Classification of Cocoa Beans Based on their Level of Fermentation using Spectral Information

Clasificación de la fermentación del grano de cacao usando información espectral

Karen Sánchez[1];
Jorge Bacca[2];
Laura Arévalo-Sánchez[3];
Henry Arguello[4];
Sergio Castillo[5];

[1] Universidad Industrial de Santander, Santander-Colombia, karen.sanchez2@correo.uis.edu.co
[2] Universidad Industrial de Santander, Santander-Colombia, jorge.bacca1@correo.uis.edu.co
[3] Universidad Industrial de Santander, Santander-Colombia, laura.arevalo4@correo.uis.edu.co
[4] Universidad Industrial de Santander, Santander-Colombia, henarfu@uis.edu.co
[5] Universidad Industrial de Santander, Santander-Colombia, scastill@uis.edu.co

Abstract

Cocoa beans are the most important raw material for the chocolate industry and an essential product for the economy of tropical countries such as Colombia. Their price mainly depends on their quality, which is determined by various aspects, such as good agricultural practices, their harvest point, and level of fermentation. The entities that regulate the international marketing of cocoa beans have been encouraging the development of new classification methods that, compared to current techniques, could save time, reduce waste, and increase the number of evaluated beans. In particular, hyperspectral images are a novel tool for food quality control. However, studies that have examined some quality parameters of cocoa using spectroscopy also involve the chemical evaluation of cocoa powder and liquor and the interior of the beans, which implies an invasive analysis, longer times, and waste generation. Therefore, in this paper, we assess the quality of cocoa beans based on their level of fermentation using a noninvasive system to obtain hyperspectral information, as well as fast image processing and spectral classification techniques. We obtained hyperspectral images of 90 cocoa beans in the range between 350 and 950 nm in an optical laboratory. In addition, each cocoa bean was classified according to its fermentation level: slightly fermented (SF), correctly fermented (CF), and highly fermented (HF). We compared this classification with that carried out by experts from the Colombia National Federation of Cocoa Growers and reported in the Colombian technical standard No. 1252. The results show that the level of fermentation of dried cocoa beans can be estimated using noninvasive hyperspectral image acquisition and processing techniques.

Keywords

Cocoa beans, level of fermentation, hyperspectral images, spectral classification, superpixel.

Resumen

Los granos de cacao son la materia prima de la industria del chocolate y un producto esencial para la economía de países tropicales como Colombia. El precio del grano depende principalmente de su calidad, determinada por diversos aspectos, tales como, buenas prácticas agrícolas, el punto de cosecha del fruto y la fermentación. Entidades que regulan el comercio internacional de granos de cacao promueven la creación de nuevas metodologías de clasificación que, en comparación con los métodos actuales, disminuyan el tiempo y los residuos y aumenten la cobertura de granos evaluados. Las imágenes hiperespectrales se han venido posicionando como una herramienta novedosa para el control de calidad de alimentos. Sin embargo, trabajos que analizan ciertos parámetros de la calidad del cacao mediante espectroscopía, también involucran etapas de estudio químico del polvo, el licor y el interior de los granos, lo que implica un análisis invasivo, así como un tiempo extenso y producción de residuos. Por lo tanto, este artículo analiza la calidad de granos de cacao a partir del parámetro estado de fermentación, usando un sistema no-invasivo de captura de información hiperespectral y técnicas rápidas de procesamiento de imágenes y clasificación espectral. Imágenes hiperespectrales de 90 granos de cacao en un rango de 350 a 950 nanómetros fueron adquiridos y se asignó una etiqueta a cada grano de cacao según su nivel de fermentación: poco, correcta y altamente fermentado. Esta clasificación se comparó con la realizada por profesionales de la federación nacional de cacaoteros a través de la norma técnica colombiana número 1252. Los resultados obtenidos muestran que es posible estimar el nivel de fermentación de granos secos de cacao usando técnicas no-invasivas de adquisición de y procesamiento de imágenes hiperespectrales.

Palabras clave

Granos de cacao, nivel de fermentación, imágenes hiperespectrales, clasificación espectral superpixel.

1.   INTRODUCTION

Cocoa beans are the dried and fully fermented seeds from the cacao tree (*Theobroma cacao*). Although this tree originated in America's rainforests, it is also grown in the tropical areas of Africa and Asia [1], [2]. Cocoa constitutes a valuable agricultural commodity for more than 40 million people around the world [3]. In addition, the chemical quality attributes of raw cocoa [4] make it highly demanded by the confectionery, aesthetics, and healthcare industries (see Table 1) [5], [6], [7].

Table 1. Chemical composition of Latin American unroasted cocoa bean samples. Source: [4].

Nutritional content of 1 kg of cocoa		
Proximates	Water (g)	59.51
	Ash (g)	40.32
	Total protein (g)	127.91
	Carbohydrate by difference (g)	337.85
	Total dietary fiber (g)	194.74
	Total fat (g)	434.56
Minerals	Calcium, Ca (mg)	1225.55
	Magnesium, Mg (mg)	3075.24
	Potassium, K (mg)	12486.21
	Phosphorus, P (mg)	4231.43
	Iron, Fe (mg)	146.47
	Manganese, Mn (mg)	21.64
	Copper, Cu (mg)	26.10
	Zinc, Zn (mg)	44.65
	Sodium, Na (mg)	261.05
	Selenium, Se (mg)	2.80

A total of 68 % of the world's cocoa beans come from Africa—the largest cocoa producer worldwide—, while only 17 % are produced by Latin American countries (Brazil, Ecuador, Mexico, Peru, Dominican Republic, and Colombia) [8]. In addition, the best and most expensive quality cocoa, known as *premium cocoa* (5 % of the world's cocoa), comes from Latin America [9], [10]. Global cocoa production has increased significantly in recent years, hitting a record of 4.85 million tons in 2019 [11]. However, it has not been enough to meet the world's demand [12]. For this reason, the International Cocoa Organization (ICCO) has suggested Latin American countries to increase their cocoa exports. To that end, less complicated and faster classification processes must be explored [13], [14].

Nowadays, in most cocoa international markets, the methods employed to classify beans consist of chemical, physical, and sensory analyses that take approximately 26 hours [15]. Additionally, this classification is often carried out with samples, that is, 100 grains per ton, to determine the quality of the load. Said analyses require tasters, technical personnel, specialized equipment, and the destruction of the samples [1], [6].

Moreover, Hyperspectral Image (HSI) acquisition and processing techniques have been increasingly used in food quality control [16] [17], [18], [19], [20]. A hyperspectral image can be represented as a three-dimensional data cube, $F \in R^{M \times N \times L}$, where M and N correspond to the spatial dimensions; and L, to the spectral dimension. Each element reflects, absorbs, and emits electromagnetic energy in different magnitudes at each specific wavelength according to its physical and chemical composition [21], [22]. Therefore, two elements with a different composition can be identified or associated through their spectral signatures [23], [24], [25], [26].

Some studies have been conducted in India, Ghana, Peru, and Germany to evaluate the quality parameters of cocoa beans using spectral information [27], [28], [29], [30], [31], [32], [33], [34]. For instance, in [33], the composition and aroma profiles of 26 cocoa beans were assessed using Mass Spectrometry (MS)-fingerprinting and Headspace-Solid Phase Micro-extraction-Gas Chromatography–Mass Spectrometry. As a result, the authors classified the beans into fine flavor cocoa, well-fermented cocoa, and low-quality cocoa. In [27], the fermentation index, pH, and polyphenol content of cocoa beans were calculated. The whole grains were used for the spectral measurements in the near-infrared range, while, for the chemical analysis, they were grounded into a fine powder. According to the results, an accuracy greater than 80 % in the fermentation index and total polyphenols was achieved. In general, these and other studies have followed chemical procedures that, although precise, involve invasive stages and require very specialized personnel; hence, they could not be generalized to all marketable beans.

Recent works have demonstrated that grouping pixels with similar characteristics within an image (called superpixels [35]) before processing HSIs makes it possible to obtain more accurate classification results and reduces the computational cost and time required by supervised classification methods [36], [37], [38], [39].

Specifically, [38] presents a multiscale HSI strategy based on superpixels to classify remote sensing images and obtain results that are up to 3 % more accurate than those provided by classification methods that do not use superpixels. Furthermore, the approach followed in [39] groups the spatial information of a Red, Green, Blue (RGB) image into superpixels and fuses such features with the spectral information of a HSI. The results show that the proposed classification method optimizes the overall accuracy and reduces the computational complexity compared to traditional approaches in which all pixels are used.

Similar results are reported in [36] and [37]. However, the superpixel technique with hyperspectral classification has not yet been used to evaluate, in a noninvasive manner, the classification of cocoa beans.

Therefore, in this study, we propose a noninvasive approach to classify cocoa beans into three categories based on their fermentation level using their hyperspectral images. In particular, the proposed classification method includes the following stages: sample preparation, acquisition of cocoa beans' spatio–spectral information in the visible and near infrared (350–950 nm) ranges, background subtraction, feature extraction with superpixels, and hyperspectral classification.

The rest of this paper is structured as follows. Section 2 summarizes the stages of the proposed method. Section 3 provides an analysis of the data and presents the results. Section 4 draws some conclusions and outlines some future lines of work.

2.   MATERIALS AND METHODS

This section details the proposed classification methodology and the data acquisition process.

2.1   Cocoa Selection

The cocoa beans were harvested in a farm located in the town of Rionegro, Santander, Colombia (7° 15' 51" N, 73° 08' 58" W). These beans were extracted from cocoa pods within the same hectare and then fermented by a local farmer. Experts from the National Cocoa Federation in Colombia selected 30 samples for each of the three fermentation levels: (i) slightly fermented (LF), (ii) correctly fermented (CF), and (iii) highly fermented (HF). In total, 90 cocoa beans were selected.

Afterwards, the beans were evaluated in an optical laboratory and their spectral data were captured using the experimental setup described below. The environmental conditions in the laboratory included a temperature of 23 °C and a humidity of 60 %.

2.2   Experimental Setup

An optical assembly was built in our laboratory to capture hyperspectral images of the cocoa beans. The built testbed is shown in Figure 1. The scene (cocoa beans) was illuminated with a tunable light source (Oriel Instruments, TLS-300 XR) that decomposes the illumination from a halogen light source in its corresponding monochromatic wavelengths, with steps of two nanometers within the spectral range between 350 and 950 nm. Such monochromatic light is propagated through a bifurcated optical fiber (Illumination Technologies, 9145HT dual 6" light line) towards two lamps that illuminate the scene.
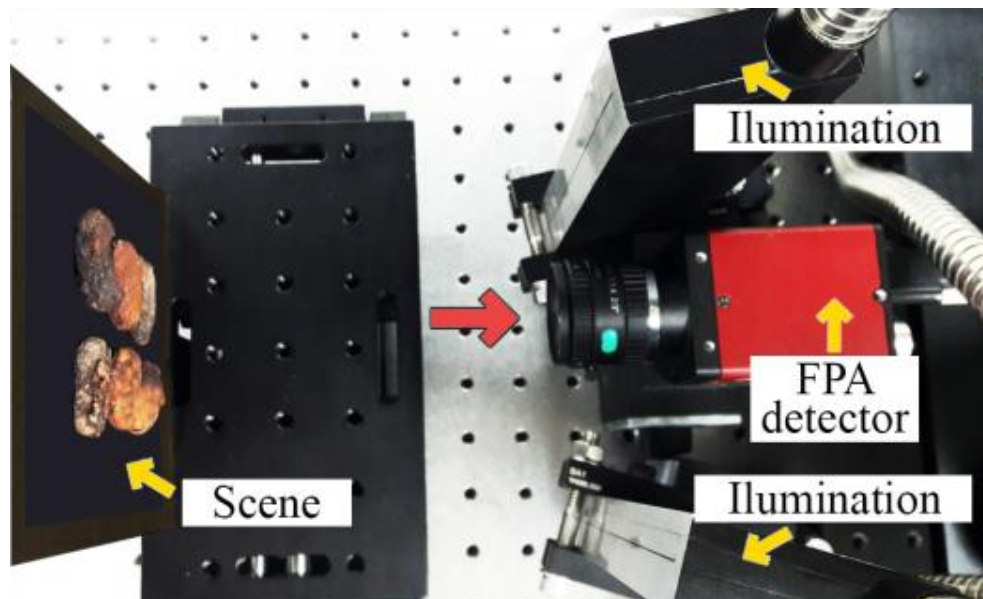

Figure 1. Top view of the testbed we used to obtain the spectral images. Source: Authors' own work.

A monochromatic sensor (AVT Stingray F-080B) captures the intensity of the light reflected by the cocoa beans ($G_{ref}(\lambda)$). Each obtained hyperspectral image exhibits 1032 x 776 pixels of spatial resolution (M x N) and 301 spectral bands (L). Also, a white scene ($G_{inc}(\lambda)$) was

acquired to calibrate the hyperspectral information as $F(\lambda) = \frac{G_{ref}(\lambda)}{G_{inc}(\lambda)}$. Six beans were organized for each scene, considering the focus and field of vision of the setup. In total, five data cubes were obtained for each of the three fermentation categories, which resulted in 15 spectral images, $F \in R^{M \times N \times L}$, each with six beans of the same category (i.e., a total of 90 cocoa beans captured).

Figure 2 presents six of the 301 spectral bands of one random bean, which were acquired following the proposed setup. In addition, it spectral band includes the spectral range (VIS by visible or NIR by Near-infrared) and wavelength of each band in nanometers. We freely published the spectral images of the 90 cocoa beans on IEEE DataPort [40]. Figure 3 plots the mean spectral responses of the three fermentation categories. After the spectral information of the cocoa beans was obtained, the spectral images were processed.
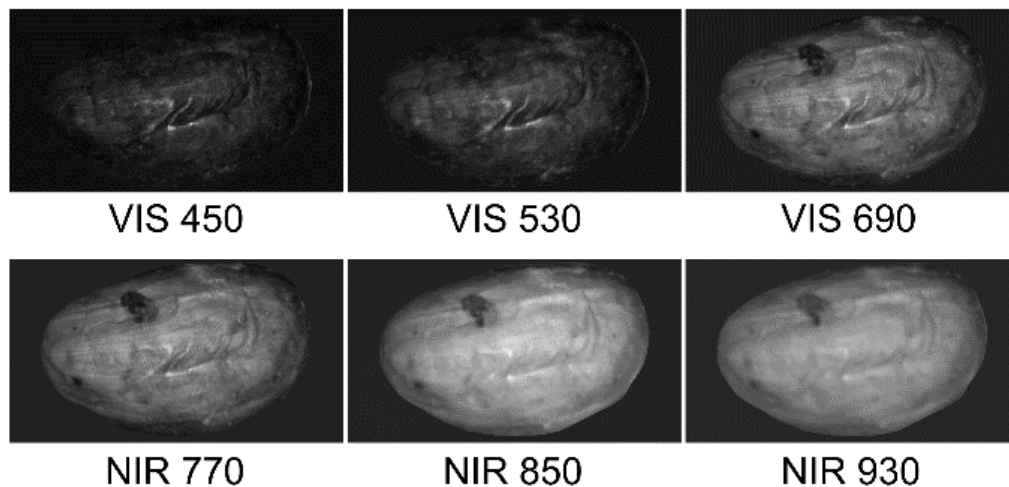

Figure 2. Examples of some spectral bands captured by the optical assembly for each bean
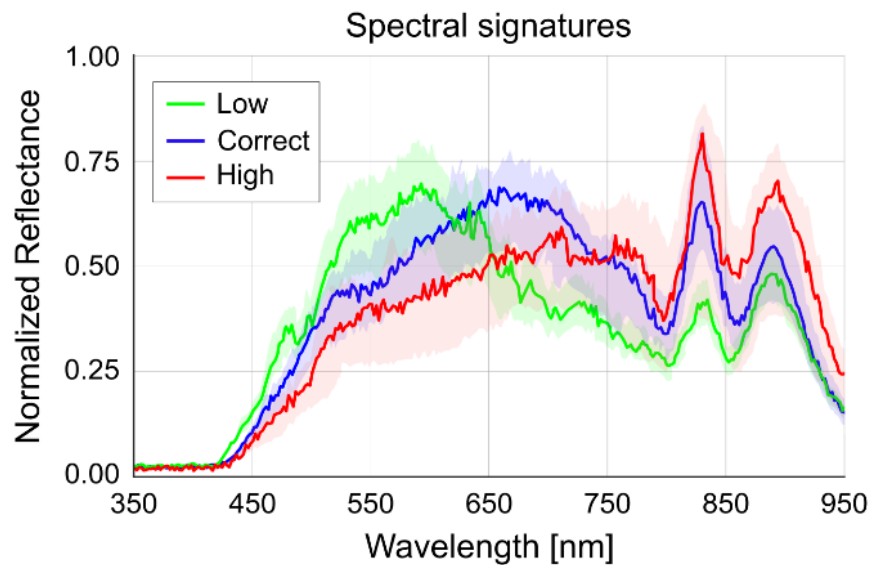Source: Authors' own work.

Figure 3. Mean spectral signatures of the 90 cocoa beans classified into three categories and their standard deviations: 30 slightly fermented beans (in green), 30 correctly fermented beans (in blue), and 30 highly fermented beans (in red). Source: Authors' own work.

2.3   Background Subtraction

The first stage in HSI processing consists in identifying and extracting the pixels containing the information of interest, that is, the beans. Since the background of the images is black, a binary mask (D) that identifies the position of the beans was created using thresholding.

Additionally, morphological operations were employed. The values of the pixels in the binary image, D, are adjusted based on the value of other pixels in its neighborhood, such that the binary result ($\underline{D} \in R^{M \times N}$) is a given closed figure (**B**). The closing operator consists of a dilation operator and an erosion operator. The dilation operator, on the one hand, adds pixels to the boundaries of the objects in the image and is mathematically expressed as (1).

$$(\mathbf{D} \oplus \mathbf{B})_{i,j} = max\{\mathbf{D}_{(i-x),(j-y)} | \, x, y \in \mathbf{B}\} \tag{1}$$

The erosion operator, on the other hand, removes the pixels in the boundaries of the objects and is expressed as (2).

$$(\mathbf{D} \ominus \mathbf{B})_{i,j} = min\{\mathbf{D}_{(i-x),(j-y)} | \, x, y \in \mathbf{B}\} \tag{2}$$

In this study, following the shape of the beans, **B** is chosen as an oval shape of 200 x 300 pixels. Then, the result of the closing operation is given by (3).

$$\bar{\mathbf{D}} = (\mathbf{D} \oplus \mathbf{B}) \ominus \mathbf{B} \tag{3}$$

Finally, the binarized image, $\underline{D}$, is multiplied with the hyperspectral image, $F \in R^{M \times N \times L}$, of each sample to obtain the reflectance values of the cocoa beans as follows (4):

$$\widetilde{\mathcal{F}} = \bar{\mathbf{D}} \circ \mathcal{F}, \tag{4}$$

where $\circ$ is the Hadamard product; and $\tilde{F}$, the HSI without background information. Figure 4 shows a HSI, F; its binary mask, $\underline{D}$, which was obtained with (3); and the respective dataset, $\tilde{F}$, without background, which was calculated using (4).
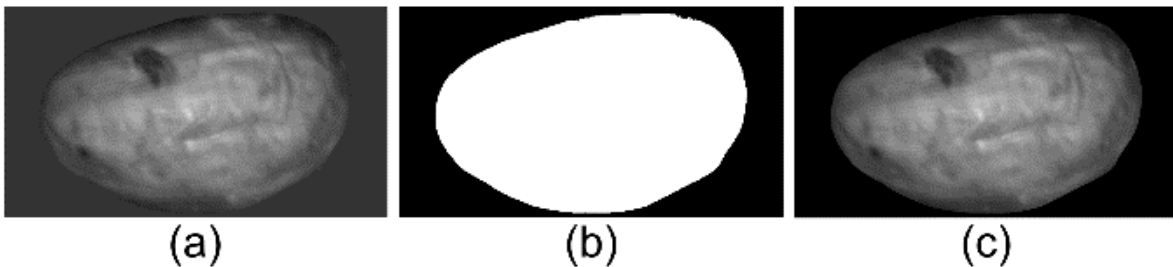

Figure 4. (a) a HSI F, (b) its binary mask $\underline{D}$, and (c) the HSI $\tilde{F}$ without background. Source: Authors' own work.

2.4   Feature Extraction

The second stage in the proposed framework involves extracting the classification features from $\tilde{F}$. Recent works have demonstrated that segmenting spectral information into spatial superpixels reduces the computational cost required by supervised classification methods and

increases accuracy [36], [37], [38], [39]. Since traditional superpixel algorithms operate on three-band images (e.g., RGB images), we propose the following to extract the classification features: first, a three-band (RGB) image of the $\tilde{F}$ hyperspectral data cube is extracted. Notice that $\tilde{F}$ is acquired from the range between 350 and 950 nm. Then, from the $\tilde{F}$ data cube, the spectral bands corresponding to 460, 530, and 670 nm are selected to form a three-band image (R, G, B). The three wavelengths represent the spectral response peaks of the blue, green, and red channels, as shown in Figure. 5.
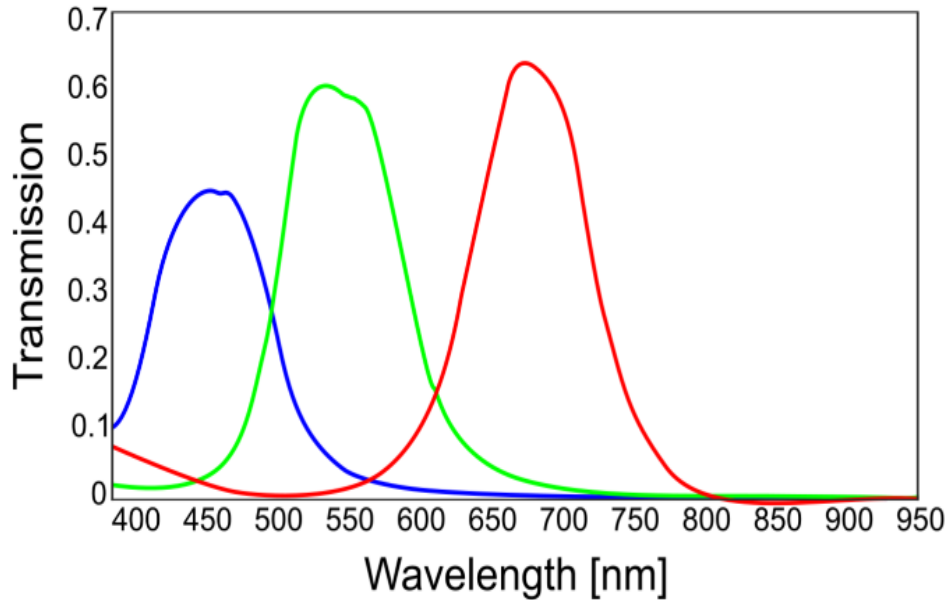

Figure 5. Theoretical spectral responses of the red, green, and blue channels. Source: Adapted from [41].

Afterwards, a segmentation algorithm is applied to the three-band image to find a superpixel map. For this purpose, we specifically use the well-known Simple Linear Iterative Clustering (SLIC) algorithm [35].

The SLIC algorithm works in the five-dimensional space, where the two coordinates (x, y) correspond to the spatial location of the superpixel and the other three components depict the RGB channels. The input variable of this algorithm is the number of desired superpixels (N_{spx}). Given N_{spx}, where the approximate size of each superpixel is MN/N_{spx}, the SLIC algorithm defines a cluster center at every grid interval, S, as follows (5):

$$S = \sqrt{MN/N_{spx}} \tag{5}$$

Hence, the algorithm chooses N_{spx} superpixel cluster centers, $C_j = \left[r_j, g_j, b_j, x_j, y_j\right]^T$ with $j = [1, N_{spx}]$.

The SLIC algorithm assumes that the pixels associated with a cluster lie in a $2S \times 2S$ area around the superpixel center on the (x, y) plane. Therefore, this is the pixel search area near each cluster center. The center is transferred to the lowest gradient position in a 3×3 neighborhood to prevent it from remaining on the edge of an object. In the next step, for each cluster center, the algorithm assigns the best-matching pixels from the search area according to the distance measure H_t defined as follows (6), (7), (8):

$$H_c = \sqrt{(r_j - r_{j\prime})^2 + (g_j - g_{j\prime})^2 + (b_j - b_{j\prime})^2} \tag{6}$$

$$H_P = \sqrt{(x_j - x_{j\prime})^2 + (y_j - y_{j\prime})^2} \tag{7}$$

$$H_t = H_c + \frac{m}{S} H_P, \tag{8}$$

where H_c is the sum of the *RGB* distance and the *xy* plane distance *normalized* by the grid interval S; and $r_j, g_j,$ and b_j denote the color of the $j-th$ superpixel cluster center, while $j\prime$ indexes each pixel, $j\prime = [1, MN]$. The value of m controls the compactness of a superpixel, which can be in the [1, 20] range. Usually, it is chosen as $= 10$ [39]. In this paper, we used $N_{spx} = 500$. Figure 6 is an RGB image segmented into superpixels with the SLIC algorithm and displayed with a false-color composite.

Finally, the spatial information of the superpixel map is matched with the hyperspectral information of the data cube
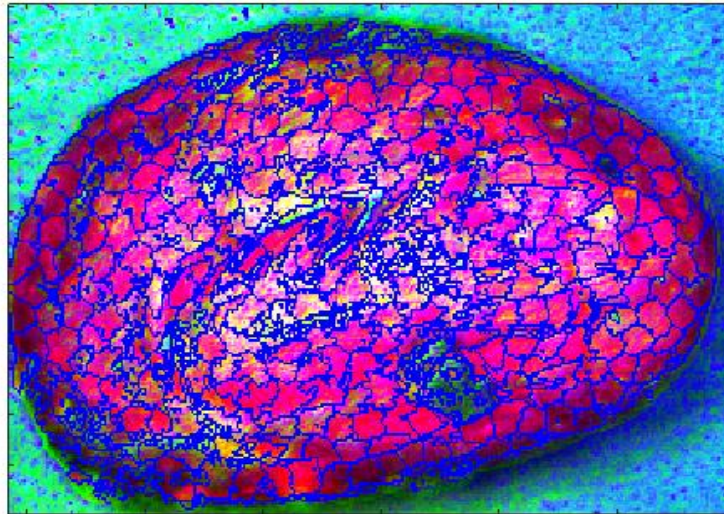

Figure 6. A cocoa bean RGB image segmented into 500 superpixels with the SLIC algorithm
Source: Authors' own work.

$\widetilde{\mathcal{F}}$ to calculate the average spectral signature of each superpixel and build an array with these spectral signatures, i.e., a matrix of size $L \times N_{spx}$, as explained below.

Let $\tilde{\mathbf{F}} \in \mathbb{R}^{L \times MN}$ be the unfolded matrix of the hyperspectral image ($\widetilde{\mathcal{F}}$) reorganized as $\tilde{\mathbf{F}} = [\tilde{\mathbf{f}}_{(1)}, \ldots, \tilde{\mathbf{f}}_{(MN)}]$, where $\tilde{\mathbf{f}}_{(k)} \in \mathbb{R}^L$ represents the spectral signature of the k-th pixel.

Mathematically, the matrix with the average spectral signatures $\mathbf{Y} \in \mathbb{R}^{L \times N_{spx}}$ is created as (9):

$$\mathbf{Y} = \tilde{\mathbf{F}}\,\mathbf{H}, \tag{9}$$

where $\mathbf{H} \in \mathbb{R}^{MN \times N_{spx}}$ is an average sorting matrix, in which the different values of zero for each column (ℓ) have a weight $1/N_s^\ell$, such that N_s^ℓ is the number of pixels grouped in a superpixel. Then, the matrix $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_{N_{spx}}]$ of spectral signatures will be classified as explained in the next subsection.

2.5 Supervised Classification

The last step in the proposed framework is to classify the spectral signatures of each superpixel in Y using the Support Vector Machine (SVM) algorithm.

For this purpose, we denote the information of n spectral signatures used in the training step as $\Theta = \{y_1, \dots, y_n\}$, and their respective class labels as $\Omega = \{\omega_1, \omega_2, \omega_3\}$, where $\omega_1, \omega_2, \omega_3 \in R^{n/3}$ represent the slight, correctly, and high fermentation levels, respectively.

2.5.1 Algorithm

The SVM algorithm was initially proposed for binary classification in order to determine a hyperplane ($\mathbf{y}^T_k\mathbf{M}+\mathbf{b}$) that optimally separates the samples of one class from those of another [42,43]. However, since this study considers three categories, multiple-class SVM should be employed [42]. However, since this study considers three categories, multiple-class SVM should be employed [43]. Specifically, said method seeks to determine d hyperplanes (in our case $d = 3$), solving the following optimization problem for the training stage (10):

$$\min_{\mathbf{M}_m, \xi_i} \frac{1}{2}\sum_{m=1}^{d} \mathbf{M}_m^T\mathbf{M}_m + \lambda \sum_{i=1}^{n} \xi_i,$$
$$s.t. \mathbf{M}_{\omega_i}^T\varphi(\mathbf{y}_i) - \mathbf{M}_t^T\varphi(\mathbf{y}_i) \geq 1 - \delta_{\omega_i,t} - \xi_i, \tag{10}$$

where $\mathbf{M}_m$ is the m-th weight vector; n, the number of training samples; $\omega_i \in \{1, \dots, d\}$, the labels of the i-th super-pixel; ξ_i, the set of slack variables that consider the nonseparability between sets belonging to different classes; λ, a regularization parameter that controls the influence of the misclassified samples; $i = 1, \dots, n$; and $t \in 1, \dots, d$. The mapping function φ projects the training data into a suitable feature space to allow for nonlinear decision surfaces, and $\delta_{i,j}$ is the Kronecker delta function with value 1 for $i{=}j$ and 0 otherwise. Finally, with trained weight vectors $\mathbf{M}_m$, the resulting decision function for any superpixel $\mathbf{y}_i$ is given by (11):

$$\overline{\omega_i} = argmax_m\mathbf{M}_m^T\varphi(\mathbf{y}_i). \tag{11}$$

Note that the result obtained in (11) is the classification of one superpixel. Therefore, to assign a label to the whole bean, the predominant label is calculated as (12)

$$\overline{\omega_J} = \arg\max_{\Omega=\{\omega_1,\omega_2,\omega_3\}} \sum_{i=1}^{N_{spx}} \delta\left(\overline{\omega_{J,i}}, \Omega\right). \tag{12}$$

3.   DATA ANALYSIS AND RESULTS

This section presents an analysis of the classification of cocoa beans in terms of their level of fermentation using the proposed method. The SVM classifier uses a Gaussian kernel and cross-validation for hyperparameter tuning, which was implemented using a Matlab toolbox.

First, the classification focuses on finding a label for each bean based on the calculation of the predominant label. Second, bean uniformity is analyzed assuming each superpixel as a sample. Finally, the influence of the number of training samples on classification accuracy is calculated.

3.1   Bean Classification Based on Predominant Label

In this subsection, we evaluate the classification proposed in (12). The objective is to label each bean with its predominant fermentation category. Table 2 lists some parameters of this classification.

Table 2. Classification parameters. Source: Authors' own work.

Number of cocoa beans	90
Superpixels per bean	500
Beans per category	30
Spectral superpixels per category	15000
Spectral superpixels in total	45000
Percentage of spectral signatures used for training the SVM classifier	15 %

In total, we classified 90 cocoa beans, each one spatially divided into 500 regions (superpixels); that is, we classified a total of 45,000 superpixels. Let us remember that each superpixel has an associated spectral signature consisting of 301 reflectance values measured between 350 and 950 nm in steps of 2 nm. For this classification, we used 15 % of the samples to train the SVM classifier. Figure 7 shows the result of classifying one of the 90 cocoa beans using the proposed method. The framework assigns a label to the spectral signature of each superpixel and, subsequently, assigns the predominant label to each bean.

As mentioned in Subsection 2.1, we also had the reference label of each bean, which was carefully assigned by a professional technical team. Therefore, after assigning a label to each bean, this method evaluated the precision of the classification proposed in this paper by comparing the results with the reference labels.
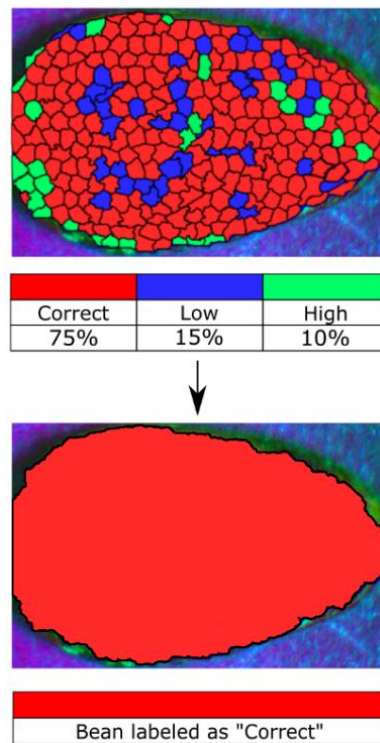

Figure 7. Visual results of the classification by superpixels (top) and predominant label (bottom)
Source: Authors' own work.

The confusion matrix in Table 3 details the number of beans correctly and incorrectly labeled with the proposed method. Each column in the matrix represents the number of predictions of each class, while each row represents the ground truth.

From the results in the confusion matrix, we calculated the values of recall, precision, truth overall, and overall classification per class as in [44]. These results are shown in Table 4.

Table 3. Confusion matrix of all the beans classified based on the predominant label (Equation 12) Source: Authors' own work.

	Slight	Correct	High
Slight	30	0	0
Correct	0	30	0
High	0	0	30

Note that all Slightly, Correctly, and Highly fermented cocoa beans were correctly labeled by the classifier (Table 3). In addition, the metrics in Table 4 show the high performance of the proposed classification approach. The overall precision of assigning the predominant label to 90 cocoa beans was 100 %.

Table 4. Classification metrics from the confusion matrix. Source: Authors' own work.

Truth overall class Slight	30	Recall class Slight	100 %
Truth overall class Correct	30	Recall class Correct	100 %
Truth overall class High	30	Recall class High	100 %
Classification overall class Slight	30	Precision class Slight	100 %
Classification overall class Correct	30	Precision class Correct	100 %
Classification overall class High	30	Precision class High	100 %
Overall Accuracy		**100 %**	

3.2 Bean Uniformity Analysis

Entities that regulate the international marketing of cocoa beans, such as The Federation of Cocoa Commerce London (FCC), require the quality of the seeds in each ton to be acceptably uniform. The objective is to guarantee the homogeneity of the products derived from cocoa.

An approach such as the one proposed here can be used to analyze the uniformity of cocoa loads. It was shown that the 90 beans used in this study were correctly classified, 30 into each one of the fermentation categories. However, note that the initial labeling of the bean (bottom, Figure 7) disregards the fact that some superpixels can be classified into other fermentation levels different from the predominant one (top, Figure 7). Therefore, classifying each superpixel in a grain separately, we can analyze how uniform each bean is.

In this subsection, we discuss a more detailed classification in which bean fermentation uniformity is calculated. This type of analysis has not yet been used commercially in the cocoa industry. Furthermore, it is rarely used in general food quality control, where uniformity is assessed at the collective level (batch uniformity).

Table 5 shows the result of classifying the average spectral signature of each one of the 45,000 superpixels. We used the general bean category as the reference label for each superpixel in the confusion matrix. Note that, compared to Table 3, the matrix in Table 5 is not diagonal. It is understood that the matrix is not diagonal because there may be regions (superpixels) with different degrees of fermentation despite being on the same grain.

Table 5. Confusion matrix of all the classified spectral superpixels (Eq. 11). Source: Authors' own work.

	Slight	Correct	High
Slight	12,915	765	1,320
Correct	435	12,360	2,205
High	210	2,610	12,180

Table 6 shows the percentage of superpixel classification in each class. The rows represent the actual labels (full-grain categories); therefore, the sum of the values in each row is equal to 100 %. In turn, the columns show the tags assigned by the classifier to the superpixels. The last row, Total, shows the percentage of superpixels labeled by the model in each fermentation level. Note that, although there are 30 beans in each category, regions of correct and high fermentation predominate, with 34.96 % and 34.89 % of the areas, respectively.

Furthermore, Table 6 shows that, in highly fermented beans, there is a considerable amount of material that is well fermented (17.4 %), and vice versa (14.7 %). In comparison, only small portions of the beans with high and correct fermentation have slight fermentation (1.4 % and 2.9 %, respectively). Regarding slightly fermented beans, there are small burned regions (high, 8.8 %) and a minority of well-fermented areas (5.1 %).

Table 6. Confusion matrix of all the classified superpixels in percentages. Source: Authors' own work

	Slight	Correct	High
Slight	86.1 %	5.1 %	8.8 %
Correct	2.9 %	82.4 %	14.7 %
High	1.4 %	17.4 %	81.2 %
Total	30.13 %	34.96 %	34.89 %

This analysis could contribute to the implementation of fermentation techniques that produce a more homogeneous drying.

3.3 Number of Training Samples and Classification Accuracy

In this subsection, we vary the number of spectral signatures per category used in the SVM classifier training in order to analyze its influence on the precision of the classification and determine how many spectral signatures are necessary to obtain acceptable results.

The training was programmed to randomly select a certain percentage of spectral samples, the same number of signatures from each category. The system was trained, and the beans were classified using the predominant label method. Specifically, the percentage of training samples was varied between 1 % and 20 % of the total signatures (45,000), as seen on the x-axis in Figure 8. A 10 % of each resulting training set was used as validation to tune the model.

Figure 8 shows the average accuracy of the test set over ten experiments. Evidently, as the number of training samples increases, the classifier improves its precision.

Note that the class most easily recognized by the classifier is Slight fermentation, followed by Correct and High, with any proportion of training samples.

Specifically, using only 2.5 % of the training samples (which is equivalent to 1,125 spectral signatures), the system correctly labeled 90 % of the cocoa beans as Slight fermentation, i.e., 27 of the 30 beans in this class. Likewise, with the same training percentage, the classifier

accurately detected 76.6 % of the beans with Correct fermentation (23 beans), and 63.3 % of those with High fermentation (19 beans).

Furthermore, in Figure 8, when the percentage of training samples exceeds 12.5 %, the overall accuracy of the classification is quite acceptable. In fact, let us remember that the results in Subsections 3.1. and 3.2. were obtained from classifications performed with 15 % of the training samples.
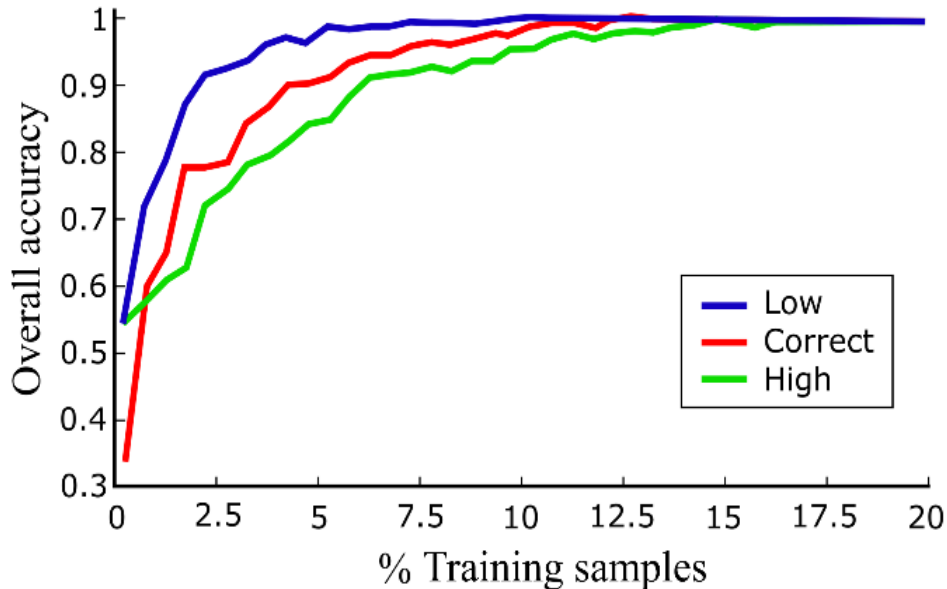

Figure 8. Overall classification accuracy as a function of the percentage of training samples
Source: Authors' own work.

4.   CONCLUSIONS

In this paper, we developed a non-invasive framework for classifying dry cocoa beans into three fermentation categories using spectral imaging and no chemical methods. The evaluation of the performance of the proposed framework showed its high accuracy in the classification of cocoa beans when 15 % or more of the samples in each category were used for training. Subsequently, the results were compared with the labels assigned to each bean by the Colombia National Federation of Cocoa Growers. The results of this study demonstrate that the use of spectral information is feasible for the noninvasive quality control of cocoa beans.

Furthermore, using the proposed classification framework, it is possible to establish the percentage of each bean that belongs to a fermentation category different from the label of the whole bean. Future studies should examine other classification methods and frameworks in order to compare their computational complexity and accuracy with those of the SVM approach.

The implementation of an analysis with this level of detail in the productive sector can help to investigate fermentation techniques that yield more uniform results. In the business sector, it can allow organizations to improve their pricing methods and have a more demanding selection of raw material for premium-quality products derived from cocoa.

5.   ACKNOWLEDGEMENTS

CONFLICTS OF INTEREST

None declared.

AUTHOR CONTRIBUTIONS

Karen Sánchez, contributed to the writing, methodology, data collection, algorithm analysis, obtaining and analysis of results.

Jorge Bacca, contributed to the algorithm development, methodology, data collection, analysis of results, validation and revision.

Laura Arévalo-Sánchez, contributed to the writing, parametric analysis of the algorithm, and validation of results.

Henry Arguello, contributed to conceptualization, methodology, revision, and supervision.

Sergio Castillo, contributed to conceptualization, revision, and supervision.

6.   REFERENCES

[1]   E. Lecumberri *et al.,* "Dietary fibre composition, antioxidant capacity and physico-chemical properties of a fibre-rich product from cocoa *(Theobroma cacao L.) Food chemistry*, vol. 104, no. 3, pp. 948-954, 2017. https://doi.org/10.1016/j.foodchem.2006.12.054

[2]   M. S. Beg; S. Ahmed; K. Jan; K. Bashir, "Status, supply chain and processing of cocoa - A review," *Trends in food science & technology*, vol. 66, pp. 108-116, Ago. 2017. https://doi.org/10.1016/j.tifs.2017.06.007

[3]   J. C. Motamayor *et al.* "Geographic and Genetic Population Differentiation of the Amazonian Chocolate Tree *(Theobroma cacao L),*" *PloS one*, vol. 3, no 10, Oct. 2008. https://doi.org/10.1371/journal.pone.0003311

[4]   M. Torres-Moreno; E. Torrescasana; J. Salas- Savadó; C. Blanch, "Nutritional composition and fatty acids profile in cocoa beans and chocolates with different geographical origin and processing conditions," *Food chemistry*, vol. 166, pp. 125-132. Jan. 2015. https://doi.org/10.1016/j.foodchem.2014.05.141

[5]   A. Wickramasuriya; J. Dunwell, "Cacao biotechnology: current status and future prospects," *Plant biotechnology journal*, vol. 16, no. 1, pp. 4-17, Jan. 2018.https://doi.org/10.1111/pbi.12848

[6]   R. Saltini; R. Akkerman; S. Frosch, "Optimizing chocolate production through traceability: A review of the influence of farming practices on cocoa bean quality," *Food control*, vol. 29, no. 1, pp. 167-187, Jan. 2013. https://doi.org/10.1016/j.foodcont.2012.05.054

[7]   C. N. Tejada-Tovar; A. Villabona-Ortíz; G. Alvarez-Bajaire; L. attin-Torres; C. Granados-Conde, "Influencia de la altura del lecho sobre el comportamiento dinámico de columna de lecho fijo en la biosorción de mercurio," *TecnoLógicas*, vol. 20, no 40, p. 71-81, Sep. 2017. https://doi.org/10.22430/22565337.706

[8]   A. Friedel Hütz; C. Huber; I. Knoke; P. Morazán; M. Mürlebach, "Strengthening the competitiveness of cocoa production and improving the income of cocoa producers in West and Central Africa," Bonn, Germany: Südwind, 2016. URL

[9]   J. Lernoud *et al.,* "The state of sustainable markets-statistics and emerging trends 2015," Report. 29694 Mar. 2016. URL

[10]   J. E. Kongor *et al.*, "Constraints for future cocoa production in Ghana," *Agroforestry Systems*, vol. 92, no. 5, pp. 1373-1385, Oct. 2018. https://doi.org/10.1007/s10457-017-0082-9

[11]   R. Swaray, "Commodity buffer stock redux: The role of International Cocoa Organization in prices and incomes," *Journal of Policy Modeling*, vol. 33, no. 3, pp. 361-369, May. 2011. https://doi.org/10.1016/j.jpolmod.2011.03.002

[12]   M. Squicciarini; J. Swinnen, *The economics of chocolate,* Oxford University Press, 2016.

[13]   Instituto Colombiano de Normalización y Certificación–ICONTEC-. "Norma Técnica Colombiana NTC 1252: Cacao en grano," 2003. URL

[14]   S. Jinap; P. S. Dimick; R. Hollender, "Flavour evaluation of chocolate formulated from cocoa beans from different countries," *Food Control*, vol. 6, no. 2, pp. 105-110, 1995. https://doi.org/10.1016/0956-7135(95)98914-M

[15]   P. C. Aculey *et al.* "Ghanaian cocoa bean fermentation characterized by spectroscopic and chromatographic methods and chemometrics," *Journal of Food Science*, vol. 75, no. 6, pp. S300-S307, Aug. 2010. https://doi.org/10.1111/j.1750-3841.2010.01710.x

[16]   A. A. Gowen; C. P. O´Donell; P. J. Cullen; G. Downey; J. M. Frias, "Hyperspectral imaging–an emerging process analytical tool for food quality and safety control," *Trends Food Sci. Technol*, vol. 18, no. 12, pp. 590-598, Dec. 2007. https://doi.org/10.1016/j.tifs.2007.06.001

[17]   J. Qin; K. Chao; M. S. Kim; R. Lu; T. F. Burks, "Hyperspectral and multispectral imaging for evaluating food safety and quality," *Journal of Food Engineering*, vol. 118, no. 2, pp. 157-171, Sep. 2013. https://doi.org/10.1016/j.jfoodeng.2013.04.001

[18]   H. Huang; L. Liu; M. O. Ngadi, "Recent developments in hyperspectral imaging for assessment of food quality and safety," *Sensors*, vol. 14, no. 4, pp. 7248-7276, Apr. 2014. https://doi.org/10.3390/s140407248

[19]   C. Garrido-Novell *et al.* "Grading and color evolution of apples using RGB and hyperspectral imaging vision cameras," *Journal of Food Engineering*, vol. 113, no. 2, pp. 281-288, Nov. 2012. https://doi.org/10.1016/j.jfoodeng.2012.05.038

[20]   J. M. Bioucas-Dias *et al.* "Hyperspectral remote sensing data analysis and future challenges," *IEEE Geoscience and remote sensing magazine*, vol. 1, no. 2, pp. 6-36, Jul. 2013. https://doi.org/10.1109/MGRS.2013.2244672

[21]   G. A. Shaw; H. K. Burke, "Spectral imaging for remote sensing," *Lincoln laboratory journal*, vol. 14, no. 1, pp. 3-28, 2003. URL

[22]   J. Bacca; C. A. Hinojosa; H. Arguello, "Kernel sparse subspace clustering with total variation denoising for hyperspectral remote sensing images," *Mathematics in Imaging*. Optical Society of America, 2017. p. MTu4C. 5. 2017. https://doi.org/10.1364/MATH.2017.MTu4C.5

[23]   C. I. Chang, "*Hyperspectral imaging: techniques for spectral detection and classification,*" Springer Science & Business Media, vol. 1, 2003.

[24]   H. Cen; Y. He, "Theory and application of near infrared reflectance spectroscopy in determination of food quality," *Trends in Food Science & Technology*, vol. 18, no. 2, pp. 72-83, Feb. 2007. https://doi.org/10.1016/j.tifs.2006.09.003

[25]   J. Pinto; H. Rueda-Chacón; H. Arguello, "Classification of Hass avocado (persea americana mill) in terms of its ripening via hyperspectral images," *TecnoLógicas*, vol. 22, no. 45, pp. 111-130, May. 2019. https://doi.org/10.22430/22565337.1232

[26] J. Bacca; H. Arguello, "Sparse Subspace Clustering for Hyperspectral Images using Incomplete Pixels." *TecnoLógicas*, vol. 22, no. 46, pp. 6-19, Sep. 2019. http://dx.doi.org/10.22430/22565337.1205

[27] S. Sunoj; C. Igathinathane; R. Visvanathan, "Nondestructive determination of cocoa bean quality using FT-NIR spectroscopy," *Computers and Electronics in Agriculture*, vol. 124, pp. 234-242, Jun. 2016. https://doi.org/10.1016/j.compag.2016.04.012

[28]   A. Veselá *et al.,* "Infrared spectroscopy and outer product analysis for quantification of fat, nitrogen, and moisture of cocoa powder," *Analytica chimica acta*, vol. 601, no. 1, pp. 77-86, Oct. 2007. https://doi.org/10.1016/j.aca.2007.08.039

[29]   C. Hue *et al.,* "Near infrared spectroscopy as a new tool to determine cocoa fermentation levels through ammonia nitrogen quantification," *Food chemistry*, vol. 148, pp. 240-245, Apr. 2014. https://doi.org/10.1016/j.foodchem.2013.10.005

[30] A. Krähmer *et al.* "Fast and neat-Determination of biochemical quality parameters in cocoa using near infrared spectroscopy," *Food Chemistry*, vol. 181, pp. 152-159, Aug. 2015. https://doi.org/10.1016/j.foodchem.2015.02.084

[31] E. Teye; X. yi- 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 research international*, vol. 55, pp. 288-293, Jan. 2014. https://doi.org/10.1016/j.foodres.2013.11.021

[32] E. Teye *et al.* "Estimating cocoa bean parameters by FT-NIRS and chemometrics analysis," *Food chemistry*, vol. 176, pp. 403-410, Jun. 2015. https://doi.org/10.1016/j.foodchem.2014.12.042

[33] P. D. Tran *et al.* "Assessing cocoa aroma quality by multiple analytical approaches," *Food Research International*, vol. 77, no. 3, pp. 657-669, Nov. 2015. https://doi.org/10.1016/j.foodres.2015.09.019

[34] N. A. Gomez; K. Sanchez; H. Arguello, "Non-Destructive Method for Classification of Cocoa Beans from Spectral Information," *2019 XXII Symposium on Image, Signal Processing and Artificial Vision (STSIVA)*, IEEE, Bucaramanga, 2019. https://doi.org/10.1109/STSIVA.2019.8730257

[35] R. Achanta *et al.* "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274-2282, May. 2012. https://doi.org/10.1109/TPAMI.2012.120

[36] H. Garcia; C. V. Correa; O. Villarreal; S. Pinilla: H. Arguello, "Multi-resolution reconstruction algorithm for compressive single pixel spectral imaging," *25th European Signal Processing Conference (EUSIPCO)*. IEEE, pp. 468-472, Kos, 2017. https://doi.org/10.23919/EUSIPCO.2017.8081251

[37] A. Jerez; H. Garcia; H. Arguello, "Spectral image fusion for increasing the spatio-spectral resolution through side information," *IEEE Colombian Conference on Applications in Computational Intelligence*. Springer, Cham, vol. 833, pp. 165-176, 2018. https://doi.org/10.1007/978-3-030-03023-0_14

[38] H. Garcia; C. V. Correa; K. Sánchez; E. Vargas; H. Arguello, "Multi-resolution coded apertures based on side information for single pixel spectral reconstruction," *26th European Signal Processing Conference (EUSIPCO)*. IEEE, pp. 2215-2219, Rome. 2018. https://doi.org/10.23919/EUSIPCO.2018.8553602

[39] K. Sanchez; C. Hinojosa; H. Arguello, "Supervised spatio-spectral classification of fused images using superpixels," *Applied optics*, vol. 58, no. 7, pp. B9-B18, 2019. https://doi.org/10.1364/AO.58.0000B9

[40] C. Hinojosa; K. Sánchez; H. García; H. Arguello, "Cocoa beans spectral image with three fermentation levels". *IEEE Dataport*, 2019. http://dx.doi.org/10.21227/esks-4b74

[41] L. Frey *et al.*, "Color filters including infrared cut-off integrated on CMOS image sensor," *Optics Express*, vol. 19, no. 14, pp. 13073-13080, 2011. https://doi.org/10.1364/OE.19.013073

[42] B. Schlkopf; A. J. Smola; F. Bach, Learning with kernels: support vector machines, regularization, optimization, and beyond, the MIT Press, 2018.

[43] G. Mountrakis; J. Im; C. Ogole, "Support vector machines in remote sensing: A review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 66, no. 3, pp. 247-259, May. 2011. https://doi.org/10.1016/j.isprsjprs.2010.11.001

[44] C. Beleites; R. Salzer; V. Sergo, "Validation of soft classification models using partial class memberships: An extended concept of sensitivity & co. applied to grading of astrocytoma tissues," *Chemometrics and Intelligent Laboratory Systems*, vol. 122, pp. 12-22, Mar. 2013. https://doi.org/10.1016/j.chemolab.2012.12.003