



OPEN

DATA DESCRIPTOR

A Multimodal hyperspectral dataset of cocoa beans with physicochemical annotation

Kebin Contreras¹, Mohamad Jouni², Mauro Dalla Mura^{2,3} & Jorge Bacca¹✉

Assessing cocoa bean quality using spectral information offers a noninvasive and objective alternative to traditional, often subjective and destructive, methods. However, progress has been limited by the lack of comprehensive datasets across multiple spectral resolutions. This work presents a new dataset capturing the spectral properties of cocoa beans at different spatio-spectral resolutions, enabling non-invasive quality assessment and scalable evaluation methodologies. It comprises 19 scenes acquired with four imaging devices under both open (invasive) and closed (non-invasive) conditions, along with corresponding physicochemical measurements. Data collection follows the Colombian standard NTC 1252:2021, which labels beans as well, partially, or poorly fermented. Global physicochemical properties—moisture, polyphenols, and cadmium—were measured using gravimetric analysis, UV-visible spectroscopy, and atomic absorption spectroscopy with microwave digestion. Hyperspectral images were obtained using four devices covering up to the 350–1000 nm spectral range. Statistical analysis shows the dataset distinguishes between cocoa quality levels under both open and closed conditions, supporting the development of automated classification methods.

Background & Summary

Cocoa, derived from the *Theobroma cacao* tree, is a grain used in the food industry, mainly grown in tropical regions of America, Africa and Asia. In Colombia, it marks a significant economic and cultural importance. For example, this grain represents 4.7% of Colombia's gross domestic product¹, and directly affects the lives of approximately 35,000 families². Furthermore, the economic impact of cocoa has the potential to increase by 10% with the export of *high-quality* cocoa beans, which attracts better prices on international markets³.

The quality of cocoa beans is influenced by key physicochemical properties such as (a) *fermentation*, (b) *polyphenol content*, (c) *presence of heavy metals*, and (d) *moisture*⁴. The Colombian standard NTC 1252:2021 standardizes the classification of cocoa by establishing criteria based on these physicochemical properties^{5,6}. Specifically, the standard NTC 1252:2021 requires selecting 100 cocoa beans, splitting each in half using the invasive “cut test”, and visually inspecting internal characteristics; specifically, color and texture. This test enables the classification of beans into three categories, as shown in Fig. 1: (a) *well-fermented*, characterized by a brown color with internal grooves and indentations; (b) *partially fermented*, showing an intermediate brown-violet coloration without defined internal structures; and (c) *poorly fermented*, with a predominantly purple color and smooth, unstructured interior.

This classification provides essential information for assessing cocoa quality and guiding post-harvest practices. This method requires a trained user to visually assess the level of fermentation by counting well- and partially fermented beans out of a total of 100. If the sum of the well- and partially fermented beans exceeds 87 beans and the moisture content is below 10%, the batch is considered high quality⁵. However, this method has notable drawbacks. It can be *subjective*, as it depends on visual inspection, and more importantly, it requires destroying the sample, making it an *invasive* procedure.

Challenges. *Hyperspectral Imaging.* *Hyperspectral imaging (HSI)* offers several advantages for evaluating the physicochemical properties of cocoa beans. Primarily:

¹Universidad Industrial de Santander, 68002, Bucaramanga, Colombia. ²Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-Lab, 38000, Grenoble, France. ³Institut Universitaire de France (IUF), 75231, Paris, France. ✉e-mail: jbacquin@uis.edu.co

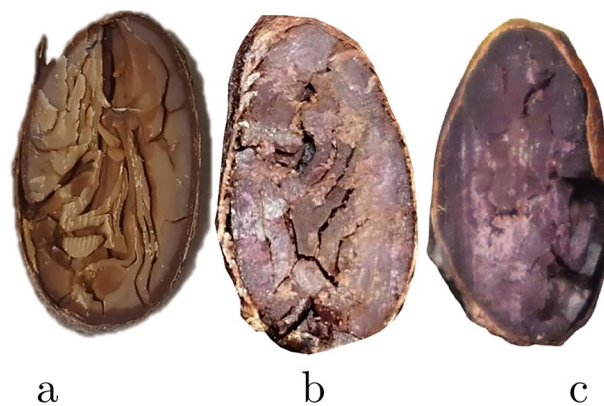


Fig. 1 Visualization of the three fermentation levels of cocoa beans: **(a)** *Well-fermented* bean: shows a light to dark brown color and has internal grooves and indentations. **(b)** *Partially fermented* bean: shows an intermediate color and less defined veining, including brown and purple hues. **(c)** *Poorly fermented* bean: shows a purple color and lacks grooves and indentations.

- **Non-destructive analysis:** HSI enables rapid acquisition of high-resolution spectral and spatial data from cocoa bean surfaces without damaging the samples. Typically, HSI systems capture subtle variations in chemical composition, such as moisture, fat, and fermentation indicators⁵.
- **Objective quality assessment:** HSI supports the development of numerical discriminant models (e.g. principal component analysis (PCA), support vector machine (SVM)) for spectral data analysis, providing an objective and reproducible quality assessment of cocoa beans.
- **Rich spectral data per sample:** Unlike single-pixel spectrometers, which produce one spectral signature per acquisition, imaging spectrometers capture spectral data across many pixels in a single image. This provides multiple spectral signatures per cocoa bean, enhancing sample characterization within a single acquisition.

However, most practical applications so far have relied heavily on datasets derived from African cocoa beans, focusing mainly on spectral signatures while neglecting spatial information^{7–10}. This dependence poses limitations for other cocoa varieties, such as those from Colombia, due to the lack of geographically diverse datasets. The scarcity of region-specific data hinders the development and deployment of robust predictive models addressing Colombian cocoa. For instance, existing datasets do not properly capture spectral variations driven by Colombia's distinct climate, soil composition, and post-harvest fermentation practices. In contrast, our work introduces the first dataset that synchronously integrates hyperspectral imaging with physicochemical measurements for Colombian cocoa beans. This integration captures unique region-specific characteristics: volcanic Andean soils influence cadmium absorption features, high-altitude tropical climates shape moisture and polyphenol development during fermentation, and traditional fermentation practices with banana leaf coverings generate internal structural variations detectable through hyperspectral analysis.

Physicochemical Properties. Moreover, there are currently no publicly available HSI databases that integrate spectro-spatial information with *physicochemical properties* such as fermentation, moisture, polyphenols, and cadmium content.

This lack of integrated spectral and physicochemical information makes it difficult to develop machine-learning models tailored to cocoa quality assessment. Therefore, it is necessary to acquire spectral information on dried cocoa beans, along with their physicochemical properties, under the specific environmental and geographical conditions of Colombia.

Contribution. To bridge this gap, we propose the Colombian Cocoa Spectral (CoCooSpec) Dataset¹¹. CoCooSpec comprises of a comprehensive database that synchronizes hyperspectral images with physicochemical properties, setting a new standard in cocoa research. The images were labeled according to fermentation levels following the Colombian NTC 1252:2021 standard. Physicochemical properties were determined using analytical methods: moisture content was measured through gravimetric analysis; polyphenol concentration was quantified using Ultraviolet-Visible (UV-Vis) spectrophotometry using the ABTS (2,2'-azino-bis (3-ethylbenzothiazoline-6-sulfonic acid) radical cation decolorization assay) and DPPH (2,2-diphenyl-1-picrylhydrazyl free radical scavenging assay) methods; and cadmium levels were analyzed via atomic absorption spectroscopy assisted by microwave digestion, following UNE-EN 15550:2018 and EU Regulation No. 488/2014. By synchronizing HSI with physicochemical properties for the first time, this database enables the development of AI-driven models that provide objective, scalable, and reproducible cocoa quality assessments. This dataset also has the potential to develop non-invasive methods of analysis, allowing cocoa bean quality to be estimated without destroying the beans. This opens the door for implementing non-invasive quality control measures directly in production lines.

CoCooSpec¹¹ comprises 19 different scenes, each corresponding to a batch of cocoa bean samples. The spatial and spectral information of each scene has been recorded by three well-known commercial hyperspectral cameras, namely, Ultris SR5, Toucan, and Specim IQ, each of which has distinct spatial, spectral, and temporal

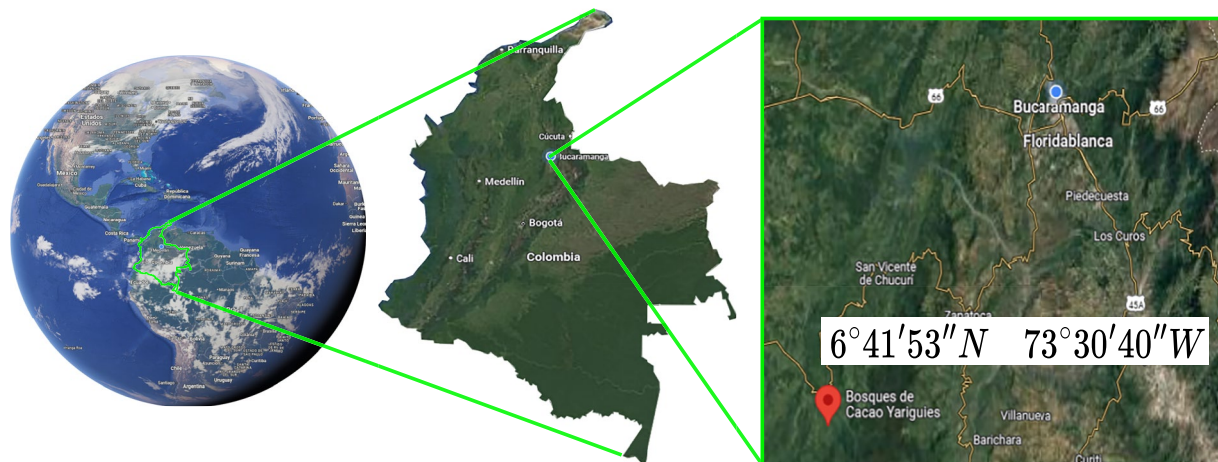


Fig. 2 Geographical location of the farm where the cocoa bean samples were collected in El Carmen de Chucurry, Santander, Colombia.



Fig. 3 Post-harvest process of cocoa pods, including (a) harvesting, (b) bean extraction, (c) fermentation, and (d) drying.

resolutions. The corresponding hyperspectral images are complemented by RGB images acquired by the conventional EOS M50 camera, providing a representation of the visual properties of the beans. Each scene exhibits two captures of the corresponding cocoa batch, one when the beans are closed (conserved) and another when they are cut open (destroyed), allowing an analysis of the visual and structural characteristics of the beans under various states of fermentation. Additionally, CoCaoSpec¹¹ includes data on the physicochemical properties of the cocoa bean batches such as fermentation, cadmium, polyphenols, and moisture. By synchronizing the HSI data with physicochemical properties for the first time, we aim to establish a new standard in cocoa research.

Furthermore, the proposed database includes metadata descriptors of the acquisition campaign, camera characteristics, cocoa bean fermentation labels carried by a field expert, cocoa bean bounding box annotations, and some interesting pixel positions in each scene. Finally, our work is accompanied with data loading and calibration utilities that can be used to preprocess the acquired hyperspectral images, such as flat-field correction and data normalization.

Methods

Cocoa Samples Collection. Cocoa beans were harvested between January 2024 and December 2024 on a farm located in the municipality of El Carmen de Chucurry, Santander, Colombia, at the geographical coordinates $6^{\circ}41'53''N-73^{\circ}30'40''W$, as shown in Fig. 2. This farm is part of the Association of Farmers Neighbors of the Serranía de Los Yariquíes National Natural Park (ASOCAPAYARI), a cocoa-producing region in Colombia. The beans were sourced from a farmer who is a member of the ASOCAPAYARI. The beans were extracted from pods within the same hectare and subsequently fermented by the farmer, who has 20 years of experience in post-harvest handling of cocoa.

Figure 3 illustrates the post-harvest process of cocoa pods from CNN-51, ICS-95, and TCS-01 varieties. In Fig. 3(a), the harvesting of the pods is shown, which are manually collected at their optimal maturity stage. In Fig. 3(b), the extraction of the cocoa beans from the pods is depicted. In Fig. 3(c), the fermentation process is presented; at this stage, the cocoa beans develop their characteristic flavor and aroma. The beans are placed in wooden boxes covered with banana leaves to facilitate fermentation. Finally, in Fig. 3(d), the drying process is shown in which the fermented beans are spread out under the sun to reduce their moisture content.

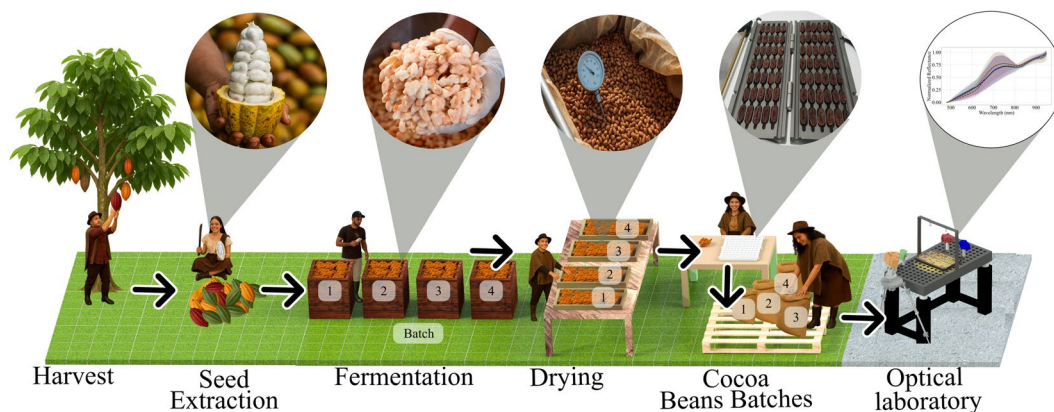


Fig. 4 Overall workflow of the dataset acquisition process, from harvesting and seed extraction to fermentation, drying, batch preparation, and final optical laboratory analysis.



Fig. 5 Cameras used in the experiment: (a) Specim IQ, (b) Toucan, (c) Ultris SR5, and (d) EOS M50. Photos obtained from the manufacturer's website.

The Fig. 4 summarizes the complete workflow for dataset acquisition. The process begins with cacao harvesting and seed extraction, followed by fermentation and drying under controlled conditions. Once dried, the beans are grouped into batches and transported to the laboratory, where the final optical analyses are performed.

Dataset Composition and Optical Setup. The HSI and RGB images of dried cocoa beans were acquired using four different cameras, covering the visible and near-infrared spectral ranges, up to [350, 1000] nm. The image acquisition process was performed using the following devices:

- EOS M50 (classic RGB camera),
- Toucan (multispectral video camera),
- Ultris SR5 (hyperspectral video camera),
- Specim IQ (hyperspectral scan-based camera),

whose characteristics are detailed below. Figure 5 illustrates the different cameras used for data acquisition.

Toucan Snapshot Multispectral Camera. The Toucan camera, developed by Silios, is a snapshot multispectral sensor covering the visible and near-infrared, i.e., covering wavelengths from 420 nm to 870 nm. The raw data is a mosaicked grayscale image, with a raw spatial resolution of 2048×2048 pixels, and a 4×4 filter array composed of 10 spectral filters. After demosaicking, a multispectral image cube is reconstructed, with dimensions $2048 \times 2048 \times 10$. The advantages of this camera is that it has a high spatial resolution, and that it is a real-time imager with a maximum frame rate of 65 Hz of raw data (without demosaicking), or 20 Hz of multispectral image cubes with online (live) demosaicking. The disadvantage is that it has a low spectral resolution, sampling only 10 bands in the domain of interest.

Ultris SR5 Snapshot Hyperspectral Camera. The Cubert Ultris SR5 camera is a snapshot hyperspectral sensor covering the visible and near-infrared, i.e., covering wavelengths from 450 nm to 850 nm. As a result, the Ultris SR5 camera captures raw hyperspectral image cubes, with dimensions $290 \times 275 \times 51$. The advantages of this camera is that it has a high spectral resolution, sampling the spectral range at 8 nm, and that it is a real-time imager with a maximum frame rate of 15 Hz of hyperspectral image cubes. The disadvantage is that it has a very low spatial resolution, and requires a relatively high exposure time, which can limit the frame rate under poorly illuminated conditions. Moreover, Ultris SR5 images require flat-field correction as a preprocessing step.

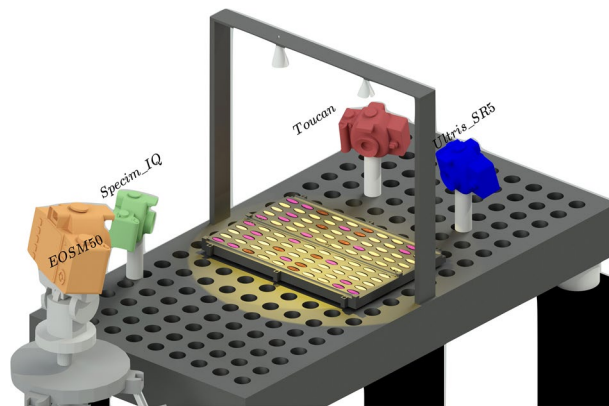


Fig. 6 Experimental setup of the cameras, lighting system, and cocoa bean guillotine.

Specim IQ Push-broom Hyperspectral Camera. The Specim IQ camera is a push-broom (scan-based) hyperspectral commercial camera, covering the visible and near-infrared, i.e., covering wavelengths from 350 nm to 1000 nm. The Specim IQ camera scans hyperspectral images of the scene with dimensions $512 \times 512 \times 204$. The advantages of this camera is that it has both a (relatively) high spatial and spectral resolution, sampling the spectral range at 7 nm, and thus provides more information on the spatial and spectral properties of the scene than the Toucan or the Ultras SR5 cameras. However, the disadvantage is that it is not a real-time snapshot imager, and may take up to 3 minutes to scan a single scene, which requires the targets to be strictly static while acquiring a single hyperspectral image.

Physicochemical Properties. The quality of cocoa beans is evaluated according to their physicochemical properties, such as the level of fermentation, moisture content, polyphenol concentration, and the presence of cadmium. These properties are analyzed using both local and global labeling approaches. At the local level, each individual cocoa bean is labeled with its specific fermentation level. At the global level, scene-based values are assigned for moisture, fermentation, polyphenol concentration, and cadmium content. Polyphenol concentration is quantified using UV-Vis spectrophotometry, employing techniques such as ABTS and DPPH assays to estimate antioxidant content based on absorbance variations. Cadmium concentration is measured with microwave digestion-assisted atomic absorption spectroscopy (AAS), which involves digesting samples with strong acids under high pressure and temperature before analyzing cadmium levels via atomic absorption spectrometry. The fermentation level is assessed through the cut test, which involves slicing the beans to examine their internal structure and color. Moisture content is determined using the gravimetric method, where cocoa samples are dried in an oven at controlled temperatures until a constant weight is reached.

Experimental Design. This section describes the experimental setup, including the configuration of HSI and RGB cameras, the cocoa bean guillotine used for the cut test, and the image acquisition procedure. Additionally, the physicochemical analysis methods and computational processing techniques applied to the dataset are detailed.

The experimental setup was designed to ensure consistent acquisition of HSI and RGB images of cocoa beans under controlled conditions. The setup, illustrated in Fig. 6, consists of the four aforementioned cameras: EOS M50, Toucan, Ultras SR5, and Specim IQ. Briefly speaking, a sample of cocoa beans is placed into the slots of a Swiss guillotine positioned at the center of the multimodal imaging platform. The guillotine is then used to cut the beans in half, transitioning their state from ‘closed’ to ‘open’.

Four batches of cocoa beans were received monthly from the farm, each weighing 1.5 kg with different fermentation times: Batch 1 with 96 hours, Batch 2 with 144 hours, Batch 3 with 192 hours, and Batch 4 with 264 hours. Each batch was divided into three sub-batches: one of 600 grams destined for the chemical laboratory, where the physicochemical properties of polyphenols, cadmium, and moisture were analyzed, and two remaining sub-batches totaling 900 grams, intended for spectral signature acquisition in the optical laboratory. Subsequently, the remaining 900 grams from each batch were divided into two additional sub-batches, which were sent to the optical laboratory for spectral signature acquisition. These sub-batches were distributed into two sets: train (70%) and test (30%), ensuring physical data separation.

For the acquisition of cocoa bean information during training and testing, the cut test was performed using a Swiss guillotine, as stipulated by the Colombian standard NTC 1252:2021. The process consists of four steps, as illustrated in Fig. 7.

Step 1: Placement of Cocoa Beans and Image Capture: In the first step, the guillotine is opened, and 50 cocoa beans are placed on one side of the guillotine. This process is repeated twice for a total of 100 beans, following the specifications of the norm. Before proceeding to the next step, hyperspectral images of the cocoa beans in their closed state are captured using the Toucan, Ultras SR5, and Specim IQ cameras.

Step 2: Cutting the Cocoa Beans: Once all the cocoa beans are placed on one side of the guillotine, it is closed, and the cutting blade is passed through the center, splitting the beans in half. This step reveals the internal characteristics and color of the beans, allowing for the identification of different cocoa classes.

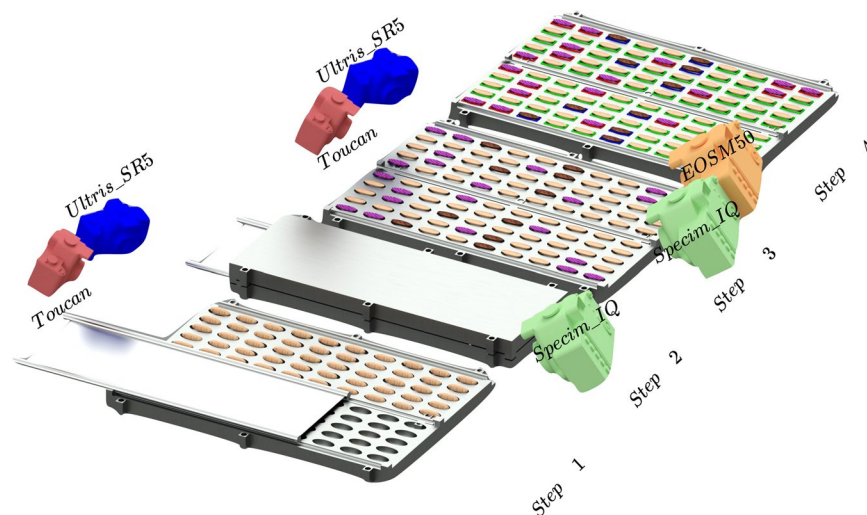


Fig. 7 Process of cutting cocoa beans using a Swiss guillotine, following the NTC 1252:2021 standard.

Step 3: Image Capture of Cocoa Beans: At this stage, hyperspectral and RGB images are taken to document the internal characteristics of the cocoa beans. The imaging process is performed using the four cameras: EOS M50, Toucan, Ultris SR5, and Specim IQ.

Step 4: Classification of Cocoa Beans and Recording of Results: After the imaging process, an expert classifies the cocoa beans into three categories according to the Colombian standard NTC 1252:2021: *Well-fermented*, *Partially fermented*, or *Poorly fermented*. The classification is based on the physical characteristics of the cocoa bean, its internal structure, and its color. This classification was performed by a cocoa expert with more than 20 years of experience in post-harvest processes, ensuring high reliability in the evaluation of fermentation degree of the cocoa samples. The percentage of fermentation (V_{fer}) is calculated according to Equation (1) as follows:

$$V_{\text{fer}} = \frac{WF + PF}{N} \times 100, \quad (1)$$

where N designates the total number of beans in the batch, WF stands for well-fermented beans, and PF denotes partially fermented beans. The three classification categories can be observed in Fig. 1.

A total of 19 scenes were captured, each containing a batch of 50 cocoa beans, with spectral images acquired in two conditions: closed beans (before the cut test) and open beans (after the cut test). Hyperspectral images in the closed state were captured using the hyperspectral cameras. After the cut test, a second acquisition process was performed, capturing hyperspectral and RGB images of the open beans to document their internal characteristics. The dataset integrates two complementary sources of information: spectral data, obtained from HSI, and physicochemical data, measured in the laboratory. Both sources are synchronized at the scene level, ensuring a direct correspondence between spectral signatures and their respective chemical properties. Each scene is labeled in the format: (Scene ID, Fermentation Level, Moisture, Polyphenols, Cadmium), where the physicochemical properties correspond to the average values measured for that specific scene. This structured dataset ensures that spectral signatures are directly linked to their corresponding physicochemical properties, enabling an integrated approach for cocoa quality assessment.

Data Preprocessing and Calibration. The dataset is released in its raw form to ensure maximum transparency and reproducibility. Alongside the raw images, we provide all necessary calibration resources ([resources/](#)), including dark-field references, flat-field images (taken with a Spectralon white reference under the same illumination conditions), and wavelength calibration files listing the spectral band centers of each device. In addition, camera-specific metadata are supplied in JSON format (e.g., bit-depth, default RGB channel indices) to facilitate normalization and visualization.

EOS M50. As a standard RGB DSLR, only the in-camera color calibration is relevant, and no additional radiometric calibration is required. This camera is used as a high-resolution spatial reference and produces 8-bit images. The raw values are normalized by the bit-depth as follows:

$$I_{\text{norm}} = \frac{I_{\text{raw}}}{2^8 - 1}, \quad (2)$$

where I_{norm} and I_{raw} denote the normalized and raw image values, respectively.

Specim IQ. The camera applies its own internal calibration, and the exported ENVI files are already corrected to reflectance values. The pixel values typically lie within [0, 1], but may slightly exceed 1 due to reference

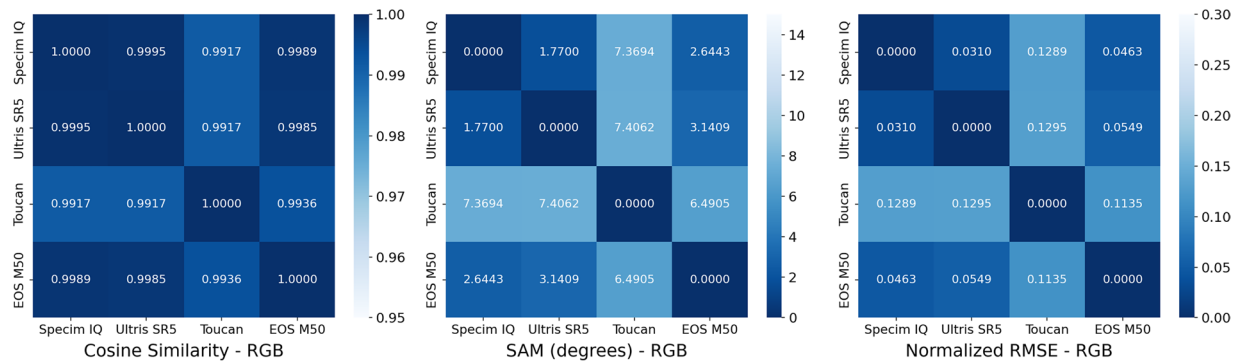


Fig. 8 Cross-device consistency in the RGB region ([450, 550, 650] nm).

imperfections. Therefore, no additional dark- or flat-field calibration is required. For consistency or visualization purposes, the data can be optionally normalized using a robust percentile-based scaling:

$$I_{\text{norm}} = \frac{I_{\text{raw}}}{P_q(I_{\text{raw}})}, \quad (3)$$

where $P_q(I_{\text{raw}})$ denotes the q -th percentile (e.g., 99th) of the raw values. This approach mitigates the effect of outliers compared to a strict maximum-based normalization. A max-based scaling,

$$I_{\text{norm}} = \frac{I_{\text{raw}}}{\max(I_{\text{raw}})}, \quad (4)$$

can also be applied, but may be less robust in the presence of noise or saturation artifacts.

Ultras SR5. Raw hyperspectral cubes are provided in TIFF format, together with calibration resources. Because this camera typically requires strong illumination to achieve high-quality images, the raw data are more susceptible to noise. Therefore, both dark-field and flat-field corrections are necessary to compensate for sensor noise, vignetting, and spectral response variations. The corrected reflectance values are computed as:

$$I_{\text{cal}} = \frac{I_{\text{raw}} - D}{F - D}, \quad (5)$$

where I_{cal} , I_{raw} , D , and F denote the calibrated image, raw image, dark reference, and flat-field reference, respectively. After correction, pixel values generally lie within [0, 1], but may slightly exceed 1 due to illumination variability or reference imperfections. Optional percentile-based normalization, as in expression (3), can be applied to rescale the data more robustly for visualization or further analysis.

Toucan. The native output is a 2048×2048 TIFF mosaic image with a 4×4 spectral filter array. Because raw acquisitions cannot be directly analyzed as hyperspectral data, we performed a minimal demosaicking step to reconstruct $2048 \times 2048 \times 10$ spectral cubes. These are distributed directly as .npz files for ease of use. Calibration resources are also included so that users can reproduce or adapt this step if desired.

Unlike the Ultras SR5, the Toucan camera is less susceptible to sensor noise and does not require dark-field correction. However, the images require flat-field correction to compensate for spectral response variations as follows:

$$I_{\text{cal}} = \frac{I_{\text{raw}}}{F}, \quad (6)$$

where I_{cal} , I_{raw} , and F denote the calibrated image, raw image, dark reference, and flat-field reference, respectively. After correction, pixel values generally lie within [0, 1], but may slightly exceed 1 due to illumination variability or reference imperfections. Optional percentile-based normalization, as in expression (3), can also be applied to rescale the data more robustly for visualization or further analysis.

Accompanying Preprocessing Code. Example Python scripts for loading, visualization, and preprocessing are provided in a separate public GitHub repository (CoCoASpec). These scripts illustrate how to use the calibration resources distributed with the dataset, while allowing users to apply their own preferred correction pipelines. This design ensures that the dataset is reproducible in its raw form, while offering practical tools to facilitate accessibility and reuse.

Cross-device consistency. To assess the cross-calibration of the imaging devices, we compared average spectra extracted from the same cocoa bean scene after preprocessing and radiometric correction. All spectra

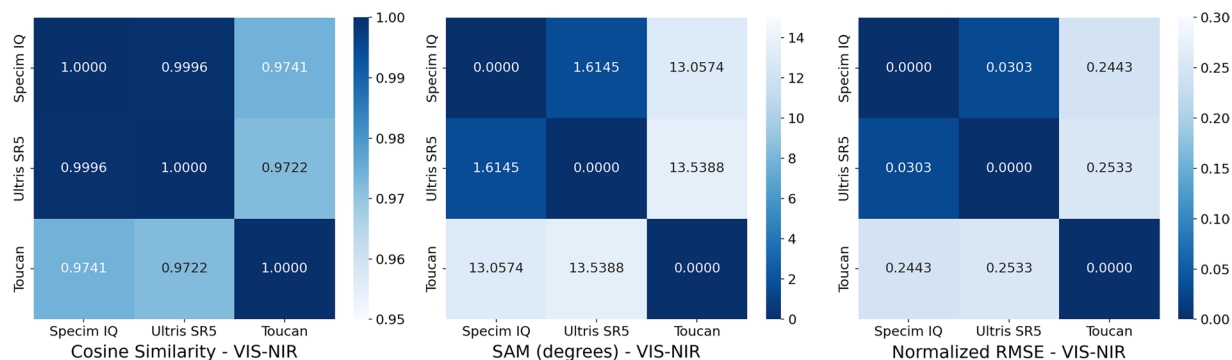


Fig. 9 Cross-device consistency in the VIS-NIR region ([450–850] nm).

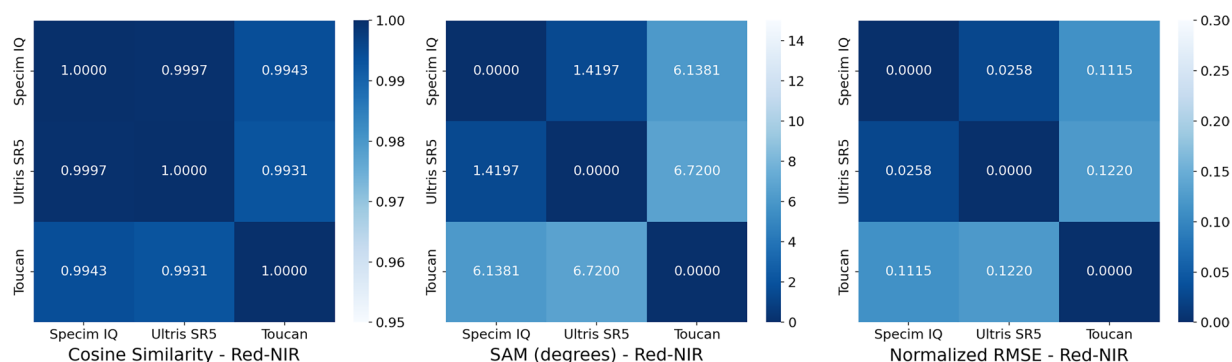


Fig. 10 Cross-device consistency in the Red-NIR region ([600–850] nm).

were interpolated to a common wavelength grid and matched in terms of mean and standard deviation. Spectral similarity was evaluated using cosine similarity, the spectral angle mapper (SAM), and normalized root mean squared error (RMSE). We considered three representative scenarios:

- **RGB region ([450, 550, 650] nm):** All four cameras (Specim IQ, Ultris SR5, Toucan, EOS M50) show excellent agreement, with cosine similarity exceeding 0.99, SAM values below 8°, and normalized RMSE values from 0.03 to 0.13 (Figure 8).
- **VIS-NIR range ([450–850] nm):** Restricting to hyperspectral devices (Specim IQ, Ultris SR5, Toucan), consistency remains high but differences are most pronounced for the Toucan, with SAM values around 13° and normalized RMSE up to 0.25 (Figure 9).
- **Red-NIR range ([600–850] nm):** Agreement improves again, with cosine similarity > 0.993, SAM values reduced to ~6°, and normalized RMSE below 0.12 (Figure 10).

These results confirm that while cross-device consistency is very strong in the RGB and Red-NIR regions, some divergence for the Toucan camera is observed in the full VIS-NIR range. Consistency remains high for the Ultris SR5 and Specim IQ cameras. This provides users with practical guidance on the reliability of cross-device analyses depending on the spectral range of interest.

Data Records

The CoCoASpec dataset is publicly available at the Hugging Face (HF) Hub (See¹¹ to download). To facilitate reproducibility and comply with Scientific Data's repository policies, the dataset is distributed under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original authors are properly credited. Moreover, the dataset is also provided as a single downloadable archive file at the root level (dataset.zip), in addition to the structured repository format, described as follows.

The dataset¹¹ is structured hierarchically, as shown in Fig. 11, where each scene corresponds to a specific cocoa bean sample. As explained in the sections “Physicochemical Properties” and “Dataset Composition and Optical Setup”, each scene includes its associated physicochemical properties, as well as the RGB and HSI acquisitions of the cocoa bean samples in their closed and open states. Additionally, each scene also includes the classification label assigned by the field expert according to the Colombian standard NTC 1252:2021. Furthermore, spatial annotations are added, indicating the location of each cocoa bean within the scene, and ensuring that each sample is accurately documented and traceable. These annotations include bounding boxes and expert-assigned fermentation categories, that is, ‘well’, ‘partially’, and ‘poorly’ fermented, for each individual bean.

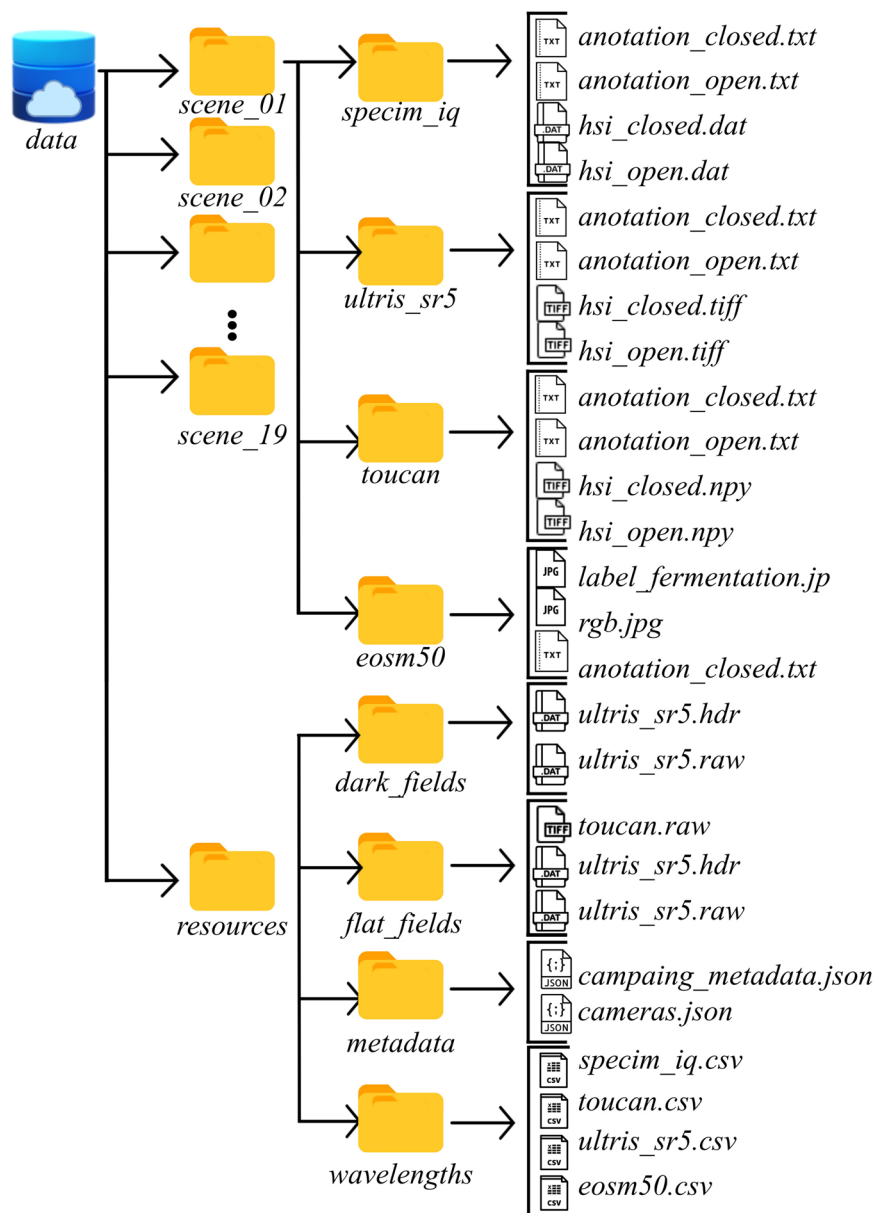


Fig. 11 Hierarchical structure of the dataset, showing the organization of scenes and imaging modalities.

Within each scene folder, subdirectories are organized by imaging device, that is, EOS M50, Toucan, Ultris SR5, and Specim IQ (Figure 12). Each device folder contains its respective image files, annotations, and image metadata. Additionally, as described in the section “Data Preprocessing and Calibration”, in the resources and metadata folders, dark-field and flat-field reference images are included to perform radiometric corrections, particularly important for the Toucan and Ultris SR5 cameras. Moreover, spectral channel wavelength centers and camera-specific metadata are provided, such as bit-depth and RGB channel indices, to help with loading, band selection, cross-calibration, and visualization.

A Python utility is also provided (<https://github.com/kebincontreras/CoCooSpec>). The dataset is accompanied with a README .md file to directly guide users of the dataset.

Technical Validation

Technical validation is presented using spectral signatures obtained from local fermentation labels and global physicochemical property labels. Principal Component Analysis (PCA), t-SNE, and UMAP methods, jointly with a correlation matrix, are included to describe the relationships between physicochemical properties and the separability of the data.

Spectral Signatures by Local Fermentation Label. Spectral measurements were performed separately for open and closed cocoa beans using the four cameras available in the experiment. Figure 13 illustrates the spectral signatures obtained, highlighting differences based on the fermentation state and the condition of the bean.

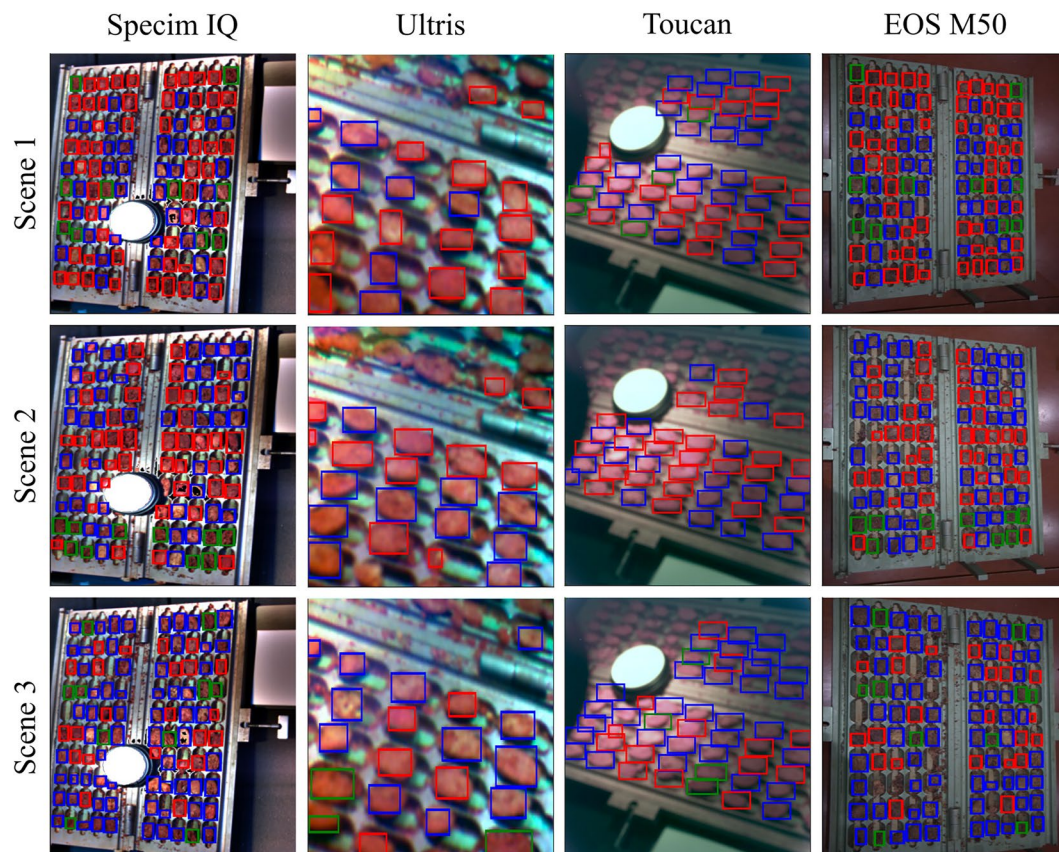


Fig. 12 Two representative scenes captured with four cameras: Specim IQ, Ultris, Toucan, and EOS M50. Green boxes indicate well-fermented beans, red boxes indicate poorly fermented beans, and blue boxes indicate partially fermented beans.

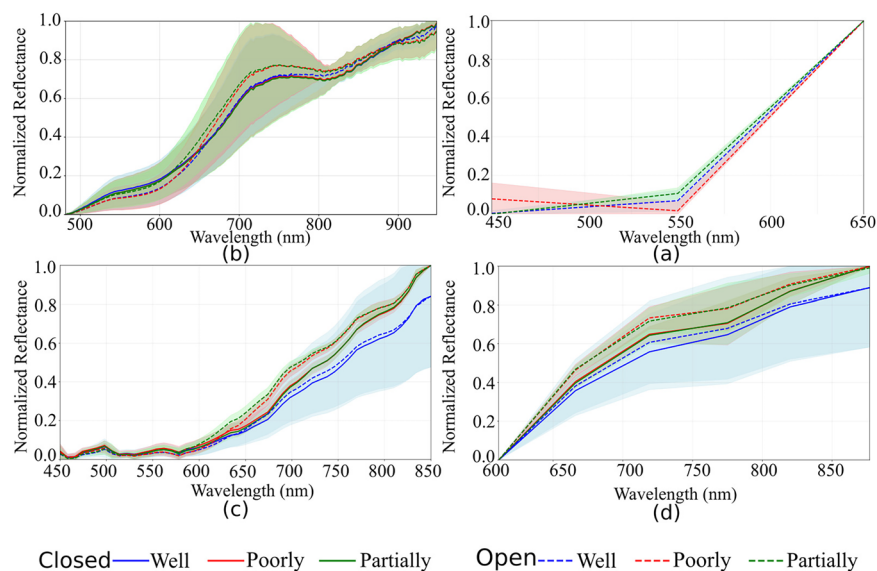


Fig. 13 Mean normalized reflectance spectra of closed (solid lines) and open (dashed lines) cocoa beans acquired with four cameras: (a) Specim_IQ, (b) EOS_M50, (c) Ultris_SR5, and (d) Toucan.

As shown in Fig. 13, the solid lines represent the mean normalized reflectance for Well (blue), Poorly (red), and Partially (green) fermentation beans under closed conditions, while the dashed lines show the same classes after opening. The shaded bands around each curve indicate ± 1 standard deviation across all samples in that class—this highlights the spectral variability both between and within fermentation levels. Under closed

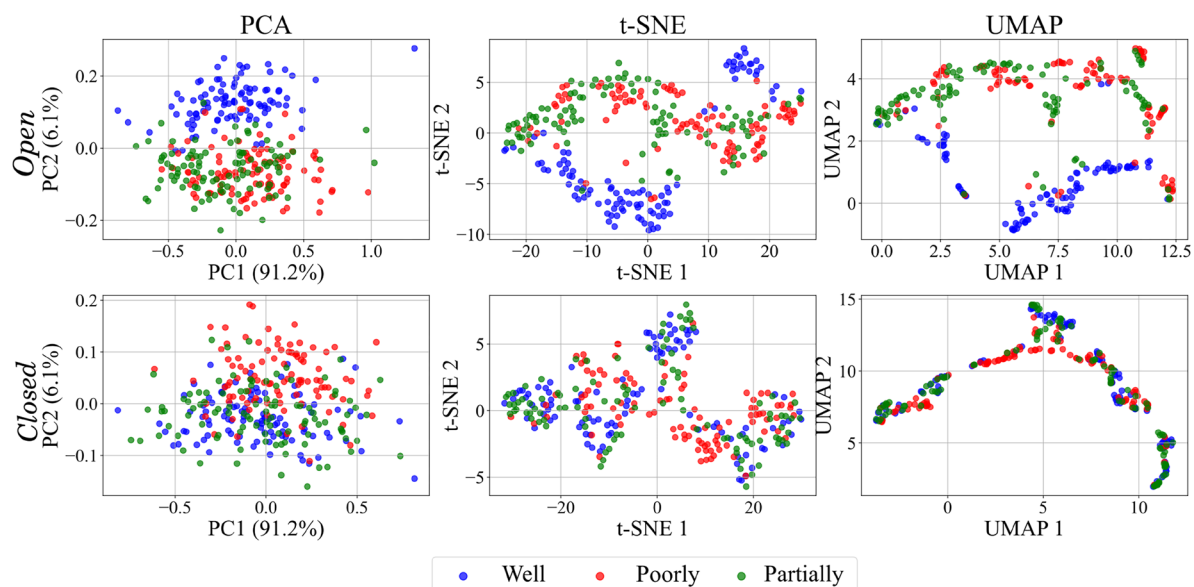


Fig. 14 Dimensionality reduction of mean normalized spectral signatures from Specim IQ for closed (bottom row) and open (top row) beans. Mean signatures were computed by averaging 50 randomly selected grains per fermentation class over 100 trials.

conditions, reflectance increases smoothly from the visible to the near-infrared, with Well beans exhibiting marginally higher values above 700 nm than Poorly or Partially beans. Opening the beans raises overall reflectance at all wavelengths (most pronounced for Poorly and Partially), likely due to changes in surface scattering and internal structure. Notably, the wider shaded region for Partially beans in the 500–650 nm range suggests greater heterogeneity in pigment distribution and moisture content.

In order to see some separability of the data, a PCA, t-SNE and UMAP analysis was carried out using the Specim IQ spectral data, which have higher spectral resolution. Figure 10 shows the mean spectral signatures for each fermentation class obtained by randomly sampling 50 grains per class and averaging their normalized spectra, repeating this procedure 100 times for closed and open conditions. The input data for PCA consisted of spectral signatures with 204 dimensions (corresponding to the 204 spectral bands of the Specim IQ camera covering 350–1000 nm), with 15,000 spectral signatures per condition analyzed (50 grains $\times$ 3 fermentation classes $\times$ 100 repetitions). These signatures were then projected into two dimensions using PCA, t-SNE, and UMAP. PCA (panel a) retains 91.2% of the variance in PC1 and 6.1% in PC2 (total 97.3%), indicating that most spectral information is captured in two axes.

Under open conditions (top row), well-fermented (blue) and poorly fermented (red) beans form two distinct groups, reflecting the greater spectral contrast between well and poorly fermented samples, while the partially fermented class (green) lies between them, its exact position influenced by expert labeling variability. In contrast, closed-bean signatures (bottom row) overlap more strongly, suggesting that non-linear models may be required to discriminate fermentation levels prior to opening, as shown in Fig. 14.

Analysis with Global Physicochemical Property Labels. In this study, spectral signatures were first extracted from individual cocoa beans using object detection annotations. These local signatures were then averaged by scene and condition (open and closed) to produce one representative signature per condition for each of the 19 scenes, yielding 38 scene-level signatures in total. For the global physicochemical analysis, the input data consisted of 204-dimensional spectral signatures from the Specim IQ camera, with 38 scene-level signatures analyzed (19 scenes $\times$ 2 conditions: open and closed). These were projected into a reduced space using PCA (Figure 15), with the first two components capturing 97.8% and 1.4% of the total variance, respectively. As shown in Fig. 15, the PCA scores illustrate how scene-level spectral signatures align with global physicochemical properties such as fermentation, moisture, cadmium, and polyphenols.

When colored by global physicochemical properties (fermentation, cadmium, moisture, and polyphenols) the PCA scores show gradual transitions along the principal component axes, indicating that the spectral transformation preserves the relative ordering and scale of these chemical indicators. In particular, a green box highlights a region where open- and closed-scene-level signatures consistently cluster across all four properties. This suggests that for scenes within this region, it may be possible to reliably estimate physicochemical attributes regardless of whether the cocoa beans are imaged in their open or closed state.

The correlation structure depicted in Fig. 16 provides additional insight into how the physicochemical properties relate to each other and influence the spectral variation captured in the PCA space. Fermentation is strongly correlated with time ($r = 0.65$), and moderately correlated with cadmium ($r = 0.40$), while showing negative associations with both moisture and polyphenols ($r = -0.48$ in both cases). In addition, polyphenols are inversely correlated with cadmium ($r = -0.48$) and time ($r = -0.43$), suggesting that prolonged fermentation may reduce antioxidant content and increase cadmium accumulation.

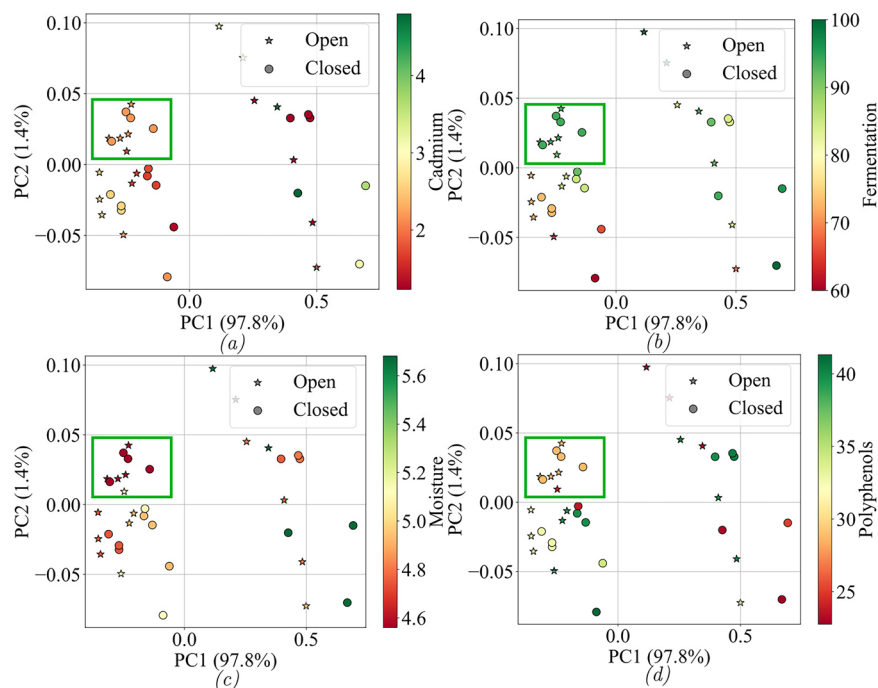


Fig. 15 PCA scores of scene level mean normalized spectral signatures from Specim_IQ, colored by physicochemical properties: (a) fermentation; (b) moisture; (c) cadmium; (d) polyphenols. The first two components capture 97.8 % and 1.4 % of the variance, respectively (total 99.2 %).

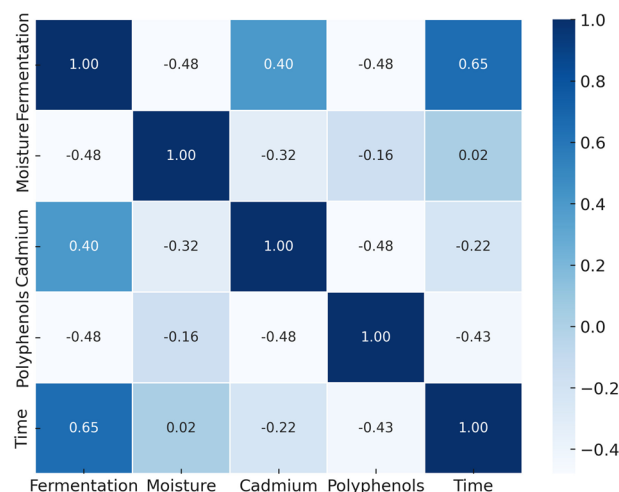


Fig. 16 Correlation matrix of the global physicochemical properties of cocoa beans, indicating the strength and direction of relationships between fermentation, moisture, cadmium and polyphenol measurements.

Usage Notes

To reproduce the analysis and visualizations presented in this work, users must clone the associated GitHub repository: <https://github.com/kebincontreras/CoCoSpec>.

The repository includes two executable scripts: `run_cocospec.bat` for Windows and `run_cocospec.sh` for Linux. These scripts automatically:

- Create and activate the required Python virtual environment,
- Install all dependencies listed in `requirements.txt`,
- Download the dataset from Hugging Face¹¹,
- Apply flat-field correction and spectral preprocessing (required for Toucan and Ultris SR5 images),
- Generate the figures included in the article, such as spectral reflectance curves, PCA/t-SNE/UMAP projections, and the correlation heatmap.

The current dataset is based on samples collected from a single farm in Colombia, which may constrain the generalization capability of models trained on this data. Future extensions of the dataset are planned to incorporate samples from multiple geographical regions, aiming to enhance variability, robustness, and applicability of machine learning models across diverse growing conditions. Furthermore, the temporal scope of data collection will also be expanded, as seasonal variations, different cocoa varieties (beyond CNN-51, ICS-95, and TCS-01), and region-specific post-harvest processing techniques may introduce additional variables that could affect model performance when deployed in different agricultural contexts. Future work will also include expanding the annotation process by involving multiple experts and incorporating complementary objective quantification methods (e.g., chemical or spectral analyses).

Data availability

The CoCaaSpc dataset is publicly available at the Hugging Face (HF) Hub¹¹, with the permanent DataCite <https://doi.org/10.57967/hf/6604>. The dataset is distributed under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original authors are properly credited.

Code availability

The source code for our custom annotation tool is publicly available at <https://github.com/kebincontreras/CoCaaSpc>.

Received: 12 August 2025; Accepted: 12 November 2025;

Published online: 27 November 2025

References

1. Ricardo, H., Bailey, K. Achieving export-led growth in Colombia. CID Working Paper Series (2008).
2. Edward Gerrard, M. World trade in cocoa. US Government Printing Office (1947).
3. Purcell, T., Martinez-Esquerre, E. & Fernandez, N. The value of rents: Global commodity chains and small cocoa producers in Ecuador. *Antipode* **50**, 641–661 (2018).
4. Myriam, R. *et al.* Physicochemical phenomena in the roasting of cocoa (*Theobroma cacao* L.). *Food Engineering Reviews* **14**, 509–533 (2022).
5. Eduard, G., Alix, E. R., Hector, F. Technology Platform for the Information Management of *Theobroma Cacao* Crops based on the Colombian Technical Standard 5811. *Engineering Letters*, vol. 30, no 1 (2022).
6. Elliot Kwaku, A. Assessing the Impact of Preharvest and Postharvest Factors on Cocoa Bean Quality Using Standard Methods and Novel Handheld Nir Spectrometer Coupled with Chemometrics. University of Cape Coast (2022).
7. Ferraris, S., Meo, R., Pinardi, S., Salis, M. & Sartor, G. Machine learning as a strategic tool for helping cocoa farmers in costa de marfil. *Sensors* **23**, 7632 (2023).
8. Ashiagbor, G. *et al.* 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).
9. Musah, S., Medeni, T. D. and Soylu, D. Assessment of role of in- novative 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, IEEE, (2019).
10. Hashimoto, J. C. *et al.* Quality control of commercial cocoa beans (*theobroma cacao* l.) by near-infrared spectroscopy. *Food Analytical Methods* **11**, 1510–1517 (2018).
11. Contreras, K., Jouni, M., Dalla Mura, M. & Bacca, J. CoCaaSpc: A multimodal hyperspectral dataset of cocoa beans with physicochemical annotation. *Hugging Face Datasets* <https://doi.org/10.57967/hf/6961>.

Acknowledgements

The authors thank Dr. Bernard Schmitt from the Institut de Planétologie et Astrophisique de Grenoble (IPAG) for providing the Specim IQ camera. Data acquisitions were carried out with the support of the Multi-camera Imaging Research and Acquisition (MIRA) Platform of GIPSA-lab. This work has been partially supported by the ECOS Nord project n. C24M01, by the French National Research Agency (ANR) under grants ANR-15-IDEX-02, ANR-20-ASTR-0006 and ANR-23-IACL-0006 and by the Institut Carnot Logiciels et Systèmes Intelligents and LabEx PERSYVAL. In addition, gratitude is due to cocoa field expert Miguel Beltran for his expertise, which contributed to the contextualization and documentation of cocoa quality parameters in this dataset. Additional thanks are given to Juan Daniel Suarez Jaimes and Juan Sebastian Espinosa Espinosa for part of the labeled processes, and the laboratory of Optics of High Dimensional Signal Processing group (HDSP) for the acquisition process.

Author contributions

Conceptualization; K.C., M.J., M.D.M., J.B. Methodology; K.C., M.J., M.D.M., J.B. Data acquisition; K.C., M.J., M.D.M., J.B. Software; K.C., M.J. Visualization; K.C., M.J., J.B. Data Curation; K.C., M.J., J.B. Writing – Original Draft; K.C., M.J., M.D.M., J.B. Writing – Review & Editing; K.C., M.J., M.D.M., J.B. Funding Acquisition; M.D.M.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.B.

Reprints and permissions information is available at www.nature.com/reprints.

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



Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025