Food Control 118 (2020) 107445

Contents lists available at ScienceDirect

Food Control

journal homepage: www.elsevier.com/locate/foodcont

Authentication of cocoa (*Theobroma cacao*) bean hybrids by NIR-hyperspectral imaging and chemometrics

J.P. Cruz-Tirado [a], Juan Antonio Fernández Pierna [b], Hervé Rogez [c], Douglas Fernandes Barbin [a,*], Vincent Baeten [b]

[a] Department of Food Engineering, School of Food Engineering, University of Campinas, Campinas, SP, Brazil
[b] Walloon Agricultural Research Centre (CRA-W), Knowledge and Valorization of Agricultural Products Department, Quality and Authentication of Products Unit, Gembloux, Belgium
[c] CVACBA (Center for Valorization of Amazonian Bioactive Compounds), UFPA (Federal University of Pará), Avenida Perimetral, 01, 66075-150, Guamá, Belém, Pará, Brazil

ARTICLE INFO

Keywords:
Genotype
Imaging
Computer vision
Near-infrared imaging

ABSTRACT

The hybridization of cocoa generates new varieties with the aim of opening new flavors, textures, and disease resistance. The objective of this study was to develop and validate classification models based on NIR hyperspectral imaging and chemometrics for the discrimination of five valuable cocoa bean hybrids. The chemometrics tools, PLS-DA and SVM, showed comparable results for two-class (hybrids) models, but SVM (3.8–23.1% prediction error) was superior to PLS-DA (4.4–34.4% prediction error) when all five classes (hybrids) were included in a model. PLS-DA maps showed a simple and informative way to discriminate hybrids, allowing a correct classification in 50–100% of cases. Finally, it can be concluded that the models created in this work could be a good and reliably alternative to the actual visual method for the discrimination of cocoa bean hybrids.

1. Introduction

According to International Trade Centre (2001), there is not a general rule to use both terms 'cacao' or 'cocoa' (*Theobroma cacao*) to refer to the bean. However, it is common to use the term 'cacao' to describe the scientific and horticultural aspects of the plant, reserving 'cocoa' for fermented and dried bean. The cocoa bean is one of the agricultural commodities highly demanded in the world for its high benefits in terms of nutrition and economics. Ivory Coast, Ghana, Indonesia, Nigeria, Cameroon and Brazil are the world's largest cocoa producer (4.6 million tons of cocoa harvested in 2016) (The International Cocoa Organization, 2018).

Hybridization is a common technology that allowed to create cocoa bean hybrids with different features such as a greater disease resistance (e.g. "witches broom disease" caused by *Moniliophthora perniciosa*), however it also can affect pod and bean yield parameters, precocity and butterfat flavor expressed after optimal fermentation. Traditionally, cocoa genotypes are classified in three varieties: Criollo, Forastero and Trinitario. Later, Motamayor et al. (2008) proposed a new sub-classification for Forastero group, including: Marañon (PA), Curaray (AGU), Iquitos (IMC), Nanay (NA), Contamana (SCA), Amelonado (BE), Purús (CAB), Nacional (MO) and Guiana (CJ). These genotypes were reported for South America, and from them, through hybridization, cocoa producers create hybrids with disease resistance, a greater yielding, and interesting flavor. Nevertheless, many times hybrids are planted together and the seeds are mixed, making it difficult to identify the purity of cocoa in relation to a variety of high economic value, in order to assure the quality of the desired final product, especially chocolate.

Various analytical techniques such as multi-element and multi-compound isotope profiling (^{13}C, ^{15}N, % C, % N) (Diomande et al., 2015), proteomic and peptidomic fingerprinting by ultra-performance liquid chromatography tandem mass spectrometry method with electrospray ionization (UHPLC-ESI-MS/MS) (Kumari et al., 2018; Scollo, Neville, Oruna-Concha, Trotin, & Cramer, 2020), nanofluidic single nucleotide polymorphism (SNP) genotyping (Fang et al., 2014) or microsatellite markers (Dinarti et al., 2015; Herrmann et al., 2015) have been developed to identify cocoa bean hybrid. It should be considered that these techniques have been developed and evolved to reduce its costs and increase applications.

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

https://doi.org/10.1016/j.foodcont.2020.107445
Received 6 April 2020; Received in revised form 30 May 2020; Accepted 22 June 2020
Available online 28 June 2020

Some non-destructive techniques such as Raman spectroscopy (Vargas Jentzsch et al., 2016) and computer vision (Jimenez et al., 2018; Mite-Baidal et al., 2019) showed a good performance to identify cocoa beans genotypes. However, in both cases the objective was to identify a specific variety (CCN-51) and two-classes model was developed. Although these techniques are accurate and reliable, some of them are time-consuming. The cocoa industry is very dynamic, with a complex supply chain, so online, rapid, minimal sample preparation and non-destructive and chemical-free analytical methods are desired.

Near infrared spectroscopy (NIRS) meets these requirements. For the cocoa industry, NIRS has proved efficient in determining protein, fat, caffeine, theobromine, (−)-Epicatechin, carbohydrates and moisture content in cocoa flour (Barbin et al., 2018; Veselá et al., 2007; Álvarez et al., 2012), to detect adulterations by substitution (e.g. carob flour) in cocoa flour (Quelal-Vásconez et al., 2019; Quelal-Vásconez, Pérez-Esteve, Arnau-Bonachera, Barat, & Talens, 2018), and geographical origin and variety of cocoa beans and shells (Mandrile et al., 2019; Teye, Huang, Dai, & Chen, 2013; Trognitz et al., 2013). However, for NIR spectra acquisition, cocoa bean samples were ground.

Near infrared hyperspectral imaging (NIR-HSI) is a NIR-based technology that yields spectral and spatial information simultaneously (Baeten, Pierna, Vermeulen, & Dardenne, 2010; Dale et al., 2013; Fernández Pierna, Baeten, & Dardenne, 2006). Previous studies have shown the ability of NIR-HSI in tandem with multivariate analysis to

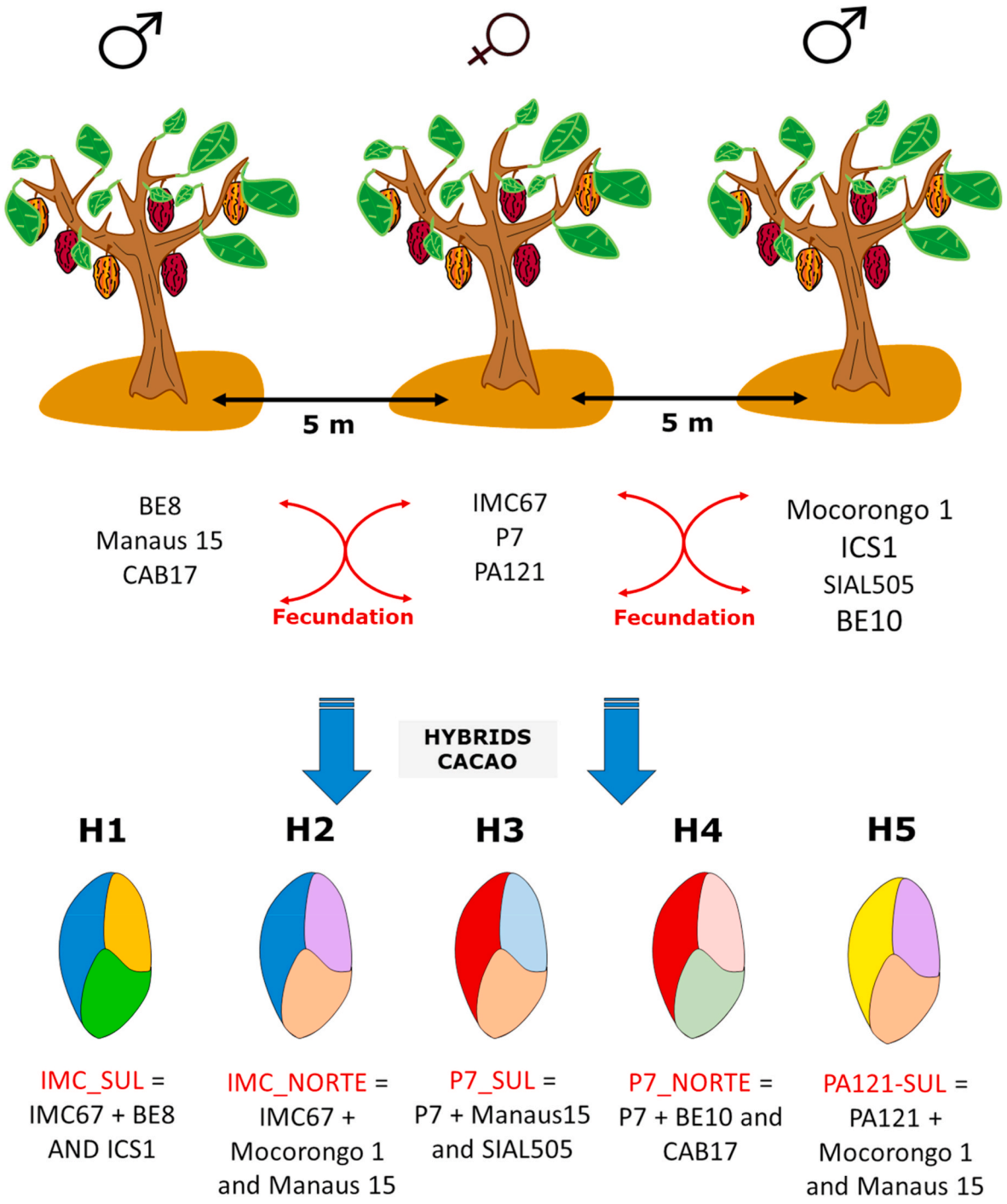

Fig. 1a. Scheme showing the process for production of cocoa bean hybrids in CEPLAC (Medicilândia, Para, Brazil).

identify hybrids of rice (X. Liu, Feng, Liu, & He, 2017), okra seeds (Nie, Zhang, Feng, Yu, & He, 2019), maize seeds (Guo, Zhu, Huang, Guo, & Qin, 2017), sweet potato (Su, Bakalis, & Sun, 2019) and soybean (Y. Liu, Wu, Yang, Tan, & Wang, 2019), among others. Regarding cocoa beans, Caporaso, Whitworth, Fowler, and Fisk (2018) showed that the spectral information obtained from NIR-HSI can be used to predict the fermentation index, total polyphenols and antioxidant activity in single peeled dried fermented cocoa beans. According to Okiyama, Navarro, and Rodrigues (2017), one kg dried cocoa bean shell is composed approximately by 504–606 g fiber, 116–181 g protein, 47–101 g moisture, ~178 g carbohydrates, 20–68 g fat and 18–58 g phenols. Previous studies reported that the chemical information obtained from the cocoa bean shell can be used to identify genetic varieties of cocoa beans (Mandrile et al., 2019).

In this study, NIR-HSI images were therefore used to extract spectral information from images of fermented and dried unpeeled cocoa beans. This information was used to discriminate cocoa beans hybrids, for quality control purposes (purity). Thus, the objective of this work was to identify and classify cocoa bean hybrids originating in the Brazilian Basin, using NIR-HSI as a non-destructive procedure.

2. Material and methods

2.1. Collection of samples

Cocoa bean hybrids samples were collected by CEPLAC (Medicilândia, Para, Brazil) in August 2018, and subsequently analyzed. Samples were collected in the same geographical zone, to avoid effects related to soil composition and climatic factors. Five cocoa bean hybrids from *Forastero* genotype were selected based on their importance for Brazilian cocoa industry and are summarized in Fig. 1a.

All samples were fermented for 7 days, under same conditions of temperature and relative humidity. After fermentation, samples were dried for 4 days under sunlight, after which samples were stored at 25 °C until analysis. Fig. 1b shows the final internal appearance of the cocoa bean hybrids, confirming the good compartmentation and typical brown colour of these beans, and also demonstrating that samples are visually similar. No additional processing such as grinding, shelling was performed on the cocoa bean hybrids.

2.2. Instrumentation

NIR hyperspectral images of cocoa bean hybrids were acquired using a line scan imaging system combined with a conveyor belt (Burger-Metrics SIA, Riga, Latvia) in room temperature at 25 °C. This device consisted of a SWIR XEVA CL2.5 320 TE4 camera (Specim Ltd., Oulu, Finland) which has a resolution of 320-pixel lines and an ImSpector N25E spectrograph (Xenics nv, Leuven, Belgium) with a spectral range of 1100–2400 nm (209 wavelengths) and a spectral resolution of 6.3 nm. Thirty-two scans per image were averaged and each pixel provided an absorbance spectrum at each point of the image. Image acquisition was performed using the HyperPro software (BurgerMetricsSIA, Riga, Latvia). More details of the device and its components can be found in Eylenbosch et al. (Eylenbosch, Bodson, Baeten, & Fernández Pierna, 2018).

Prior to sample image acquisition, the NIR-hyperspectral imaging system was calibrated with a dark image (by shutting off the lens entrance) and a white image (background) was collected from a standard white reference board (empty Teflon plate). The spectra were then automatically corrected. The cocoa beans were then placed on a conveyor belt (with a speed of 1.1 mm/s) in groups of 10 beans for each image. Image acquisition was performed in room conditions (25 °C, 50% relative humidity).

$$I = \frac{(I_0 - B)}{(W - B)} * 100 \quad (1)$$

where I_0 is the original hyperspectral image; B is the dark image and W is the white image.

2.3. Spectral data collection

In total, 50 fermented and dried cocoa beans from each hybrid were chosen for the construction of the models. Additionally, a second set of 200 cocoa beans was used for pixel-to-pixel external validation. The number of cocoa bean hybrids chosen was sufficient to create robust models, since the experiment was controlled in (1) soil type (geographical origin), (2) same fermentation and drying process and (3) same harvest date and (4) same pre-analysis storage time. The mean spectrum (ROI) was extracted from whole cocoa beans for each hybrid (both sides). For this purpose, a mask was built to isolate the cocoa beans from the background using the difference in intensity. After ROI segmentation, spectral libraries for cocoa bean hybrids were compiled by calculating mean spectra for each of the 50 cocoa beans per hybrid, on both sides, finally resulting in 100 mean spectra for each variety (library with 500 spectra). In addition, a new set of cocoa bean hybrids containing 200 beans (200 images) was processed under the same conditions. This new dataset was acquired in order to demonstrate that the models created can predict on a completely independent set of data, enhancing their robustness for industrial applications.

2.4. Data treatment

Chemometric data analysis was carried out using the PLS Toolbox from Eigenvector Research, Inc. (Manson, WA, USA) for Matlab R2017 (Mathworks, Natick, USA). First, the reflectance signal is converted to absorbance prior to spectral data treatments. The mean spectra of cocoa bean hybrids were pre-processed using standard normal variate (SNV) and first derivative (Savitzky Golay, filter width 15 and a polynomial order of 2) to remove random shift of the baseline offset, light scattering interferences and noise, avoiding external factors to affect the spectra (Vidal & Amigo, 2012).

Principal component analysis (PCA) was carried out to investigate systematic differences between samples and to find and remove outliers (Sendin, Manley, Baeten, Fernández Pierna, & Williams, 2019). For this purpose, spectral data were mean centered, after which PCA modeling was performed using the singular value decomposition (SVD) algorithm with a confidence level of 0.95 for Q and T^2.

For discrimination purposes, Partial Least Squares discriminant analysis (PLS-DA) and Support Vector Machines (SVM) were applied to the data set. PLS-DA is a supervised lineal method widely used for food quality control and authentication. SVM is a supervised learning technique that works through the searching of hyperplanes (high dimensional feature space) by the use of kernel functions and penalization criteria, allowing both linear and non-linear classifications (Pierna, Baeten, Renier, Cogdill, & Dardenne, 2004).

Two approaches have been tested in this work. In the first approach, the aim was to construct PLS-DA and SVM models to discriminate between two cocoa bean hybrids. The spectral data was constituted of 200 mean spectra for each model, of which 80% (160 mean spectra) were randomly selected for the calibration and cross-validation of the models and the remaining 20% (40 spectra) was used as test set to assess the discriminative capacity of the models. In the second approach, the aim was to construct one PLS-DA and SVM models to discriminate between the five cocoa bean hybrids. Spectral data were randomly divided into two subsets: calibration and cross-validation set (80% or 400 mean spectra) and test set (20% or 100 mean spectra). In both cases, leave-one-out cross validation was applied when building the PLS-DA models, and the number of latent variables was chosen through the evaluation of sensitivity (Eq. (1)), specificity (Eq. (2)) and classification error (Eq. (3)) of cross-validation and prediction (Nie et al., 2019). For SVM models, the penalty parameters of cost (c) and kernel function

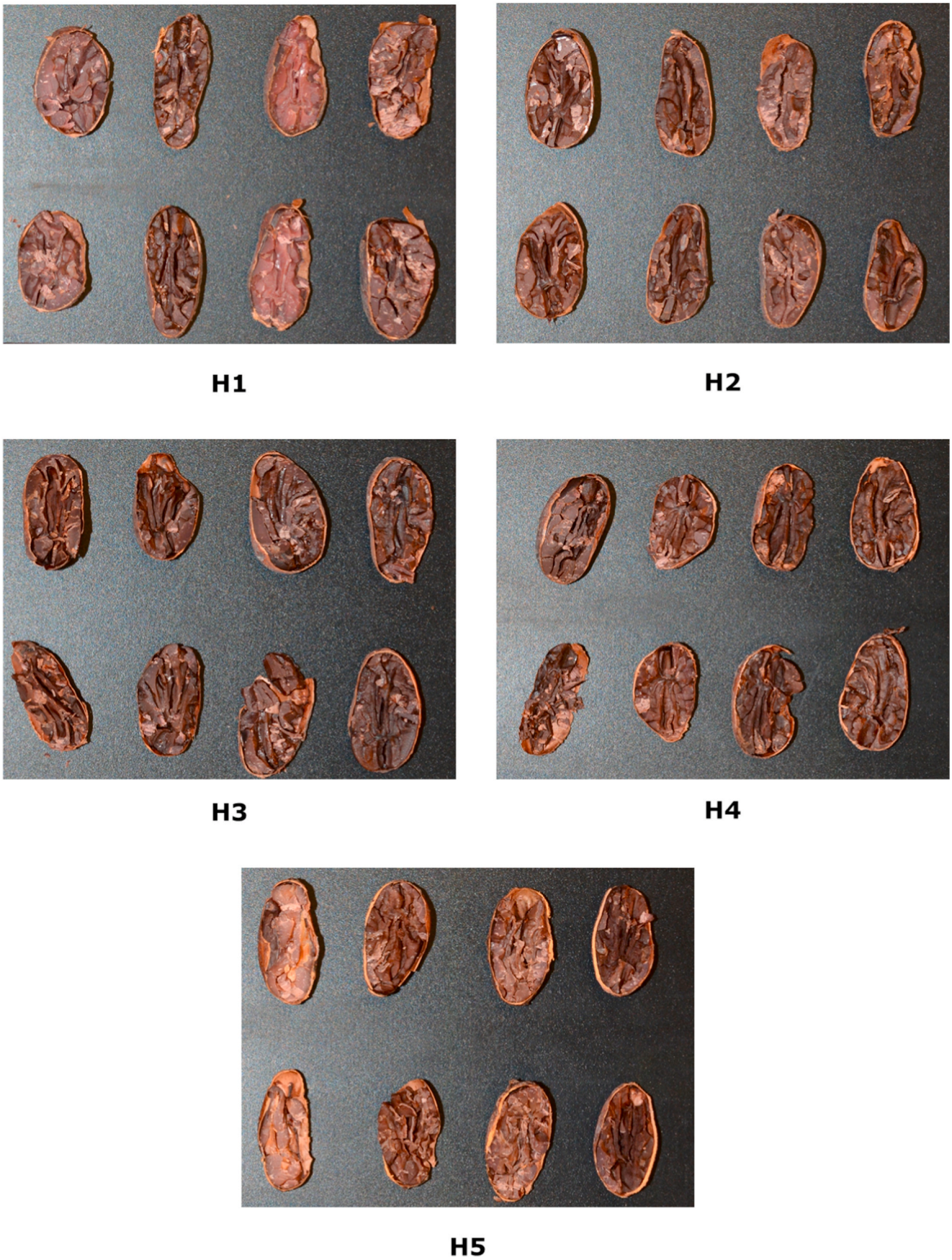

Fig. 1b. Photo of the visual aspect of the internal quality of cocoa bean hybrids.

parameters gamma (g) were optimized using a grid search (Fernández Pierna et al., 2012; Pierna et al., 2011). The model performance was also statistically evaluated according to sensitivity, specificity and classification error.

$$Sensitivity(\%) = \frac{TP}{TP + FN} \quad (2)$$

$$Specisficity(\%) = \frac{TN}{TP + FN} \quad (3)$$

$$Error(\%) = \frac{FP + FN}{TP + FP + TN + FN} \quad (4)$$

Where TP: true positive (positive samples correctly classify), TN: true negative (negative sample correctly classify), FP: false positive (positive samples incorrectly classify), FN: false negative (negative samples incorrectly classify).

For external validation of the models and to develop classification maps, a new sample test set (new cocoa bean hybrids not considered in the calibration or in the initial validation of the models) was used. The new test set consisted of 200 samples of the 5 cocoa bean hybrids (20% of each hybrid) and were divided as follows on two subsets: SET 1 contained hybrids of the same class and SET 2 contained the 5 hybrids (20% of each hybrid) placed in known spatial positions in the image. The discriminative capacity of the PLS-DA models for two classes and for five classes were evaluated based on the correct classification rate (% CCR) for each hybrid in the new test set. Predictive ability was assessed on hyperspectral images (pixel-to-pixel) and measured as the % CCR using an algorithm with 4 approaches: (1) prediction using raw model, (2) prediction using majority vote, (3) filtered model by deleted samples no possible to classify (difference between 2 classes with more probability < 65 pixel), and 4) PLS-DA model applied to pixel with mean spectra value. All PLS-DA maps were constructed using their own program developed in Matlab R2017 (Mathworks, Natick, USA).

3. Results and discussion

3.1. Spectra profile

The raw spectra and pre-processed spectra of five cocoa bean hybrids are presented in Fig. 2A and B, respectively. The mean spectra of all hybrids had similar pattern of absorbance, but their relative absorbance was different in some spectral regions. Similarities in the shape of the spectrum are inherent to the specie (*Theobroma cacao*), while differences in absorbance are related to variations in the composition of the shell of cocoa beans (Quelal-Vásconez et al., 2019). Because all hybrids came from the same geographic area and the postharvest process (fermentation and drying conditions), it is possible to assume that variations in composition, and therefore spectral similarities and differences, are a consequence of hybrid genetics.

Absorption bands of 1181 and 1426 nm correspond to the second overtone of O–H stretching and O–H deformation (Teye et al., 2013), which is associated with water and fiber content (Okiyama et al., 2017). In addition, the band at 1426 nm can also be associated with the first overtone of the N–H stretching vibration. Therefore this band is associated with the –CONH– structure (peptide) and related to protein content in cocoa beans shells (Mandrile et al., 2019; Quelal-Vásconez et al., 2019; Veselá et al., 2007). An absorption band of 1263 nm is associated with the second overtone of C–H stretching (–CH_3 or –CH_2), due to the presence of fibers and carbohydrates (Okiyama et al., 2017; Osborne, Fearn, Hindle, & Osborne, 1993). The absorption bands of 1533, 1577 and 1650–1780 nm are associated with functional groups like (−)-epicatechin (flavanol), theobromine and caffeine (alkaloids), proteins, volatile and non-volatile acids (Teye et al., 2013; Veselá et al., 2007; Álvarez et al., 2012). Also, absorption band at 1715 nm is mainly related to fat content and fatty acid. The spectral region at 1900–1950 nm is mainly associated with O–H combinations, influenced by moisture content in cocoa bean. It has also been reported that the band at 1916 nm is related to the second overtone of C═O and it is associated with the content of (−)-epicatechin (Álvarez et al., 2012) and lignin content, while the band at 1990 nm is related to O–H combinations and with asymmetric stretching of N–H and amide II group, strongly influenced by protein content. Bands at 2092 and 2280 nm are also related to second overtone of CH═CH and CH_3 combination respectively, and they are characteristics of lignin, aromatics, polyphenols and fatty acids in cocoa beans (Sunoj, Igathinathane, & Visvanathan, 2016; Teye et al., 2015). However, the peak at 2092 nm is associated with protein content (Caporaso et al., 2018; Veselá et al., 2007). The absorption bands at 2199 and 2375 nm could be attributed to the combination of C–H (Ma, Wang, Chen, Cheng, & Lai, 2017) and to the stretching and rocking vibrations of C–H and C–C of cellulose and hemicellulose (Okiyama et al., 2017; Wang et al., 2018), respectively. These bands were previously reported for cocoa bean shell (Mandrile et al., 2019; Quelal-Vásconez et al., 2019).

3.2. PCA

PCA modeling was performed using pre-processed data (Fig. 2C) to identify possible clusters and to evaluate the effect of the pre-processing (SNV + 1st derivate).

PCA shows the possibility, although not easily, to observe groups of hybrids of the same class. H1 hybrid (purple symbol) was characterized by negative scores on PC1 and by negative scores on PC2. H2 (red symbol) and H4 (green symbol) hybrids were located on positive scores on PC1. While H2 hybrid presents greater variability on PC2, the H5 hybrid were characterized by positive scores on PC2. H3 (greenlight blue symbol) hybrid are characterized by negative scores on PC2, although they have great variability throughout PC1. Besides, H1 hybrids overlap on H2 hybrids, with whom they share a common ancestor (Fig. 1). The results of the PCA show that for the same hybrid there is great variability between the samples, while there is an overlap between the hybrids of different classes. This could occur due to two situations: (1) the cocoa bean shell has fiber and protein as the majority components (Mandrile et al., 2019) and, (2) the five hybrids share genetic ancestry, so the concentration of fiber and protein can be very similar. Therefore, the overlap may be associated with the similarity in its major components, while the clustering would be associated with the difference in minor components such as fatty acid tryptamides (FATs, more abundant in shell) (Quelal-Vásconez et al., 2018), phenolic compounds and aromatic compounds (Okiyama et al., 2017) developed during fermentation and drying for each class of hybrid.

3.3. Discriminant analysis

PLS-DA models were created using the full spectral region (1100–2400 nm), however, the results (data no show) showed little sensitivity to discriminate between the 5 hybrids. Therefore, in this work, PLS-DA and SVM was carried out in two approaches using spectral data in the range 1369–2054 nm. This spectral region was more significant for discrimination (lower % classification error) than using the full spectrum. PLS-DA and SVM discrimination models were calibrated and validated using the mean spectrum of the cocoa bean hybrids. Discrimination models were created to identify a specific hybrid (genetic cross-linking) within a heterogeneous batch of cocoa beans, where the provenance of each bean was known. These models were created to determine the purity of a batch of cocoa beans, more necessarily to associate it with the origin of a specific tree, since cocoa trees are hermaphrodites, and can contain more than two hybrids to the same time.

3.3.1. IMC vs P7 classes

Here, we investigated the possibility of constructing PLS-DA and SVM classification models to discriminate between hybrids according to

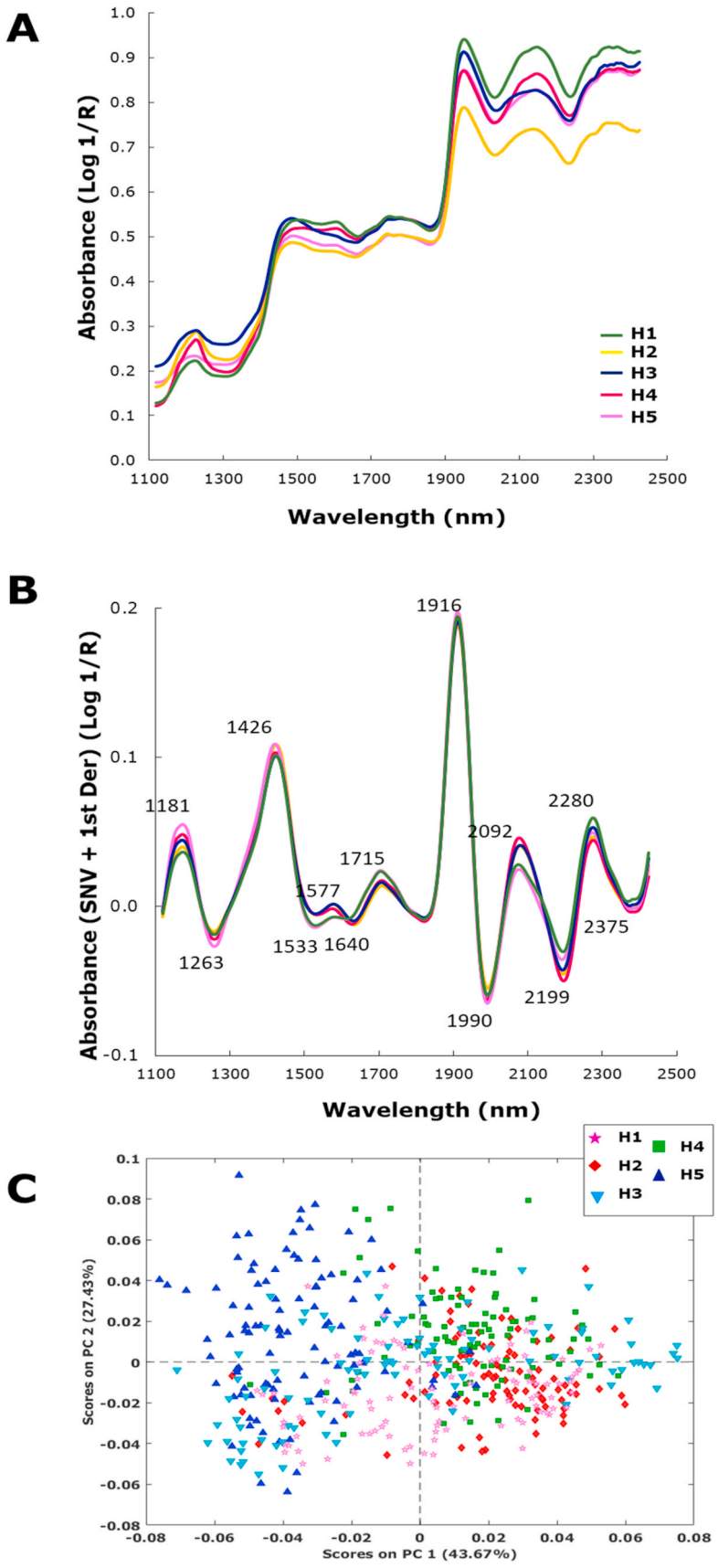

Fig. 2. Spectral profile of raw spectra (A), spectra treated with SNV + first derivative (B) and PCA scores based on spectra treated with SNV + first derivate (C).

their "mother", being IMC 67 for H1 and H2, and P7 for H3 and H4. The PLS-DA model obtained a better result than the SVM model (11.3% *vs* 17.5% error, Table 1), although with a large number of latent variables (14). Recently, Scollo et al. (2020) reported that the IMC 67 hybrid had a different protein profile from other types of hybrids (e.g. amino-hydrolase content), which allowed it to be clearly classified. Loading plot (Fig. 1A, supplementary material) shows the importance of spectral region 1900–1990 nm for PLS-DA model, which is strongly associated with O–H combinations, mainly water, but it also with functional groups of proteins, polyphenols, fatty acid and aromatics (Caporaso et al., 2018; Teye et al., 2013).

Fig. 3 shows the classification results (based on mean spectrum) and the pixel-to-pixel allocation of each hybrid class (IMC *vs.* P7). The classification results show as five beans of IMC hybrids and four beans of P7 hybrids are misclassified. For the pixel-to-pixel classification, using majority vote (Fig. 3b), all grains are correctly assigned to the ICM class (100%), while for class P7 only 80% of the correct classification was reached (16/20). The approach c (Fig. 3c) shows that cocoa bean hybrids from P7 misclassified as IMC always had a larger number of pixels like the IMC class (class 1 - class 2 < 65 pixels). This is also visible in Fig. 3a. The higher discrimination capability may be associated with the ancestry of the "mothers" of the hybrids. IMC 67 is a variety from the genotype "Iquitos", while P7 probably comes from the "Namay" or "Marañon" groups (Motamayor et al., 2008). Satisfactorily, the P7 hybrids classification improved when the average spectrum value was applied to each pixel for classification (Fig. 3d), reaching 95% correct classification (19/20). Because the shell is maternally derived from the integuments of the ovary, its composition is preferably dominated by "mother" genetic. In this sense, success in the discrimination of hybrids is probably associated with the particularities in composition of each "mother" species.

3.3.2. *Two-class models*

Here, two-classes PLS-DA and SVM models were built to discriminate between two specific cocoa bean hybrids, their performance is shown in Table 1. In general, the PLS-DA and SVM models showed a good performance with a % sensitivity of 60–100% and a % error of 0–22.5%. In most cases, the% error for SVM models was less than or equal to that for PLS-DA models, except for models H1 *vs* H2 and H3 *vs* H4. This is probably because the SVM algorithm is more complex and can work linearly and non-linearly way. However, generally speaking, a significant improvement was probably not observed for SVM models due to the complexity of the hybrids. The classification models were more sensitive for H1 in all cases (Table 1), which allows their clear distinction from any other hybrid.

For all PLS-DA models, the loadings (Fig. 1B–K, supplementary material) showed the importance of the spectral region around 1650 nm, associated with proteins and functional groups like (−)-epicatechin, theobromine and caffeine content, and the band at 1715 is associated with fatty acid in cocoa beans, as previously reported by Alvarez et al. (2012). Also, the spectral region at 1900–1990 nm is associated with asymmetric stretching of N–H and amide II group present in the protein, which is second larger component in cocoa shell (116–181 g protein/kg dried cocoa shell) (Okiyama et al., 2017). Also, each hybrid has a particular protein profile (Scollo et al., 2020), allowing the formation of flavor precursors in cocoa beans, however, not all protein is degraded or transformed, since the proteolytic processes are different in each hybrid (Moreira, Vilela, Santos, Lima, & Schwan, 2018).

The H1 *vs* H2 model allowed, in the new external validation set (Fig. 4A) to correctly classify all cocoa bean hybrids (100%). Fig. 4A, (approach *a*) shows the model sensitivity for the H1 class, where almost 100% of the pixels corresponding to H1 hybrid were assigned to that class. That sensitivity was lower for H2, although not sufficient to incorrectly classify hybrids in any of the approaches (Fig. 4A, approach *b-d*). For H1 hybrid, the classification rate could be related to protein content and profile, which are quite specific and distinctive in their predecessors IMC 67 and ICS 1 (fine cocoa) (Scollo et al., 2020).

By contrast, the model H3 *vs* H4 instead showed less sensitivity for prediction pixel-to-pixel (Fig. 4B, approach *a*). Using majority vote (Fig. 4B, approach *b*), the hybrids of class H4 were correctly classified at 100%, while class H3 was correctly classified at 70%. However, the best prediction was achieved using the third approach, where H3 and H4 hybrids were correctly classified at 100% and 90% (Fig. 4B, d). Here, parents SIAL505 (H4) and BE10 (H3) (Fig. 1) belong to the genotype "Amelonado" (Motamayor et al., 2008). This would explain why some hybrids of the H3 class are confused with hybrids of the H4 class.

Fig. 4C (approach *a*) shows that H5 hybrids show a definite assignment of the class, with almost all pixels correctly assigned. This

Table 1
Performance of two-classes PLS-DA and SVM classification models for the hybrids of cocoa beans obtained by HSI in the spectral region 1369–2054 nm.

Hybrid	Model	Parameter[a]	Sensitivity		Specificity		Error	
			CV	Pred	CV	Pred	CV	Pred
IMC *vs* P7	PLS-DA	14	0.969	0.875	0.963	0.900	0.034	0.113
	SVM	(1; 0.0316)	0.956	0.775	0.963	0.875	0.041	0.175
H1 *vs* H2	PLS-DA	9	0.975	1.000	0.975	1.000	0.025	0.000
	SVM	(10; 0.0316)	0.988	1.000	0.963	0.750	0.025	0.125
H1 *vs* H3	PLS-DA	6	0.975	0.700	0.925	0.900	0.069	0.200
	SVM	(0.3; 0.0316)	0.975	0.600	0.975	0.950	0.025	0.225
H2 *vs* H4	PLS-DA	6	0.963	0.650	0.975	0.900	0.031	0.225
	SVM	(1; 0.0316)	1.000	0.650	0.988	0.950	0.006	0.200
H2 *vs* H5	PLS-DA	4	0.975	0.750	0.975	0.950	0.025	0.150
	SVM	(100; 0.00031)	0.975	0.700	1.000	1.000	0.125	0.150
H1 *vs* H3	PLS-DA	4	0.963	1.000	0.938	0.850	0.050	0.075
	SVM	(31.62; 0.001)	0.963	1.000	0.925	8.000	0.050	0.050
H1 *vs* H4	PLS-DA	8	1.000	1.000	0.975	9.000	0.013	0.050
	SVM	(1; 0.1)	1.000	0.988	1.000	0.950	0.006	0.025
H1 *vs* H5	PLS-DA	5	0.988	1.000	1.000	0.950	0.006	0.025
	SVM	(100; 0.00316)	1.000	1.000	0.988	0.950	0.006	0.025
H3 *vs* H4	PLS-DA	12	0.963	0.900	0.988	0.900	0.025	0.100
	SVM	(3.16; 0.0316)	0.875	0.900	0.950	0.850	0.088	0.125
H3 *vs* H5	PLS-DA	8	0.938	0.800	0.950	0.900	0.056	0.150
	SVM	(3.16; 0.001)	0.963	0.950	0.975	0.900	0.031	0.075
H4 *vs* H5	PLS-DA	8	0.988	0.900	0.975	1.000	0.019	0.050
	SVM	(10; 0.01)	1.000	0.900	1.000	1.000	0.000	0.050

PLS-DA: partial least square discriminant analysis; SVM: support vector machine discriminant analysis; CV: cross-validation; Pred: prediction.
[a] Parameter for PLS-DA model is the optimal number of LVs; for SVM model is different penalty parameters: cost (c) and kernel function parameters gamma (g).

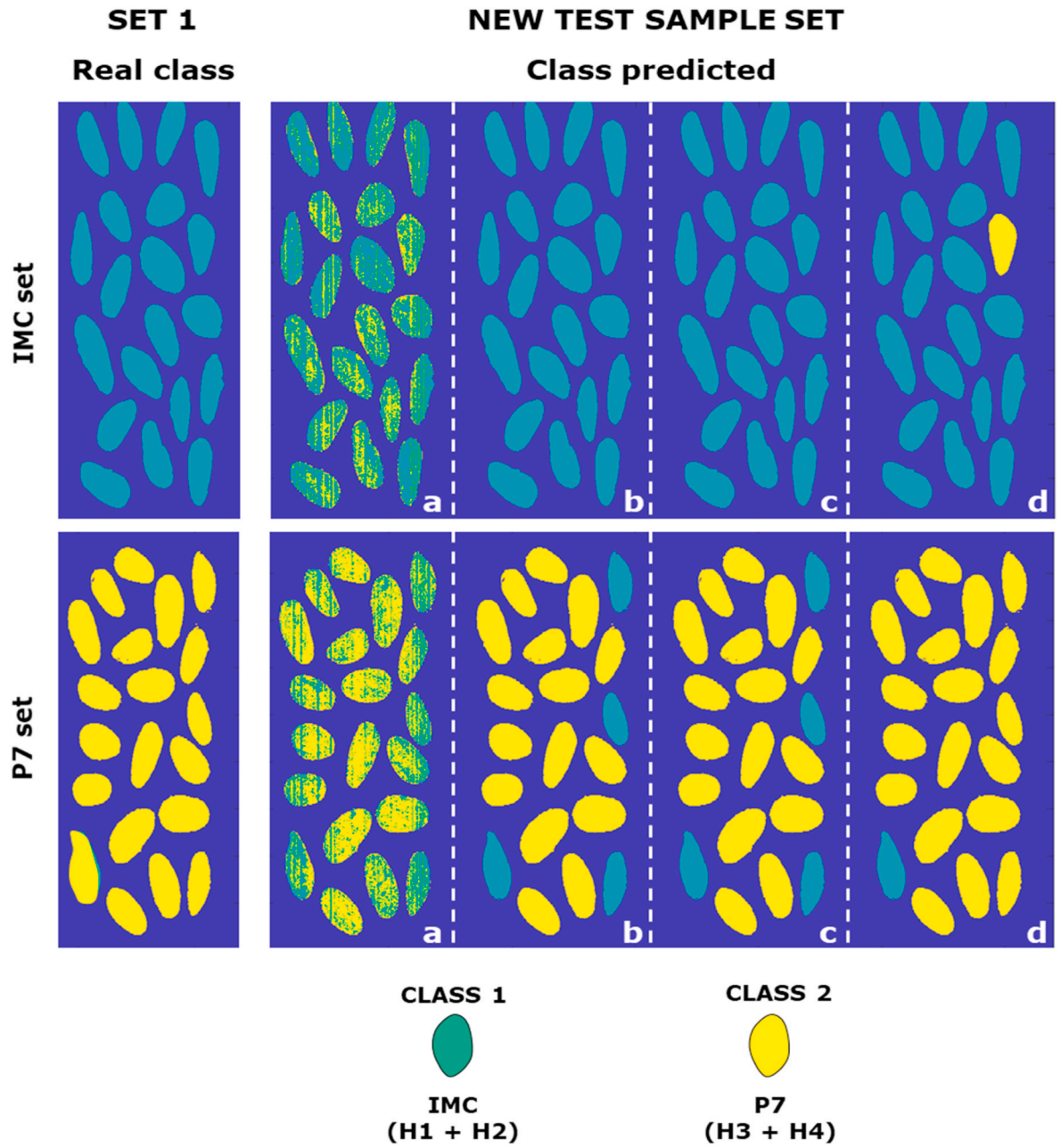

Fig. 3. PLS-DA analysis for two classes of hybrids of cocoa beans according to "mother": IMC *vs* P7. PLS-DA maps for the most probable class assigned to the cocoa bean hybrids were produced using: a) PLS-DA model; b) model applying majority vote; c) filtered model by deleted samples no possible to classify (difference between two classes with more probability < 65 pixels); d) PLS-DA model applied to pixels with mean spectra value.

facilitated their authentication, as any of the three approaches (Fig. 4C, b-d) achieved correct classification at a rate of 100% in the new test set of H5 and 90% in the new test set of H2 (data not shown). Probably, the success in the classification of the hybrids H2 and H5 is related to the difference in the genotype to which their mothers belonged, i.e. IMC 67 (mother of H2) belonging to the genotype "Iquitos" and PA121 (mother of H5) belonging to the genotype "Marañon (Parinari I)" (Motamayor et al., 2008).

The differentiation between the H5 and H4 hybrids was more complicated (Fig. 4D), with a 70% correct classification rate for the H5 hybrid and 90% for the H4 hybrid. A similar result with 80% correct classification rate for H5 and 100% for H3 (data not shown) was achieved. None of the three approaches (Fig. 4D, b-d) led to an improvement in the prediction of cocoa bean hybrids. Hypothetically, the greater sensitivity of the model for identifying the P7 samples suggests that these hybrids have a distinctive composition of residual compounds (e.g. protein fractions) which may be quantity and type than in H5 hybrids.

From the results presented in this section, we can conclude the offspring hybrids of IMC 67 (H1 and H2) presumably have a particularly distinctive composition of *fine cocoa* than the other hybrids, especially the H1 hybrid. This may be associated with its ancestor ICS1, which is a hybridization of cocoa with a fine aroma (Criollo and Trinitario) (Castro-Alayo, Idrogo-Vásquez, Siche, & Cardenas-Toro, 2019; Kongor et al., 2016; Scollo et al., 2020). The performance of classification models was similar to reported to discriminate two-classes of Ecuadorian cocoa bean genotypes using Raman spectroscopy (accuracy = 91.8%) (Vargas Jentzsch et al., 2016) and computer vision (until 98%) (Jimenez et al., 2018), even reaching an error of 0% for some pairs of hybrids that was not previously reported.

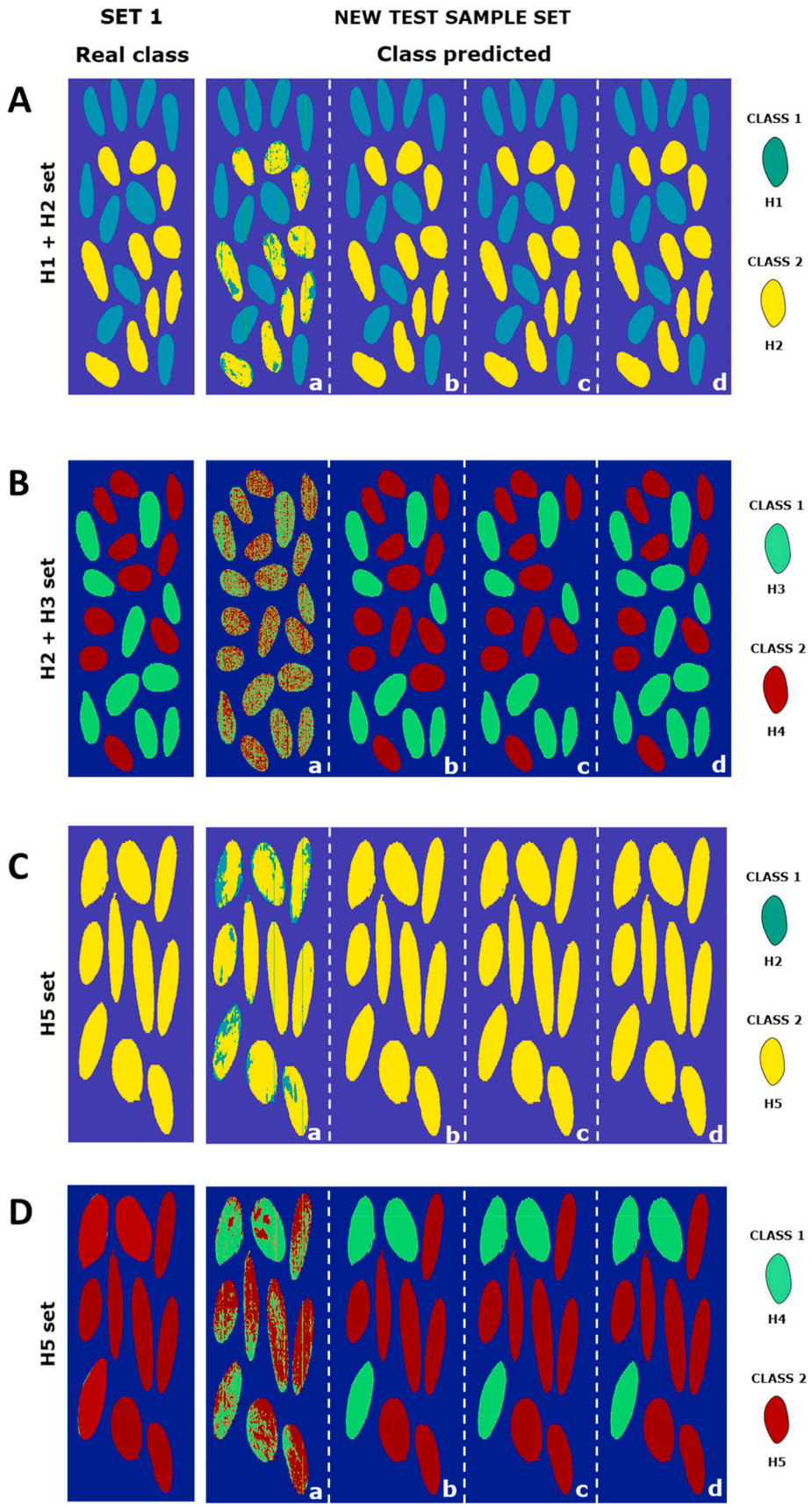

Fig. 4. PLS-DA analysis for two classes of hybrids of cocoa beans. A) Model H1 *vs* H2; B) Model H3 *vs* H4; C) Model H1 *vs* H5; D) Model H3 *vs* H5. PLS-DA maps for the most probable class assigned to new test sample set were produced using: a) PLS-DA model; b) model applying majority vote; c) filtered model by deleted samples no possible to classify (difference between two classes with more probability < 65 pixels); d) PLS-DA model applied to pixels with mean spectra value.

3.3.3. Five-classes model

In this step, one PSL-DA and one SVM model for the five cocoa bean hybrids. Table 2 shows the performance evaluated in terms of sensitivity, specificity, and error for PLS-DA and SVM models. Both PLS-DA and SVM showed good performance in discriminating between the five types of cocoa hybrids (Fig. 1). The lowest sensitivity and the highest error in cross-validation and prediction set was recorded for the hybrids H2 and H3 for both PLS-DA and SVM models. However, the prediction error for the test set was lower (18.1% for H2 and 23.1% for H3) for SVM models compared to PLS-DA models (23.1% for H2 and 34.4% for H3). In contrast, both models showed a good performance (lower error in test set) for the H1 hybrid, suggesting that this cocoa hybrid could have a rather particular composition compared to the other hybrids. Therefore, its classification is more reliable, which is clearly shown in the low prediction error in test set (4.4% for PLS-DA and 3.8% for SVM).

In Fig. 5 it can be seen how many of the hybrids presented pixels of all classes. More clearly, PLS-DA maps (approach *a*) show how for the same hybrid, pixels of all classes can exist. This is mainly associated with shared genetic information, for example because the same mother is shared (1/2 genetic information, especially for H3 and H4 hybrids), especially "mother" genetic that is responsible for shell composition. Meanwhile, it is likely that not all compounds present in the shell have been transformed after the fermentation and drying process, similar pixels in each cocoa bean are therefore associated with fiber, alkaloids (e.g. theobromine (Biehl & Ziegleder, 2003)) and some proteins (e.g. albumins (Dodo, Fritz, & Furtek, 1992)) that do not degrade or transformed completely during fermentation.

Fig. 6 shows the five-classes PLS-DA map created using a new data set (SET 2), which had samples with known spatial position and the same number of samples for each hybrid (10). For new data set (SET 2), Fig. 6 (approach *a*) reflects the genetic complexity of hybrids, showing how some cocoa beans can be assigned pixels of 2 or more classes. However, as in two-class models, the H1 (Class 1) hybrid clearly differs from the other hybrids. Using approaches *b* and *c*, H1 correct classification reached 90%, but it reached 100% when the average spectrum value is assigned to each pixel (Fig. 6, d). This further reinforces the theory that hybrids descended from IMC 67 are compositionally particular. While for the H3 samples, the correct classification was 80% in all cases. Also, Fig. 6c shows that two cocoa beans (1 from H3 and 1 H4) were delighted because there was no significant number of pixels assigned to a single class (difference between number of pixels in class H3 and class H4 was <65 pixels).

For its part, H4 achieved a 100% correct classification using any of the proposed approaches. The greatest difficulty for the five-class model was to identify the H5 and H3 hybrids, which were confused with the H4 hybrid (correct classification 40–60%), using any approach. This behavior was also previously observed in the two-class models. This could indicate that these hybrids are likely to be compositionally very similar, so they could have developed the same flavor compounds in cocoa.

Finally, using five-classes models we reached a correct prediction was 60–100% for the cocoa beans set that contained a single hybrid and between 40 and 100% for cocoa beans set containing all hybrids. Previously, it was reported that using SSRs markets, the classification error rate of cocoa germplasm was 15–44% (Motilal & Butler, 2003). We can therefore say that our five-classes model, in addition to being non-destructive and chemical-free and requiring a minimum sample preparation, is reliable for the discrimination of cocoa hybrids.

4. Conclusion

In this paper we describe, for the first time, the development and robust validation of a method based on hyperspectral images for identifying and classifying cocoa bean hybrids has been proposed. The results indicated that comparable results are obtained for PLS-DA and SVM for the two-class models, whereas for the model that included the five classes, the SVM models showed a significant improvement reducing prediction error. The H1 hybrid was the most distinguishable in all cases. A second validation using a new data set (SET 1 and SET 2) was performed for pixel-to-pixel prediction of hybrid classes. The maps show the reliability of the two-class models for classifying all the hybrids correctly (70–100% CCR), while prediction from the image using the five-class model leads to 100% correct classification of the H1, H2 and H4 hybrids. Future studies should investigate the composition of these hybrids as well as that of their ancestors. In addition, the feasibility of the method in other hybrids and with a larger number of samples should be investigated.

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.

CRediT authorship contribution statement

J.P. Cruz-Tirado: Software, Validation, Formal analysis, Investigation, Writing - original draft, Visualization. **Juan Antonio Fernández Pierna:** Software, Methodology, Validation, Writing - original draft, Writing - review & editing. **Hervé Rogez:** Conceptualization, Data curation, Writing - review & editing, Resources, Supervision, Project administration, Funding acquisition. **Douglas Fernandes Barbin:** Writing - review & editing, Resources, Supervision, Project administration, Funding acquisition. **Vincent Baeten:** Conceptualization, Data curation, Writing - review & editing, Resources, Supervision, Project administration, Funding acquisition.

Table 2
Performance of the five-classes PLS-DA and SVM classification model for the hybrids of cocoa beans obtained by HSI in the spectral region of 1369–2054 nm.

Hybrid	Model	Parameter[a]	Sensitivity		Specificity		Error	
			CV	Pred	CV	Pred	CV	Pred
H1	PLS-DA	12	0.963	1.000	0.944	0.912	0.047	0.044
H2			0.938	0.600	0.897	0.938	0.083	0.231
H3			0.787	0.500	0.806	0.813	0.203	0.344
H4			0.925	0.850	0.881	0.813	0.097	0.169
H5			0.938	0.950	0.956	0.925	0.048	0.053
H1	SVM	(10; 0,01)	0.963	1.000	0.994	0.925	0.022	0.038
H2			0.950	0.650	0.984	0.988	0.033	0.181
H3			0.875	0.600	0.975	0.938	0.075	0.231
H4			0.975	0.900	0.978	0.938	0.023	0.081
H5			0.925	0.900	0.991	0.975	0.042	0.063

PLS-DA: partial least square discriminant analysis; SVM: support vector machine discriminant analysis; CV: cross-validation; Pred: prediction.
[a] Parameter for PLS-DA model is the optimal number of LVs; for SVM model is different penalty parameters: cost (c) and kernel function parameters gamma (g).

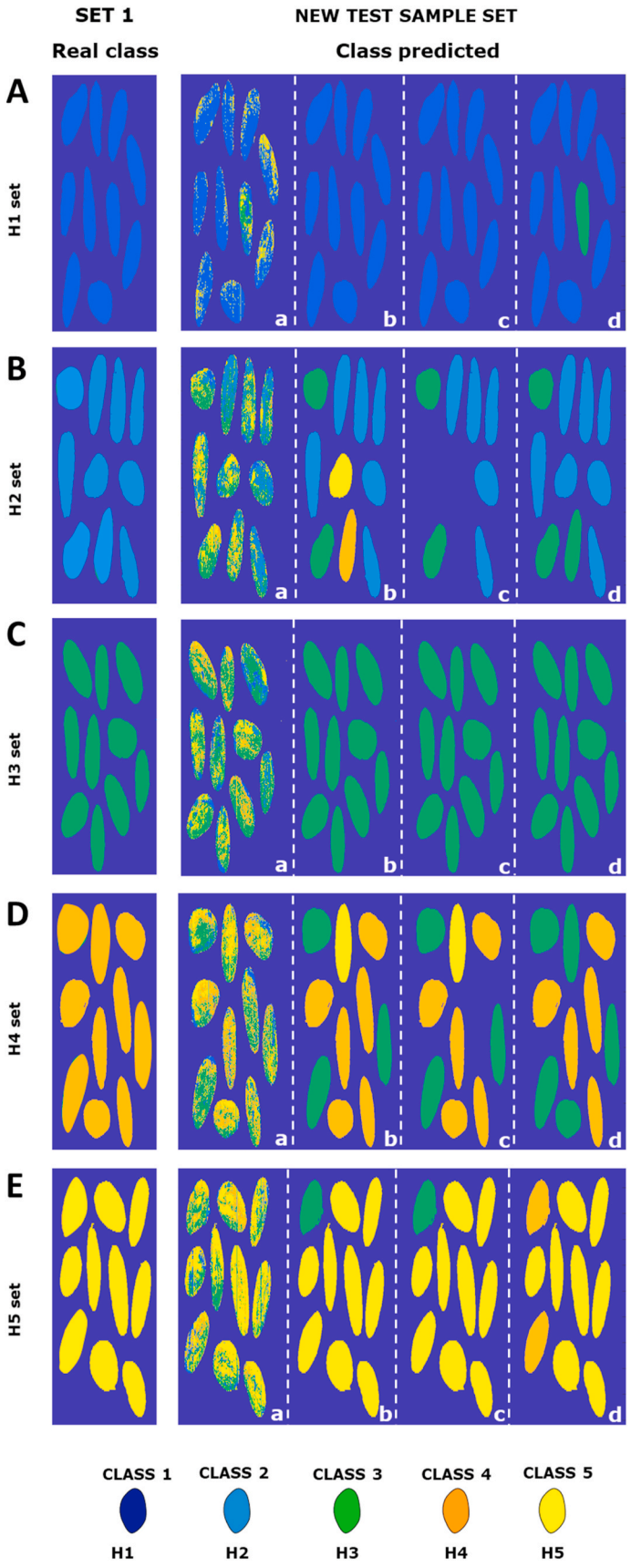

Fig. 5. PLS-DA analysis for five classes of hybrids of cocoa beans. A) H1; B) H2; C) H3; D) H4; E) H5. PLS-DA maps for the most probable class assigned to new test sample set were produced using: a) PLS-DA model; b) model applying majority vote; c) filtered model by deleted samples no possible to classify (difference between two classes with more probability < 65 pixels); d) PLS-DA model applied to pixels with mean spectra value.

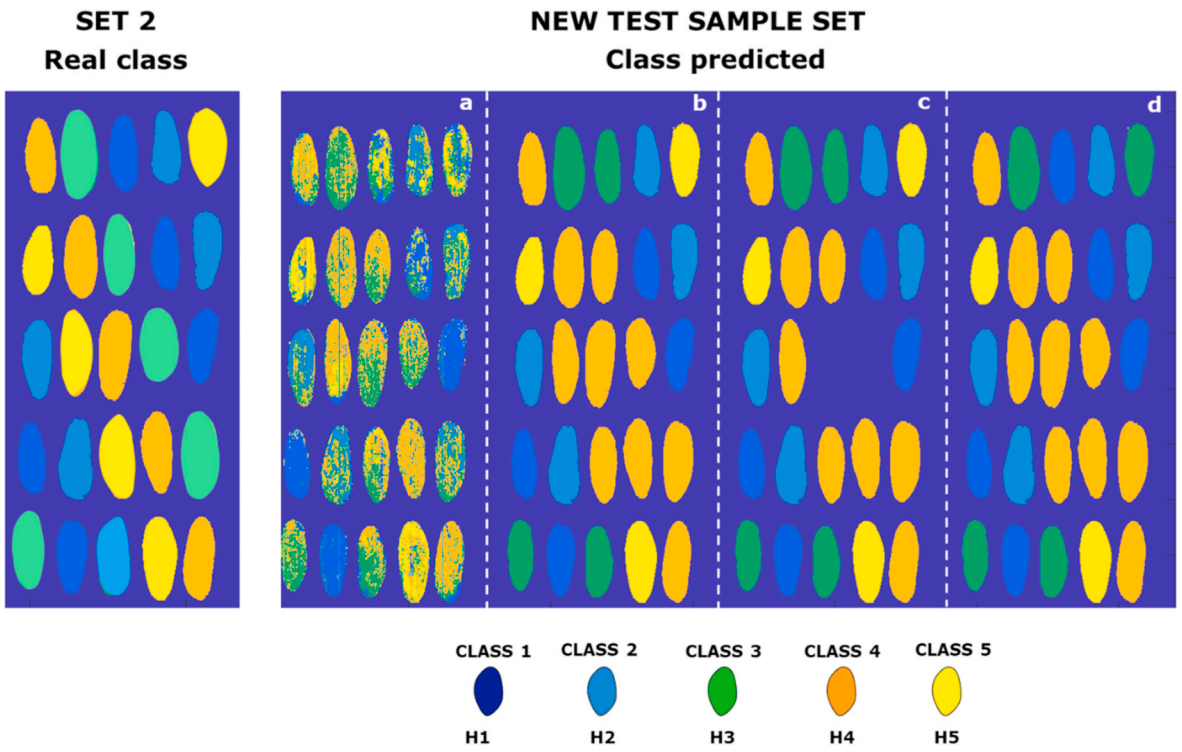

Fig. 6. PLS-DA maps for the most probable class assigned to new test sample set. (Set 2) with hybrids with known spatial position. PLS-DA maps for the most probable class assigned to new test sample set were produced using: a) PLS-DA model; b) model applying majority vote; c) filtered model by deleted samples no possible to classify (difference between two classes with more probability < 65 pixels); d) PLS-DA model applied to pixels with mean spectra value.

Acknowledgements

This study was financed in part by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001; and São Paulo Research Foundation (FAPESP) (project number 2019/04833-3, 2018/02500-4, 2015/24351-2). This study was also financed by the BELCOBRA 'Optimization and implementation of analytical methods for the traceability and authenticity of cocoa and chocolate' (Brazil-Belgium cooperation CAPES-WBI/2017-2019/361394). Authors are grateful to CEPLAC (Medicilandia, PA, Brazil) for the mature fruits, to Teixeira L.E.O. and Freitas A.L. (UFPA) for their technical assistance for fermentation and drying process and to N. Kayoka (CRA-W) for hi technical assistance in NIR-HSI measurements. J. P. Cruz-Tirado acknowledges scholarship funding from FAPESP/BEPE, grant n° 2019/04833- 3.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodcont.2020.107445.

References

Álvarez, C., Pérez, E., Cros, E., Lares, M., Assemat, S., Boulanger, R., et al. (2012). The use of near infrared spectroscopy to determine the fat, caffeine, theobromine and (−)-Epicatechin contents in unfermented and sun-dried beans of Criollo cocoa. *Journal of Near Infrared Spectroscopy, 20*(2), 307–315. https://doi.org/10.1255/jnirs.990.

Baeten, V., Pierna, J. A. F., Vermeulen, P., & Dardenne, P. (2010). NIR hyperspectral imaging methods for quality and safety control of food and feed products: Contributions to four European projects. *NIR News, 21*(6), 10–13. https://doi.org/10.1255/nirn.1200.

Barbin, D. F., Maciel, L. F., Bazoni, C. H. V., Ribeiro, M. da S., Carvalho, R. D. S., Bispo, E. da S., … Hirooka, E. Y. (2018). Classification and compositional characterization of different varieties of cocoa beans by near infrared spectroscopy and multivariate statistical analyses. *Journal of Food Science & Technology, 55*(7), 2457–2466. https://doi.org/10.1007/s13197-018-3163-5.

Biehl, B., & Ziegleder, G. (2003). In *COCOA | Chemistry of processing (B. B. T.-E. of F. S. and N.)* (Second E. Caballero). https://doi.org/10.1016/B0-12-227055-X/00261-3.

Caporaso, N., Whitworth, M. B., Fowler, M. S., & Fisk, I. D. (2018). Hyperspectral imaging for non-destructive prediction of fermentation index, polyphenol content and antioxidant activity in single cocoa beans. *Food Chemistry, 258*, 343–351. https://doi.org/10.1016/j.foodchem.2018.03.039.

Castro-Alayo, E. M., Idrogo-Vásquez, G., Siche, R., & Cardenas-Toro, F. P. (2019). Formation of aromatic compounds precursors during fermentation of Criollo and Forastero cocoa. *Heliyon, 5*(1), Article e01157. https://doi.org/10.1016/j.heliyon.2019.e01157.

Dale, L. M., Thewis, A., Boudry, C., Rotar, I., Dardenne, P., Baeten, V., et al. (2013). Hyperspectral imaging applications in agriculture and agro-food product quality and safety control: A review. *Applied Spectroscopy Reviews, 48*(2), 142–159. https://doi.org/10.1080/05704928.2012.705800.

Dinarti, D., Susilo, A. W., Meinhardt, L. W., Ji, K., Motilal, L. A., Mischke, S., et al. (2015). Genetic diversity and parentage in farmer selections of cacao from Southern Sulawesi, Indonesia revealed by microsatellite markers. *Breeding Science, 65*(5), 438–446. https://doi.org/10.1270/jsbbs.65.438.

Diomande, D., Antheaume, I., Leroux, M., Lalande, J., Balayssac, S., Remaud, G. S., et al. (2015). Multi-element, multi-compound isotope profiling as a means to distinguish the geographical and varietal origin of fermented cocoa (Theobroma cacao L.) beans. *Food Chemistry, 188*, 576–582. https://doi.org/10.1016/j.foodchem.2015.05.040.

Dodo, H. W., Fritz, P. J., & Furtek, D. B. (1992). *A cocoa 21 kilodalton seed protein has trypsin inhibitory activity*. Cafe Cacao The (France).

Eylenbosch, D., Bodson, B., Baeten, V., & Fernández Pierna, J. A. (2018). NIR hyperspectral imaging spectroscopy and chemometrics for the discrimination of roots and crop residues extracted from soil samples. *Journal of Chemometrics, 32*(1), Article e2982. https://doi.org/10.1002/cem.2982.

Fang, W., Meinhardt, L. W., Mischke, S., Bellato, C. M., Motilal, L., & Zhang, D. (2014). Accurate determination of genetic identity for a single cacao bean, using molecular markers with a nanofluidic system, ensures cocoa authentication. *Journal of Agricultural and Food Chemistry, 62*(2), 481–487. https://doi.org/10.1021/jf404402v.

Fernández Pierna, J. A., Baeten, V., & Dardenne, P. (2006). Screening of compound feeds using NIR hyperspectral data. *Chemometrics and Intelligent Laboratory Systems, 84*(1), 114–118. https://doi.org/10.1016/j.chemolab.2006.03.012.

Fernández Pierna, J. A., Vermeulen, P., Amand, O., Tossens, A., Dardenne, P., & Baeten, V. (2012). NIR hyperspectral imaging spectroscopy and chemometrics for the detection of undesirable substances in food and feed. *Chemometrics and Intelligent Laboratory Systems, 117*, 233–239. https://doi.org/10.1016/j.chemolab.2012.02.004.

Guo, D., Zhu, Q., Huang, M., Guo, Y., & Qin, J. (2017). Model updating for the classification of different varieties of maize seeds from different years by hyperspectral imaging coupled with a pre-labeling method. *Computers and Electronics in Agriculture, 142*, 1–8. https://doi.org/10.1016/j.compag.2017.08.015.

Herrmann, L., Felbinger, C., Haase, I., Rudolph, B., Biermann, B., & Fischer, M. (2015). Food fingerprinting: Characterization of the ecuadorean type CCN-51 of theobroma cacao L. Using microsatellite markers. *Journal of Agricultural and Food Chemistry, 63* (18), 4539–4544. https://doi.org/10.1021/acs.jafc.5b01462.

Jimenez, J. C., Amores, F. M., Solórzano, E. G., Rodríguez, G. A., La Mantia, A., Blasi, P., et al. (2018). Differentiation of Ecuadorian National and CCN-51 cocoa beans and their mixtures by computer vision. *Journal of the Science of Food and Agriculture, 98* (7), 2824–2829. https://doi.org/10.1002/jsfa.8790.

Kongor, J. E., Hinneh, M., de Walle, D. Van, Afoakwa, E. O., Boeckx, P., & Dewettinck, K. (2016). Factors influencing quality variation in cocoa (Theobroma cacao) bean flavour profile — a review. *Food Research International, 82*, 44–52. https://doi.org/10.1016/j.foodres.2016.01.012.

Kumari, N., Grimbs, A., D'Souza, R. N., Verma, S. K., Corno, M., Kuhnert, N., et al. (2018). Origin and varietal based proteomic and peptidomic fingerprinting of Theobroma cacao in non-fermented and fermented cocoa beans. *Food Research International, 111*, 137–147. https://doi.org/10.1016/j.foodres.2018.05.010.

Liu, X., Feng, X., Liu, F., & He, Y. (2017). Identification of hybrid rice strain based on near-infrared hyperspectral imaging technology. *Transactions of the Chinese Society of Agricultural Engineering, 33*(22), 189–194.

Liu, Y., Wu, T., Yang, J., Tan, K., & Wang, S. (2019). Hyperspectral band selection for soybean classification based on information measure in FRS theory. *Biosystems Engineering, 178*, 219–232. https://doi.org/10.1016/j.biosystemseng.2018.12.002.

Mandrile, L., Barbosa-Pereira, L., Sorensen, K. M., Giovannozzi, A. M., Zeppa, G., Engelsen, S. B., et al. (2019). Authentication of cocoa bean shells by near- and mid-infrared spectroscopy and inductively coupled plasma-optical emission spectroscopy. *Food Chemistry, 292*, 47–57. https://doi.org/10.1016/j.foodchem.2019.04.008.

Ma, H., Wang, J., Chen, Y., Cheng, J., & Lai, Z. (2017). Rapid authentication of starch adulterations in ultrafine granular powder of Shanyao by near-infrared spectroscopy coupled with chemometric methods. *Food Chemistry, 215*, 108–115. https://doi.org/10.1016/j.foodchem.2016.07.156.

Mite-Baidal, K., Solís-Avilés, E., Martínez-Carriel, T., Marcillo-Plaza, A., Cruz-Ibarra, E., & Baque-Bustamante, W. (2019). In R. Valencia-García, G. Alcaraz-Mármol, J. del Cioppo-Morstadt, N. Vera-Lucio, & M. Bucaram-Leverone (Eds.), *Analysis of computer vision algorithms to determine the quality of fermented cocoa (theobroma cacao): Systematic literature review BT - ICT for agriculture and environment.* Cham: Springer International Publishing.

da V Moreira, I. M., de F Vilela, L., Santos, C., Lima, N., & Schwan, R. F. (2018). Volatile compounds and protein profiles analyses of fermented cocoa beans and chocolates from different hybrids cultivated in Brazil. *Food Research International, 109*, 196–203. https://doi.org/10.1016/j.foodres.2018.04.012.

Motamayor, J. C., Lachenaud, P., da Silva e Mota, J. W., Loor, R., Kuhn, D. N., Brown, J. S., et al. (2008). Geographic and genetic population differentiation of the amazonian chocolate tree (theobroma cacao L). *PloS One, 3*(10), Article e3311. https://doi.org/10.1371/journal.pone.0003311. Retrieved from.

Motilal, L., & Butler, D. (2003). Verification of identities in global cacao germplasm collections. *Genetic Resources and Crop Evolution, 50*(8), 799–807. https://doi.org/10.1023/A:1025950902827.

Nie, P., Zhang, J., Feng, X., Yu, C., & He, Y. (2019). Classification of hybrid seeds using near-infrared hyperspectral imaging technology combined with deep learning. *Sensors and Actuators B: Chemical, 296*, 126630. https://doi.org/10.1016/j.snb.2019.126630.

Okiyama, D. C. G., Navarro, S. L. B., & Rodrigues, C. E. C. (2017). Cocoa shell and its compounds: Applications in the food industry. *Trends in Food Science & Technology, 63*, 103–112. https://doi.org/10.1016/j.tifs.2017.03.007.

Osborne, B. G., Fearn, T., Hindle, P. H., & Osborne, B. G. (1993). *Practical NIR spectroscopy with applications in food and beverage analysis* (Vol. 2). Longman Scientific & Technical Harlow.

Pierna, J. A. F., Baeten, V., Renier, A. M., Cogdill, R. P., & Dardenne, P. (2004). Combination of support vector machines (SVM) and near-infrared (NIR) imaging spectroscopy for the detection of meat and bone meal (MBM) in compound feeds. *Journal of Chemometrics, 18*(7-8), 341–349. https://doi.org/10.1002/cem.877.

Pierna, J. A. F., Lecler, B., Conzen, J. P., Niemoeller, A., Baeten, V., & Dardenne, P. (2011). Comparison of various chemometric approaches for large near infrared spectroscopic data of feed and feed products. *Analytica Chimica Acta, 705*(1), 30–34. https://doi.org/10.1016/j.aca.2011.03.023.

Quelal-Vásconez, M. A., Lerma-García, M. J., Pérez-Esteve, É., Arnau-Bonachera, A., Barat, J. M., & Talens, P. (2019). Fast detection of cocoa shell in cocoa powders by near infrared spectroscopy and multivariate analysis. *Food Control, 99*, 68–72. https://doi.org/10.1016/j.foodcont.2018.12.028.

Quelal-Vásconez, M. A., Pérez-Esteve, É., Arnau-Bonachera, A., Barat, J. M., & Talens, P. (2018). Rapid fraud detection of cocoa powder with carob flour using near infrared spectroscopy. *Food Control, 92*, 183–189. https://doi.org/10.1016/j.foodcont.2018.05.001.

Scollo, E., Neville, D. C. A., Oruna-Concha, M. J., Trotin, M., & Cramer, R. (2020). UHPLC–MS/MS analysis of cocoa bean proteomes from four different genotypes. *Food Chemistry, 303*, 125244. https://doi.org/10.1016/j.foodchem.2019.125244.

Sendin, K., Manley, M., Baeten, V., Fernández Pierna, J. A., & Williams, P. J. (2019). Near infrared hyperspectral imaging for white maize classification according to grading regulations. *Food Analytical Methods, 12*(7), 1612–1624. https://doi.org/10.1007/s12161-019-01464-0.

Su, W.-H., Bakalis, S., & Sun, D.-W. (2019). Chemometrics in tandem with near infrared (NIR) hyperspectral imaging and Fourier transform mid infrared (FT-MIR) microspectroscopy for variety identification and cooking loss determination of sweet potato. *Biosystems Engineering, 180*, 70–86. https://doi.org/10.1016/j.biosystemseng.2019.01.005.

Sunoj, S., Igathinathane, C., & Visvanathan, R. (2016). Nondestructive determination of cocoa bean quality using FT-NIR spectroscopy. *Computers and Electronics in Agriculture, 124*, 234–242. https://doi.org/10.1016/j.compag.2016.04.012.

Teye, E., Huang, X., Dai, H., & Chen, Q. (2013). Rapid differentiation of Ghana cocoa beans by FT-NIR spectroscopy coupled with multivariate classification. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 114*, 183–189. https://doi.org/10.1016/j.saa.2013.05.063.

Teye, E., Huang, X., Sam-Amoah, L. K., Takrama, J., Boison, D., Botchway, F., et al. (2015). Estimating cocoa bean parameters by FT-NIRS and chemometrics analysis. *Food Chemistry, 176*, 403–410. https://doi.org/10.1016/j.foodchem.2014.12.042.

The International Cocoa Organization. (2018). Cocoa producing and cocoa consuming countries. Retrieved February 1, 2018, from https://www.icco.

Trognitz, B., Cros, E., Assemat, S., Davrieux, F., Forestier-Chiron, N., Ayestas, E., … Hermann, M. (2013). Diversity of cacao trees in Waslala, Nicaragua: Associations between genotype spectra, product quality and yield potential. *PloS One, 8*(1), Article e54079. https://doi.org/10.1371/journal.pone.0054079. Retrieved from.

UNCTAD/WTO, I. T. C. (2001). *Cocoa: A guide to trade practices*. United Nations Publications.

Vargas Jentzsch, P., Ciobotă, V., Salinas, W., Kampe, B., Aponte, P. M., Rösch, P., … Ramos, L. A. (2016). Distinction of Ecuadorian varieties of fermented cocoa beans using Raman spectroscopy. *Food Chemistry, 211*, 274–280. https://doi.org/10.1016/j.foodchem.2016.05.017.

Veselá, A., Barros, A. S., Synytsya, A., Delgadillo, I., Čopíková, J., & Coimbra, M. A. (2007). Infrared spectroscopy and outer product analysis for quantification of fat, nitrogen, and moisture of cocoa powder. *Analytica Chimica Acta, 601*(1), 77–86. https://doi.org/10.1016/j.aca.2007.08.039.

Vidal, M., & Amigo, J. M. (2012). Pre-processing of hyperspectral images. Essential steps before image analysis. *Chemometrics and Intelligent Laboratory Systems, 117*, 138–148.

Wang, J., Zhang, X., Sun, S., Sun, X., Li, Q., & Zhang, Z. (2018). Online determination of quality parameters of dried soybean protein–lipid films (Fuzhu) by NIR spectroscopy combined with chemometrics. *Journal of Food Measurement and Characterization, 12* (3), 1473–1484. https://doi.org/10.1007/s11694-018-9762-z.