



Research article

Band mapping in hyperspectral images for predicting physicochemical parameters in black tea (*Camellia sinensis*)

André Rodríguez-León^{1,*} and Jimmy Oblitas²

¹ Facultad de Ciencias Agrarias, Universidad Nacional de Cajamarca, Av. Atahualpa N° 1050, Cajamarca 06002, Perú

² Facultad de Ingeniería, Universidad Privada del Norte, Vía de Evitamiento s/n cuadra 15, Cajamarca 06002, Perú

* **Correspondence:** Email: arodriguezl@unc.edu.pe; Tel: +51962721910.

Abstract: Black tea is one of the most widely consumed infusions worldwide, whose quality changes during storage due to physicochemical and color changes. However, quantitative prediction of key physicochemical parameters, such as pH and water activity (A_w), during storage has been scarcely explored using hyperspectral imaging and machine learning. This study evaluated visible–near infrared (VIS–NIR) spectra acquired by hyperspectral imaging (HSI) from samples stored under different conditions to predict pH and A_w . A reproducible machine learning pipeline (second-order Savitzky–Golay derivative + standard normal variant (SNV), train–test baseline evaluation followed by stratified cross-validation) was implemented, and PLS, elastic net, random forest, support vector regression (SVR)-radial basis function (RBF), and XGBoost were compared. SVR-RBF showed the best performance (pH $R^2 \approx 0.94$, A_w $R^2 \approx 0.82$). Consensus-based spectral attribution identified relevant bands in the visible (390–550 nm), associated with polyphenolic chromophores, and in the short NIR (870–960 and 980–1000 nm), linked to O–H and C–H overtones. Leave-one-band-out (LOBO) validation revealed minimal variations ($\Delta R^2 \approx 0$), confirming informational redundancy and robustness against band exclusion. In conclusion, VIS–NIR combined with machine learning is a potentially effective and portable tool for monitoring critical quality parameters in black tea, reducing the need for recalibration and facilitating the implementation of compact sensors in at-line and on-line applications.

Keywords: black tea; VIS–NIR spectroscopy; spectral attribution; Elastic Net; LOBO; quality control

Abbreviations

ANOSIM: Analysis of similarities

Aw: Water activity

BH-FDR: Benjamini–Hochberg false discovery rate

EN: Elastic net

HSI: Hyperspectral imaging

LOBO: Leave-one-band-out

PCA: Principal component analysis

PERMANOVA: Permutational multivariate analysis of variance

PLS: Partial least squares

RF: Random forest

RMSE: Root mean square error

R²: Coefficient of determination

SG: Savitzky–Golay

SNV: Standard normal variate

SVR: Support vector regression

VIP: Variable importance in projection

VIS–NIR: Visible–near infrared

XGB: XGBoost

1. Introduction

Black tea (*Camellia sinensis*) is one of the most widely consumed infusions worldwide, and its quality depends on physicochemical parameters that change during storage due to oxidation and phenolic degradation processes. These transformations affect the product's stability, sensory acceptance, and shelf life [1]. Although conventional analytical methods allow accurate characterization of these parameters, their destructive nature, long analysis times, and lack of adaptability to production environments limit their use for online monitoring. In this context, visible–near infrared (VIS–NIR) spectroscopy, combined with machine learning algorithms, emerges as a fast, non-invasive, and process-adaptable alternative [2].

However, its implementation faces major challenges related to the reproducibility and robustness of spectral models. The literature highlights limitations such as preprocessing dependency, risk of overfitting, and low interpretability of multivariate models, which hinder their transferability to industrial scale [3]. Moreover, environmental variability, instrumental drift, and the inherent heterogeneity of foods reduce the stability of spectral signals, generating uncertainty in predictive performance. In Latin American countries, the lack of optical infrastructure, specialized personnel, and validation under real-world conditions aggravates this issue, restricting the adoption of advanced optical technologies in the food industry [4,5].

Several international studies have demonstrated the potential of VIS–NIR spectroscopy coupled with machine learning for non-destructive evaluation of food matrices. In Fu brick tea, the HSI–PKO–SVR combination enabled prediction of polyphenol content with $R^2 \approx 0.92$ [6], while in green tea, partial least squares (PLS)–DA models discriminated storage periods with accuracies of 96%–98%, and catechin regressions achieved $R_p > 0.90$ [7]. Likewise, in moisture quantification using HSI, $R^2 > 0.94$ values were obtained [1], confirming the sensitivity of the technique. Current literature, however, features a significant imbalance, as research predominantly concentrates on major bioactive

compounds like polyphenols and catechins, leaving critical physicochemical indicators such as pH and water activity (Aw) largely unexplored. Furthermore, most investigations evaluate calibrations under controlled or static conditions aimed primarily at quality grading or authentication, without addressing model transferability or the continuous tracking of degradation dynamics over time.

The present study addresses these limitations by integrating VIS–NIR spectroscopy obtained through HSI with machine learning algorithms to analyze black tea under accelerated storage. This study specifically focuses on monitoring the aging process and predicting time-dependent variations through the quantification of pH and Aw, establishing a clear distinction from conventional quality testing designed for adulteration detection. A reproducible pipeline is proposed based on the comparison of linear and nonlinear models, consensus-based band selection (VIP, elastic net [EN], RF, and XGBoost [XGB]), and leave-one-band-out (LOBO) validation. This approach seeks to quantify the contribution of stable and chemically interpretable spectral regions, ensuring model accuracy and transferability. In doing so, it aims to strengthen quality control and the competitiveness of the Peruvian food sector through modern optical tools.

2. Materials and methods

2.1. Samples and experimental design

Special loose-leaf black tea (*Camellia sinensis*) was used as the raw material, harvested in the province of Huayopata, Cusco, and artisanally processed by Machu Picchu Tea Company (Peru Special Products S.R.L., Lima, Peru). The samples were acquired through a purchase order, consisting of a 2-kilogram batch packaged under controlled conditions and transported to the city of Cajamarca. Upon arrival, the batch was stored in airtight, food-grade polyethylene containers, protected from light and humidity until experimental processing.

From the original 2-kilogram bulk batch, individual representative samples were sub-sampled and packaged into small, food-grade airtight bags to establish the experimental units. A full factorial experimental design was implemented under accelerated storage conditions, involving two main factors: storage time (6 levels, including day 0) and temperature (3 levels: 30, 40, and 50 °C). This combination resulted in 18 distinct experimental conditions.

To ensure high statistical reliability and capture spectral variability, 18 independent replicates were analyzed for each time-temperature combination, yielding a total dataset of 324 spectral and physicochemical observations. All subsequent chemical measurements and hyperspectral image acquisitions were performed systematically across these replicates.

Black tea samples were subjected to different storage conditions over a 30-day period. For each condition, VIS–NIR spectra were recorded in the range of 390–1050 nm (resolution ≈ 2.14 nm), together with reference physicochemical parameters: pH and Aw. In addition, categorical factors associated with the experimental design (i.e., day and temperature) were documented, including 18 replicates for each combination of levels. Thus, each row of the dataset integrated a spectrum with its corresponding analytical measurements.

2.2. Spectral acquisition and data collection

As can be seen in Figure 1, the methodological process starts with the measurements of samples. Spectral measurements were performed using a Resonon Pika IR-Plus hyperspectral imaging system (Resonon Inc., Bozeman, MT, USA), configured in reflectance mode with a spectral

range of 390–1050 nm and a nominal spectral resolution of 2.14 nm per channel. The system was equipped with a line-scanning (pushbroom) hyperspectral camera, diffuse halogen illumination lamps, and a Teflon® reflective reference panel (Spectralon®) for white/dark calibration prior to each sampling session [1].

Water activity and pH measurements were obtained using an AQUALAB 4TE and a Phs-3e digital pH meter, respectively. Both instruments were previously calibrated with their standard reference samples, and calibration was performed before each measurement to ensure consistent results [9,10].

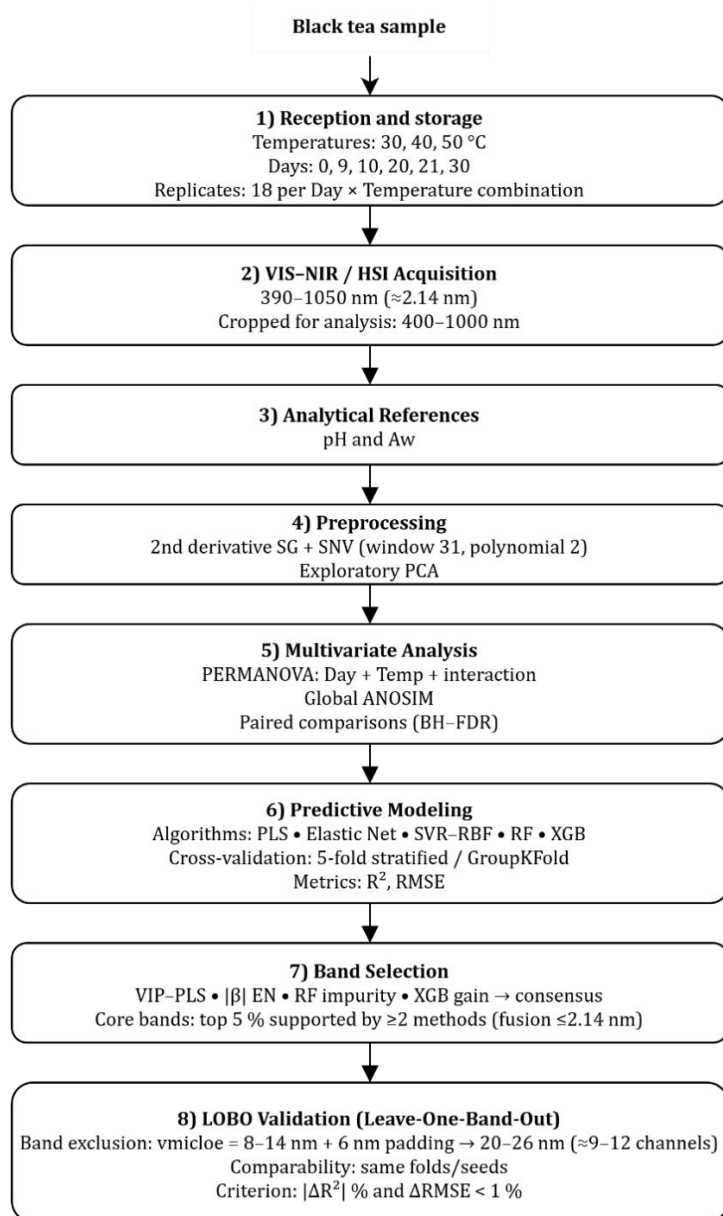


Figure 1. Experimental workflow adopted for VIS–NIR hyperspectral acquisition, preprocessing, multivariate analysis, predictive modeling, consensus band selection and LOBO validation. Black tea samples were stored at three temperatures (30, 40, and 50 °C) during six storage periods (0–30 days). For each experimental condition, hyperspectral images and physicochemical reference measurements (pH and Aw) were obtained before spectral preprocessing and machine learning analysis.

2.3. Spectral preprocessing

The spectra were trimmed to the 400–1000 nm interval to avoid regions with low signal-to-noise ratios at the extremes. A standard preprocessing routine based on the second-order Savitzky–Golay derivative (window = 31, polynomial = 2) followed by standard normal variate (SNV) was applied, implemented within pipelines to prevent information leakage during validation. For additional exploratory analyses, a simplified SNV-only preprocessing was also evaluated. In the case of principal component analysis (PCA) a smaller smoothing window (15 points) was used to highlight visual patterns without overfitting noise.

To reduce intra-replicate variability, an averaged matrix (X) was also constructed by averaging all spectra per repetition and associating the corresponding physicochemical parameters, yielding approximately 18 representative observations per treatment.

2.4. Exploratory and statistical analysis

The influence of the factors day and temperature on the spectral structure was evaluated through dissimilarity analysis. To this end, Euclidean distances were computed over the preprocessed spectra, and the following tests were applied:

- Multifactorial PERMANOVA.
- Global ANOSIM, to measure the strength of separation between groups.
- Pairwise comparisons adjusted for multiple testing using the Benjamini–Hochberg (BH-FDR) method.

2.5. Predictive modeling

For each response variable (pH and A_w), five representative algorithms were trained and compared:

- PLS;
- EN, with L1/L2 regularization;
- SVR with RBF kernel;
- Random Forest (RF);
- XGB.

The pipeline integrated two blocks: (i) One-hot encoding of the day and temperature factors and (ii) preprocessed spectra (SG second derivative + SNV, or variant). For each fold, preprocessing, one-hot encoding, model fitting, and prediction were performed entirely within a scikit-learn pipeline to avoid information leakage. The models were evaluated using 5-fold stratified cross-validation, reporting R^2 and root mean square error (RMSE) as performance metrics [11,12]. Additionally, a metadata-only ridge regression model, trained exclusively with the experimental factors (day and temperature), was included as a regularized linear baseline to quantify the incremental predictive value provided by the VIS–NIR spectra [6].

2.6. Consensus band selection

Spectral attribution was conducted by integrating four complementary variable-selection approaches: VIP (PLS), absolute coefficients (EN), mean impurity (FR), and gain (XGB). Each method generated an independent wavelength ranking, and subsequently, an average importance ranking was computed.

Adjacent positions with an interval ≤ 2.14 nm were merged into consensus bands, allowing the definition of continuous spectral intervals supported by multiple algorithms. Additionally, core bands

were established by considering the top 5% indicators from each method and retaining those supported by at least two algorithms, using the median number of supports as a stability indicator.

This approach aimed to minimize dependence on a single algorithm and to provide chemically plausible and statistically stable spectral regions, thereby reducing the risk of overfitting due to spurious signals.

2.7. LOBO validation

To assess the robustness of the core bands, the LOBO protocol was applied. For each identified band, the entire pipeline was retrained, excluding an interval defined as:

$$\lambda_0 \pm (W_{\text{core}} + W_{\text{padding}})$$

where:

- W_{core} corresponded to the band width (≈ 8 – 14 nm; median = 12 nm).
- $W_{\text{padding}} = 6$ nm, resulting in effective exclusions of 20 – 26 nm (≈ 9 – 12 channels).

Identical validation folds and random seeds were maintained relative to the base model to ensure comparability. The reported metrics were the relative variation of R^2 and RMSE with respect to the complete model, considering invariance when $|\Delta R^2| < 10^{-3}$ and the relative change in RMSE $< 1\%$.

This analysis determined whether prediction relied on isolated spectral peaks or whether the information was distributed across redundant spectral regions—an essential aspect for transferability across different instruments or experimental conditions.

To ensure statistical validity, all transformation steps (one-hot encoding, scaling, SG–SNV, PCA) were executed within cross-validation pipelines in scikit-learn, thus preventing information leakage between training and testing sets.

The hyperparameters of the selected models were optimized using an internal cross-validation scheme with compact search grids, restricting the exploration space to reduce the risk of overfitting.

Regarding statistical tests, p-values were adjusted using the Benjamini–Hochberg (BH-FDR) procedure to control the false discovery rate.

2.8. Implementation and reproducibility

This set of measures ensured that the reported metrics reflected generalizable performance rather than fitting to experimental noise or artificial replications.

Data processing and analysis were implemented in Python, using the free Google Colab environment, with the libraries pandas, numpy, scikit-learn, xgboost, and matplotlib. A random_state value of 42 was set to guarantee the reproducibility of the experiments.

The following were stored as research outputs:

- Analysis notebooks.
- Performance tables by variable and model.
- LOBO exclusion results (CSV).
- Summary figures by variable.

3. Results

3.1. Physicochemical data

During accelerated storage, black tea samples exhibited a consistent decline in both pH and Aw, as shown in Figure 2. At the beginning of the experiment, pH values averaged approximately 5.11 ± 0.02 across all temperature conditions.

As storage progressed, a steady acidification was observed, with final pH values near 4.70 ± 0.04 after 30 days. The decrease followed a quasi-linear trend, with samples stored at higher temperatures (50 °C) showing a slightly faster decline during the first 10 days. Nevertheless, all temperature profiles converged at the end of the storage period, suggesting a chemically driven, temperature-dependent acidification process.

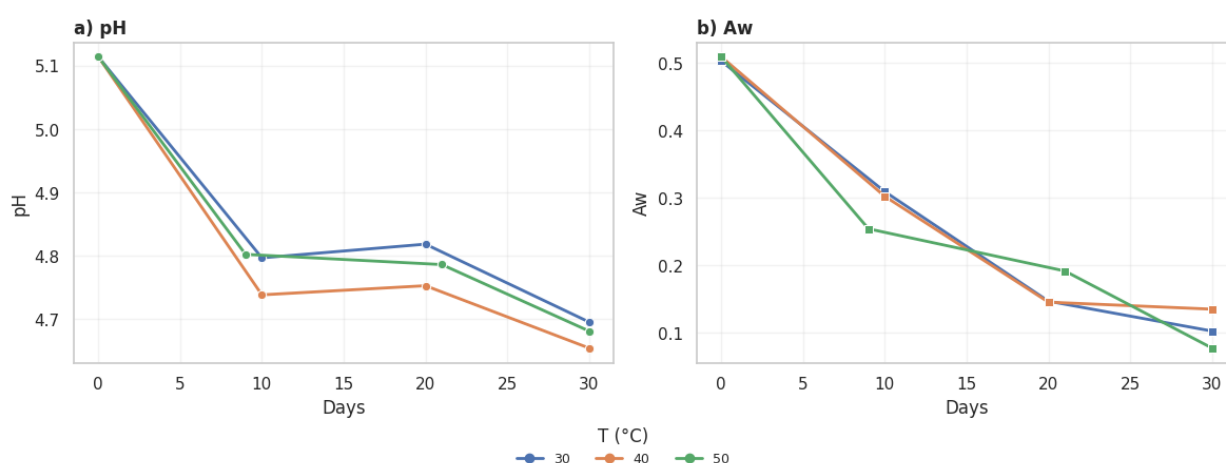


Figure 2. Evolution of pH (a) and Aw (b) of black tea stored under accelerated conditions (30, 40, and 50 °C) over 30 days. Curves represent the mean values obtained from 18 independent replicates per day–temperature combination.

Similarly, Aw exhibited a marked reduction throughout storage (Figure 2(b)). Initial measurements recorded Aw values around 0.50 ± 0.01 , which progressively dropped to approximately 0.10 ± 0.02 by day 30. The three temperature treatments displayed parallel downward trajectories, with minor fluctuations during intermediate storage days (around days 10–20), particularly at 40 °C. By the end of the experiment, all samples stabilized around similar moisture activity levels, indicating pronounced loss or redistribution of free water across the evaluated conditions.

3.2. Initial exploration of spectra (PCA)

The PCA applied to the preprocessed VIS–NIR spectra revealed a structured latent space associated with the experimental conditions (Figure 3). The first two principal components explained 60.2% of the total spectral variance, with PC1 accounting for 44.7% and PC2 for 15.5%. Although this percentage is lower than that observed in highly homogeneous datasets, it captures the dominant sources of spectral variability and reveals a well-defined organization of samples according to storage conditions. Fresh samples (D0) were clearly separated from the remaining groups, whereas intermediate and late storage conditions occupied progressively distinct regions of the score space,

indicating systematic modifications in the spectral signatures during accelerated storage.

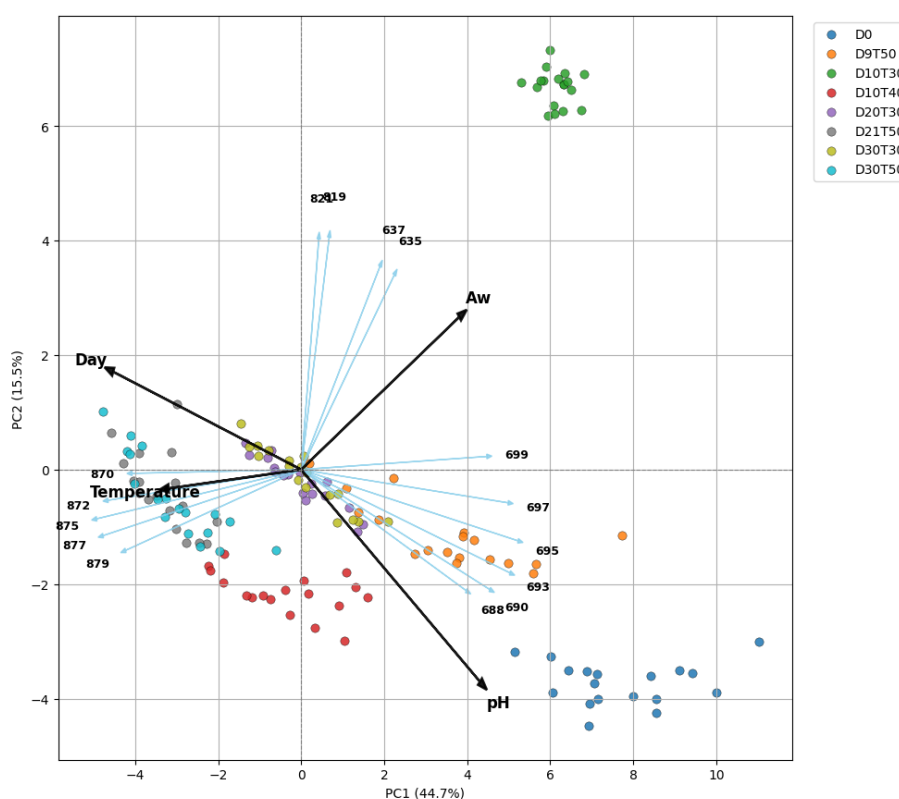


Figure 3. PCA biplot obtained from preprocessed VIS–NIR spectra (SG second derivative + SNV). Sample scores are colored according to storage condition, whereas arrows represent physicochemical variables (pH and Aw), experimental factors (day and temperature) and the wavelengths with the highest loadings.

The PCA biplot further demonstrated coherent relationships between the spectral data and the measured physicochemical parameters. The pH vector was oriented toward the positive direction of PC1 and the negative direction of PC2, whereas Aw was projected toward the positive quadrant of both components. This configuration indicates that both variables contribute positively to the main source of spectral variation represented by PC1, while PC2 differentiates their respective contributions. Conversely, the vectors corresponding to storage day and temperature were closely aligned toward the negative direction of PC1, suggesting that these experimental factors were positively associated with each other and inversely related to the physicochemical parameters. This opposite orientation is consistent with the progressive decrease in pH and water activity as storage time and temperature increased.

The loading vectors revealed that the spectral discrimination was primarily driven by specific wavelength groups rather than isolated variables. Bands around 635–637 nm and 819–821 nm contributed predominantly to the positive direction of PC2, whereas wavelengths between approximately 688 and 699 nm were mainly associated with positive PC1 values. In contrast, the 870–879 nm region exhibited negative PC1 loadings, contributing to the separation of samples subjected to advanced storage conditions. These loading patterns indicate that both visible and short-wave near-infrared regions participate in the multivariate discrimination of black tea during storage, supporting the subsequent feature-selection analyses and reinforcing the chemical interpretability of the predictive models.

3.3. Differences by experimental design

Multivariate statistical tests confirmed that the clustering observed in the PCA corresponded to significant effects of the experimental factors day and temperature on the spectral data.

The multifactorial PERMANOVA (Table 1) detected strong and statistically significant contributions from both storage time and temperature. The factor day exhibited the highest effect (pseudo-F = 346.8; $p_{\text{perm}} = 0.0005$), followed by temperature (pseudo-F = 244.6; $p_{\text{perm}} = 0.0005$). A significant interaction effect between day $\times$ temperature was also identified (pseudo-F = 49.0; $p_{\text{perm}} = 0.0005$), indicating that storage temperature modulated the time-dependent evolution of spectral patterns in black tea.

Similarly, the global ANOSIM test supported these findings, yielding R values between 0.50 and 0.55 ($p_{\text{perm}} \leq 0.0001$) for both factors. The overall PERMANOVA confirmed this trend, with robust pseudo-F values of 405.4 for day and 367.5 for temperature, both with $p_{\text{perm}} = 0.0001$.

The pairwise comparisons, adjusted using BH-FDR, further revealed systematic differences between storage days. For instance, the contrast between D20 and D0 reached a pseudo-F of 1121.94 ($p_{\text{adj}} = 0.001$), and between D21 and D0 a pseudo-F of 1104.96 ($p_{\text{adj}} = 0.001$). Other groups such as D10 vs D30 (pseudo-F = 299.2; $p_{\text{adj}} = 0.001$) also exhibited notable divergences, confirming that the spectral differences were non-random and reflected progressive degradation processes.

Table 1. Results of multifactorial PERMANOVA, global ANOSIM and pairwise comparisons performed on preprocessed VIS–NIR spectra.

Factor/Comparison	Pseudo-F	p_{perm}
Multifactorial PERMANOVA		
Day	346.8	0.0005
Temperature	244.6	0.0005
Day x Temp	49	0.0005
Global ANOSIM / PERMANOVA		
Day	405.4	0.0001
Temperature	367.5	0.0001
Pairwise Comparisons (BH-FDR Adjusted)		
D20 vs D21	8.96	0.0005
D20 vs D30	20.03	0.0005
D20 vs D10	387.29	0.0005
D20 vs D9	517.48	0.0005
D20 vs D0	1121.94	0.0005
D21 vs D10	224.25	0.0005
D21 vs D9	413.67	0.0005
D21 vs D0	1104.96	0.0005
D10 vs D9	246.62	0.0005
D30 vs D10	299.2	0.0005

*Note: Pseudo-F values correspond to permutation-based multivariate analysis. Pairwise p-values were adjusted using the Benjamini–Hochberg false discovery rate (BH-FDR).

3.4. Explanatory modeling and algorithm comparison

The predictive models developed for pH and Aw using spectral and metadata features were first evaluated through a baseline train–test split (Table 2), revealing markedly different performance depending on both the target variable and the algorithm.

For pH prediction, the SVR with RBF kernel achieved the best performance with $R^2 \approx 0.94$ and RMSE ≈ 0.029 , indicating a robust predictive relationship between the VIS–NIR spectrum and this parameter after preprocessing. RF and XGB also reached similar levels of accuracy ($R^2 \approx 0.935$ – 0.934 ; RMSE ≈ 0.029 – 0.030), suggesting that ensemble-based models are effective at capturing the variability related to pH.

In contrast, the prediction for Aw showed moderate performance. SVR remained the best-performing model ($R^2 \approx 0.82$; RMSE ≈ 0.087), followed by RF and XGB, which yielded R^2 values close to 0.79. This difference suggests that the relationship between spectrum and water activity is more complex and potentially nonlinear.

The PLS regression model consistently demonstrated lower performance for both variables ($R^2 = 0.42$ – 0.78), confirming that fixed latent-space projections are less effective in capturing spectroscopic complexity compared to adaptive or nonlinear methods.

Table 2. Baseline predictive performance obtained using the train–test split.

Variable	Model	R^2	RMSE
pH	SVR (RBF)	0.937849	0.028655
	RandomForest	0.93516	0.02909
	XGBoost	0.934338	0.029509
	ElasticNet	0.902084	0.035898
	PLS	0.777004	0.054313
Aw	SVR (RBF)	0.820928	0.087327
	RandomForest	0.799276	0.093905
	XGBoost	0.787424	0.096753
	ElasticNet	0.786646	0.095987
	PLS	0.42277	0.167439

*Note: Higher R^2 and lower RMSE indicate better predictive performance.

Table 3. Cross-validation performance of the best predictive models.

Variable	Model	R^2 ($\pm$ EE)	RMSE ($\pm$ EE)
Aw	Meta-only Ridge	0.572 ± 0.029	0.144 ± 0.005
	SVR (spectrum + metadata)	0.831 ± 0.030	0.090 ± 0.008
pH	Meta-only Ridge	0.666 ± 0.022	0.066 ± 0.003
	SVR (spectrum + metadata)	0.945 ± 0.003	0.027 ± 0.001

*Note: Meta-only ridge corresponds to a model trained exclusively with metadata (day and temperature), whereas SVR combines metadata and VIS–NIR spectral variables. Values are reported as mean $\pm$ standard error (SE) across the five validation folds.

To evaluate the stability and generalization of the optimized models, stratified five-fold cross-validation was performed. Table 3 reports the mean predictive performance and its associated standard error across the validation folds, providing an estimate of model variability.

To determine the additional predictive value provided by the VIS–NIR spectra, the best-performing spectral model (SVR) was compared against a baseline ridge regression model trained exclusively with the experimental metadata (Day and temperature). Ridge was selected as the reference model because it provides a robust regularized linear baseline, enabling the incremental predictive value of the spectral information to be quantified relative to the experimental design variables alone. Five-fold cross-validation showed that incorporating spectral information substantially improved predictive performance for both physicochemical parameters. For pH, R^2 increased from 0.666 to 0.945, whereas for Aw it increased from 0.572 to 0.831. These results demonstrate that the spectral information captures chemical variability not explained solely by the experimental design variables.

These results demonstrate the advantage of integrating spectral information into the prediction of critical quality parameters, reinforcing its applicability in at-line and on-line monitoring systems.

3.5. Spectral attribution

To identify the most informative spectral regions for predicting pH and Aw, we integrated the results from four feature selection techniques: VIP–PLS, SVR coefficients, RF impurity-based importance, and XGB gain. This consensus-based approach enables the detection of relevant spectral intervals consistently highlighted across multiple algorithms.

Figures 4(a),(b) show the consensus maps for pH and Aw, respectively, illustrating the regions where the majority of methods converge in identifying high relevance.

For pH prediction (Figure 4(a)), two key blocks were detected:

- Visible range (390–550 nm): linked to polyphenolic chromophores responsible for tea's coloration.
- Short NIR (870–960 and 980–1000 nm): associated with O–H and C–H vibrational overtones.

Within these intervals, core bands—narrow peaks supported by at least two methods—were found at: 649–650 nm, 874–886 nm, 923–926 nm, 937 nm, and 955 nm. These signals, combining chromatic and vibrational responses, support the high predictive power observed for pH ($R^2 \approx 0.94$).

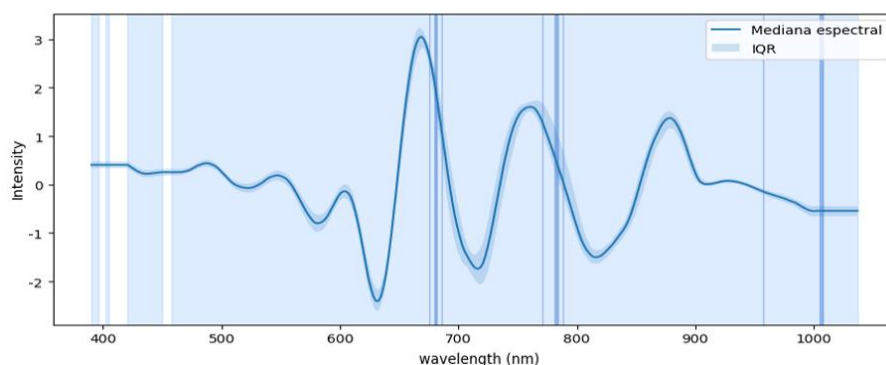
For Aw prediction (Figure 4(b)), sensitivity was observed across the spectrum, with two key regions standing out:

- 675–690 nm, associated with water–matrix interactions.
- 957–1007 nm, attributed to O–H overtones directly related to water content.

Core bands were concentrated at: 675.3 nm, 679.6–681.8 nm, 686.1 nm, 770.8 nm, 781.8–783.9 nm, 788.3–788.9 nm, 957.5 nm, and 1005–1007 nm, supporting the role of water distribution and structural dynamics in Aw prediction.

Figure 4 and Table 4 collectively show how VIS–NIR spectroscopy captures both surface pigment responses and structural water dynamics, enabling accurate modeling of black tea deterioration.

a) pH



b) Aw

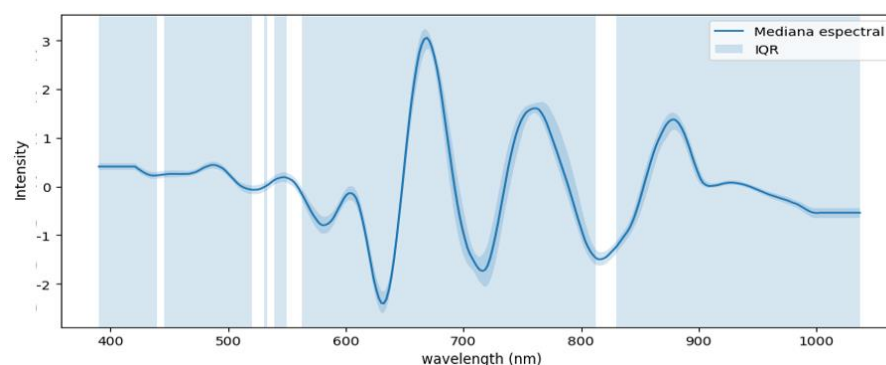


Figure 4. Consensus spectral relevance maps for pH (a) and Aw (b). The blue curve represents the median spectral importance obtained from the integrated feature-selection methods, while the shaded region corresponds to the interquartile range (IQR). White vertical regions indicate the consensus spectral intervals identified as relevant across multiple algorithms.

Table 4. Relevant and core bands.

Variable	Model	R ²	RMSE
pH	Consensus	390–550	Chromophores (polyphenols, color)
	Core	649–650	Visible absorption (related to color/pigments)
	Core	874–886	O–H / C–H overtones
	Core	923–926	O–H vibrations (water)
	Core	937	Combined O–H
	Core	955	C–H overtone / matrix interaction
	Core	1005–1007	O–H / C–H combinations
Aw	Consensus	675–690	Water–matrix interaction
	Core	675.3	Hydrated chromophore
	Core	679.6–681.8	Visible O–H interaction
	Core	686.1	Chromatic variation linked to moisture
	Core	770.8	Water–matrix transition band
	Core	781.8–783.9	O–H vibrational signals
	Core	788.3–788.9	Matrix–water interaction
	Core	957.5	O–H overtones (free water)
	Core	1005–1007	O–H / C–H combinations
	Core	1005–1007	O–H / C–H combinations

*Note: Consensus bands correspond to spectral intervals identified by multiple feature-selection algorithms, whereas core bands represent the highest-confidence wavelengths supported by at least two methods.

For pH (Figure 5(a)), the reference line at $VIP = 1$ delineated critical zones: in the visible range (400–550 nm), peaks attributable to polyphenolic chromophores stood out; between 650–700 nm, signals associated with pigment oxidation emerged; and between 870–960 nm, the highest peaks ($VIP > 1.4$) were concentrated, corresponding to O–H and C–H absorptions. The 950–1000 nm region maintained elevated values, reinforcing the role of these signals in predicting the parameter.

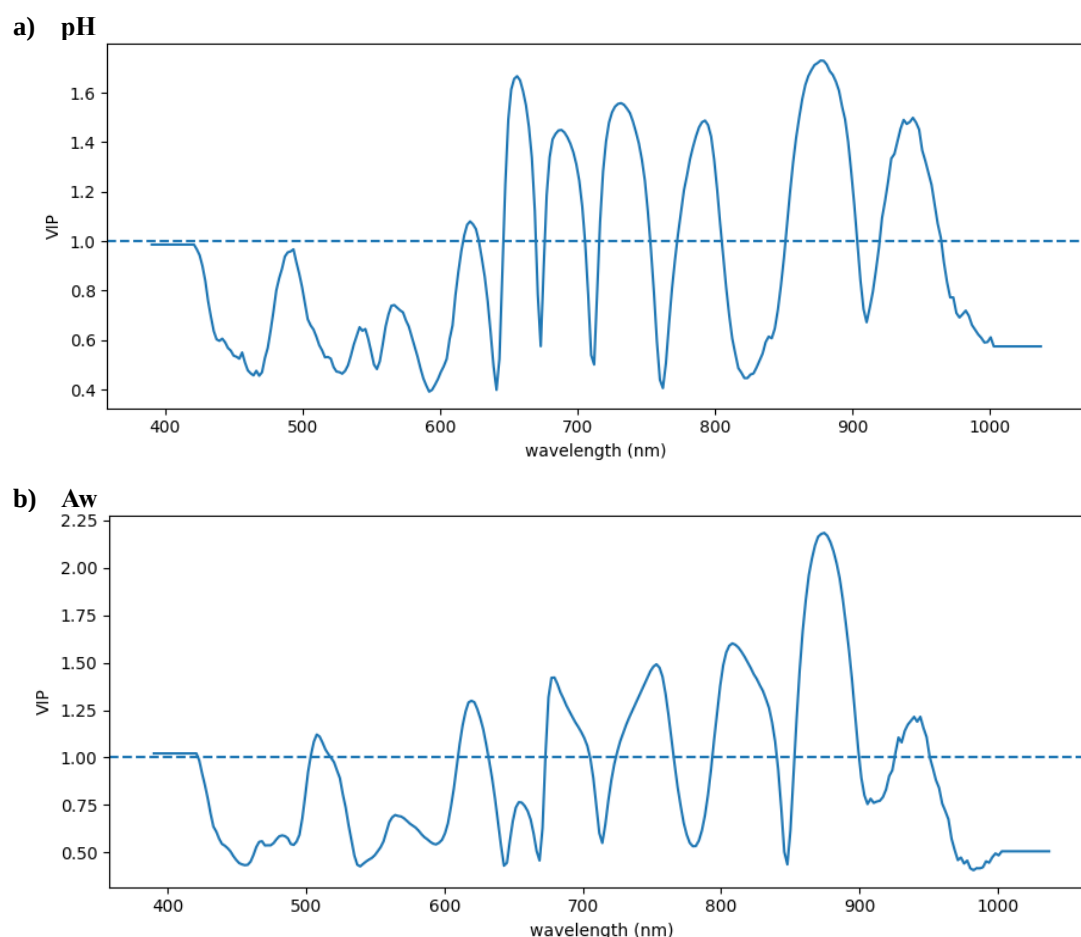


Figure 5. Variable importance in projection (VIP) profiles for pH (a) and Aw (b). The dashed horizontal line ($VIP = 1$) indicates the conventional threshold above which wavelengths are considered influential in the PLS model.

3.6. Validation by band exclusion (LOBO)

The LOBO protocol evaluated the robustness of the best-performing models for pH and Aw through the systematic exclusion of each core band (Table 5). In each case, the entire pipeline was retrained while maintaining the same validation folds, and the resulting metrics were compared against the baseline model.

Table 5. LOBO validation for pH and Aw.

Variable	Model	R ² (±EE)	RMSE (±EE)
pH	(EN/RF)	0.93–0.94	0.0006
Aw	(EN/RF/SVR)	0.79–0.82	0.0001

*Note: Values summarize the range obtained after excluding each consensus band while maintaining identical validation folds.

The results showed that, after removing any band, the change in model fit was practically negligible:

- Variations in R² remained close to zero ($|\Delta R^2| < 10^{-3}$).
- Relative changes in RMSE were below 1%, effectively indistinguishable from 0.

These findings suggest that the information useful for predicting pH and Aw is not concentrated in a single spectral peak but rather distributed across contiguous windows within the visible and short-NIR regions. In practice, this implies that the models are resilient to instrumental variations (e.g., lamp drift or calibration shift) and can be transferred to VIS–NIR instruments with different discretizations without performance loss.

Similarly, to further strengthen the robustness assessment, the LOBO analysis was repeated using the effective exclusion windows generated by each core band together with the predefined ± 6 nm padding, corresponding to approximately 20–26 nm-wide spectral intervals. This complementary analysis evaluated whether removing broader contiguous regions affected model performance.

Figure 6 shows the variation in the coefficient of determination (ΔR^2) when each band is excluded, where values close to zero indicate minimal impact. For pH (blue bars) and Aw (orange bars), most exclusions produced negligible changes ($\approx 10^{-4}$), confirming spectral redundancy. Only for pH, around 890 nm, a slight drop was observed ($\Delta R^2 \approx -0.001$), without compromising the overall model fit.

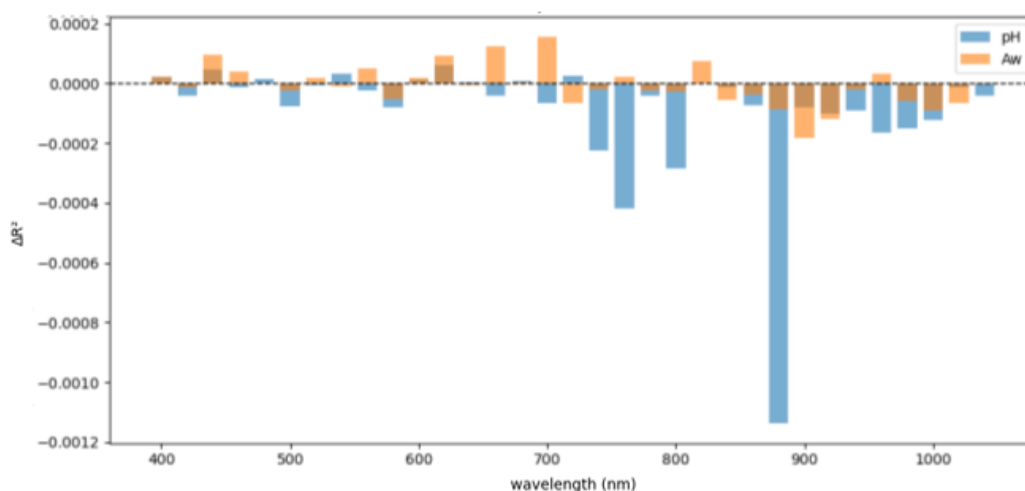


Figure 6. Variation in model performance (ΔR^2) after exclusion of each consensus spectral band during LOBO validation. Blue bars correspond to pH and orange bars to Aw. Values close to zero indicate high spectral redundancy and robustness of the predictive models.

4. Discussion

During accelerated storage, the progressive decline in pH and Aw reflects dual processes: acidification and moisture loss. This finding is consistent with catechins and production of low-pKa

organic acids reported by [14], while the reduction in free water echoes the observations of Pujapanda et al. [15] on plant matrices. Together, these changes suggest that black tea's stability is compromised through both chemical transformation and water redistribution. The stable acidification seen across temperatures indicates that these reactions may be more dependent on internal matrix chemistry than external thermal variations.

The PCA results indicate that the spectral variability of black tea is jointly governed by storage time, temperature, and the associated physicochemical changes, with the first two principal components explaining 60.2% of the total variance. Although this proportion is lower than that reported for more homogeneous systems, it adequately captures the dominant sources of variation and reveals a structured organization of samples according to the experimental conditions. Similar multivariate patterns have been reported for food matrices undergoing progressive storage-related transformations, where spectral responses reflect the combined influence of environmental factors rather than a single degradation process [16]. The biplot further showed that both pH and water activity contributed positively to PC1, while their separation along PC2 suggests partially distinct responses to storage. Conversely, the opposite orientation of the day and temperature vectors indicates that the progressive physicochemical deterioration is closely associated with the experimental storage conditions. The distribution of wavelength loadings across the visible and short-wave near-infrared regions further supports that both chromophoric changes and water-related absorptions contribute to the observed spectral evolution, consistent with concurrent polyphenol oxidation and moisture redistribution during black tea storage [17].

Multifactorial PERMANOVA and ANOSIM analyses highlight significant effects of both storage day and temperature on spectral patterns. High pseudo-F values (e.g., day: 346.8; temperature: 244.6) and robust R-values (≈ 0.50 – 0.55) indicate that these variables impart strong, independent influences on tea spectra. The significant day $\times$ temperature interaction (pseudo-F = 49.0) points to synergistic effects: higher temperatures accelerate chemical changes, while time extends these alterations. Pairwise comparisons further show pronounced shifts between early and late storage days (D20 vs D0), consistent with progressive oxidation and the formation of complex, higher-molecular-weight compounds [18].

Across algorithm comparisons, SVR, RF, and XGB yield superior predictive performance for pH ($R^2 \approx 0.94$) and Aw ($R^2 \approx 0.82$), while PLS lags behind. These findings align with Torres de Barros et al. [19], who observed PLS limitations in adulterated creatine analysis, and with Elamshity and Alhamdan [20], whose quality index for dried fruits favored neural networks over PLS. The results suggest that nonlinear relationships and ensemble learning better capture complex spectral patterns in black tea, underscoring the importance of flexible modeling frameworks for accurate predictions in plant-based matrices.

Cross-validation results demonstrate that models utilizing both spectral and metadata variables outperform those relying solely on metadata. The combined approach yields a higher coefficient of determination for both pH and Aw ($R^2 = 0.945$ and 0.831) compared with the meta-only Ridge models. This finding emphasizes that spectral data provide critical chemical insights beyond basic sample descriptors. It echoes the conclusions of Elamshity and Alhamdan [20], who noted that including spectral information significantly boosts predictive accuracy. Thus, spectroscopic data are essential for capturing subtle changes in tea chemistry that metadata alone cannot explain.

The consensus band analysis indicates that pH is influenced by both visible and short-NIR regions: 400–550 nm relates to chromophoric (polyphenolic) signals, while 870–960 nm and 980–1000 nm involve O–H and C–H overtones. These results are consistent with known chemical pathways in tea degradation, where oxidation of polyphenols and changes in water structure drive shifts in pH and

color [17]. Conversely, Aw correlates more strongly with short-NIR O–H bands (675–690 nm and 957–1007 nm), reflecting the redistribution of free and bound water in the matrix. This duality supports the notion that both phenolic oxidation and moisture dynamics govern tea quality [21].

LOBO analysis reveals that excluding individual core bands has negligible effects on model performance ($|\Delta R^2| < 10^{-3}$), confirming that useful information is distributed across contiguous windows rather than concentrated in specific peaks. This redundancy ensures that the models remain stable despite instrumental variations or minor data loss. Similar patterns of resilience have been reported in other spectral studies [22,23], indicating that robust predictive models benefit from capturing cumulative regional signals rather than relying on isolated wavelengths. The only minor sensitivity observed was a slight drop at ~890 nm for pH, which did not compromise overall robustness.

The robust performance under LOBO conditions and the validation by 20 nm interval exclusions imply that the developed models could be transferred to different VIS–NIR devices or applied in diverse processing environments. This supports the idea that targeted band selection and multialgorithm consensus are critical for creating generalizable models [24]. The collective findings suggest that the detection of black tea deterioration can be reliably monitored in real time, facilitating industrial implementation and enhancing product quality control. Integrating such robust models across processing lines would allow producers to maintain consistency while adapting to inevitable instrument or environmental variability.

5. Conclusions

Under a reproducible pipeline based on second-order Savitzky–Golay derivative and SNV preprocessing, followed by stratified validation, SVR-RBF consistently outperformed PLS, EN, RF, and XGB in predicting the evaluated physicochemical parameters. The reproducibility of the predictive performance across validation folds indicates that the relationship between the VIS–NIR spectra and the target properties remained stable and highly predictable after preprocessing.

The fusion of importance metrics (VIP–PLS, absolute coefficients from SVR, impurity from RF, and gain from XGB) delineated consensus bands in the visible and short-NIR regions (e.g., ~675–690, 870–960, and 980–1000 nm). LOBO validation (band exclusion with fixed padding) showed ΔR^2 and ΔRMSE values near zero, indicating local informational redundancy and confirming that prediction does not depend on isolated peaks but on neighboring windows that capture the same vibrational overtones/combinations.

Multivariate analysis (PCA) together with dissimilarity tests (PERMANOVA/ANOSIM) demonstrated that the spectral evolution of black tea was jointly influenced by storage time, temperature, and the associated physicochemical changes. The agreement between unsupervised spectral organization and statistical hypothesis testing confirms that VIS–NIR spectroscopy effectively captures the progressive aging of the tea matrix. Furthermore, the baseline evaluation followed by stratified five-fold cross-validation ensured that the reported predictive performance reflects robust within-study generalization rather than overfitting to the experimental dataset.

LOBO invariance and redundancy across contiguous windows support the feasibility of compact VIS–NIR sensors (400–1000 nm) with different discretizations and slight spectral shifts while maintaining performance for pH and Aw. This enables at-line/on-line implementations with lower recalibration frequency and transferability between instruments through spectral standardization procedures.

Future work should assess the sensitivity of LOBO to bandwidth and padding, as well as perform external validation by supplier and cross-instrument transfer.

Author contributions

André Rodríguez-León: Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft, writing—review & editing, visualization, supervision, project administration, funding acquisition; Jimy Oblitas: Methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft, Writing—review & editing, visualization, supervision, project administration, funding acquisition. All authors have read and approved the final version of the manuscript for publication.

Use of AI tools declaration

Generative AI tools were used only to improve the clarity and readability of some paragraphs. All content was critically reviewed and validated by the authors.

Acknowledgments

The authors declare that no specific funding was received for this study. The authors also thank the Universidad Nacional de Cajamarca for providing institutional support.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Zhang Y, Liao B, Gouda M, et al. (2025) Visualization of moisture distribution in stacked tea leaves on process flow line using hyperspectral imaging. *Foods* 14: 1551. <https://doi.org/10.3390/foods14091551>
2. Luo X, Sun C, He Y, et al. (2023) Cross-cultivar prediction of quality indicators of tea based on VIS-NIR hyperspectral imaging. *Ind Crop Prod* 202: 117009. <https://doi.org/10.1016/j.indcrop.2023.117009>
3. Li D, Chen Q, Ouyang Q, et al. (2025) Advances of VIS/NIRS and imaging techniques assisted by AI for tea processing. *Crit Rev Food Sci* 65: 7575–7593. <https://doi.org/10.1080/10408398.2025.2474183>
4. Feng L, Wu B, Zhu S, et al. (2021) Application of visible/infrared spectroscopy and hyperspectral Imaging with machine learning techniques for identifying food varieties and geographical origins. *Front Nutr* 8. <https://doi.org/10.3389/fnut.2021.680357>
5. Ye W, Xu W, Yan T, et al. (2022) Application of near-infrared spectroscopy and hyperspectral imaging combined with machine learning algorithms for quality inspection of grape: A review. *Foods* 12: 132. <https://doi.org/10.3390/foods12010132>
6. Gong J, Chen G, Deng Y, et al. (2024) Non-destructive detection of tea polyphenols in Fu Brick tea based on hyperspectral imaging and improved PKO-SVR method. *Agriculture* 14: 1701. <https://doi.org/10.3390/agriculture14101701>
7. Li F, Shen J, Yang Q, et al. (2024) Monitoring quality changes in green tea during storage: A hyperspectral imaging method. *Food Chem X* 23: 101538. <https://doi.org/10.1016/j.fochx.2024.101538>

8. Rodríguez-León A, Oblitas J, Quevedo-Olaya JL, et al. (2026) Non-destructive detection of *Elasmopalpus lignosellus* infestation in fresh asparagus using VIS–NIR hyperspectral imaging and machine learning. *Foods* 15: 355. <https://doi.org/10.3390/foods15020355>
9. Campbell GS, Galloway M, Campbell Z (2022) Measurement of water activity in the presence of high volatile concentrations using a tunable diode laser—single laboratory validation, first action 2021.04. *JAOAC Int* 105: 649–656. <https://doi.org/10.1093/jaoacint/qsac003>
10. Delgado M, Vanegas M, Delgado G (2007) Metrología química I: Calibración de un pHmetro y control de calidad. *Universitas* 1: 14–20. <https://doi.org/10.5377/universitas.v1i1.1627>
11. Felizzato G, Bagnulo E, Tapparo G, et al. (2026) Methodological challenges in machine learning and deep learning applied to food analysis: A critical review. *J Chromatogr A* 1783: 467176. <https://doi.org/10.1016/j.chroma.2026.467176>
12. Ban Y, Ren Y, Dong W, et al. (2026) Deep neural network frameworks (DNNFs) for enhanced non-destructive spectroscopic quantification in foods: From hybrid pipelines to end-to-end modelling. *Trends Food Sci Tech* 175: 105834. <https://doi.org/10.1016/j.tifs.2026.105834>
13. Gillariose J, Joseph J, Chesneau C (2026) Lasso and Ridge regression: A comprehensive review of applications and developments in machine learning. *Int J Data Sci Anal* 21: 7. <https://doi.org/10.1007/s41060-025-00957-y>
14. Zheng Q, Li W, Gao X (2021) The effect of storage time on tea polyphenols, catechin compounds, total flavones and the biological activity of Ya'an Tibetan tea (*Camellia sinensis*). *J Food Process Pres* 45. <https://doi.org/10.1111/jfpp.16004>
15. Pujapanda R, Bashir O, Shams R, et al. (2025) Influence of carrier agents on nutritional, chemical, and antioxidant profiles of fruit juices during drying processes: A comprehensive review. *DIFO* 5: 228. <https://doi.org/10.1007/s44187-025-00504-5>
16. Saputra A, Herpandi H, Supriadi A, et al. (2025) Spectroscopic and chemometric characterization of storage-induced changes in pempek Palembang using FT-IR and PCA. *Int J Design Nat Ecodynamics* 20: 1323–1332. <https://doi.org/10.18280/ijдне.200612>
17. Giorgini E, Notarstefano V, Foligni R, et al. (2023) First ATR-FTIR characterization of black, green and white teas (*Camellia sinensis*) from European tea gardens: A PCA analysis to differentiate leaves from the In-Cup infusion. *Foods* 13: 109. <https://doi.org/10.3390/foods13010109>
18. Falcioni R, Moriwaki T, Sversut M (2022) Classification and prediction by pigment content in lettuce (*Lactuca sativa* L.) varieties using machine learning and ATR-FTIR spectroscopy. *Plants* 11: 3413. <https://doi.org/10.3390/plants11243413>
19. de Barros AT, Santos J, Silveira G, et al. (2025) Partial least squares methods in open-source software for identification of adulterated creatine using infrared spectroscopy. *Food Chem* 483: 143998. <https://doi.org/10.1016/j.foodchem.2025.143998>
20. Elamshity M, Alhamdan A (2025) Development and prediction of a non-destructive quality index (Qi) for stored date fruits using VIS–NIR spectroscopy and artificial neural networks. *Foods* 14: 3060. <https://doi.org/10.3390/foods14173060>
21. Liu J, Luan X, Liu F (2024) *Full wavelength selection and deep spectra feature extraction network for near-infrared spectroscopy quantitative analysis*, In: 2024 IEEE 13th Data Driven Control and Learning Systems Conference (DDCLS), IEEE, 1350–1355. <https://doi.org/10.1109/DDCLS61622.2024.10606701>

22. Malegori C, Muncan J, Mustorgi E, et al. (2022) Analysing the water spectral pattern by near-infrared spectroscopy and chemometrics as a dynamic multidimensional biomarker in preservation: rice germ storage monitoring. *Spectrochim Acta A* 265: 120396. <https://doi.org/10.1016/j.saa.2021.120396>
23. Visnupriyan R, Flanagan BM, Harper KJ, et al. (2024) Near infrared spectroscopy combined with chemometrics as tool to monitor starch hydrolysis. *Carbohydr Polym* 324: 121469. <https://doi.org/10.1016/j.carbpol.2023.121469>
24. Qi W, Jiang Q, Ma T, et al. (2025) Near-Infrared spectroscopy combined with chemometrics for liquor product quality assessment: A review. *Foods* 14: 2992. <https://doi.org/10.3390/foods14172992>



AIMS Press

© 2026 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0>)