD**	**228**	**8**	**1.03**	**6.79**	**0.98**	**19.33**	**1.79**	**3.52**	**0.93**	**8.40**
	SD	228	9	0.80	8.59	0.99	24.94	1.86	3.41	0.96	8.04
	MSC	228	9	1.10	6.40	0.98	18.10	1.64	3.75	0.94	9.14
	Raw	228	6	1.16	6.12	0.97	17.31	1.45	4.16	0.96	10.31
Cocoa stem fragments	**FD**	**228**	**5**	**1.04**	**6.82**	**0.98**	**19.17**	**1.29**	**4.73**	**0.97**	**11.62**
	SD	228	5	0.91	7.61	0.98	21.94	1.34	4.49	0.97	11.23
	MSC	228	5	1.24	5.70	0.97	16.11	1.37	4.49	0.96	10.98
	Raw	228	5	0.70	10.07	0.99	28.63	0.92	5.94	0.98	16.33
Cocoa nibs	FD	228	4	0.60	11.77	0.99	33.46	0.85	6.26	0.98	17.61
	SD	**228**	**4**	**0.49**	**14.06**	**0.99**	**41.11**	**0.82**	**6.36**	**0.98**	**18.40**
	MSC	228	6	0.60	11.71	0.99	33.11	0.95	6.02	0.97	15.86

VIP, Variable importance in projection; FD, first derivative; SD, second derivative; MSC, multiplicative scatter correction; Bands, wavelengths used for model development; LVs, latent variables; RMSEC, root mean square error of calibration; RPDC, ratio of prediction deviation of calibration; R2C, coefficient of determination of calibration; RERC, range error ratio of calibration; RMSEP, root mean square error of prediction; RPDP, ratio of prediction deviation of prediction; R2P, coefficient of determination of prediction; RERP, range error ratio of prediction. The overall best models are highlighted in bold.

inaccuracy models. On the other hand, the chemical information obtained from selected wavelengths might be more efficient than the full spectra [43]. Therefore, it can be interesting selecting the more prominent variables and excluding the non-informative ones.

The spectral responses of cocoa shells and impurities are closely related to their chemical compositions since their main components are polysaccharides, including fibers such as cellulose and lignin, and fats in small proportions (except in cocoa nibs) [44]. These considerations on the chemical composition lay the foundations for correctly interpreting the regression models implemented with the NIRS data. However, Brown and Green observed that the spectral interpretation of PLSR models should not only rely on the regression vectors, since they are dependent on the samples in the calibration/training set, the implicit covariance of the components, and the signal to noise ratio of the data [45]. Thus, an important tool for the spectral interpretation of PLSR models is the variable selection approach that is a critical step in multivariate analysis to improve the model's predictive performance and enhance the model's interpretability with parsimonious representation [33].

In this study, VIP scores was used to select the most informative wavelengths from the best PLSR models based on the full spectral range. The selected important wavelengths in each impurity model for both equipment are listed in Table 3. Even though cocoa shell and impurities have different compositions, there is not a considerable variation in the selected variables, as shown in the visual comparison plot in Figs. 3 and 4. For all models, the dominant spectral regions for the benchtop and portable device were 1420–2460 nm and 1160–1680 nm, respectively. All models have selected at least 6 wavelengths in these ranges, excluding cocoa nibs' models. Although these different NIR regions contain bands related to combination modes, only in the spectral range of the benchtop device it is possible to highlight the characteristic bands of organic acids due to O–H stretching combined with the C–O stretching around 1890, 2285 and 2456 nm. Figs. 3 and 4 show that two chemically meaningful spectral windows were selected in all models using the benchtop (1420–1890 nm

Table 2
Performance of the PLSR models at full NIR spectral range using the benchtop device.

Impurity	Preprocessing	##Bands	##LVs	Calibration				Prediction			
				$RMSE_C$	RPD_C	R^2_C	RER_C	$RMSE_P$	RPD_P	R^2_P	RER_P
	Raw	610	5	0.63	10.86	0.99	31.68	0.37	15.13	1.00	40.66
Cocoa leaf	**FD**	**610**	**5**	**0.30**	**22.95**	**1.00**	**66.57**	**0.69**	**7.36**	**0.99**	**21.70**
	SD	610	5	0.35	20.54	1.00	57.78	0.65	7.88	0.99	23.09
	MSC	610	5	0.31	22.80	1.00	64.53	1.04	4.96	0.99	14.49
	Raw	610	5	0.32	22.18	1.00	62.21	1.14	4.90	0.97	13.18
Cocoa pod	**FD**	**610**	**6**	**0.22**	**31.48**	**1.00**	**91.92**	**0.71**	**7.99**	**0.99**	**21.15**
	SD	610	4	0.35	20.07	1.00	56.76	1.21	4.58	0.96	12.43
	MSC	610	4	0.24	29.05	1.00	82.21	0.95	5.58	0.97	15.76
	Raw	610	4	0.35	19.81	1.00	57.03	1.25	3.77	0.99	12.02
Cocoa stem fragments	**FD**	**610**	**4**	**0.21**	**33.02**	**1.00**	**96.26**	**0.47**	**12.08**	**1.00**	**31.98**
	SD	610	4	0.29	23.79	1.00	68.55	0.93	6.48	1.00	16.20
	MSC	610	4	0.19	36.70	1.00	106.07	1.10	4.54	1.00	13.58
	Raw	610	5	0.23	30.30	1.00	86.89	0.36	15.79	1.00	41.40
Cocoa nibs	**FD**	**610**	**5**	**0.15**	**43.84**	**1.00**	**134.74**	**0.24**	**23.61**	**1.00**	**63.06**
	SD	610	4	0.18	37.89	1.00	108.58	0.44	13.09	1.00	34.16
	MSC	610	5	0.15	45.70	1.00	135.11	0.25	21.84	1.00	60.15

FD, first derivative; SD, second derivative; MSC, multiplicative scatter correction; Bands, wavelengths used for model development; LVs, latent variables; RMSEC, root mean square error of calibration; RPDC, ratio of prediction deviation of calibration; R2C, coefficient of determination of calibration; RERC, range error ratio of calibration; RMSEP, root mean square error of prediction; RPDP, ratio of prediction deviation of prediction; R2P, coefficient of determination of prediction; RERP, range error ratio of prediction. The overall best models are highlighted in bold.

Table 3
Selected wavelengths using VIP scores algorithm for both devices.

Equipment	Impurity	#Bands	Selected wavelengths (nm)
DLP^{R} NIRscan™ Nano	Cocoa leaf	8	1166, 1184, 1233, 1375, 1402, 1433, 1449, 1671
	Cocoa pod	9	953, 1166, 1184, 1233, 1375, 1402, 1433, 1640, 1671
	Cocoa stem	7	1143, 1184, 1233, 1371, 1402, 1490, 1662
	Cocoa nibs	4	1210, 1243, 1436, 1662
Tango FT-NIR	Cocoa leaf	6	1424, 1715, 1891, 2285, 2336, 2456
	Cocoa pod	6	1424, 1701, 2042, 2137, 2285, 2456
	Cocoa stem	6	1715, 1739, 1962, 2137, 2285, 2456
	Cocoa nibs	5	1715, 1739, 1776, 2285, 2456

Bands, wavelengths used for PLSR models development.

and 2280–2460 nm) and portable (1160–1240 nm and 1370–1490 nm) device. In these spectral regions, the major absorption usually occurs at 1210, 1424, 1715, 2285 and 2456 nm corresponding to the combination of O–H, C–H and CH_2 bonds related to water, polysaccharides and fat [19]. These five wavelengths (or very adjacent) were selected in all models, excluding the cocoa leaf's models, and were also prominent in the original raw spectra.

In general, all PLSR models obtained for both equipment from the variable selection method showed similar performances to those obtained with the full spectral range (Table 4). Fig. 1S and 2S (supplementary material) show the correlation of predicted and actual values for calibration and prediction to the best PLSR models of the benchtop and portable device, respectively. Regarding portable NIR data (Table 4) it is interesting to underline that, in terms of prediction, the accuracy of the model for cocoa nibs (R^2_P = 0.99 and RMSEP = 0.74) and cocoa stem fragments (R^2_P = 0.97 and RMSEP = 1.35) was higher compared to the model for cocoa pods (R^2_P = 0.89 and RMSEP = 2.04) and cocoa leaves (R^2_P = 0.94 and RMSEP = 1.53). As per PLS regression using benchtop device (Table 4), the R^2_P and RMSEP were similar and therefore, acceptable for calibration in all PLSR models. On the other hand, for prediction, the accuracy of the model for cocoa nibs (R^2_P = 1.00 and RMSEP = 0.21) and the cocoa leaves (R^2_P = 1.00 and RMSEP = 0.43) were higher compared to the model for cocoa pods (R^2_P = 1.00 and RMSEP = 0.43) and cocoa stem fragments (R^2_P = 1.00 and RMSEP = 0.39). Therefore, the results showed that the VIP scores wavelength selection method was efficient and improved the calibration performance of the PLSR models. Generally, when variable selection was used it led to a more robust model [33]. Most of the selected wavelengths have been previously described in literature in the prediction of several compounds, such as fat, polysaccharides, moisture, polyphenols in cocoa beans and derived products [15,18,46].

Overall, after wavelengths selection, the benchtop device data were reduced from 949 to 12 variables (1424, 1701, 1715, 1739, 1776, 1891, 1962, 2042, 2137, 2285, 2336 and 2456 nm), combining the four sets of variables selected from the impurity models, with a total data size reduction close to 99 %. On the other hand, using the portable device, the data were reduced from 228 to 17 variables (953, 1143, 1166, 1184, 1210, 1233, 1243, 1371, 1375, 1402, 1433, 1436, 1449, 1490, 1640, 1662 and 1671 nm) also by combination the four sets of variables selected from the impurity models (93 % reduction). This reduction of the data sets in smaller subsets is interesting from an operational point of view. The subsets of 12 and 17 wavelengths for the benchtop and portable device, respectively, can provide the basis to design a low-cost, fast, and accurate spectral system for predicting real-time impurities in cocoa shell. These wavelengths can be used as bandpass filters instead of a spectrometer. A spectral system with a few wavelengths will reduce the cost of the device and increase the speed in the production line, which might fulfill the requirement of the food industry.

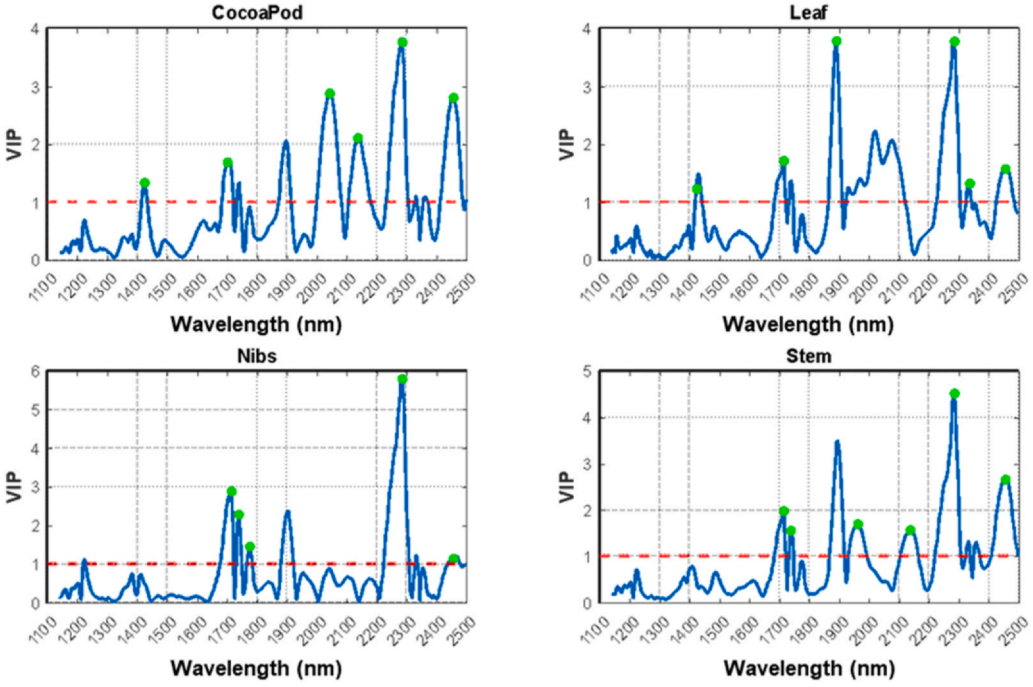

Fig. 3. Variable importance in projection (VIP) scores of the PLSR models constructed from benchtop NIR spectrometer (bands were selected from PLSR VIP created using FD data).

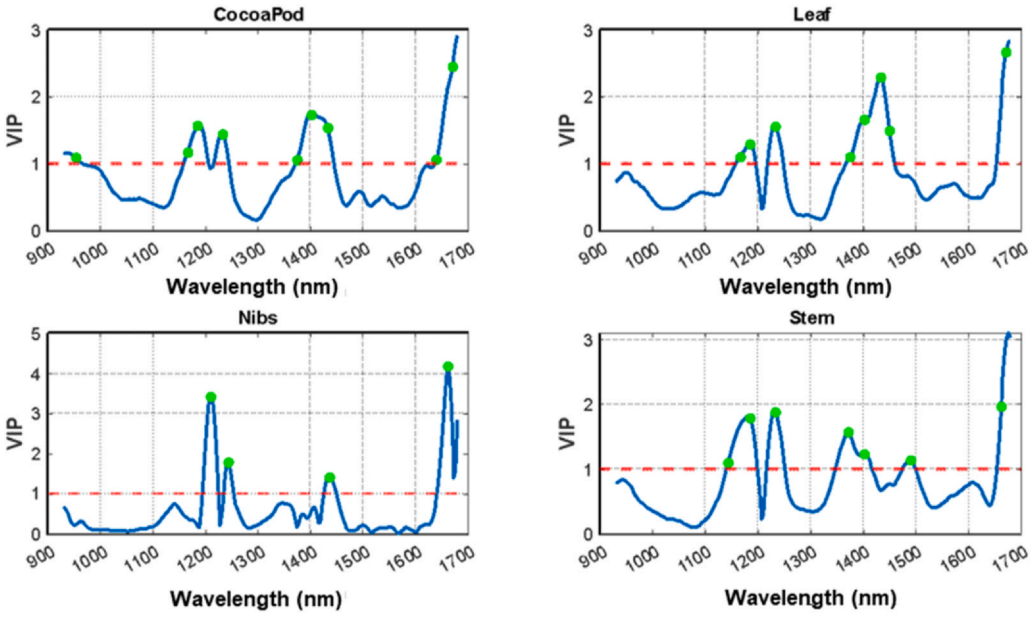

Fig. 4. Variable importance in projection (VIP) scores of the PLSR models constructed from portable NIR spectrometer (bands were selected from PLSR VIP created using FD data, excluding Nibs). For Nibs, SD data were used from good model using full bands PLSR model.

Table 4
Effect VIP variable selection algorithm on PLSR models.

Equipment	Impurity	L#LVs	Calibration				Prediction			
			$RMSE_C$	RPD_C	R^2_C	RER_C	$RMSE_P$	RPD_P	R^2_P	RER_P
DLP[R] NIRscan™ Nano	Cocoa leaf	7	1.50	4.41	0.95	13.37	1.53	3.94	0.94	9.78
	Cocoa pod	8	1.95	3.60	0.92	10.25	2.04	2.85	0.89	7.34
	Cocoa stem fragments	5	1.28	5.56	0.97	15.62	1.35	4.65	0.97	11.13
	Cocoa nibs	4	0.49	14.01	0.99	40.97	0.74	7.16	0.99	20.37
Tango FT-NIR	Cocoa leaf	5	0.64	10.74	0.99	31.14	0.43	12.44	1.00	35.13
	Cocoa pod	6	0.30	22.98	1.00	67.09	0.70	8.34	1.00	21.49
	Cocoa stem fragments	3	0.28	24.99	1.00	71.45	0.39	14.61	1.00	38.90
	Cocoa nibs	5	0.20	32.75	1.00	100.66	0.21	26.07	1.00	69.94

LVs, latent variables; RMSEC, root mean square error of calibration; RPDC, ratio of prediction deviation of calibration; R2C, coefficient of determination of calibration; RERC, range error ratio of calibration; RMSEP, root mean square error of prediction; RPDP, ratio of prediction deviation of prediction; R2P, coefficient of determination of prediction; RERP, range error ratio of prediction.

The multivariate calibration proposed in this study using a benchtop and a portable spectrometer demonstrated that the portable is a robust device for industrial applications, despite its reduced wavelength range. The results showed that the portable device is as competitive as the costly benchtop from a prediction performance perspective. One of the advantages lies upon that the former presents a smaller size, thus offering solution on the go as opposed to the latter [47]. The combination of simple devices with PLSR modeling may offer a very interesting and reliable tool for predicting impurities in cocoa shell directly in the processing industry. In addition, if this approach is applied throughout the cocoa supply chain, could improve, in a sustainable way, the quality of the products that reach consumers' tables in everyday life such as chocolate and chocolate powder.

However, there are also some limitations in using portable NIR devices, such as when acquiring spectra data, the samples need to be rotated manually according to the operator's interference. Moreover, during measurements, typical noises occur regularly on a specific wavelength range, particularly in the beginning and the end of spectra. This is due to instability and overheating electronic components inside the portable NIRS instrument [48]. Therefore, studying the predictive capacity of different NIR devices is extremely important for choosing the right device to match the market needs, mainly from the operational aspect of the industries.

4. Conclusions

NIR spectroscopy in tandem with the PLSR statistical method provided models with appropriate predictive and generalization capacity to predict impurities content in cocoa shell powders. The applied VIP scores variable selection method and spectral preprocessing further improved the performance of the developed models. The best PLSR calibration models built using the benchtop (RPD > 8.34) and portable (RPD > 2.85) device showed good prediction performances. The results indicate that NIR spectroscopy is an adequate tool for identifying and quantifying impurities in cocoa shells, which might help the quality control in the cocoa industry, as well as granting its authenticity for further industrial applications.

CRediT authorship contribution statement

Marciano Marques de Oliveira: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing – original draft. **Marcus V. S. Ferreira**: Methodology, Software, Data curation, Writing – review & editing. **Mohammed Kamruzzaman**: Methodology, Software, Data curation, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Douglas Fernandes Barbin**: Methodology, Software, Data curation, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This study was financed in part by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001, CAPES PrInt 88887.694897/2022-00, and São Paulo Research Foundation (FAPESP) (project number 2015/24351–2). Prof. Douglas Fernandes Barbin is CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brazil) research fellow (308260/2021-0).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jpbao.2023.100015.

References

[1] R. Campos-Vega, B.D. Oomah, H.A. Vergara-Castañeda, Food Wastes and By-products: Nutraceutical and Health Potential, first ed..,, John Wiley & Sons,, 2020.
[2] ICCO (The International Cocoa Organisation), Cocoa Stat., 43. 〈https://www.icco.org/〉, 2022 (accessed 13 December 2022).
[3] D.C. Okiyama, S.L. Navarro, C.E. Rodrigues, Cocoa shell and its compounds: applications in the food industry, Trends Food Sci. Technol. 63 (2017) 103–112, https://doi.org/10.1016/j.tifs.2017.03.007
[4] O. Rojo-Poveda, L. Barbosa-Pereira, G. Zeppa, C. Stévigny, Cocoa bean shell—a by-product with nutritional properties and biofunctional potential, Nutrients 12 (4) (2020) 1123, https://doi.org/10.3390/nu12041123
[5] Z.S. Vásquez, D.P. de Carvalho Neto, G.V. Pereira, L.P. Vandenberghe, P.Z. de Oliveira, P.B. Tiburcio, H.L.G. Rogez, A.G. Neto, C.R. Soccol, Biotechnological approaches for cocoa waste management: a review, Waste Manag. 90 (2019) 72–83, https://doi.org/10.1016/j.wasman.2019.04.030
[6] O.A. Lessa, I.M. de Carvalho Tavares, L.O. Souza, L.G.P. Tienne, M.C. Dias, G.H.D. Tonoli, E.V.B. Vilas Boas, S.G.F. Leite, M.L.E. Gutarra, M. Irfan, M. Bilal, M. Franco, New biodegradable film produced from cocoa shell nanofibrils containing bioactive compounds, J. Coat. Technol. Res. 18 (6) (2021) 1613–1624, https://doi.org/10.1007/s11998-021-00519-4
[7] T.F. Soares, M.B.P. Oliveira, Cocoa by-products: characterization of bioactive compounds and beneficial health effects, Molecules 27 (5) (2022) 1625, https://doi.org/10.3390/molecules27051625
[8] M. Rebollo-Hernanz, S. Cañas, Y. Aguilera, V. Benitez, A. Gila-Díaz, P. Rodriguez-Rodriguez, I.M. Cobeta, A.L.L. de Pablo, M.C. Gonzalez, S.M. Arribas, M.A. Martin-Cabrejas, Validation of cocoa shell as a novel antioxidant dietary fiber food ingredient: nutritional value, functional properties, and safety, 773-773, Curr. Dev. Nutr. 4 (Supplement_2) (2020), https://doi.org/10.1093/cdn/nzaa052_042
[9] P. Manda, D.S. Dano, J.H. Kouadio, A. Diakite, B. Sangare-Tigori, M.J.M. Ezoulin, A. Soumahoro, A. Dembele, G. Fourny, Impact of industrial treatments on ochratoxin A content in artificially contaminated cocoa beans, Food Addit. Contam. 26 (7) (2009) 1081–1088, https://doi.org/10.1080/02652030902894397
[10] M.V. Copetti, B.T. Iamanaka, J.C. Frisvad, J.L. Pereira, M.H. Taniwaki, Mycobiota of cocoa: from farm to chocolate, Food Microb. 28 (8) (2011) 1499–1504, https://doi.org/10.1016/j.fm.2011.08.005
[11] N. Cain, C. Marji, K. von Wuthenau, T. Segelke, M. Fischer, Food targeting: determination of the cocoa shell content (*Theobroma cacao L.*) in cocoa products by LC-QqQ-MS/MS, Metabolites 10 (3) (2020) 91, https://doi.org/10.3390/metabo10030091
[12] M.M. Oliveira, J.P. Cruz-Tirado, J.V. Roque, R.F. Teófilo, D.F. Barbin, Portable near-infrared spectroscopy for rapid authentication of adulterated paprika powder, J. Food Compos. Anal. 87 (2020) 103403, https://doi.org/10.1016/j.jfca.2019.103403
[13] D.F. Barbin, A.T. Badaro, D.C. Honorato, E.Y. Ida, M. Shimokomaki, Identification of turkey meat and processed products using near infrared spectroscopy, Food Control 107 (2020) 106816, https://doi.org/10.1016/j.foodcont.2019.106816
[14] Q. Wu, M.M. Oliveira, E.M. Achata, M. Kamruzzaman, Reagent-free detection of multiple allergens in gluten-free flour using NIR spectroscopy and multivariate analysis, J. Food Compos. Anal. 120 (2023) 105324, https://doi.org/10.1016/j.jfca.2023.105324
[15] A. Veselá, A.S. Barros, A. Synytsya, I. Delgadillo, J. Čopíková, M.A. Coimbra, Infrared spectroscopy and outer product analysis for quantification of fat, nitrogen, and moisture of cocoa powder, Anal. Chim. Acta 601 (1) (2007) 77–86, https://doi.org/10.1016/j.aca.2007.08.039
[16] D.F. Barbin, L.F. Maciel, C.H.V. Bazoni, M.D.S. Ribeiro, R.D.S. Carvalho, E.D.S. Bispo, M.P.S. Miranda, E.Y. Hirooka, Classification and compositional characterization of different varieties of cocoa beans by near infrared spectroscopy and multivariate statistical analyses, J. Food Sci. Technol. 55 (7) (2018) 2457–2466, https://doi.org/10.1007/s13197-018-3163-5
[17] C. Álvarez, E. Pérez, E. Cros, M. Lares, S. Assemat, R. Boulanger, F. Davrieux, The use of near infrared spectroscopy to determine the fat, caffeine, theobromine and (−)-epicatechin contents in unfermented and sun-dried beans of Criollo cocoa, J. Infrared Spec. 20 (2) (2012) 307–315, https://doi.org/10.1255/jnirs.990
[18] A. Krähmer, A. Engel, D. Kadow, N. Ali, P. Umaharan, L.W. Kroh, H. Schulz, Fast and neat–determination of biochemical quality parameters in cocoa using near infrared spectroscopy, Food Chem. 181 (2015) 152–159, https://doi.org/10.1016/j.foodchem.2015.02.084
[19] M.A. Quelal-Vásconez, M.J. Lerma-García, É. Pérez-Esteve, A. Arnau-Bonachera, J.M. Barat, P. Talens, Fast detection of cocoa shell in cocoa powders by near infrared spectroscopy and multivariate analysis, Food Control 99 (2019) 68–72, https://doi.org/10.1016/j.foodcont.2018.12.028
[20] E. Teye, X. Huang, H. Dai, Q. Chen, Rapid differentiation of Ghana cocoa beans by FT-NIR spectroscopy coupled with multivariate classification, Spectrochim. Acta A 114 (2013) 183–189, https://doi.org/10.1016/j.saa.2013.05.063
[21] E. Teye, X.Y. Huang, W. Lei, H. Dai, Feasibility study on the use of Fourier transform near-infrared spectroscopy together with chemometrics to discriminate and quantify adulteration in cocoa beans, Food Res. Int. 55 (2014) 288–293, https://doi.org/10.1016/j.foodres.2013.11.021
[22] C.W. Huck, New trend in instrumentation of NIR spectroscopy—miniaturization, Near-infrared Spectroscopy, first ed..,, Springer,, Singapore, 2021, pp. 193–210.
[23] K.B. Beć, J. Grabska, C.W. Huck, Principles and applications of miniaturized near-infrared (NIR) spectrometers, Chem.–Eur. J. 27 (5) (2021) 1514–1532, https://doi.org/10.1002/chem.202002838
[24] L. Rodriguez-Saona, D.P. Aykas, K.R. Borba, A. Urtubia, Miniaturization of optical sensors and their potential for high-throughput screening of foods, Curr. Opin. Food Sci. 31 (2020) 136–150, https://doi.org/10.1016/j.cofs.2020.04.008
[25] K.B. Beć, J. Grabska, C.W. Huck, Miniaturized NIR spectroscopy in food analysis and quality control: promises, challenges, and perspectives, Foods 11 (10) (2022) 1465, https://doi.org/10.3390/foods11101465
[26] E.K. Anyidoho, E. Teye, R. Agbemafle, Nondestructive authentication of the regional and geographical origin of cocoa beans by using a handheld NIR spectrometer and multivariate algorithm, Anal. Methods-UK 12 (33) (2020) 4150–4158, https://doi.org/10.1039/D0AY00901F
[27] E.K. Anyidoho, E. Teye, R. Agbemafle, C.L. Amuah, V.G. Boadu, Application of portable near infrared spectroscopy for classifying and quantifying cocoa bean quality parameters, J. Food Process. Pres. 45 (5) (2021) e15445, https://doi.org/10.1111/jfpp.15445
[28] L. Mandrile, L. Barbosa-Pereira, K.M. Sorensen, A.M. Giovannozzi, G. Zeppa, S.B. Engelsen, A.M. Rossi, Authentication of cocoa bean shells by near-and mid-infrared spectroscopy and inductively coupled plasma-optical emission spectroscopy, Food Chem. 292 (2019) 47–57, https://doi.org/10.1016/j.foodchem.2019.04.008
[29] C. Hernández-Hernández, V.M. Fernández-Cabanás, G. Rodríguez-Gutiérrez, A. Bermúdez-Oria, A. Morales-Sillero, Viability of near infrared spectroscopy for a rapid analysis of the bioactive compounds in intact cocoa bean husk, Food Control 120 (2021) 107526, https://doi.org/10.1016/j.foodcont.2020.107526
[30] W. Wu, B. Walczak, D.L. Massart, K.A. Prebble, I.R. Last, Spectral transformation and wavelength selection in near-infrared spectra classification, Anal. Chim. Acta 315 (3) (1995) 243–255, https://doi.org/10.1016/0003-2670(95)00347-3
[31] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemom. Intell. Lab. 2 (1–3) (1987) 37–52, https://doi.org/10.1016/0169-7439(87)80084-9
[32] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. 58 (2) (2001) 109–130, https://doi.org/10.1016/S0169-7439(01)00155-1
[33] I.G. Chong, C.H. Jun, Performance of some variable selection methods when multicollinearity is present, Chemom. Intell. Lab. 78 (1–2) (2005) 103–112, https://doi.org/10.1016/j.chemolab.2004.12.011
[34] Y.H. Yun, H.D. Li, B.C. Deng, D.S. Cao, An overview of variable selection methods in multivariate analysis of near-infrared spectra, Trac-Trend Anal. Chem. 113 (2019) 102–115, https://doi.org/10.1016/j.trac.2019.01.018
[35] B.G. Osborne, T. Fearn, P.H. Hindle, Practical NIR spectroscopy with applications in food and beverage analysis, Longman Scientific and Technical, first ed., Harlow, UK, 1993.
[36] J. Workman, L. Weyer, Practical Guide to Interpretive Near Infrared Spectroscopy, CRC Press,, New York, 2008, pp. 239–287.
[37] J.S. Ribeiro, M.M. Ferreira, T.J.G. Salva, Chemometric models for the quantitative descriptive sensory analysis of Arabica coffee beverages using near infrared spectroscopy, Talanta 83 (5) (2011) 1352–1358, https://doi.org/10.1016/j.talanta.2010.11.001
[38] P.C. Williams, K.H. Norris, Near-infrared Technology in the Agricultural and Food Industries, AACC,Inc, St.Paul, Minnesota, 1987, pp. 145–169.
[39] J.C. Hashimoto, J.C. Lima, R. Celeghini, A.B. Nogueira, P. Efraim, R.J. Poppi, J.A. Pallone, Quality control of commercial cocoa beans (*Theobroma cacao L.*) by near-infrared spectroscopy, Food Anal. Method. 11 (5) (2018) 1510–1517, https://doi.org/10.1007/s12161-017-1137-2
[40] Å. Rinnan, F. Van Den Berg, S.B. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, Trac-Trend Anal. Chem. 28 (10) (2009) 1201–1222, https://doi.org/10.1016/j.trac.2009.07.007
[41] H. Cen, Y. He, Theory and application of near infrared reflectance spectroscopy in determination of food quality, Trends Food Sci. Tech. 18 (2) (2007) 72–83, https://doi.org/10.1016/j.tifs.2006.09.003
[42] M.M. Oliveira, A.T. Badaró, C.A. Esquerre, M. Kamruzzaman, D.F. Barbin, Handheld and benchtop vis/NIR spectrometer combined with PLS regression for fast prediction of cocoa shell in cocoa powder, Spectrochim. Acta A 298 (2023) 122807, https://doi.org/10.1016/j.saa.2023.122807

[43] M. Kamruzzaman, D. Kalita, M.T. Ahmed, G. ElMasry, Y. Makino, Effect of variable selection algorithms on model performance for predicting moisture content in biological materials using spectral data, Anal. Chim. Acta 1202 (2022) 339390, https://doi.org/10.1016/j.aca.2021.339390

[44] M.S. Fowler, F. Coutel, Cocoa beans: from tree to factory, Beckett's Ind. Choc. Manuf. Use (2017) 9–49.

[45] C.D. Brown, R.L. Green, Critical factors limiting the interpretation of regression vectors in multivariate calibration, Trac-Trend Anal. Chem. 28 (4) (2009) 506–514, https://doi.org/10.1016/j.trac.2009.02.003

[46] X. Huang, E. Teye, L.K. Sam-Amoah, F. Han, L. Yao, W. Tchabo, Rapid measurement of total polyphenols content in cocoa beans by data fusion of NIR spectroscopy and electronic tongue, Anal. Methods-UK 6 (14) (2014) 5008–5015, https://doi.org/10.1039/C4AY00223G

[47] B. Giussani, G. Gorla, J. Riu, Analytical chemistry strategies in the use of miniaturised NIR instruments: an overview, Crit. Rev. Anal. Chem. (2022) 1–33, https://doi.org/10.1080/10408347.2022.2047607

[48] C. Pasquini, Near infrared spectroscopy: a mature analytical technique with new perspectives–a review, Anal. Chim. Acta 1026 (2018) 8–36, https://doi.org/10.1016/j.aca.2018.04.004