Learning-based Spectral Regression for Cocoa Bean Physicochemical Property Prediction

Kebin Contreras[1†], Emmanuel Martinez[2†], Brayan Monroy[2],
Sebastian Ardila[2], Cristian Ramirez[2], Mariana Caicedo[2],
Hans Garcia[1], Tatiana Gelvez-Barrera[3], Juan Poveda-Jaramillo[2],
Henry Arguello[2], Jorge Bacca[2*]

[1]Physics School and [2]Department of Computer Science, Universidad
Industrial de Santander, Carrera 27 Calle 9, Bucaramanga, 680001,
Santander, Colombia.
[3]CNRS, Inserm, CREATIS UMR 5220, Université de Lyon, Université
Claude Bernard Lyon 1, UJM-Saint Etienne, Street, Lyon, 69000,
Auvergne-Rhône-Alpes, France.

*Corresponding author(s). E-mail(s): jbacquin@uis.edu.co;
†These authors contributed equally to this work.

Abstract

Cocoa bean quality assessment is essential for ensuring compliance with commercial standards, protecting consumer health, and increasing the market value of the cocoa product. The quality assessment estimates key physicochemical properties, such as fermentation level, moisture content, polyphenol concentration, and cadmium content, among others. This assessment has traditionally relied on the accurate estimation of these properties via visual or sensory evaluation, jointly with laboratory-based physicochemical analyses, which are often time-consuming, destructive, and difficult to scale. This creates the need for rapid, reliable, and noninvasive alternatives. Spectroscopy, particularly in the visible (VIS: 400–700 nm) and near-infrared (NIR: 700–2500 nm) ranges, offers a noninvasive alternative by capturing the molecular signatures associated with these properties. Therefore, this work introduces a scalable methodology for evaluating the quality of cocoa beans by predicting key physicochemical properties from the spectral signatures of cocoa beans. This approach utilizes a conveyor belt system integrated with a VIS-NIR spectrometer, coupled with learning-based regression models. Furthermore, a dataset is built using cocoa bean batches from

1

Santander, Colombia. Ground-truth reference values were obtained through standardized laboratory analyses and following commercial cocoa quality regulations. To further evaluate the proposed methodology's generalization, performance is tested on samples collected from other Colombian regions and from Cusco, Peru. Experimental results show that the proposed models achieved $\mathcal{R}^2$ scores exceeding **0.98** across all physicochemical properties, and reached **0.96** accuracy on geographically independent samples. This non-destructive approach represents a suitable and scalable alternative to conventional laboratory methods for quality assessment across the cocoa production chain.

Keywords: Cocoa Beans, Physicochemical Properties, Spectral Imaging, Data Regression, Near-Infrared Spectroscopy, Agriculture

Introduction

Cocoa beans (*Theobroma cacao* L.) rank among the world's leading agricultural products, with global production estimated at 4.368 million tons during the 2023/2024 crop season according to the International Cocoa Organization (ICCO, 2025). Cultivation remains concentrated in West Africa ($\approx 75\%$), Asia and Oceania ($\approx 5\%$), and Latin America (20%) (Kongor et al., 2024). In Latin America and the Caribbean, cocoa occupies over 1.7 million hectares and contributes close to one-fifth of global supply, with major producing countries including Ecuador, Brazil, Peru, Colombia, the Dominican Republic, and Mexico(ICCO, 2025; Huetz-Adams, 2022). The region is also the world's leading source of fine-flavor cocoa, accounting for approximately 90% of global exports in this premium segment, particularly through Ecuador, the Dominican Republic, and Peru (ICCO, 2023).

Critically, cocoa farming supports the livelihoods of approximately 5 to 6 million smallholder households worldwide, often operating low-input farms of 2 to 5 hectares, and contributes 60% to 90% of their household income, boosting food security, education, and rural economies (Kongor et al., 2024; Huetz-Adams, 2022). For these farmers, participation in fine-flavor cocoa markets offers access to higher premiums at origin, although the small size of this niche market (only 12% of global exports) and strict quality requirements present barriers (ICCO, 2023). Nonetheless, cocoa farming remains a culturally rooted and economically vital activity, particularly in Latin America, offering smallholder farmers a path to resilience, rural development, and greater inclusion in sustainable and differentiated value chains.

The quality of cocoa beans is determined by genetic factors, post-harvest practices, and geographic conditions, among others, which play a crucial role in shaping their final attributes (Kongor et al., 2016). Key post-harvest processes, such as, fermentation, drying, and roasting, directly influence the sensory profile and physicochemical characteristics of the beans (de Brito et al., 2001). These processes affect critical attributes like flavor, aroma, and nutritional content, which in turn determine the product's marketability and economic value (De Vuyst and Weckx, 2016). Therefore,

Fig. 1: Visual classification of cocoa beans based on the cut test method. The top row shows closed beans being difficult to classify, and the bottom row shows the corresponding internal appearance after cutting. Beans are categorized as Premium, Standard, or Ordinary based on internal and external appearance, considering color, texture, and uniformity.

the ability to assess and determine these quality factors is essential for maintaining consistency in production and meeting the standards and requirements of both national and international markets.

In Latin America, countries such as Colombia, Ecuador, Brazil, and Peru have established national standards to regulate cocoa quality by international trade and food safety requirements. These regulations define criteria such as fermentation level, moisture content, cadmium concentration, and physical defects. Colombia applies the NTC 1252:2021, which classifies dry cocoa beans by quality (ICONTEC, 2021). Ecuador enforces NTE INEN 176:2021, its official standard for fermented cocoa beans (INEN, 2021). Brazil regulates quality through MAPA Normative No. 38/2018, which sets classification parameters for cocoa beans (MAPA, 2018). In Peru, cocoa quality control relies on NTP-ISO 2292:2019, based on international sampling guidelines (INACAL, 2019).

Fermentation remains a key determinant of cocoa quality across the region. It is traditionally evaluated through the cut test, a destructive visual inspection method rooted in ancestral knowledge (ICONTEC, 2021; INEN, 2021; MAPA, 2018; INACAL, 2019). Although widely used, this technique is inherently subjective, relying on human interpretation of internal and external bean characteristics. As illustrated in Fig. 1, cocoa beans are visually classified into premium, standard, or ordinary categories based on external appearance and internal characteristics revealed by the cut test. In particular, ordinary beans display a dark purple interior, indicative of low fermentation. In contrast, standard and premium beans show more developed brown coloration and well-formed internal structures, reflecting higher levels of fermentation, with premium beans typically exhibiting the most uniform and desirable traits (ICONTEC, 2021; INEN, 2021; MAPA, 2018; INACAL, 2019).

In addition to visual inspection, physicochemical properties such as moisture content, polyphenol concentration, and cadmium levels are important indicators of cocoa bean quality (Samanta et al., 2022; Abt and Robin, 2020). Moisture affects shelf life

and the risk of mold growth, while polyphenols are valued for their antioxidant properties but are often degraded during fermentation and roasting (Samanta et al., 2022). Cadmium, which is influenced by soil composition, presents a food safety concern, particularly in beans from some Latin American regions where soil levels are naturally higher (Abt and Robin, 2020). These factors impact both the nutritional and sensory qualities of cocoa and have significant implications for regulatory standards and international trade (Samanta et al., 2022; Abt and Robin, 2020).

Cocoa quality assessment typically relies on laboratory-based quantitative techniques such as atomic absorption spectroscopy (AAS) (Araujo et al., 2020), gas chromatography (GC) (Ducki et al., 2008), and mass spectrometry (MS) (Cain et al., 2019). While these methods provide detailed insights into the internal composition and physicochemical properties of cocoa beans, they are invasive and destructive. Additionally, they require specialized infrastructure, trained personnel, and incur high operational costs, making them largely inaccessible to rural communities and small-scale agriculture (Niemenak et al., 2014). The extended processing times also limit their applicability in scenarios where rapid decision-making is needed. Consequently, many small-scale producers are unable to access such analyses, hindering their ability to optimize cocoa quality and remain competitive in the market.

Recent research has shown that visible spectrum (VIS) ranged from 400-700 nm and near-infrared (NIR) range from 700-2500 nm spectroscopy are effective, reliable, and non-invasive techniques for predicting the cocoa quality (Araujo et al., 2020; Sánchez et al., 2021; Hashimoto et al., 2018; Suarez et al., 2025; Pinto et al., 2024; Diaz-Delgado et al., 2025; Teye et al., 2020). These methods collect data reflecting complex molecular vibrations that reveal the physicochemical properties of cocoa beans (Sandorfy et al., 2007). In the VIS range, the reflected radiation from the material is observed, while the absorbed radiation induces electronic transitions in the valence electrons of the constituent molecules. In the NIR range, the overtones of molecular bond vibrations become detectable. In particular, spectral information has been used to predict quality-related attributes such as fermentation level, polyphenol content, and antioxidant activity, with strong correlations reported between polyphenol levels and fermentation (Sánchez et al., 2021; Gomez et al., 2019; Caporaso et al., 2018; Alvarado et al., 2023; Diaz-Delgado et al., 2025). Most recent studies on cocoa quality assessment have been conducted in African countries, where regional environmental conditions and cocoa varieties significantly influence the analysis and outcomes, often requiring the destruction of the beans, since conventional laboratory protocols depend on grinding or chemically processing the samples to obtain accurate measurements (Hashimoto et al., 2018; Ferraris et al., 2023; Ashiagbor et al., 2020; Musah et al., 2019).

Therefore, this study presents a comprehensive framework for estimating the quality of Colombian cocoa beans using noninvasive spectral analysis. It introduces a standardized cocoa bean acquisition protocol and a custom-designed spectral optical system that predicts key physicochemical properties such as fermentation level, moisture content, cadmium concentration, and polyphenol concentration directly from spectral data. Unlike traditional methods that rely on destructive sampling or laboratory processing, this approach preserves the physical integrity of the beans and is

suitable for post-harvest analysis. A spectral dataset was collected from cocoa beans in Santander, Colombia, covering wavelengths from 400 to 2500 nanometers with a spectral resolution of 3648 digitized points. Ground truth labels were obtained through standardized laboratory analyses.

To benchmark predictive performance, a comparative framework was developed to evaluate state-of-the-art machine learning and deep learning regression models in a supervised setting. This enables accurate, scalable, and noninvasive quality assessment for the Colombian cocoa industry. To assess model generalization, additional samples were collected from other regions of Colombia such as Putumayo, Huila and Santander, as well as from Cusco in Peru. The experimental evaluation demonstrated that the proposed methods consistently delivered highly reliable predictions across all physicochemical properties, maintaining strong performance even when tested on geographically distinct samples. This non-destructive strategy therefore provides a practical and scalable alternative to conventional laboratory techniques for quality assessment throughout the cocoa production chain.

Materials and Methods

Physicochemical Cocoa Bean Labeling Settings

To assess cocoa bean quality, this study considers key physicochemical attributes: fermentation level, moisture content, polyphenol concentration, and cadmium content. These properties are critical indicators of both product quality and compliance with safety standards. Each batch of cocoa beans was analyzed following the proposed protocol, and labeled accordingly based on these measurements. Fermentation level was evaluated visually through cut tests as established in national standards. Moisture, polyphenol content, and cadmium concentration were quantified through laboratory methods conducted by the Grupo de Ciencia y Tecnología de Alimentos (CICTA) (CICTA, 2025) at the UIS. These analyses follow validated procedures and relevant technical standards, including national (NTC) and international norms. Table 1 summarizes the relevance of each property and the analytical methods employed for their determination.

Spectral Acquisition System

The system used to acquire spectral signatures of cocoa beans consists of a halogen light source (HL-200) with a spectral coverage of 340–2400 nm, connected to a bifurcated optical fiber (R200-7-VIS-NIR). One arm of the bifurcated fiber redirects the concentrated light beam from the halogen source to the cocoa bean, while the other arm collects the reflected light to be dispersed and integrated by linear sensors inside the Vis-NIR (FLAME-S-VIS-NIR-ES) and NIR-SWIR (NIRQUEST+2.5) spectrometers. The beans are transported through the optical path by a custom-designed conveyor belt powered by a NEMA17 stepper motor and controlled by an Arduino UNO using an A4988 stepper motor driver. The bifurcated fiber is fixed 14 cm (measured from its output edge) above the conveyor belt surface. A 3D-printed funnel and pinion system ensures proper alignment and movement of the cocoa beans throughout

Table 1: Physicochemical properties, analytical methods and regulations for cocoa bean quality assessment.

Property	Importance	Method / Regulation
Fermentation Level	Influences flavor, aroma, and bean color. Evaluated externally (uniformity, mold) and internally via "cut test." Reddish-brown beans indicate proper fermentation and purple implies incomplete processing.	Cut test (visual) *NTC 1252:2021* (ICONTEC, 2021)
Moisture	Reflects drying efficiency and influences shelf life. Elevated moisture content increases the risk of microbial growth, particularly molds. Determined by gravimetric analysis through weight loss after oven-drying at 103±2°C.	Gravimetric method. *GOMESL.01 V07, 2023-06-26* (CICTA, 2025); *NTC 1252:2021* (ICONTEC, 2021)
Polyphenols	Key indicators of antioxidant capacity and contributors to sensory attributes. Quantified via their radical-scavenging activity against DPPH and ABTS, monitored by the decrease in absorbance.	UV-Vis spectrophotometry (Kus et al., 1996). *GOMEPT.01 V01, 2021-09-23* (CICTA, 2025)
Cadmium Content	Toxic heavy metal subject to strict regulatory limits due to its bioaccumulative nature. Typically absorbed from contaminated soils and evaluated using microwave-assisted digestion followed by atomic absorption spectroscopy (AAS).	*NTC-EN 14084:2021* (ICONTEC, 2021).

the spectral acquisition process. The system captures at least 30 spectral signatures per bean as it moves at a speed of 45.3 mm/s.

Dataset Construction

The dataset contains VIS and NIR spectral signatures along with their corresponding physicochemical labels. Reflectance values were calculated relative to white and black references obtained from a Spectralon. To eliminate noisy boundary regions, the spectral data were cropped to the (500–800 nm) range for VIS and the (1100–2000 nm) range for NIR.

Spectral signatures unrelated to cocoa beans (e.g., background regions) were discarded using a threshold of 0.25 based on the SAM method (Kruse et al., 1993). The physicochemical labels correspond to a subset of spectral signatures obtained from the same physical batch of cocoa beans.

To reduce variance and improve robustness, a bootstrapping (Zoubir and Iskandler, 2007) resampling procedure was applied to each cocoa batch. This involved sampling K random subsets of size s from the available signatures and computing the mean for each subset. Specifically, for VIS data, 1000 realizations were generated from an initial set of 1000 signatures, and for NIR, 2000 realizations were produced from 500 original signatures. In both cases, the sample size was fixed at $s = 50$.

Model Training and Evaluation

All experiments were conducted on a workstation equipped with an AMD Ryzen 7 5700X 8-core processor operating at 3.40 GHz, 64 GB of RAM, and an NVIDIA GeForce RTX 4070 GPU with 12 GB of VRAM.

The evaluated regression models included Support Vector Regression (SVR)(Smola and Schölkopf, 2004), Random Forest(Breiman, 2001), CNN (Liu et al., 2019), Long Short-Term Memory (LSTM)(Hochreiter and Schmidhuber, 1997) networks, Spectral-Net (Martins et al., 2022), and Transformer-based architectures(Vaswani et al., 2017). Models were trained separately using input features derived from either the visible (500–800 nm) or near-infrared (1100–2000 nm) spectral ranges. Evaluation metrics comprised the coefficient of determination ($\mathcal{R}^2$) and Mean Squared Error (MSE), calculated for each physicochemical property.
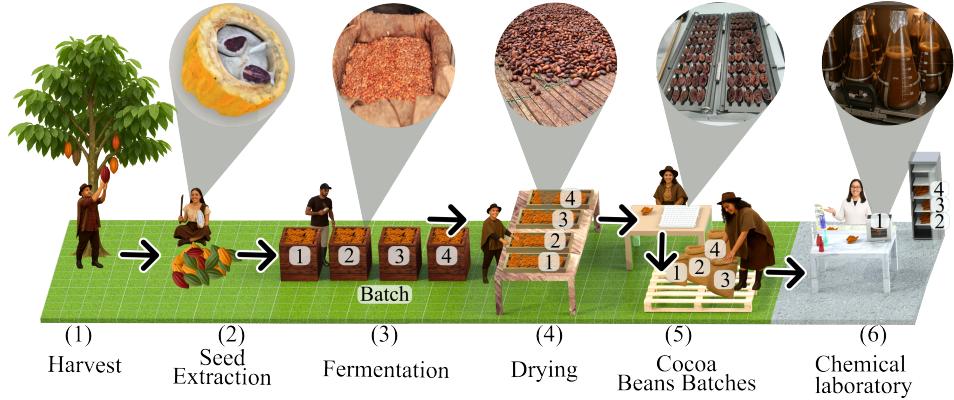
Hyperparameter optimization was performed using grid search and cross-validation on the training set. Deep learning models were implemented in PyTorch and trained using the Adam optimizer, with early stopping criteria based on validation loss convergence.

Results and Discussion

Cocoa Bean Acquisition Protocol

The acquisition and labeling protocol of cocoa beans was structured into seven sequential stages, as illustrated in Fig. 2. This workflow was designed to ensure a reproducible and standardised pipeline from field collection to chemical and spectral laboratory analysis. A batch is defined as the experimental unit, consisting of 1.5 kg of dried cocoa beans. All procedures were carried out independently for each batch as described below:

Stage 1 − Harvest: Cocoa pods were manually harvested at optimal ripeness from a farm located in El Carmen de Chucurí, Santander, Colombia (coordinates $6°41'53''$N, $73°30'40''$W). The farm is part of *"La Asociación de Campesinos Vecinos del Parque Natural Nacional Serranía Los Yariguíes"* (ASOCAPAYARI), and the beans were supplied by a farmer affiliated with the Federación Nacional de Cacaoteros (FEDECACAO) in Colombia. All pods mainly correspond to three commercial clones

Fig. 2: End-to-end pipeline for cocoa bean acquisition and labeling: (1) Harvest of ripe pods; (2) Manual seed extraction; (3) Controlled fermentation in wooden boxes with banana leaves; (4) Sun-drying to stable weight; (5) Guillotine cut test for visual fermentation grading and packaging; (6) Batch split for laboratory analyses.

widely cultivated in Colombia: CNN-51, ICS-95, and TCS-01 (Rosas-Patiño et al., 2025; Rodriguez-Medina et al., 2019).

Stage 2 − Seed Extraction: The harvested pods were opened, and cocoa beans were manually extracted by trained personnel to maintain consistency across batches.

Stage 3 − Fermentation: The extracted beans were placed in wooden boxes lined with banana leaves and fermented under controlled temperature conditions, following regional fermentation practices that incorporate natural microbial activity and periodic mixing to ensure uniformity. To ensure variation in fermentation levels across batches, four different durations were used: 96, 144, 192, and 264 hours. This variation was critical for assessing the influence of fermentation on physicochemical properties.
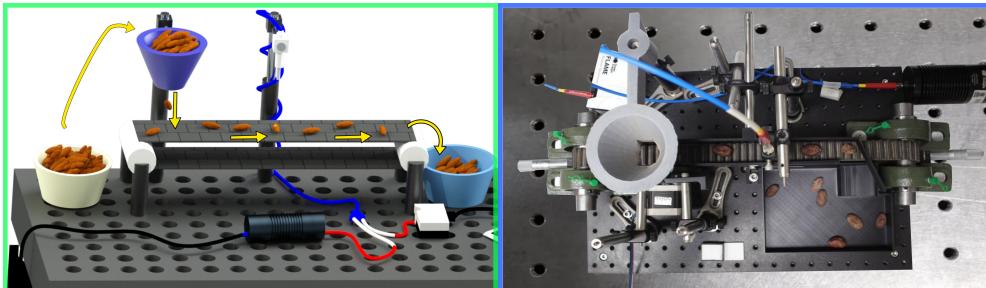
Stage 4 − Drying: After fermentation, the beans were sun-dried until reaching a constant weight. Drying was performed under controlled conditions in accordance with Colombian regulation NTC 1252:2021, which stipulates a maximum moisture content of 10% for dried cocoa beans.

Stage 5 − Guillotine Cut Test: To assess fermentation level, each batch were evaluated using the NTC 1252:2021 standard. A Swiss guillotine was used to bisect 100 beans per batch. The beans were classified into three categories: premium (P, well fermented), standard (S, partially fermented), and ordinary (O, underfermented). The fermentation ratio was calculated as $V_{\text{fer}} = (P + S)/N$, where N is the total number of beans, $P, S \geq 0$, and $P + S \leq N$. This allowed labeling of each batch with a ground-truth fermentation level.

Stage 6 − Batch Division: Posteriorly, each 1.5 kg batch was divided into two portions: 600 g were set aside for chemical analysis, while 900 g were allocated for spectral signature acquisition. The spectral portion was further split into training and testing subsets (70% and 30%, respectively), ensuring physical separation to avoid data leakage.

Stage 7 – Chemical Analysis: The 600 g subsample reserved for chemical analysis was processed at the *"Ciencia y Tecnología de Alimentos"* (CICTA, 2025) research group from the Universidad Industrial de Santander (UIS). Moisture content was measured using a gravimetric oven-drying method, cadmium concentration was determined via microwave-assisted atomic absorption spectroscopy, and polyphenol content was quantified using the Folin-Ciocalteu colorimetric assay as explained in the methodology part.

Spectral Optical System

Fig. 3: Automated system for spectral acquisition of cocoa beans. The left panel shows the system design, and the right panel shows its actual construction viewed from above: the conveyor belt, the light source, and the spectrometer for noninvasive data acquisition.

A dedicated optical system was designed for the spectral acquisition of Colombian cocoa beans, operating at a throughput of up to 113 beans per minute under controlled lighting and acquisition parameters, as shown in Fig. 3. Each bean is positioned within a fixed measurement zone, where it is illuminated by a halogen lamp and scanned in both visible (VIS: 400–700 nm) and near-infrared (NIR: 700–2500 nm) spectral ranges.
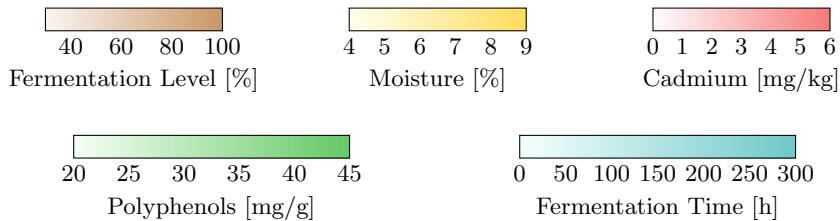
The system employs a bifurcated optical fibre that simultaneously delivers illumination and collects reflected light. The fibre is mounted at a fixed vertical distance of 14 cm from the conveyor belt. This configuration results in an average surface illumination of 59%, with stable acquisition geometry across measurements. The spectrometer connection is modular, allowing rapid switching between VIS and NIR devices without requiring adjustments to the optical alignment or acquisition setup.

Cocoa Bean Dataset

To build the cocoa dataset, we processed 20 cocoa bean batches following the presented protocol. These batches were collected over nine months, where four batches were gathered per sampling date. A detailed summary of the processed batches' physicochemical properties is provided in Table 2. The table presents the fermentation level, moisture content, cadmium concentration, polyphenol content, and fermentation time for each

Table 2: Summary of physicochemical properties of Colombian cocoa bean samples grouped by date of receipt. The table includes fermentation level, moisture content, cadmium concentration, polyphenol content, and fermentation time for each lot. Grouping by date highlights variations across reception periods, potentially reflecting differences in post-harvest handling, origin, or environmental conditions.

#	Fermentation Level [%]	Moisture [%]	Cadmium [mg/kg]	Polyphenols [mg/g]	Fermentation Time [h]
Date of Receipt: 15/04/2024					
1	60	5.12	2.14	41.30	96
2	66	4.93	1.29	34.24	144
3	84	4.80	1.25	40.38	264
4	92	4.75	1.23	39.81	264
Date of Receipt: 27/06/2024					
5	73	4.79	2.57	32.85	144
6	85	4.94	1.69	39.75	110
7	94	4.56	2.19	28.78	216
8	96	5.09	1.73	23.74	252
Date of Receipt: 22/10/2024					
9	66	5.82	5.55	27.66	96
10	94	5.68	4.80	23.05	144
11	96	5.67	3.65	25.09	216
12	100	5.67	3.14	22.76	252
Date of Receipt: 22/11/2024					
13	30	6.68	<0.09	35.41	30
14	45	6.60	<0.09	37.29	45
15	70	6.87	<0.09	36.48	70
16	70	8.44	<0.09	25.90	70
Date of Receipt: 18/01/2025					
17	44	4.78	2.65	39.16	30
18	70	4.88	2.72	16.69	45
19	87	5.01	2.24	35.77	70
20	96	4.16	1.7	37	70

40 60 80 100
Fermentation Level [%]

4 5 6 7 8 9
Moisture [%]

0 1 2 3 4 5 6
Cadmium [mg/kg]

20 25 30 35 40 45
Polyphenols [mg/g]

0 50 100 150 200 250 300
Fermentation Time [h]

batch, organized by the date of receipt. This grouping highlights potential variations across different reception periods, which may reflect differences in post-harvest handling, environmental conditions, or the origin of the beans. The cocoa batches with the date of receipt 18/01/2025 are used only for evaluation purposes, while other batches are used for training regression-based models.
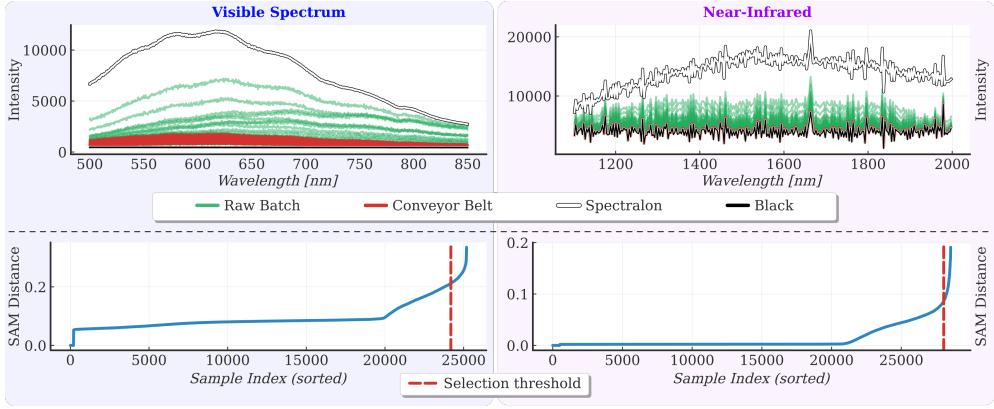
Table 3: Summary of physicochemical properties of cocoa bean samples grouped by regions. The table includes fermentation level, moisture content, cadmium concentration, and polyphenol content for each lot. Grouping by region highlights variations across Latin America zones, potentially reflecting differences in post-harvest handling, origin, or environmental conditions.

Date of Receipt	Region	Country	Fermentation Level [%]	Moisture [%]	Cadmium [mg/kg]	Polyphenols [mg/g]
01/06/2025	Santander	Colombia	96	4.16	1.7	37
06/06/2025	Huila	Colombia	60	5.02	0.84	35.21
10/06/2025	Putumayo	Colombia	96	5.02	<0.09	28.80
12/06/2025	Cusco	Peru	100	5.36	2.52	33.94

Following the same experimental protocol, three additional cocoa bean batches were collected from Huila and Putumayo in southern Colombia and from the Cusco region in Peru, as shown in Tab. 3. These additional batches were introduced in the test dataset to assess cocoa bean quality under diverse harvest and post-harvest conditions, such as differences in climate, fermentation practices, and drying methods. This variability is expected to enrich the analysis and improve the generalizability of the learning-based spectral regression models for cocoa bean quality assessment.

After acquisition with the proposed optical system, the spectral signatures were analyzed to identify the wavelength ranges of interest and remove regions affected by noise extremes. The VIS spectrum was defined from 500–850 nm, and the NIR spectrum from 1100–2000 nm. To isolate cocoa bean spectral signatures from background signals introduced by the conveyor belt, as shown in Fig. 4, a Spectral Angle Mapper (SAM) (Kruse et al., 1993) distance metric was computed between known conveyor belt spectra and all raw samples in the intensity domain. This allowed quantitative discrimination between sample and background spectra. A threshold-based selection strategy was applied to retain only the most distinct spectral signatures. Specifically, the n samples with the highest SAM distances were selected: $n = 1000$ for VIS and $n = 500$ for NIR. This filtering step aimed to exclude the most representative conveyor belt contributions and retain the cocoa-specific signatures for further analysis.

Spectral variability within the selected cocoa samples was assessed across acquisition batches. The principal Component Analysis (PCA) revealed substantial spectral overlap between batches, as shown at the left of Fig. 5, indicating a high degree of correlation and limited spectral separation. Such redundancy is undesirable for robust downstream modeling of physicochemical properties. Hence, to reduce variance and enhance discriminability, a bootstrapping (Zoubir and Iskandler, 2007) approach was adopted, consistent with Colombian standard NTC 1252:2021. For each acquisition batch, 50 spectral signatures were randomly selected and averaged spatially. This process was repeated 1000 times for VIS and 2000 times for NIR, generating representative mean profiles while attenuating intra-batch variability. The resulting datasets exhibited markedly improved spectral differentiation across batches, particularly in the 500–550 nm and 1950–2000 nm ranges for VIS and NIR, respectively,

Fig. 4: **Raw spectral dataset.** Visible (left) and near-infrared (right) intensity spectra for cocoa (green), background (red), and reference targets (black, white). Bottom: SAM distances to conveyor references; red dashed line indicates selection threshold.

as shown at the right of Fig. 5. Corresponding PCA projections with 2 components confirmed enhanced clustering and separation, facilitating more robust interpretation and subsequent modeling.
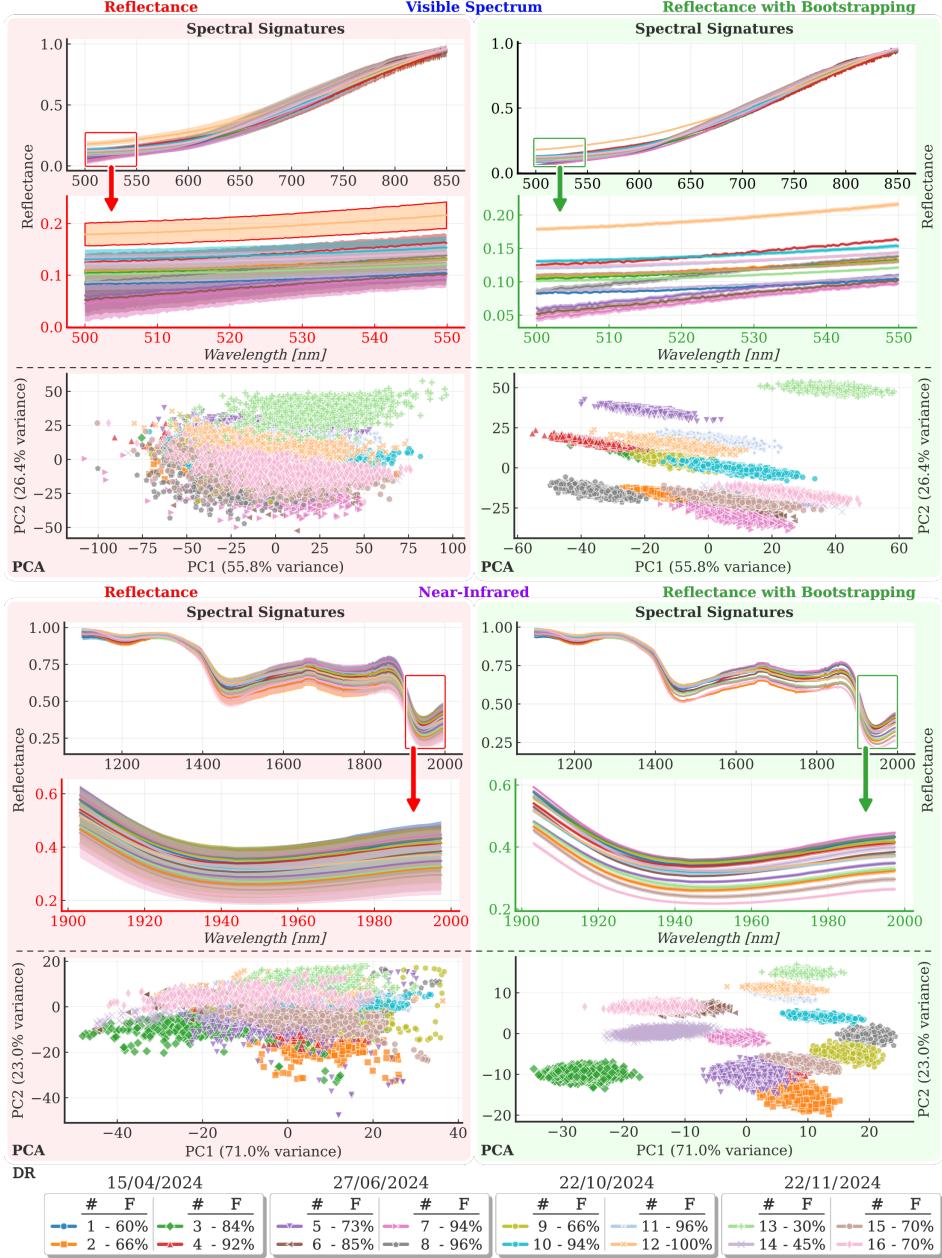
Regression-based Cocoa Bean Quality Assessment

A variety of machine learning and deep learning models (Smola and Schölkopf, 2004; Breiman, 2001; Hochreiter and Schmidhuber, 1997; Vaswani et al., 2017) were assessed to predict four physicochemical attributes of cocoa beans: cadmium concentration, moisture content, fermentation level, and polyphenol content. The models were trained and evaluated using spectral data from the VIS and NIR regions. Table 4 presents a summary of the results in terms of coefficient of determination ($\mathcal{R}^2$) and mean squared error (MSE).

Machine learning models such as SVR, RFR, and KNNR produced lower and more variable predictive results. SVR and RFR yielded $\mathcal{R}^2$ values below 0.7 for most targets. KNNR showed better performance for cadmium and moisture using NIR data, with $\mathcal{R}^2$ values of 0.9081 and 0.9740, respectively. However, their results were less consistent than those obtained with deep learning models.

Generalization on Geographically Independent Samples

To evaluate model generalization, performance was tested on external cocoa batches sourced from various regions as shown in Table 3. Reference values were obtained from standardized laboratory analyses. As shown in Table 5, predictions were computed using the best model of ML and DL models with both VIS and NIR spectra. In the international batch from Peru, the most accurate polyphenol prediction was obtained with VIS-DL (29.43 mg/g, reference: 28.80 mg/g). Cadmium was best estimated using VIS-DL (0.08 mg/kg, reference: 0.09 mg/kg). Fermentation was also best predicted by

Fig. 5: **Raw reflectance vs. reflectance with bootstrapping.** Top: reflectance spectra; bottom: PCA projections of cocoa samples from four 2024 dates. Raw data (left) show overlap, especially at 500–550 nm. Bootstrapped averaging (right) reveals clearer batch clustering. DR: Date of Receipt.

Table 4: **Comparison of $\mathcal{R}^2$ and MSE for Deep Learning and Machine Learning Models using the designed dataset.** The best and second-best performances per column are highlighted in green (bold) and blue (underline), respectively, per model category (Deep Learning and Machine Learning), regardless of spectral range.

Method	Range	$\mathcal{R}^2$				MSE			
		Cadmium	Moisture	Ferment.	Polyph.	Cadmium	Moisture	Ferment.	Polyph.
Deep Learning									
S-Net	VIS	**0.9605**	0.9665	0.8457	0.8684	**0.045**	0.032	0.054	0.043
	NIR	0.9449	0.9522	0.8718	0.8619	0.050	0.034	0.052	0.041
CNN	VIS	0.8308	0.4032	0.8128	0.5289	0.134	0.128	0.097	0.115
	NIR	0.8232	0.7697	0.7863	0.7180	0.140	0.100	0.095	0.098
LSTM	VIS	0.9070	0.9555	**0.8718**	0.8205	0.088	0.038	**0.050**	0.045
	NIR	0.8696	0.9681	0.8429	0.8457	0.098	0.036	0.055	0.047
Transf.	VIS	0.9558	0.9824	0.7965	0.8900	0.048	0.025	0.062	0.039
	NIR	0.9590	**0.9926**	0.7671	**0.9006**	0.046	**0.022**	0.065	**0.037**
Machine Learning									
SVR	VIS	0.6321	0.3034	0.6168	0.3214	0.250	0.270	0.190	0.230
	NIR	0.6729	0.3167	0.6058	0.3407	0.240	0.265	0.185	0.225
RFR	VIS	0.6020	0.8592	0.6670	0.7992	0.270	0.070	0.140	0.120
	NIR	0.6175	0.8701	0.6707	**0.8005**	0.265	0.068	0.138	0.118
KNNR	VIS	0.8901	0.9365	0.7232	0.7689	0.090	0.050	0.110	0.098
	NIR	**0.9081**	**0.9740**	**0.7372**	0.8002	**0.085**	**0.045**	**0.108**	**0.095**

VIS-DL (93.64%, reference: 96.00%). For moisture, NIR-DL yielded the closest result (4.97%, reference: 5.02%).

Fig. 6 illustrates the prediction results for each region and configuration. These findings indicate that the models trained with data from Santander were able to produce accurate estimations when applied to batches from other origins, under varying post-harvest conditions. This result suggests that the spectral and compositional variability of Santander cocoa encompasses representative patterns that are also present in beans from other regions, enabling the models to generalize effectively. In addition, the consistency of predictions across different physicochemical properties demonstrate the robustness of the proposed approach. Although some fluctuations appear depending on the attribute and the spectral domain (VIS or NIR), the overall trend confirms that both machine learning and deep learning models capture key features that remain stable across origins.
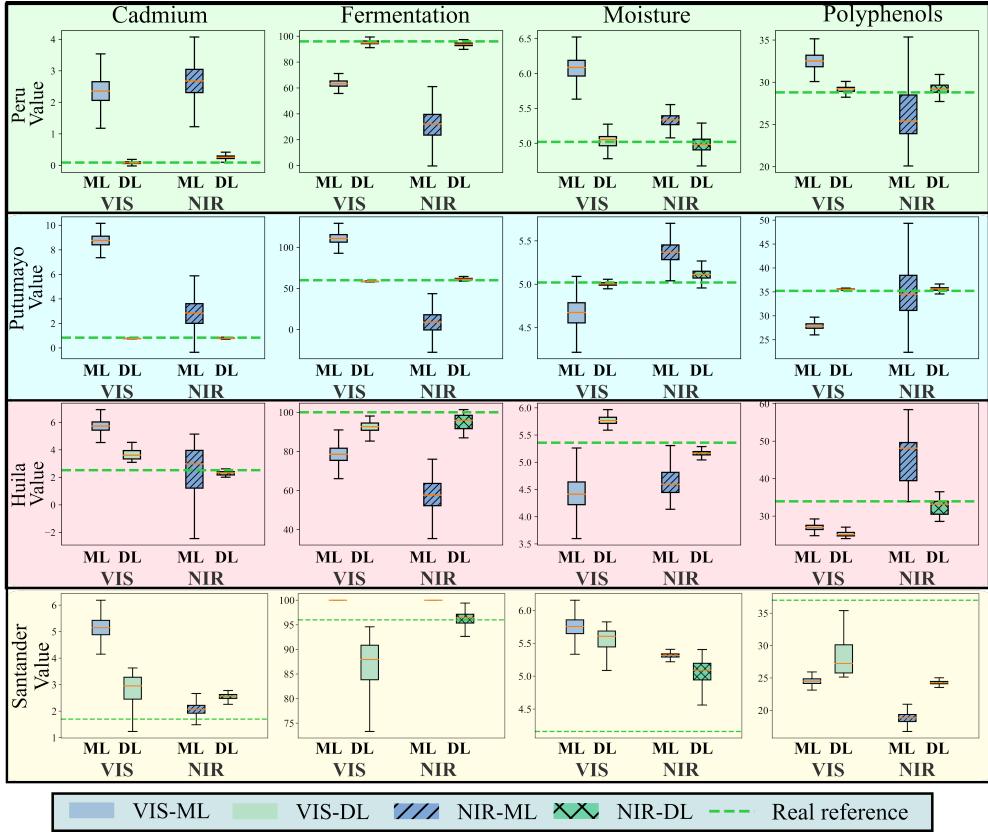
Conclusion

This study presents a scalable and non-destructive approach for assessing cocoa bean quality using VIS and NIR spectroscopy with learning-based regression models. The proposed models demonstrated predictive performance for key physicochemical properties, outperforming traditional ML techniques. These findings underscore the

Table 5: **Generalization on different region cocoa batches.** The closest prediction to the real lab value is highlighted in green, the second closest in blue.

Batch	Property	Chemical Labels	VIS ML	NIR ML	VIS DL	NIR DL
					Predicted of our model	
Peru	Cadmium	0.09	2.36	2.68	**0.08**	0.26
Peru	Fermentation	96.00	63.46	31.06	**93.64**	93.43
Peru	Moisture	5.02	6.08	5.33	5.07	**4.97**
Peru	Polyphenols	28.80	32.55	26.39	29.43	**29.28**
Putumayo	Cadmium	0.84	8.76	2.76	0.77	**0.80**
Putumayo	Fermentation	60.00	99.91	8.68	**58.44**	61.67
Putumayo	Moisture	5.02	4.67	5.36	**5.00**	5.11
Putumayo	Polyphenols	35.21	27.89	**35.02**	35.58	35.54
Huila	Cadmium	2.52	5.73	**2.59**	3.66	2.30
Huila	Fermentation	100.00	78.53	57.17	92.49	**95.19**
Huila	Moisture	5.36	4.44	4.62	5.77	**5.16**
Huila	Polyphenols	33.94	27.01	45.66	25.22	**32.55**
Santander	Cadmium	1.70	2.07	2.79	2.54	**1.70**
Santander	Fermentation	96.00	99.98	86.59	**96.01**	100.00
Santander	Moisture	4.16	5.75	5.31	5.54	**5.07**
Santander	Polyphenols	37.00	**28.17**	24.33	24.52	28.07

potential of DL architectures, particularly Transformer and SpectralNet, for modeling spectral data in agricultural applications. Beyond overall outperformance, the study identifies property–range leaders: Transformer achieves accuracy for moisture and polyphenols using NIR spectra; SpectralNet provides the best cadmium estimation using VIS spectra; and LSTM attains the strongest fermentation prediction in VIS, enabling a property-specific deployment strategy. Moreover, the proposed acquisition and spectral curation pipeline, combining background discrimination and intra-batch bootstrapped averaging, improves inter-batch separability and stabilizes regressors, constituting an enabler for in-line quality control.

The integration of VIS–NIR spectroscopy with learning-based regression enables rapid, non-invasive evaluation of cocoa quality. This approach provides an alternative to destructive laboratory analyses, reducing time, cost, and resource demands while maintaining compliance with international quality standards. Generalization tests conducted on cocoa samples from regions such as Huila and Putumayo in Colombia, and Cusco in Peru, confirmed that models trained on Santander data retain performance across varied geographic, environmental, and post-harvest conditions. These results validate the transferability and reliability of the models in real-world scenarios. Crucially, the cross-regional evaluation evidences geographic transfer with errors competitive relative to local laboratory references, indicating that a model trained in one producing area can be ported to distinct agroecological contexts with minimal recalibration. This dataset will be made available to support future research aimed at improving the prediction of cocoa bean physicochemical properties from spectral

Fig. 6: Predicted versus laboratory values for external cocoa batches using ML and DL models with VIS and NIR spectra. The green dashed line indicates the reference value.

information. Collectively, these findings position VIS–NIR plus DL as a viable, scalable substitute for destructive assays in post-harvest quality control, enabling faster decision cycles and broader coverage at lower operational cost.

Despite these outcomes, certain challenges must be addressed to enable large-scale industrial adoption. Future work should focus on integrating spectral acquisition systems into on-site processing environments and developing hybrid models that combine spectral and image-based features. These enhancements could further improve model accuracy and adaptability, supporting the implementation of sustainable, data-driven quality control systems within the cocoa industry. Further research should also quantify calibration-transfer requirements across devices and sites, and assess active learning schemes to maintain accuracy under drift in post-harvest practices and seasonal shifts.

Acknowledgements

Declarations

- Conflict of interest/Competing interests: The authors declare no competing interests.
- Code and Data availability: The code and associated dataset are publicly available at https://github.com/PIgroupUIS/SpecCocoa_Regression_Physicochemical_Properties.git
- Author contribution: Author initials used in the table above correspond to the following full names: KC = Kebin Contreras, EM = Emmanuel Martinez, BM = Brayan Monroy, SA = Sebastian Ardila, CR = Cristian Ramirez, MC = Mariana Caicedo, HG = Hans Garcia, TG = Tatiana Gelvez-Barrera, JPJ = Juan Poveda-Jaramillo, HA = Henry Arguello, JB = Jorge Bacca.

Contribution	Authors
Conceptualization	KC, EM, BM, SA, CR, MC, HG, TG, JPJ, HA, JB
Methodology	KC, EM, SA, MB, JB
Data acquisition	KC, EM, BM, HG, SA
Software	KC, EM
Visualization	KC, EM, JB
Data Curation	KC, EM, BM
Writing – Original Draft	KC, EM, SA
Writing – Review & Editing	BM, TG, JB
Regulatory Standards Analysis	CR, MC
Supervision / General Review	HG, JPJ, JB

References

ICCO: International Cocoa Organization. May 2025 Quarterly Bulletin of Cocoa Statistics (2025). https://www.icco.org/may-2025-quarterly-bulletin-of-cocoa-stati

stics/

Kongor, J.E., Owusu, M., Oduro-Yeboah, C.: Cocoa production in the 2020s: Challenges and solutions. CABI Agriculture and Bioscience **5**(1), 102 (2024)

Huetz-Adams, F.: Cocoa Barometer Latin American Baseline. Südwind. Technical Report (2022). https://www.cocoabarometer.org/

ICCO: International Cocoa Organization. Fine or Flavour Cocoa (2023). https://www.icco.org/fine-or-flavor-cocoa/

Kongor, J.E., Hinneh, M., Walle, D., Afoakwa, E.O., Boeckx, P., Dewettinck, K.: Factors influencing quality variation in cocoa (theobroma cacao) bean flavour profile—a review. Food Research International **82**, 44–52 (2016)

Brito, E.S., García, N.H.P., Gallão, M.I., Cortelazzo, A.L., Fevereiro, P.S., Braga, M.R.: Structural and chemical changes in cocoa (theobroma cacao l) during fermentation, drying and roasting. Journal of the Science of Food and Agriculture **81**(2), 281–288 (2001)

De Vuyst, L., Weckx, S.: The cocoa bean fermentation process: from ecosystem analysis to starter culture development. Journal of Applied Microbiology **121**(1), 5–17 (2016)

ICONTEC: Instituto Colombiano de Normas Técnicas y Certificación. Cacao en grano: Especificaciones y requisitos de calidad. Norma Técnica Colombiana, NTC 1252 (2021). https://tienda.icontec.org/gp-cacao-en-grano-especificaciones-y-requisitos-de-calidad-ntc1252-2021.html

INEN: Instituto Ecuatoriano de Normalización. NTE INEN 176:2021 – Cacao en grano seco fermentado – Requisitos. Norma Técnica Ecuatoriana (2021). https://anecacao.com/wp-content/uploads/2024/04/NTE-INEN-176-SEXTA-REVISION-1.pdf

MAPA: Ministério da Agricultura, Pecuária e Abastecimento. Instrução Normativa nº 38, de 1º de outubro de 2018: Classificação do cacau em amêndoas. Norma Técnica Brasileira (2018). https://sistemasweb.agricultura.gov.br/sislegis/action/detalhaAto.do?method=visualizarAtoPortalMapa&chave=250964455

INACAL: Instituto Nacional de Calidad del Perú. NTP-ISO 2292:2019: Granos de cacao — Muestreo (5ª edición). Norma Técnica Peruana basada en ISO 2292:2017 (2019). https://camcafeperu.com.pe/admin/recursos/normas/NTP-ISO%202292_2019%20Granos%20de%20cacao.%20Muestreo.%205%C2%AA%20Edici%C3%B3n.pdf

Samanta, S., Sarkar, T., Chakraborty, R., Rebezov, M., Shariati, M.A., Thiruvengadam, M., Rengasamy, K.R.: Dark chocolate: An overview of its biological activity, processing, and fortification approaches. Current Research in Food Science **5**,

1916–1943 (2022)

Abt, E., Robin, L.P.: Perspective on cadmium and lead in cocoa and chocolate. Journal of agricultural and food chemistry **68**(46), 13008–13015 (2020)

Araujo, L.S., Tapia, W., Ortiz, A.V.: Verification of the atomic absorption spectroscopy with graphite furnace analytical method for the quantification of cadmium in cocoa almonds (theobroma cacao). La Granja **31**(1), 56 (2020)

Ducki, S., Miralles-Garcia, J., Zumbé, A., Tornero, A., Storey, D.M.: Evaluation of solid-phase micro-extraction coupled to gas chromatography–mass spectrometry for the headspace analysis of volatile compounds in cocoa products. Talanta **74**(5), 1166–1174 (2008)

Cain, N., Alka, O., Segelke, T., Wuthenau, K., Kohlbacher, O., Fischer, M.: Food fingerprinting: Mass spectrometric determination of the cocoa shell content (theobroma cacao l.) in cocoa products by hplc-qtof-ms. Food chemistry **298**, 125013 (2019)

Niemenak, N., Eyamo, J., Onomo, P., Youmbi, E.: Physical and chemical assessment quality of cocoa beans in south and center regions of cameroon. Journal of Syllabus Reviews Science Series **5**(2014), 27–33 (2014)

Sánchez, K., Bacca, J., Arévalo-Sánchez, L., Arguello, H., Castillo, S.: Classification of cocoa beans based on their level of fermentation using spectral information. TecnoLógicas **24**(50), 172–188 (2021)

Hashimoto, J.C., Lima, J.C., Celeghini, R.M., Nogueira, A.B., Efraim, P., Poppi, R.J., Pallone, J.A.: Quality control of commercial cocoa beans (theobroma cacao l.) by near-infrared spectroscopy. Food analytical methods **11**, 1510–1517 (2018)

Suarez, J., Espinosa, J., Contreras, K., Bacca, J.: Automated classification of cocoa bean fermentation levels using computer vision. In: 2025 XXV Symposium of Image, Signal Processing, and Artificial Vision (STSIVA), pp. 1–5 (2025). IEEE

Pinto, A., Deryck, A., Lima, G.V., Oliveira, A.C., Moura, F.G., Barbin, D.F., Pierna, J.A.F., Baeten, V., Rogez, H.: Advances in the individual authentication of cocoa beans: Vis/nir spectroscopy as a tool to distinguish fermented from unfermented beans and classify genotypes in the eastern amazonia. Food Control **164**, 110559 (2024)

Diaz-Delgado, L.C., Monroy, B., Bacca, J., Arguello, H.: Deep gaussian optical bandpass filter design for fermentation index estimation in cocoa beans. In: 2025 XXV Symposium of Image, Signal Processing, and Artificial Vision (STSIVA), pp. 1–5 (2025). IEEE

Teye, E., Anyidoho, E., Agbemafle, R., Sam-Amoah, L.K., Elliott, C.: Cocoa bean

and cocoa bean products quality evaluation by nir spectroscopy and chemometrics: A review. Infrared Physics & Technology **104**, 103127 (2020)

Sandorfy, C., Buchet, R., Lachenal, G.: Principles of molecular vibrations for near-infrared spectroscopy. Near-Infrared Spectroscopy in Food Science and Technology; Ozaki, Y., McClure, WF, Christy, AA, Eds, 11–46 (2007)

Gomez, N.A., Sanchez, K., Arguello, H.: Non-destructive method for classification of cocoa beans from spectral information. In: 2019 XXII Symposium on Image, Signal Processing and Artificial Vision (STSIVA), pp. 1–5 (2019). IEEE

Caporaso, N., Whitworth, M.B., Fowler, M.S., Fisk, I.D.: Hyperspectral imaging for non-destructive prediction of fermentation index, polyphenol content and antioxidant activity in single cocoa beans. Food chemistry **258**, 343–351 (2018)

Alvarado, M.C., Sanchez, P.D.C., Polongasa, S.G.N.: Emerging rapid and non-destructive techniques for quality and safety evaluation of cacao: recent advances, challenges, and future trends. Food Production, Processing and Nutrition **5**(1), 40 (2023)

Diaz-Delgado, L.C., Monroy, B., Bacca, J., Arguello, H.: Deep gaussian optical bandpass filter design for fermentation index estimation in cocoa beans. In: 2025 XXV Symposium of Image, Signal Processing, and Artificial Vision (STSIVA), pp. 1–5 (2025). IEEE

Ferraris, S., Meo, R., Pinardi, S., Salis, M., Sartor, G.: Machine learning as a strategic tool for helping cocoa farmers in côte d'ivoire. Sensors **23**(17), 7632 (2023)

Ashiagbor, G., Forkuo, E.K., Asante, W.A., Acheampong, E., Quaye-Ballard, J.A., Boamah, P., Mohammed, Y., Foli, E.: Pixel-based and object-oriented approaches in segregating cocoa from forest in the juabeso-bia landscape of ghana. Remote Sensing Applications: Society and Environment **19**, 100349 (2020)

Musah, S., Medeni, T.D., Soylu, D.: Assessment of role of innovative technology through blockchain technology in ghana's cocoa beans food supply chains. In: 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), pp. 1–12 (2019). IEEE

CICTA: Grupo de Investigación en Ciencia y Tecnología de Alimentos. Universidad Industrial de Santander. Ficha en la plataforma ScienTI (GrupLAC). Clasificación "A" por Minciencias (2025). https://scienti.minciencias.gov.co/gruplac/jsp/visualiza/visualizagr.jsp?nro=00000000003053

Kus, S., Marczenko, Z., Obarski, N., *et al.*: Derivative uv-vis spectrophotometry in analytical chemistry. Chem. Anal **41**(6), 889–927 (1996)

ICONTEC: Instituto Colombiano de Normas Técnicas y Certificación. NTC-EN

14084:2021. Productos alimenticios. Determinación de plomo, cadmio, cinc, cobre y hierro mediante espectrometría de absorción atómica (EAA) tras digestión en microondas. Norma técnica NTC-EN 14084:2021, Instituto Colombiano de Normas Técnicas y Certificación (March 2021). Adopción idéntica de la norma EN 14084:2003; Comité técnico 052: Cacao, Chocolate y Confitería. https://tienda.icontec.org/gp-productos-alimenticios-determinacion-de-elementos-traza-determinacion-de-plomo-cadmio-cinc-cobre-y-hierro-mediante-espectrometria-de-absorcion-atomica-eaa-despues-de-digestion-en-microondas-ntc-1en14084-2021.html

Kruse, F.A., Lefkoff, A.B., Boardman, J.W., Heidebrecht, K.B., Shapiro, A., Barloon, P., Goetz, A.F.: The spectral image processing system (sips)—interactive visualization and analysis of imaging spectrometer data. Remote sensing of environment **44**(2-3), 145–163 (1993)

Zoubir, A.M., Iskandler, D.R.: Bootstrap methods and applications. IEEE Signal Processing Magazine **24**(4), 10–19 (2007)

Smola, A.J., Schölkopf, B.: A tutorial on support vector regression. Statistics and computing **14**(3), 199–222 (2004)

Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)

Liu, Q., Cai, J., Fan, S.-Z., Abbod, M.F., Shieh, J.-S., Kung, Y., Lin, L.: Spectrum analysis of eeg signals using cnn to model patient's consciousness level based on anesthesiologists' experience. IEEE Access **7**, 53731–53742 (2019)

Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)

Martins, J., Guerra, R., Pires, R., Antunes, M., Panagopoulos, T., Brázio, A., Afonso, A., Silva, L., Lucas, M., Cavaco, A.: Spectranet–53: A deep residual learning architecture for predicting soluble solids content with vis–nir spectroscopy. Computers and Electronics in Agriculture **197**, 106945 (2022)

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

Rosas-Patiño, G., Sánchez-Castillo, V., Patiño-Torres, C.: Response of the cocoa (theobroma cacao) ccn-51 clone to fertilization under agroclimatic conditions of the colombian amazon trapezium. Revista UDCA Actualidad & Divulgación Científica **28**(1) (2025)

Rodriguez-Medina, C., Arana, A.C., Sounigo, O., Argout, X., Alvarado, G.A., Yockteng, R.: Cacao breeding in colombia, past, present and future. Breeding Science **69**(3), 373–382 (2019)