Infrared Physics & Technology 141 (2024) 105482

Contents lists available at ScienceDirect

Infrared Physics and Technology

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

Mid-infrared spectroscopy and machine learning as a complementary tool for sensory quality assessment of roasted cocoa-based products

Gentil A. Collazos-Escobar [a,c,*], Yeison Fernando Barrios-Rodríguez [b,c,**], Andrés F. Bahamón-Monje [c], Nelson Gutiérrez-Guzmán [c]

[a] Grupo de Análisis y Simulación de Procesos Agroalimentarios (ASPA), Instituto Universitario de Ingeniería de Alimentos–FoodUPV, Universitat Politècnica de València, Camí de Vera s/n, Edificio 3F, 46022 Valencia, Spain
[b] i-Food, Instituto Universitario de Ingeniería de Alimentos–FoodUPV, Universitat Politécnica de València, Camí de Vera s/n, 46022 Valencia, Spain
[c] Centro Surcolombiano de Investigación en Café (CESURCAFÉ), Departamento de Ingeniería Agrícola, Universidad Surcolombiana, Neiva, Colombia

ARTICLE INFO

Keywords:
Mid-infrared
Functional groups
Quality monitoring
Non-destructive testing
Machine learning
Artificial intelligence
Optimization

ABSTRACT

Monitoring sensory quality in cocoa-based products is time-consuming and requires expert panelists. Integrating Mid-infrared (MIR) spectroscopy and chemometric models is a promising tool for real-time quality inspection. This study evaluated machine learning (ML) models based on the latent relationship between spectral and sensory information to predict the overall quality of roasted cocoa. Fifty-four roasted cocoa samples were analyzed using ATR–FTIR in the 4000–650 cm^{-1} range and sensory evaluated by four trained panelists. Spectral data were preprocessed using Multiplicative Scatter Correction (MSC) and combined with sensory data. Subsequently, the block-scale Principal Component Analysis (PCA) was performed. Secondly, a PCA was calibrated only on the spectral data to obtain uncorrelated regressors as input to the supervised ML techniques. Supported Vector Machine Regression Model (SVM_R) and the Random Forest Regression Model (RF_R) were used to predict the overall quality of roasted cocoa samples. The training (75 %) and validation (25 %) of the ML techniques were performed 1000 times, and the hyperparameters optimization of each method was assessed via multifactor Analysis of Variance (ANOVA). According to the tasting panel results, the cocoa beans from different growing areas, initially appeared to have similar sensory characteristics. However, using PCA, a distinction was identified in the northern beans. The SVM_R and RF_R models demonstrated an outstanding ability to describe the overall quality of roasted cocoa samples. Further, the statistical results revealed the potential of MIR coupled with SVM_R as a reliable and robust tool for the rapid (CT < 0.02 s) and accurate prediction (MRE < 2 %, R^2 > 99.9 %) of the overall quality of roasted cocoa-based products. This work demonstrates that it is possible to implement artificial intelligence tools to support decisions in cocoa evaluation, ensuring compliance with quality standards and allowing segmentation according to origin and characteristics.

1. Introduction

Cocoa (*Theobroma cacao* L.) is a globally traded commodity, widely used as a raw material with significant relevance in various industries, including confectionery, functional foods, and beverages (cocoa and chocolate derivatives) [1]. Cocoa is popularly appreciated and consumed worldwide by people of all ages for its flavor, color, and health benefits [2]. The chemical composition of cocoa beans has been extensively investigated due to its potential cardiovascular health. Cocoa beans are mainly composed of fat (>40 %), proteins (12–13 %), fiber (11–19 %), and carbohydrates (>32 %) [3]. Further, the beans are also rich in polyphenols and methylxanthines such as theobromine and caffeine, which have been described as having a crucial role in reducing overall serum cholesterol, enhancing lipoprotein levels and insulin sensitivity, as well as providing protection against cognitive decline, Alzheimer's, and Parkinson's disease [4].

* Corresponding author at: Grupo de Análisis y Simulación de Procesos Agroalimentarios (ASPA), Instituto Universitario de Ingeniería de Alimentos–FoodUPV, Universitat Politècnica de València, Camí de Vera s/n, Edificio 3F, 46022 Valencia, Spain.
** Corresponding author at: i-Food, Instituto Universitario de Ingeniería de Alimentos–FoodUPV, Universitat Politécnica de València, Camí de Vera s/n, 46022 Valencia, Spain.
E-mail addresses: gencoles@etsiamn.upv.es (G.A. Collazos-Escobar), yfbarrod@upv.es (Y.F. Barrios-Rodríguez).

https://doi.org/10.1016/j.infrared.2024.105482
Received 11 May 2024; Received in revised form 23 July 2024; Accepted 6 August 2024
Available online 8 August 2024

The cocoa industry is facing increased global demand for cocoa beans. Moreover, consumers are increasingly concerned about the origin of cocoa products and the supply chain behind cocoa production. These facts create the need to increase production and the intensification of research in this field [5]. In this sense, the worldwide demand for cocoa-based products acclaims higher demands regarding its quality assessment from different aspects such as sensory, physical, chemical, and nutritional, among others [6].

The quality of cocoa beans is highly influenced by several factors such as geographical origin, edaphoclimatic conditions, genotype, and postharvest activities on fermentation, drying, roasting, and storage [7]. Therefore, quality standardization is critical as cocoa beans are cultivated mainly by many independent farmers who applied different agronomical practices and postharvest activities, resulting in heterogeneity of cocoa batches [1]. Thus, the use of a larger dataset that contains the description of cocoa commodities in detail (origin, growing practices, postharvest activities) and its combination with a robust multivariate model could be a reliable tool to trace the authenticity and the quality of cocoa products for accomplishing the growing interest of consumers in food quality terms [8]. To achieve this, collecting data on the quality of cocoa beans from different growing areas and origins could be an effective strategy to cluster cocoa bean samples based on their sensory attributes [9].

Quality assessment of cocoa liquor and cocoa-based products in the food industry is commonly done by organoleptic sensory tasting. The sensory evaluation is time-consuming and requires trained expert tasters, which hinders its inline industrial application as a routine analysis for quality monitoring. In the framework of Food Quality 4.0 and big data, the real-time quality inspection of all manufactured food products and processes is a big challenge at all factories [10]. So, the cocoa industry must develop non-destructive and non-invasive tools to support and complement sensory analysis information in real-time.

Integrating current technologies, such as vibrational near-infrared (NIR), fluorescence spectroscopy, chromatography, nuclear magnetic resonance, and X-ray with chemometrics models, could complement the sensory analysis of food products [11]. In this way, vibrational spectroscopy methods such as Fourier Transform Infrared spectroscopy (FTIR) have emerged as a promising technology for the quality assessment of foodstuffs [12]. This technology enables a fast, easy, and non-destructive testing mode, adequate for process characterization, quality inspection, and detection of adulterant food matrices, and its combination with multivariate data analysis has improved quality control in food science [13].

The large volume of data obtained from FTIR technology needs to be analyzed using robust chemometrics models. In this sense, data mining (DM) and machine learning (ML) enable the mathematical modeling of complex chemical-based datasets by extracting meaningful features driven by the data (pattern recognition) [14]. Additionally, ML allows the prediction of learned features based on latent patterns (non-evident information) presented in datasets [15]. Pattern recognition DM-based methods are split into two main groups: the unsupervised techniques, which include the Principal Component Analysis (PCA) and Hierarchical Cluster Analysis (HCA). At the same time, the supervised approach consists of the k-Nearest Neighbours (kNN), Partial Least Squares (PLS) and soft independent modeling by class analogy (SIMCA), among others. Further, in the ML field, the supervised techniques include Support Vector Machines (SVM), Decision Trees (DT), and Random Forests (RF), among others [14].

Several works have reported the feasibility of combining FTIR and unsupervised/supervised chemometric models for the quality assessment of various food products [16,17,18,13]. Regarding the applications of FTIR made on cocoa beans and cocoa-based products, Batista et al. [19] and Hu et al. [20] have satisfactorily calibrated a PLS model that used the infrared spectral data for the accurate prediction of total antioxidant capacity and total phenolic compounds in cocoa beans and chocolate. Nevertheless, there is an essential gap in the literature about integrating FTIR and chemometrics models to understand the non-evident fundamental relationship between infrared spectral data and sensory attributes of roasted cocoa beans. Furthermore, there is no information about a robust multivariate chemometric model for the quality assessment of roasted cocoa products in a non-destructive way by using an infrared spectrum. This knowledge supplies trustworthy information on the specific chemical markers of cocoa-based products, and their correlation with sensory attributes provides the basis to develop an automatic intelligence system for support and complement the sensory decision-making of incoming cocoa batches and cocoa-based products at the industrial level [1].

Therefore, the main aim of this work was to develop a multivariate model based on the latent relationship between spectral and sensory information for rapid and accurate prediction of the overall quality of roasted cocoa beans from different growing areas in Huila-Colombia. For this purpose: (i) experimentally determined the Mid-infrared spectra and sensory profile of roasted cocoa beans cultivated in different areas, (ii) explored the latent relationship between spectral and sensory information via PCA-coupled RF models, and finally (iii) addressed the computer modeling of spectral and sensory information using ML techniques for the rapid prediction of the overall quality of roasted cocoa beans using FTIR technology.

2. Materials and methods

2.1. Reagents

The following reagents were used in the study: theobromine 99 % (CAS 83-67-0, Sigma-Aldrich, USA), caffeine 99 % (CAS 58-08-2, Sigma-Aldrich, USA), acetic acid (CAS 64-19-7, Merck, Germany) and methanol (CAS 67-56-1, Merck, Germany).

2.2. Cocoa sample processing

Fifty-four cocoa samples (*Theobroma cacao* L.) of 60 kg were collected directly from cocoa farmers in the Huila region of Colombia. The samples were processed in controlled conditions in the Centro Surcolombiano de Investigación en Café (CESURCAFÉ). Cocoa samples were obtained from three different growing areas, namely the northern region (Colombia and Tello municipalities), the northwestern region (Palermo, Santa Maria, Teruel, Íquira and Nátaga), and the central region (Gigante, Algeciras, Campoalegre and Rivera).

Regarding cocoa pod processing, cocoa beans were extracted from their fruits and fermented immediately in wooden crates for seven days. During the first three days, the raw cocoa mass reached 45 °C in anaerobic mode to achieve germ death, and then aerobic conditions were guaranteed to complete the fermentation process. Subsequently, the samples were sun-dried until they reached a moisture content between 6 and 7 % on a wet basis (% w.b.). During the cocoa drying process, the moisture content of beans was monitored with a grain moisture tester (G600, Gehaka AGRA, Brazil). Only the healthy dried cocoa beans without infection or physical damage were considered for the analysis.

2.3. Genotype characterization of dried cocoa samples

To characterize the genotype of cocoa samples, the relationship between theobromine/caffeine was determined. The analysis was achieved in triplicate following the methodology reported by Collazos-Escobar et al. [21]. For this purpose, aqueous extractions were carried out by using 200 mg of dried cocoa previously ground in a rotary knife (FPSTFP3322, Oster®, Colombia) in 50 mL of Milli-Q hot water for 25 min at 85 °C in a water bath (WNE 45, Memmert, Germany). Then, the samples were independently stirred in the magnetic plate at 800 rpm for 10 min.

An Agilent 1260 Infinity II series liquid chromatograph HPLC

instrument (Agilent Technologies. Santa Clara. CA. USA) with a Poroshell 120-C18 (2.7 μm, 4 μm − 4.6 × 150 mm) column was used for the analysis. An isocratic elution with methanol and water with 0.2 % acetic acid (20:80 v/v) for 10 min was employed to detect theobromine and caffeine.

2.4. Roasting and grinding conditions

The roasting process was carried out with dried cocoa samples (110 g) at a temperature of 120 ± 2 °C for 27 ± 2 min using laboratory rotatory equipment (TC-150R, Quantik, Colombia). The roasting program was followed according to Collazos-Escobar et al. [22]. Subsequently, the roasted cocoa samples were manually dehulled and ground in a rotary knife (FPSTFP3322, Oster®, Colombia) to obtain cocoa nibs.

2.5. Initial moisture content and water activity (a_w)

The moisture content of roasted cocoa nibs was determined in triplicate by drying 10 g of samples in an oven (UF55, Memmert GmbH+Co. KG, Germany) at 105 °C until constant weight (24 h approximately). The a_w was measured in triplicate using a vapor sorption analyzer (VSA, Aqualab Decagon Devices-Inc. Pullman, USA). In every test, 5 g of roasted cocoa sample was used to measure a_w.

2.6. Fourier-transform infrared (FTIR) spectroscopy

Mid-infrared (MIR) spectral information of roasted cocoa samples was acquired using an FTIR spectrophotometer (Cary 630, Agilent Technologies, Santa Clara, CA) coupled with a horizontal ATR sampling accessory (Diamond ATR). The infrared spectra were obtained in the wavenumber range of 4000–650 cm^{-1}, using 4 cm^{-1} resolution, scan rate 16, and background correction [23]. Each spectrum for the roasted cocoa samples consisted of 900 wavenumbers. To remove any bias linked to the experimental acquisition of MIR data, all the infrared spectra were preprocessed via baseline correction followed by Multiplicative Scatter Correction (MSC) [21]. Data preprocessing was performed by R statistical software (version 4.2.3–2023, R statistics, St. Louis, MO, USA) using the *ChemoSpec* R-function [24].

2.7. Sensory analysis of cocoa liquor

The preparation of cocoa liquor for sensory analysis started with the fine-grinding of roasted cocoa beans in a Bunn G3HD milling device (Bunn Coffee Mill, Springfield, II. USA). Then, cocoa liquor samples were tempered in a water bath at 50 ± 2 °C during the sensory analysis session. Three highly trained panelists from CESURCAFÉ and one expert panelist from the Sensory Analysis Laboratory of the Universidad de Antioquia evaluated the cocoa liquor samples. The sensory panelists were previously trained under the standard (GTC 280, 2017). To achieve that, all panelists received five days of orientation, 2 h per session daily, to familiarize themselves with the cocoa samples and the evaluation process. The panelists collected a reference framework using aromatic standards for cocoa and its different products, such as powders, chocolates, and beverages, to obtain the list of descriptors that would allow obtaining the maximum sensory information of the samples. Furthermore, to prevent any bias from excessive sample analysis, the analysis was limited to a maximum of eight samples per day, according to recommendations by Barrientos et al. [6].

Ten sensory attributes were assessed: acid, aroma, astringent, bitter, cocoa/chocolate, floral, fruity, green/raw, malt/candy, and nuts. These attributes were selected according to the Equal Exchange and TCHO technical team's tasting guide for the sensory analysis of cocoa [25]. The sum of individual scores for all sensory attributes represented the overall quality of cocoa samples, as this cumulative metric quantifies the sensory balance of a cocoa sample.

An ordinal scale of 0 to 10 was used to quantify the attribute's intensities. The score of 0 indicated the absence of characteristics in the sample, with 1–2 low intensity, 3–5 medium intensity, 6–8 high intensity, and 9–10 very high intensity. The sensory analysis was conducted in individual room cubicles with ambient conditions of 26 ± 2 °C and 50 to 60 % relative humidity, and the liquor samples were given to the panelists at the temperature of 50 ± 2 °C.

Sensory attributes (acid, aroma, astringent, bitter, cocoa/chocolate, floral, fruity, green/raw, malt/candy, nuts, and overall quality) were analyzed using a Generalized Linear Mixed (GLM) model. A GLM model was used to assess the influence of cocoa growing areas and trained testers on the different sensory attributes. Cocoa growing areas were entered as a fixed factor, and tasters were included as a random factor within each model. Mean pairwise comparisons were performed using Fisher's Least Significant Difference (LSD) intervals to statistically determine whether the mean values of sensory attributes were significantly ($p < 0.05$) affected by cocoa growing areas. This model was selected due to its suitability for analyzing response variables with different distributions, such as normal, binomial, Poisson, gamma, and others. Thus, the use of a GLM model represents a robust approach applicable to a wide range of data, including continuous, binary, count data, and proportions [26]. The residual validation of the fitted GLM models was conducted by performing several tests on the model's residuals to examine their normality (Shapiro-Wilk's test and q-q plot), independence (Ljung-Box's test), and homoscedasticity (linear regression-MLR on square residuals). Hypothesis tests and statistical assumptions were assessed at a confidence level of 95 %. The statistical analysis was carried out using STATGRAPHICS Centurion XVIII (Manugistics, Inc., Rockville, MD, USA).

2.8. Explorative analysis

A PCA was performed to explore the latent relationship between MIR spectral information and sensory quality attributes of roasted cocoa samples. This approach was conducted to elucidate the latent relationship between MIR spectral information and sensory data, thereby enabling differentiation between cocoa-growing regions. The latent variables calculated via this strategy revealed common sources of variability between the spectral and sensory datasets. Consequently, these latent variables were utilized as exploratory tools to assess the distinguishability of different regions based on their latent structures.

For this approach, the spectral (54 samples × 900 wavenumbers of MIR-spectra, section 2.6) and sensory (54 samples × 11 quality attributes, section 2.7) datasets were combined in the same framework (54 samples × 911 original variables) to model the non-evident information based on the latent structure of the PCA technique.

Firstly, the multi-block scaling strategy was carried out to avoid model bias and to balance the effect caused by differences in the scales and number of variables of both combined data sets [27]. This approach consisted firstly of mean-centered and scale independently of the spectral and sensory datasets to have unit variance. Then, each scaled dataset was divided into the root square of its number of variables. Further, they were integrated with the same framework (54 samples × 911 original variables) for PCA analysis.

The PCA model used the Singular Value Decomposition (SVD) algorithm to extract the orthogonal latent eigenspace by compressing the spectral and sensory information into a reduced number of Latent Variables (LVs) via a linear combination of the original variables [28]. A total of 53 LVs, which account for the 100 % variability of the original data set, were extracted. The SVD calculations led to obtaining uncorrelated scores (t, 54 samples × 53 LVs) and the PCA loadings (p, 53 LVs × 911 original variables) corresponding to the weight of the original variables in the framework to explain the variability of the original space. Moreover, to detect and remove outlier samples from the experimental data and to validate the PCA model, multivariate control statistics such as the Residual Sum of Squares (RSS) and the Hotelling t-squared (T^2) were employed [29].

Secondly, to analyze the influence of the growing region over the latent structure of spectral and sensory information, the RF algorithm in multi-classification mode (RF_C) was fitted to select the most critical LVs to faithfully differentiate the cocoa growing areas [21]. This strategy consisted of calibrating RF_C using the scores-t as model inputs and the column vector of cocoa growing areas as a response variable. Computing modeling was carried out 1000 times employing 100 random trees, and the RF calculations were performed using the *RandomForest* R-package [30]. Then, the RF_C algorithm's Mean Decrease Accuracy (MDA) criterion was employed to rank the most relevant LVs to accurately classify the cocoa samples belonging to each growing region.

The variable importance score assigned to a LV during the training process of an RF_C or Random Forest Regression Model (RF_R) model assesses the informativeness of each feature. This score is used as the criterion for stratified sampling of the feature subspace during forest construction. This method of feature selection effectively leverages the more informative features while not entirely disregarding the less informative ones [31].

2.9. Machine learning-based model for sensory quality assessment

The Supported Vector Machine Regression Model (SVM_R) and the RF_R were used to mathematically describe the overall quality scores of the roasted cocoa samples (section 2.7) based on the infrared information obtained by FTIR technology. For this purpose, a PCA model (different from the PCA model tuned in section 2.8) was first calibrated using spectral data (54 samples × 900 wavenumbers of MIR-spectra, section 2.6) to obtain the uncorrelated scores, which explained all the infrared spectral data (t_{sp}). In this model, all of the LVs (53 components) were extracted (summarizing 100 % of experimental variability), and the PCA model's validation was also achieved via RSS and T^2 multivariate statistics.

Afterward, the computer modeling procedure was performed, considering t_{sp} as model regressors and the overall quality as a response. The training and statistical validation of the ML techniques were performed following the strategy reported by Sanchez-Jimenez et al. [29]. This consisted of randomly dividing the experimental data 1000 times in two data sets; 75 % was used for model training, and the remaining 25 % was used for model validation.

To optimize the hyperparameters of each ML technique, different independent multilevel factorial designs (DOEs) were formulated as the basis to find the best combination of hyperparameters to maximize the accuracy in the prediction of the overall quality of roasted cocoa samples by the ML model. This strategy was implemented to statistically quantify the impact of combining different hyperparameters on the predictive accuracy and capability of each ML model. In this way, the SVM_R technique was assessed by a DOE ($2^1 3^2$), considering two kernel functions (KF, *rbfdot* and *laplacedot*), three types (*esp-bsvr*, *esp-svr*, and *nu-svr*) and regularization parameter (C: 1, 500.5, and 1000). Further, in all cases, the epsilon parameter (ε) was set to 0.1.

Regarding the RF_R algorithm, a DOE (4^1) was formulated using different numbers of trees (100, 550, 1000 and 10,000) while keeping constant the number of predictors sampled for splitting at each node (Mtry) as the square root of the number of LVs, resulting in Mtry = 7.34 [21]. Computational procedures were conducted on R statistical software using different R-packages: the *kernlab* for SVM_R [32]and the *RandomForest* for the RF_R [30].

Additionally, an ensemble-learning feature selection strategy was proposed to effectively select the LVs during the tuning process of a ML model, aiming to maximize its predictive power. This was achieved by utilizing the percentage increase in Mean Square Error (MSE, %) criterion derived from the optimized RF_R as a variable selection tool to identify the most significant LVs. This criterion quantifies the increase in MSE caused by removing an explanatory variable from the model, thereby revealing the importance of each regressor in calibrating the predictive model [31]. These selected LVs were then used to train a new SVM_R.

In each partition of the experimental data set, the formulated DOEs were simultaneously executed to train (75 %) and validate (25 %) the ML using each hyperparameter combination. The optimal configuration of each ML technique was found using two multiway Analysis of Variance (ANOVA) models. Both considered the hyperparameters of each ML technique and the random partitions of the experimental data set as model factors and the mean relative error (MRE), Eq. (1) and coefficient of determination (R^2), Eq. (2) as variable responses to be independently minimized and maximized, respectively.

$$\mathrm{MRE}\,(\%) = \frac{100}{N}\sum_{i=1}^{N}\frac{\left|OV_{exp} - OV_{pred}\right|}{OV_{exp}} \tag{1}$$

$$R^2(\%) = 100 - \frac{\sum_{i=1}^{N}\left(OV_{exp} - OV_{pred}\right)^2}{\sum_{i=1}^{N}\left(\overline{OV_{exp}} - OV_{pred}\right)^2} \tag{2}$$

where OV_{exp} and OV_{pred} are the experimental and predicted overall quality values and N is the number of experimental data. Low MRE figures and $R^2 > 98$ % reflect a reasonably satisfactory fitting of a mathematical model [33].

Furthermore, the computation times (CT, s) for training and validation processes were recorded (*system time* R-function was employed). This information was considered to quantify the time necessary to train and validate the SVM_R and RF_R techniques and their computational cost. Computation analyses were run on an Intel Core i7 processor, working at 2.2 GHz and with 16 GB RAM.

Finally, the multifactor ANOVA models employed to optimize the hyperparameters of SVM_R and RF_R were residually validated by examining the normality (Shapiro-Wilk's test and q-q plot), independence (Ljung-Box's test), and homoscedasticity (linear regression-MLR on square residuals). These statistical procedures were also carried out using STATGRAPHICS Centurion XVIII (Manugistics, Inc., Rockville, MD, USA).

3. Results and discussion

3.1. Sensory and spectra characterization of cocoa samples by region

Roasted cocoa samples presented an initial moisture content of 0.02 ± 3.01 × 10^{-3} kg water/kg dry matter (kg water/kg d.m., equal to 1.90 ± 0.31 % wet basis) and a_w of 0.30 ± 0.02. Therefore, this result allowed us to classify the roasted cocoa samples as low-moisture food [34]. These values were quite similar to those reported by Collazos-Escobar et al. [22] who claimed that such figures are characteristic of properly roasted cocoa samples. The characterization of cocoa samples in terms of moisture content and water activity is crucial for accurate infrared spectral analysis. Controlling water content in food samples is essential because water's strong absorption bands in the infrared region can bias FTIR analysis. Moreover, excess water content in food products can significantly impact their sensory and physical properties, including texture, flavor, and aroma. Therefore, to ensure the reliability of infrared spectral-based investigations, the initial determination of moisture content and water activity was imperative.

Sensory analysis evidenced scores close to zero in the attributes associated with defects (over-fermented and earthy/moldy), indicating that the cocoa beans were fermented correctly (Table 1). The regions only presented statistically significant differences ($p < 0.05$) in the aroma attribute, with the northwestern region showing higher values (6.5 ± 0.35), while the central region gave lower values (6.2 ± 0.36).

The flavor is a fundamental characteristic for differentiating cocoa hybrids, with compounds such as 2,3-butanediol, linalool, β-myrcene, cis/ trans −β-ocimene, 2-nonanone, 2-nonanol, 2-heptanol, methyl acetate, acetophenone being those that have been shown to differentiate between different cocoa genotypes [35]. On the other hand, the fact that

Table 1
Mean scores of sensory attributes of cocoa samples from three regions.

Sensory attribute	Central	North	Northwestern
Acid	3.9 ± 0.23[a]	4.10 ± 0.20[a]	4.10 ± 0.24[a]
Aroma	6.20 ± 0.36[a]	6.40 ± 0.20[b]	6.50 ± 0.34[ab]
Astringent	3.65 ± 0.23[a]	3.50 ± 0.16[a]	3.50 ± 0.19[a]
Bitter	3.70 ± 0.15[a]	3.70 ± 0.14[a]	3.70 ± 0.22[a]
Cocoa/chocolate	4.40 ± 0.25[a]	4.60 ± 0.21[a]	4.40 ± 0.25[a]
Floral	4.30 ± 0.32[a]	4.20 ± 0.29[a]	4.20 ± 0.44[a]
Fruity	4.20 ± 0.30[a]	4.20 ± 0.30[a]	4.30 ± 0.41[a]
Green/raw	3.10 ± 0.37[a]	3.10 ± 0.39[a]	2.90 ± 0.34[a]
Malt/Candy	3.40 ± 0.26[a]	3.40 ± 0.14[a]	3.35 ± 0.29[a]
Nuts	4.10 ± 0.40[a]	4.10 ± 0.13[a]	4.30 ± 0.29[a]
Overall quality	6.10 ± 0.49[a]	6.10 ± 0.39[a]	6.30 ± 0.55[a]

Samples: Central (n = 15), North (n = 16), Northwestern (n = 23). Results are expressed as mean ± standard deviation (M±SD). Different letters in the row indicate significant differences between regions ($p < 0.05$).

there are no differences in most attributes indicates that the cocoa produced in the different regions of Huila-Colombia are very homogeneous in their sensory characteristics, and the existing differences are challenging to detect by a sensory tasting panel. Regarding the genotype, in the growing areas of Huila-Colombia, the cocoa farmers plant several genotypes indiscriminately in the same growing area. Due to this reason, it was necessary to establish the theobromine/caffeine relationship between the cocoa variety of roasted samples [36]. The relation of theobromine/caffeine values between 3–9 defines the Trinitario variety, and the Forastero variety can be classified with values higher than 9 (Fig. 1).

These genotype results were interesting because of their contribution to explaining the behavior of the sensory data, as they demonstrate the contribution of all genotypes in all regions. The cocoa crops of Colombia have a high genetic variability due to the ecotypes generated from crosses between the Forastero and Trinitario clones. The infrared spectra and the primary vibrations were obtained from the different roasted cocoa by region (Fig. 2). The wavenumber at 3400 cm^{-1} has been related to the (O—H), high-intensity peaks of 2924 and 2855 cm^{-1} corresponded with symmetric and asymmetric group vibrations (C—H) due to modification in the alkenes, lipids, and olefins that are typical of roasting cocoa beans, while 1745 cm^{-1} corresponding to the vibration of (C═O) esters group [37]. In other studies, bands at 2922–2855 cm^{-1} have been associated with vibrations of the C—H bonds of caffeine and lipid molecules [18]. The bands referred to at 1663 and 1620 cm^{-1} can be attributed to stretching vibrations of alkenes (C═C), while the axial deformation of the group (N—H) in the aromatic ring of possible alkaloids such as caffeine and theobromine shows signals between 1750⁻1600 cm^{-1} [21]. Carbohydrates such as sucrose, glucose, and fructose generally offer absorption bands between 1400 and 900 cm^{-1}.

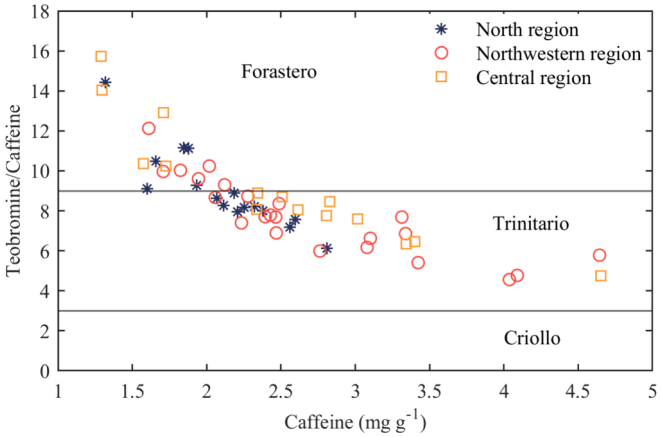

Fig. 1. Determination of cocoa genotype using the theobromine-caffeine ratio.

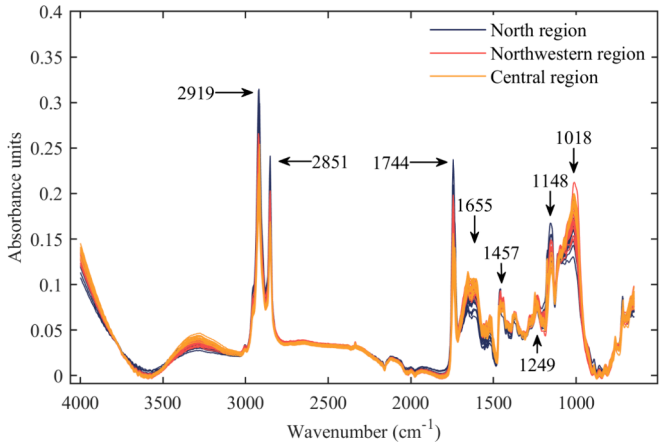

Fig. 2. Infrared spectrum and primary vibrations between 4000–650 cm^{-1} of the roasted cocoa samples.

In chocolates, the bands 909 cm^{-1}, 1239 cm^{-1}, 1472 cm^{-1}, 1734 cm^{-1}, 2917 cm^{-1}, and 2850 cm^{-1} were associated with lipids, 989 cm^{-1}, 1051 cm^{-1}, and 1067 cm^{-1} was assigned to carbohydrates, proteins in 1178 cm^{-1} and 1650 cm^{-1} is associated with the possible presence of caffeine [20]. These results indicated that a simple qualitative analysis of the infrared spectrum and sensory characteristics did not allow us to distinguish the differences between the roasted cocoa; it is necessary to perform a deeper analysis of the information to demonstrate the behavior of the samples by region.

3.2. Multivariate exploratory analysis by PCA and RF_C

A PCA was performed to understand if any pattern allowed a better characterization of the roasted cocoa samples based on sensory and spectral information. This analysis permitted extracting the main underlying trends and patterns, filtering out noise, and improving the data quality. Scatterplot scores of PC1 and PC2 explained 43.8 % of the variability (Fig. 3A).

The scatter of observations revealed a clear separation between samples from different regions. Despite some overlap, a general trend of clustering by region emerged, indicating that samples from different regions had distinct characteristics (Fig. 3A). It was observed that the central region had a different spatial location compared to the northern region. Additionally, PC3 (13.5 %) showed different clustering patterns, making it possible to separate the northwest region from the other two regions (Fig. 3B). This result manifested that both PC2 and PC3 contributed to the differentiation between regions, although PC2 appears to have a greater impact on the separation based on the explained variability.

From the PCA results, it is possible to infer that the samples present spectral and sensory characterization differences. This allows us to observe how PCA helps to eliminate redundancies and retain only the most informative features.

Fig. 3C showed the importance of each principal component in the RF_C model measured as mean decrease precision. This method measures each variable's impact on the accuracy of the model, providing an intuitive way to identify which variables are most critical to the prediction. In this regard, PC1 has the highest importance, with a decrease in accuracy of approximately 4 % (Fig. 3C), indicating that it is the model's most relevant component for classification. Other principal components also have some importance, but significantly less than PC1. This reinforces the interpretation of the PCA plots, where PC1 explained most of the variability. Thus, it was evident that some components were more critical for grouping the samples by regions (Fig. 3C), highlighting the contribution of PC1, PC3, and PC2.

It should be noted that for PC1, there is a high contribution of the

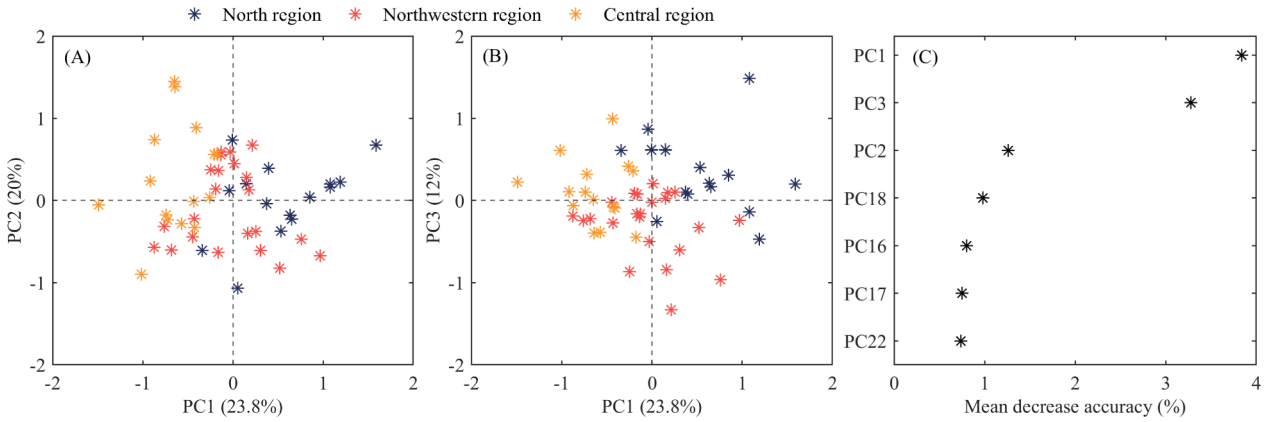

Fig. 3. Clustering trends by regions obtained by PC1 vs. PC2 (A), PC1 vs. PC3 (B) and the importance of each component in separating the samples according to accuracy in the RF_C (C). PC (Principal Components), RF_C (Random Forest Classification Model).

sensory variables green/raw, acid fruity, and floral (Fig. 4A). This indicates that samples located along this component (central and northern region) were more closely related to these variables (Fig. 3A and 3B). Also noteworthy is the influence of spectral bands in the range 2800–3000 cm^{-1}. This may be due to the power of the stretching of the functional groups present in the lipids (asymmetric and symmetric CH_2), which have been described at 2930–2920 cm^{-1} and 2860–2840 cm^{-1} [19]. PC3 was another component that evidenced great importance for separating the cocoa samples (Fig. 3C). This component contributes to the separation of samples from the northwest region (Fig. 3B), possibly due to the influence of floral, bitter, and chocolate/cocoa sensory variables and wavelengths between 2500–2700 cm^{-1} (Fig. 4B). Finally, the importance of the sensory variables is highlighted in most of the PCs, especially in PC2, where an essential contribution of the variables green/raw, astringent, bitter, acid, cocoa/chocolate, floral, nuts, aroma, fruity and overall quality is evidenced (Fig. 4C). Wavelengths between 650–750 cm^{-1} and 1300–1400 cm^{-1} also show evidence of significant loading in PC18, PC16 and PC17 (Fig. 4). Some vibrations between 650–900 cm^{-1} in the infrared spectrum bands have been associated with some polysaccharides such as galactan and fructose β-d-fructose [38], originating from valence vibration of the C—O bond and stretching of the C—O bond. On the other hand, carbohydrate vibration attributed to angular deformation of the aromatic ring C\H can occur in the region between 1400 to 1200 cm^{-1} [18].

3.3. Machine learning for cocoa quality assessment

As Materials and Methods (section 2.9) explained, the mathematical modeling of the overall quality of roasted cocoa samples as a function of spectral information was performed using two ML techniques. The statistical results of SVM_R and RF_R are presented separately in Table 2 for the training (75 %) and validation (25 %) data set.

At the same time, the assessment of correspondence between the experimental overall quality and predicted values by the trained/validated ML models is depicted in Fig. 5. Thus, the training of the ML modeling provided MRE ranging from 3.30×10^{-3} % to 2.20 %, R^2 varied from 99.77 % to 99.99 %, and CT ranged between 0.02 s to 0.07 s (Table 2). In the case of validation, the MRE figures obtained varied from 1.60 % to 3.1 %, the R^2 was between 99.74 % to 99.85 %, and in the interval between 2.65×10^{-3} s to 3.34×10^{-3} s for the CT.

In general, the high figures of both goodness of fit metrics and lower computational time in the model's training and prediction of the validation data set revealed that SVM_R and RF_R techniques exhibited a noticeable ability to describe the overall quality of roasted cocoa

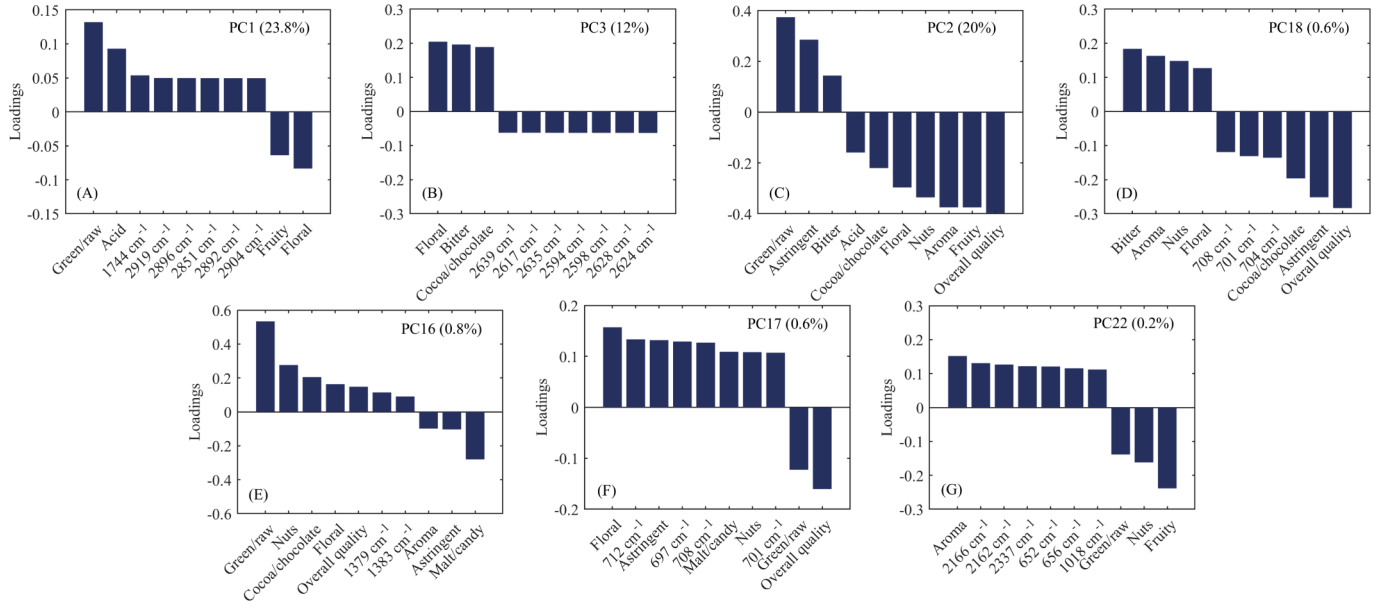

Fig. 4. Loadings for PCs (Principal Components) critical for classifying cocoa samples by region according to RF_C (Random Forest Classification Model) analysis.

Table 2
The goodness of fit of trained Machine Learning (ML) models (75 %) and their statistical results for the validation data set (25 %).

ML technique	Hyperparameters	MRE (%)	R^2 (%)	CT (s)
SVM_R	KF: *laplacedot* Type: *nu-svr* C=500.5 $\varepsilon = 0.1$	Training $3.30 \times 10^{-3} \pm 5.50 \times 10^{-4}$ Validation 1.60 ± 1.00	99.99 ± 0.01 99.74 ± 0.01	$0.02 \pm 8.44 \times 10^{-3}$ $2.65 \times 10^{-3} \pm 6.30 \times 10^{-3}$
RF_R	NRT=1000 Mtry = 7.34	Training 2.20 ± 0.20 Validation 3.10 ± 0.80	99.85 ± 0.02 99.77 ± 0.03	0.07 ± 0.02 $3.34 \times 10^{-3} \pm 6.86 \times 10^{-3}$

SVM_R (Supported Vector Machine Regression Model) and RF_R (Random Forest Regression Model), KF (Kernel function), C (Regularization parameter), ε (epsilon), NRT (number of regression trees), Mtry (number of predictors sampled for splitting at each node). MRE (mean relative error), R^2 (coefficient of determination), and CT (computation time). Results are expressed as mean ± standard error.

samples as a function of the infrared spectral information. Further, both methods accomplish the criteria to select a mathematical model for practical applications (lower MRE figures and $R^2 > 98$ %, section 2.9). However, assessing the agreement between the experimental data and prediction by the ML models is highly recommended. In this sense, both SVM_R and RF_R showed a noticeable linear agreement between experimental and predicted overall quality values for the training (Fig. 5A and 5C) and validation (Fig. 5B and 5D) data sets, respectively. The larger correspondence between the experimental overall quality of roasted cocoa samples and the predicted by the ML models indicated that the infrared spectral data summarized into all LVs extracted from the original space and considering them (t_{sp}) as model regressors of ML techniques, were able to explain the variability of the overall quality of roasted cocoa samples. Thus, the statistical results obtained in Table 2 and the closely aligned cocoa sensory quality assessment by ML techniques (Fig. 5) revealed that both trained/validated models could relate the latent relationship between infrared spectra and sensory information. Nonetheless, the statistical results also allowed us to detect a slightly more remarkable ability of SVM_R than RF_R for making predictions of the overall quality of roasted cocoa samples due to the lower MRE and CT and higher R^2 values of SVM_R.

The description of the overall quality of roasted cocoa samples was fundamentally a data-driven task. Given the variability of industrial-level challenges, there is no one-size-fits-all ML model or set of predefined hyperparameters that can address all problems [39]. Each issue must be individually analyzed with hyperparameters tailored to the specific context. Consequently, investigating different ML techniques to better describe the quality score of roasted cocoa based on infrared spectra is of significant research interest. This necessitates exploring various models and optimizing their hyperparameters, as detailed in section 2.9.

To gain a comprehensive understanding of the performance differences between SVM_R and RF_R, it is essential to examine the fundamental principles behind each algorithm. SVM_R depicts a sophisticated mathematical approach that uses vector spaces and margin optimization to find hyperplanes for classifying data. This involves maximizing the margin between classes while minimizing the predicted error by solving a quadratic optimization problem. Furthermore, SVM_R demonstrates versatility by adeptly handling both linear and nonlinear systems,

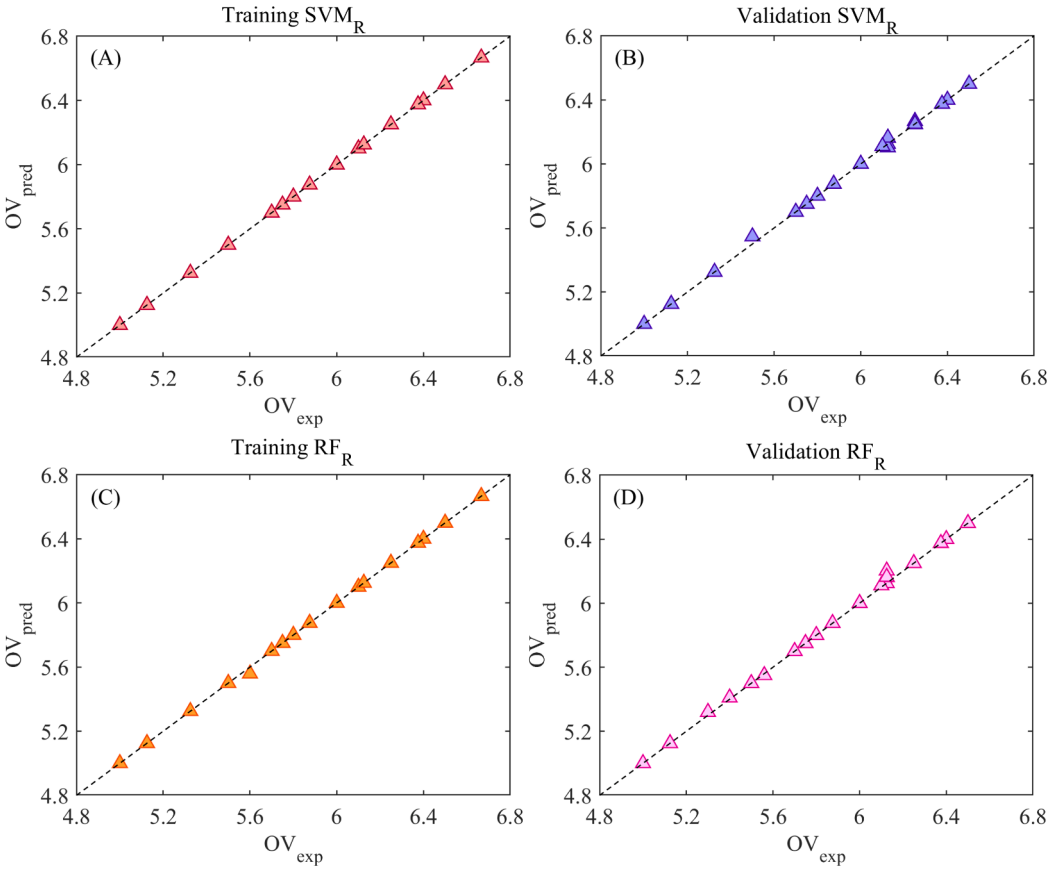

Fig. 5. Experimental and predicted OV (overall accuracy) values via SVM_R and RF_R for training (75 %) and validation (25 %) data sets. SVM_R (Supported Vector Machine Regression Model) and RF_R (Random Forest Regression Model).

achieved by employing KFs to transform data into higher-dimensional spaces [40]. From the statistical outcomes for the trained/validated SVM ($R^2 < 99.9$ and MRE < 2 %, Table 2), it can be observed that the *laplace*dot kernel function, *nu-svr*, and C=500.5 with a constant $\varepsilon = 0.1$ were the optimal hyperparameters for this technique to accurately describe the overall quality of roasted cocoa samples as a function of infrared spectra. This result can be attributed to the combination of reduced outlier sensitivity provided by the *laplacedot* kernel (compared to *rbfdot*), the flexible error control of *nu-svr* (respect to *esp-bsvr*, *esp-svr*, and the balanced regularization parameter C (lower values tend to underfitting and higher promote overfitting), which together enhance the model's robustness and generalization capability [41,42].

RF_R ensembles several randomly selected regression trees (NRT, Table 2) based on the bagging principle to improve the model's diversity and uses mathematical averaging to avoid overfitting. This approach reduces the model's sensitivity to variations in the training data [43]. Each regression tree splits the data into subsets based on feature conditions to minimize the error between predicted and experimental values [44]. According to the statistical results (section 2.9 and Table 2), using 1000 NRT was sufficient to accurately describe the overall quality of roasted cocoa samples. In contrast, using only 100 or 550 NRT could lead to increased underfitting, while 10,000 NRT would substantially increase computational costs without significant improvements in prediction accuracy. Therefore, 1000 NRT depicted an optimal balance between computational efficiency and model accuracy [31].

Although solving non-linear problems via SVM_R requires a lot of informatics resources [14], the CT was lower than the one used via RF_R to ensemble 1000 NRT to solve the same mathematical task. Thus, SVM_R could be considered the best ML model for assessing the sensory quality of roasted cocoa samples. The feasibility of SVM_R has also been successfully used in many applications, such as the mathematical description of the water adsorption process of Achira biscuits [39], the discrimination of Extra Virgin Olive Oils [45], for the determination of coconut maturity based acoustical signals [46], for detection of oil yield using NIR [47]and quantification of butter yellow adulteration in mustard oil [48].

Finally, the statistical results of the ensemble-learning feature selection strategy based MSE criterion of RF_R for training a new SVM_R are illustrated in Fig. 6.

As seen in Fig. 6A, the most relevant LVs for maximizing the agreement between the experimental and predicted overall quality values of roasted cocoa samples were identified for both training (MRE < 0.1 %, $R^2 > 99.9$ %, and CT=0.01 s, Fig. 6B) and validation (MRE < 2 %, $R^2 > 99.0$ %, and CT=1×10^{-3} s, Fig. 6C) datasets. These LVs were ranked according to the MSE criteria of the optimized RF_R (1000 NRT, Table 2). The results demonstrated that using PC7, PC32, PC4, PC37, PC36, PC24, PC51, PC5, PC40, PC48, PC6, PC18, PC47, and PC17 (Fig. 6D to 6R) in the tuning of the SVM_R allows for reliable prediction of cocoa quality. These LVs revealed the contribution of key infrared spectral bands correlated with the sensory scores of samples, as seen in the loading's plots (Fig. 6D to 6R). The ensemble-learning strategy was effective not only for accurately describing overall quality as a function of infrared spectra but also for significantly reducing the CT, thereby calibrating a rapid and parsimonious predictive tool for real-time quality inspection.

Integrating infrared spectral data with sensory information into ML models posed several challenges. The main issue was the high dimensionality and complexity of the spectral data, combined with the high heterogeneity of the sensory data [49]. These factors complicated the mathematical modeling, affecting the time and cost of computational calculations required to predict the overall quality of roasted cocoa samples. To address these challenges, our research implemented several strategies. The use of infrared spectral preprocessing, dimensionality reduction techniques using PCA, and the training of advanced ML latent variables-based models allowed for the reduction of the overfitting risks and enhanced computational efficiency. Both are essential for an inline industrial implementation and provide the basis to develop further studies in this field.

This is the first step in developing data-driven computer chemometric models, which would be able to include more regressors in the mathematical modeling. For instance, different cocoa varieties, post-harvest processing activities such as fermentation, drying, storage, roasting, and non-invasive acquisition techniques (FTIR, hyperspectral imaging, ultrasound, among others) to describe not only their influence on sensory quality but also polyphenols, methylxanthines, antioxidants, rheological properties, among others. These computer tools could help support real-time decision-making at the industrial level and facilitate

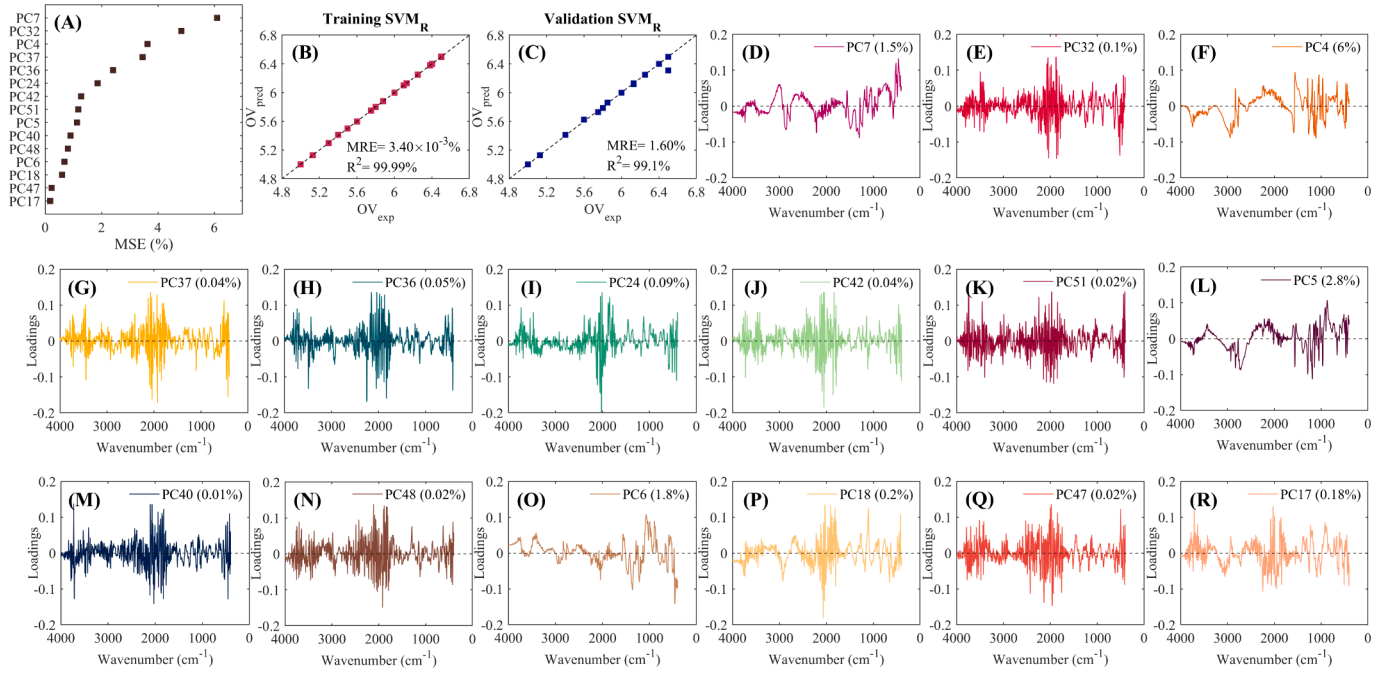

Fig. 6. Ensemble-learning strategy for latent-variable selection in the training (75 %) and validation (25 %) process of the Support Vector Machine Regression (SVM_R) model.

the implementation of non-invasive intelligent systems to screen cocoa batches with unpleasant sensory profiles that do not meet quality standards. Furthermore, they can help segment batches according to their origin and quality features.

4. Conclusions

This work showed that the incorporation of ML models such as RF_R and SVM_R depicted an excellent option to reliably establish the relationship between infrared spectral data and sensory attributes of roasted cocoa beans. Additionally, the incorporation of latent variables revealed important information about the effect of the region on the sensory quality of cocoa, which was impossible to reveal through sensory analysis or the description of the spectrum.

The infrared spectral data comprised of all latent variables extracted from the original space and considered as regressors of the ML techniques model effectively explained the variability in the overall quality of the roasted cocoa samples. Supported Vector Machine Regression Model permitted a precise description of the overall quality of roasted cocoa samples (MRE < 2 % and $R^2 > 99.9$ %) within lower computational time–cost (CT < 0.02 s), allowing the real-time monitoring of the overall quality of roasted cocoa beans and cocoa-based products as an intelligent tool to support and complement sensory quality assessment in the screening of incoming cocoa batches at industrial level.

Finally, this study highlights the importance of developing more complete chemometric models that can incorporate a variety of regressors in the mathematical modeling (cocoa varieties, fermentation, drying, storage, roasting) and various non-invasive acquisition techniques (FTIR, FTNIR, hyperspectral imaging, and ultrasound). This integrated approach allows an understanding of the influence of multiple factors associated with cocoa cultivation and processing on sensory quality and other aspects such as polyphenol, methylxanthine, and antioxidant content.

CRediT authorship contribution statement

Gentil A. Collazos-Escobar: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Yeison Fernando Barrios-Rodríguez:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Andrés F. Bahamón-Monje:** Writing – original draft, Methodology. **Nelson Gutiérrez-Guzmán:** Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgments

The authors sincerely thank the Centro Surcolombiano de Investigación en Café (CESURCAFÉ) from the Universidad Surcolombiana for their invaluable support, which was essential for completing this work. This study was supported by the funding for open access charge: Universitat Politècnica de València.

References

[1] E. Bagnulo, C. Scavarda, C. Bortolini, C. Cordero, C. Bicchi, E. Liberto, Cocoa quality: Chemical relationship of cocoa beans and liquors in origin identitation, Food Res. Int. 172 (2023), https://doi.org/10.1016/j.foodres.2023.113199.
[2] V. Barišić, N.C. Icyer, S. Akyil, O.S. Toker, I. Flanjak, Đ. Ačkar, Cocoa based beverages – composition, nutritional value, processing, quality problems and new perspectives, Trends Food Sci. Technol. (2023), https://doi.org/10.1016/j.tifs.2022.12.011.
[3] E. Fanning, G. Eyres, R. Frew, B. Kebede, Linking cocoa quality attributes to its origin using geographical indications, Food Control (2023), https://doi.org/10.1016/j.foodcont.2023.109825.
[4] F. Mariatti, V. Gunjević, L. Boffa, G. Cravotto, Process intensification technologies for the recovery of valuable compounds from cocoa by-products, Innov. Food Sci. Emerg. Technol. (2021), https://doi.org/10.1016/j.ifset.2021.102601.
[5] N.N. Suh, E.L. Molua, Cocoa production under climate variability and farm management challenges: Some farmers' perspective, J Agric Food Res 8 (2022), https://doi.org/10.1016/j.jafr.2022.100282.
[6] L.D.P. Barrientos, J.D.T. Oquendo, M.A.G. Garzón, O.L.M. Álvarez, Effect of the solar drying process on the sensory and chemical quality of cocoa (Theobroma cacao L.) cultivated in Antioquia, Colombia, Food Res. Int. 115 (2019) 259–267, https://doi.org/10.1016/j.foodres.2018.08.084.
[7] J.E. Kongor, M. Hinneh, D. Van de Walle, E.O. Afoakwa, P. Boeckx, K. Dewettinck, Factors influencing quality variation in cocoa (Theobroma cacao) bean flavour profile - a review, Food Res. Int. (2016), https://doi.org/10.1016/j.foodres.2016.01.012.
[8] M. Perez, A. Lopez-Yerena, A. Vallverdú-Queralt, Traceability, authenticity and sustainability of cocoa and chocolate products: a challenge for the chocolate industry, Crit. Rev. Food Sci. Nutr. (2021), https://doi.org/10.1080/10408398.2020.1819769.
[9] C.S. Siow, E.W.C. Chan, C.W. Wong, C.W. Ng, Antioxidant and sensory evaluation of cocoa (Theobroma cacao L.) tea formulated with cocoa bean hull of different origins, Future Foods 5 (2022), https://doi.org/10.1016/j.fufo.2021.100108.
[10] A. Hassoun, S. Jagtap, G. Garcia-Garcia, H. Trollman, M. Pateiro, J.M. Lorenzo, M. Trif, A.V. Rusu, R.M. Aadil, V. Šimat, J. Cropotova, J.S. Câmara, Food quality 4.0: From traditional approaches to digitalized automated analysis, J. Food Eng. (2023), https://doi.org/10.1016/j.jfoodeng.2022.111216.
[11] A.C. de Oliveira, A. Marien, J. Hulin, Y. Muhovski, V. Baeten, E. Janssen, G. Berben, H. Rogez, F. Debode, Development of real-time PCR methods for cocoa authentication in processed cocoa-derived products, Food Control 131 (2022), https://doi.org/10.1016/j.foodcont.2021.108414.
[12] A. Tan, J. Zhao, Y. Zhao, X. Li, H. Su, Determination of microplastics by FTIR spectroscopy based on quaternion parallel feature fusion and support vector machine, Chemom. Intel. Lab. Syst. 243 (2023), https://doi.org/10.1016/j.chemolab.2023.105018.
[13] Y.F. Barrios-Rodríguez, C.A. Rojas Reyes, J.S. Triana Campos, J. Girón-Hernández, J. Rodríguez-Gamir, Infrared spectroscopy coupled with chemometrics in coffee post-harvest processes as complement to the sensory analysis, LWT 145 (2021), https://doi.org/10.1016/j.lwt.2021.111304.
[14] A.M. Jiménez-Carvelo, A. González-Casado, M.G. Bagur-González, L. Cuadros-Rodríguez, Alternative data mining/machine learning methods for the analytical evaluation of food quality and authenticity – A review, Food Res. Int. (2019), https://doi.org/10.1016/j.foodres.2019.03.063.
[15] C.K. Tanui, S. Karanth, P.M.K. Njage, J. Meng, A.K. Pradhan, Machine learning-based predictive modeling to identify genotypic traits associated with Salmonella enterica disease endpoints in isolates from ground chicken, LWT 154 (2022), https://doi.org/10.1016/j.lwt.2021.112701.
[16] V. Cortés, P. Talens, J.M. Barat, M.J. Lerma-García, Discrimination of intact almonds according to their bitterness and prediction of amygdalin concentration by Fourier transform infrared spectroscopy, Postharvest Biol. Technol. 148 (2019) 236–241, https://doi.org/10.1016/j.postharvbio.2018.05.006.
[17] R. Jamwal, S. Kumari, B. Balan, A.S. Dhaulaniya, S. Kelly, A. Cannavan, D.K. Singh, Attenuated total Reflectance-Fourier transform infrared (ATR–FTIR) spectroscopy coupled with chemometrics for rapid detection of argemone oil adulteration in mustard oil, Lwt 120 (2020) 108945, https://doi.org/10.1016/j.lwt.2019.108945.
[18] V. Belchior, B.G. Botelho, S. Casal, L.S. Oliveira, A.S. Franca, FTIR and chemometrics as effective tools in predicting the quality of specialty coffees, Food Anal. Methods 13 (2020) 275–283, https://doi.org/10.1007/s12161-019-01619-z.
[19] N.N. Batista, D.P. de Andrade, C.L. Ramos, D.R. Dias, R.F. Schwan, Antioxidant capacity of cocoa beans and chocolate assessed by FTIR, Food Res. Int. 90 (2016) 313–319, https://doi.org/10.1016/j.foodres.2016.10.028.
[20] Y. Hu, Z.J. Pan, W. Liao, J. Li, P. Gruget, D.D. Kitts, X. Lu, Determination of antioxidant capacity and phenolic content of chocolate by attenuated total reflectance-Fourier transformed-infrared spectroscopy, Food Chem. 202 (2016) 254–261, https://doi.org/10.1016/j.foodchem.2016.01.130.
[21] G.A. Collazos-Escobar, Y.F. Barrios-Rodriguez, A.F. Bahamón-Monje, N. Gutiérrez-Guzmán, Uses of mid-infrared spectroscopy and chemometric models for differentiating between dried cocoa bean varieties, Revista Brasileira De Engenharia Agricola e Ambiental 27 (2023) 803–810, https://doi.org/10.1590/1807-1929/agriambi.v27n10p803-810.
[22] G.A. Collazos-Escobar, N. Gutiérrez-Guzmán, H.A. Váquiro-Herrera, C. M. Amorocho-Cruz, Water dynamics adsorption properties of dried and roasted cocoa beans (theobroma cacao L.), Int. J. Food Prop. 23 (2020) 434–444, https://doi.org/10.1080/10942912.2020.1732408.
[23] Y. Barrios-Rodriguez, Y. Devia-Rodriguez, N. Gutierrez Guzmán, Detection of adulterated coffee by Fourier-transform infrared (FTIR) spectroscopy associated

with sensory analysis, Coffee Sci. 17 (2022) 1–12, https://doi.org/10.25186/.v17i.1970.
[24] B. Hanson, M. Bostock, M. Keinsley, T. Gupta, Type Package Title Exploratory Chemometrics for Spectroscopy (2024).
[25] Equal Exchange. (2018). Guía de cata (Edición JUNIO 2018). https://equalexchange.coop/sites/default/files/Tasting-Guide_vF-JUNIO2018-ESP.pdf.
[26] K.P. Dunn, Generalized linear models. International Encyclopedia of Education (Fourth Edition). (2023) 583-589. https://doi.org/10.1016/B978-0-12-818630-5.10077-6.
[27] P. Mishra, J.M. Roger, D. Jouan-Rimbaud-Bouveresse, A. Biancolillo, F. Marini, A. Nordon, D.N. Rutledge, Recent trends in multi-block data analysis in chemometrics for multi-source data integration, TrAC - Trends Anal. Chem. (2021), https://doi.org/10.1016/j.trac.2021.116206.
[28] O.M.O. Kruse, J.M. Prats-Montalbán, U.G. Indahl, K. Kvaal, A. Ferrer, C. M. Futsaether, Pixel classification methods for identifying and quantifying leaf surface injury from digital images, Comput. Electron. Agric. 108 (2014) 155–165, https://doi.org/10.1016/j.compag.2014.07.010.
[29] V. Sanchez-Jimenez, G.A. Collazos-Escobar, A. González-Mohino, T.E. Gomez Alvarez-Arenas, J. Benedito, J.V. Garcia-Perez, Non-invasive monitoring of potato drying by means of air-coupled ultrasound, Food Control 148 (2023), https://doi.org/10.1016/j.foodcont.2023.109653.
[30] A. Liaw, M. Wiener, Classification and Regression by randomForest, R News 2 (2002).
[31] F. Wang, S. Ma, G. Yan, A PLS-based random forest for NOx emission measurement of power plant, Chemom. Intel. Lab. Syst. 240 (2023), https://doi.org/10.1016/j.chemolab.2023.104926.
[32] A. Karatzoglou, S. Alex, H. Kurt, Title Kernel-Based Machine Learning Lab, 2023.
[33] M. Edrisi Sormoli, T.A.G. Langrish, Moisture sorption isotherms and net isosteric heat of sorption for spray-dried pure orange juice powder, LWT Food Sci. Technol. 62 (2015) 875–882, https://doi.org/10.1016/j.lwt.2014.09.064.
[34] J.C. Acuff, J.S. Dickson, J.M. Farber, E.M. Grasso-Kelley, C. Hedberg, A. Lee, M. J. Zhu, Practice and progress: updates on outbreaks, advances in research, and processing technologies for low-moisture food safety, J. Food Prot. (2023), https://doi.org/10.1016/j.jfp.2022.11.010.
[35] S.P. Akoa, R. Boulanger, P. Effa Onomo, M. Lebrun, M.L. Ondobo, M.C. Lahon, S. A. Ntyam Mendo, N. Niemenak, P.F. Djocgoue, Sugar profile and volatile aroma composition in fermented dried beans and roasted nibs from six controlled pollinated Cameroonian fine-flavor cocoa (Theobroma cacao L.) hybrids, Food Biosci. 53 (2023), https://doi.org/10.1016/j.fbio.2023.102603.
[36] L.C. Carrillo, J. Londoño-Londoño, A. Gil, Comparison of polyphenol, methylxanthines and antioxidant activity in Theobroma cacao beans from different cocoa-growing areas in Colombia, Food Res. Int. 60 (2014) 273–280, https://doi.org/10.1016/j.foodres.2013.06.019.
[37] S.M. Rojas, F. Chejne, H. Ciro, J. Montoya, Roasting impact on the chemical and physical structure of Criollo cocoa variety (Theobroma cacao L), J. Food Process Eng. 43 (2020), https://doi.org/10.1111/jfpe.13400.
[38] S. Türker-Kaya, C.W. Huck, A review of mid-infrared and near-infrared imaging: Principles, concepts and applications in plant tissue analysis, Molecules (2017), https://doi.org/10.3390/molecules22010168.
[39] G.A. Collazos-Escobar, N. Gutiérrez-Guzmán, H.A. Váquiro-Herrera, J. Bon, J. A. Cárcel, J.V. García-Pérez, Model-based investigation of water adsorption in Achira (Canna edulis K.) biscuits, LWT 189 (2023) 115472, https://doi.org/10.1016/j.lwt.2023.115472.
[40] Aasim, S.N. Singh, A. Mohapatra, Data driven day-ahead electrical load forecasting through repeated wavelet transform assisted SVM model, Appl. Soft Comput. 111 (2021), https://doi.org/10.1016/j.asoc.2021.107730.
[41] C.E. da Silva Santos, R.C. Sampaio, L. dos Santos Coelho, G.A. Bestard, C.H. Llanos, Multi-objective adaptive differential evolution for SVM/SVR hyperparameters selection, Pattern Recognition 110 (2021) 107649, https://doi.org/10.1016/j.patcog.2020.107649.
[42] P. Tsirikoglou, S. Abraham, F. Contino, C. Lacor, G. Ghorbaniasl, A hyperparameters selection technique for support vector regression models, Appl. Soft Comput. J. 61 (2017) 139–148, https://doi.org/10.1016/j.asoc.2017.07.017.
[43] D.A. Otchere, Fundamental error in tree-based machine learning model selection for reservoir characterisation, Energy Geoscience (2023), https://doi.org/10.1016/j.engeos.2023.100229.
[44] I.K. Nti, A. Zaman, O. Nyarko-Boateng, A.F. Adekoya, F. Keyeremeh, A predictive analytics model for crop suitability and productivity with tree-based ensemble learning, Decision Anal. J. 8 (2023), https://doi.org/10.1016/j.dajour.2023.100311.
[45] C. Scatigno, G. Festa, FTIR coupled with machine learning to unveil spectroscopic benchmarks in the Italian EVOO, Int. J. Food Sci. Technol. 57 (2022) 4156–4162, https://doi.org/10.1111/ijfs.15735.
[46] J.A. Caladcad, S. Cabahug, M.R. Catamco, P.E. Villaceran, L. Cosgafa, K. N. Cabizares, M. Hermosilla, E.J. Piedad, Determining Philippine coconut maturity level using machine learning algorithms based on acoustic signal, Comput. Electron. Agric. 172 (2020), https://doi.org/10.1016/j.compag.2020.105327.
[47] F. Zhang, J. Liu, J. Lin, Z. Wang, Detection of oil yield from oil shale based on near-infrared spectroscopy combined with wavelet transform and least squares support vector machines, Infrared Phys. Technol. 97 (2019) 224–228, https://doi.org/10.1016/j.infrared.2018.12.036.
[48] R. Amsaraj, S. Mutturi, Support vector machine-based rapid detection and quantification of butter yellow adulteration in mustard oil using NIR spectra, Infrared Phys. Technol. 129 (2023), https://doi.org/10.1016/j.infrared.2023.104543.
[49] Z.Y. Algamal, M.K. Qasim, M.H. Lee, H.T. Mohammad Ali, Improving grasshopper optimization algorithm for hyperparameters estimation and feature selection in support vector regression, Chemom. Intel. Lab. Syst. 208 (2021), https://doi.org/10.1016/j.chemolab.2020.104196.