IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 3, June 2026, pp. 2300~2313
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i3.pp2300-2313

Comparing global and variety-specific ensemble models for avocado maturity prediction with near-infrared

Christian Ovalle, Jhonn Garcia Jimenez, Jose Briones Zuñiga
Department of Engineering, Faculty of Engineering, Universidad Tecnológica del Perú, Lima, Perú

Article Info

Article history:

Received Sep 19, 2025
Revised Feb 16, 2026
Accepted Apr 20, 2026

Keywords:

Avocado ripeness
Ensemble stacking
Machine learning
Near-infrared spectroscopy
Non-destructive quality control
Precision agriculture
Variety-specific models

ABSTRACT

Ensuring accurate, non-destructive maturity classification of avocados is critical to supply chain optimization in agro-industrial systems. This study presents a predictive framework that integrates near-infrared (NIR) spectroscopy with ensemble stacking machine learning (ML) to enhance the precision of avocado ripeness assessment. The proposed methodology compares global versus variety-specific models for 'Hass' and 'Fuerte' avocado types, leveraging spectral data (900–1,700 nm) and multiple base classifiers, including random forest (RF), gradient boosting (GB), support vector machines (SVMs), decision trees (DT), k-nearest neighbors (KNN), and categorical boosting (CatBoost), combined via linear regression as a meta-learner. Experimental results revealed that the stacking models outperformed individual learners, with variety-specific GB model achieving the highest performance (Matthews correlation coefficient (MCC) =0.679, area under the curve (AUC) =0.931). These findings highlight the critical importance of varietal specificity in model calibration and demonstrate how ensemble strategies can improve robustness, scalability, and interpretability in intelligent agricultural systems. The proposed model provides a computationally efficient solution for real-time quality control and supports the deployment of AI-powered systems within agricultural supply chains in developing regions.

Corresponding Author:

Christian Ovalle
Department of Engineering, Faculty of Engineering, Universidad Tecnológica del Perú
Lima, Peru
Email: dovalle@utp.edu.pe

1. INTRODUCTION

In recent years, the avocado industry has consolidated its position as one of the most economically valuable and strategically relevant sectors within Peruvian agriculture [1], [2]. Peru currently ranks among the world's leading avocado exporters, consistently occupying a position within the top three global producers, driven by sustained international demand and the increasing preference for high-quality fruit in global markets [3], [4]. This context highlights the urgent need for efficient, reliable, and scalable methods capable of assessing internal quality attributes that determine commercial maturity, shelf life, and postharvest performance [5].

Conventional maturity assessment techniques, including firmness testing and dry matter determination (coefficients of determination or R^2), are widely used but present significant limitations. These methods are destructive, labor-intensive, time-consuming, and unsuitable for large-scale or real-time industrial applications [6]–[8]. As a response to these constraints, near-infrared (NIR) spectroscopy has emerged as a promising non-destructive alternative, enabling rapid estimation of internal quality parameters such as dry matter, sugar content, and lipid concentration without damaging the fruit [9], [10]. Owing to its rapid acquisition, portability, and reagent-free operation, NIR spectroscopy has been successfully deployed in both laboratory environments and commercial packing lines, positioning it as a practical tool for quality control in the agri-food industry [11]–[13].

Extensive research has demonstrated the applicability of NIR spectroscopy across a wide range of fruits, including mango, apple, kiwi, grape, and banana [14]–[17]. Traditionally, regression-based techniques such as partial least squares (PLS) and its variants, including di-PLS, have been employed to model fruit maturity. However, their predictive performance is often influenced by contextual factors such as environmental conditions, seasonal variability, and cultivar-specific characteristics [18], [19]. In the specific case of avocado, previous studies using NIR spectroscopy have reported promising predictive capabilities for maturity-related parameters, particularly dry matter, firmness, and lipid content, with R^2 commonly ranging between 0.70 and 0.80 [20]–[27]. Although some works have achieved higher accuracy levels ($R^2 \approx 0.98$) through optimized calibration and cross-validation strategies [28], these findings also emphasize the sensitivity of NIR-based models to measurement conditions, sampling protocols, and seasonal effects [29].

More recently, advances in machine learning (ML) have contributed to further improvements in predictive performance. Algorithms such as convolutional neural networks (CNNs) and support vector machines (SVMs) have demonstrated enhanced feature extraction and classification capabilities in Hass avocado datasets [30]. Despite these advances, most existing studies focus on single-variety datasets or develop generalized models without explicitly addressing the trade-off between global models trained across multiple varieties and variety-specific models. This gap is particularly relevant in real-world production systems, where heterogeneous cultivars coexist and require robust yet adaptable predictive solutions [20], [21].

Within this framework, and with the objective of promoting the adoption of non-destructive technologies in the Peruvian agro-export sector, this study proposes an approach based on NIR reflectance spectroscopy combined with supervised ML algorithms. The main objective is to develop a non-invasive system for avocado maturity prediction and to systematically evaluate performance differences between global models and variety-specific models. By doing so, the study aims to optimize maturity classification and postharvest management processes, ultimately contributing to improved operational efficiency and competitiveness in international markets.

The novelty of this research lies in the application of ensemble stacking as a maturity classification strategy, an approach that remains underexplored in NIR-based avocado assessment. Unlike individual predictive models, stacking integrates the complementary strengths of multiple supervised learning algorithms through a meta-learning framework, resulting in improved predictive accuracy and greater model stability. This methodological contribution not only enhances the robustness of the proposed system but also facilitates its integration into digital agribusiness platforms, including IoT-based field monitoring systems, digital traceability solutions in supply chains, and Agriculture 4.0 applications. Furthermore, the comparative analysis between global and variety-specific models provides practical guidelines for large-scale implementation in commercial classification systems. Finally, this study contributes to the technological advancement of developing countries by proposing computationally efficient, scalable, and industry-oriented solutions that support export competitiveness and foster innovation in the agricultural sector.

2. METHOD

The methodology of this study was structured following the phases of the cross-industry standard process for data mining (CRISP-DM) framework [31]. The adoption of this framework, a widely recommended practice for ensuring robustness in ML projects, was adapted here to address a chemometric classification problem. Adhering to a systematic workflow aligns with best practices in spectral data analysis, where a rigorous process is required from data preparation to final model validation [32], [33]. As detailed in Figure 1, the process begins with reflectance capture (900–1,700 nm) of 70 samples of the Hass and Fuerte varieties. In an initial phase, the data undergo a first preparation pipeline (Savitzky-Golay (SG) and standard normal variate (SNV)).

Subsequently, a principal component analysis-based outlier detection filter (PCA-OD) with a 90% confidence level and five principal components is applied individually. In this step, scans identified as outliers are discarded, and the remaining spectra are then averaged for each measurement point. Subsequently, a second chemometric pipeline is applied globally to all spectra, integrating SNV and minimum noise fraction (MNF) transformations, a second SG-derivative (SG-D2) with a spectral clipping to 1,070–1,627 nm, and Gram-Schmidt (GS) to enhance critical chemical patterns. The system's success lies in correlating these data with the reference variable of dry matter (%DM), obtained by thermal drying at 70 °C for 24 hours, allowing maturity to be categorized into three thresholds: <24.5%, 24.5–26.5%, and ≥26.5%. Finally, the data feeds the ML modeling block, where classifiers such as adaptive boosting (AdaBoost), random forest (RF), gradient boosting (GB), SVM, and k-nearest neighbors

(KNN) are deployed, optimized through a 20-fold cross-validation. Robustness is verified using receiver operating characteristic (ROC) curves and confusion matrices, and performance is evaluated with metrics such as area under the curve (AUC), classification accuracy, F1-score, accuracy, recall, and Matthews correlation coefficient (MCC).

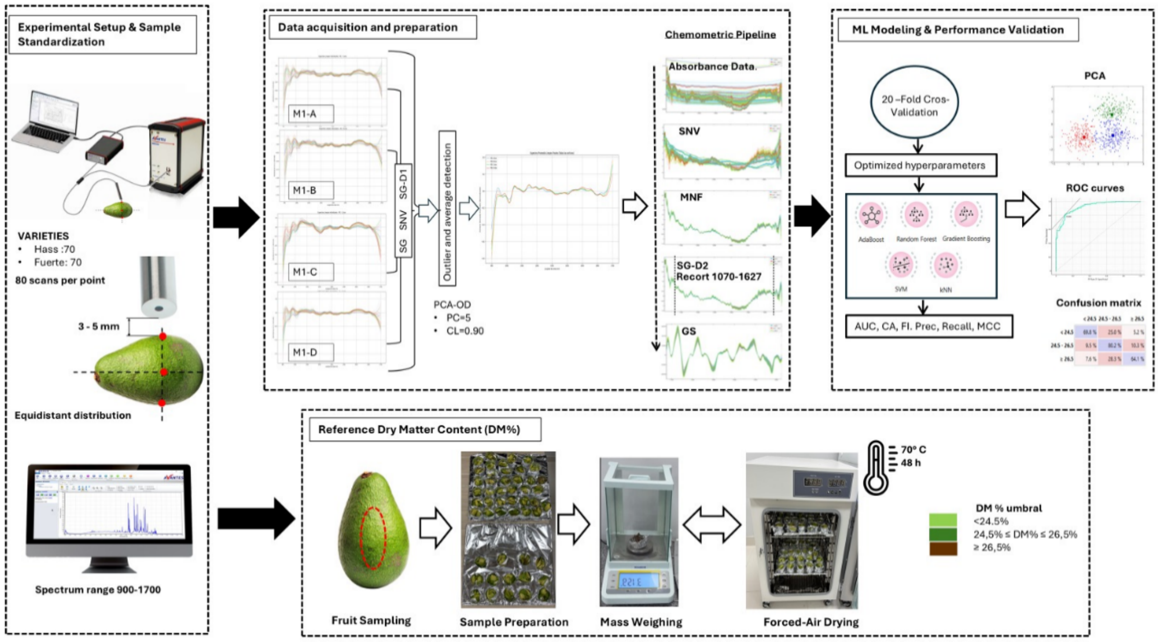


Figure 1. Methodological workflow for the non-destructive classification of avocado maturity using NIR spectroscopy and ML

2.1. Sample collection and selection

This study was conducted using Hass and Fuerte avocado samples collected from a certified commercial farm located in the Sayan–Huara area, representative of Peru's export-oriented avocado production. Fruits were intentionally selected at different maturity stages, preliminarily determined through visual inspection (peel color), tactile firmness, and postharvest condition. This intentional stratification ensured adequate variability and representativeness of the dataset for both spectral analysis and classification modeling. The distribution of samples by variety and origin is summarized in Table 1.

Table 1. Description of samples used

Variety	Region	Zone	Quantity
Hass	File	Sayan	70
Strong	File	Sayan	70

2.2. Spectral acquisition system

Spectral data were acquired using an AvaSpec-NIR-1.7 spectrometer operated through AvaSoft 8.16 software. The system covered a wavelength range of 900–1,700 nm, with an integration time of 5 ms and an average of 80 scans per measurement and an avocado distance of 3–5 mm, providing sufficient spectral resolution to capture the main absorption bands associated with water, lipids, carbohydrates and dry matter in avocado pulp [29], [34]. All measurements were recorded in absorbance mode.

To ensure measurement consistency and minimize variability, a fiber-optic probe equipped with an integrated halogen light source was positioned perpendicularly to the fruit surface using a mechanical holder that maintained a fixed acquisition distance. Prior to spectral acquisition, the system was calibrated using a white reference standard (Teflon) and a dark reference to correct for environmental and instrumental interferences. The experimental setup and instrumentation are illustrated in Figure 2.

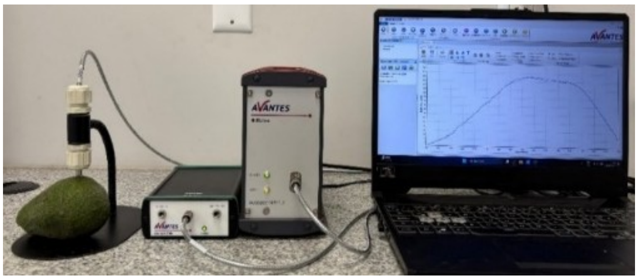
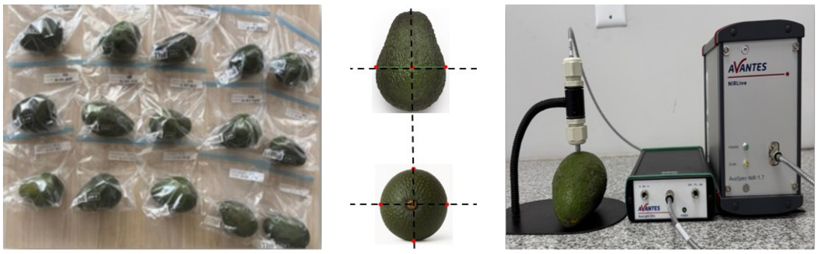


Figure 2. AvaSpec-NIR-1.7 NIR device

2.3. Sampling and spectral acquisition

A standardized sampling protocol was designed to capture the representative spectral variability of each fruit. Four equidistant measurement points were marked along the equatorial region of each avocado. At each point, spectra were acquired using the AvaSpec-NIR-1.7 spectrometer operating in diffuse reflectance mode and coupled to a fiber-optic probe. To maximize the signal-to-noise ratio (SNR) and ensure robust spectral measurements, each final spectrum was obtained by averaging 80 consecutive scans at the same point. This procedure resulted in four high-quality representative spectra per fruit, which were subsequently used for chemometric analysis and model development. All spectral data were collected within the 900–1,700 nm range. The sample preparation and spectral acquisition process is shown in Figure 3, where Figure 3(a) shows the label of samples for identification, Figure 3(b) shows the marking of the area of analysis of the samples, and Figure 3(c) shows the acquisition of spectra.



(a) (b) (c)

Figure 3. Preparation and acquisition of spectra in avocado samples of (a) label of samples for identification, (b) marking of the area of analysis of the samples, and (c) acquisition of spectra

2.4. Baseline analysis and maturity classification

For model training and validation, %DM was selected as the reference variable to define avocado maturity stages. This parameter is widely recognized in the scientific literature and by the avocado industry as a reliable indicator of both physiological and edible maturity [35], [36]. Dry matter content was determined using the standard gravimetric method. Approximately 10 g of homogenized pulp from each fruit were weighed using an analytical balance with a precision of ±0.001 g and subsequently dried in a forced-air convection oven at 70 °C for 24 hours until constant weight was achieved. The complete drying and analysis process is illustrated in Figure 4, where Figure 4(a) shows the preparation of fresh samples, Figure 4(b) shows the heat treatment in oven, and Figure 4(c) shows the dehydrated samples. Finally, for the development of the classification models, each fruit was assigned to one of three categories 'immature,' 'optimal maturity,' or 'overripe' based on its %DM value. The thresholds used for this categorization, derived from commercial standards of the export industry [37], [38], are detailed in Table 2.

Table 2. Avocado classification thresholds according to the %DM

Maturity category	%MS umbral
Unripe	<24.5%
Optimal maturity	24.5% ≤%DM ≤26.5%
Overripe	≥26.5%

(a) (b) (c)

Figure 4. Drying process of avocado samples for physical-chemical analysis of (a) preparation of fresh samples, (b) heat treatment in oven, and (c) dehydrated samples

2.5. Preprocessing of spectral data

Raw NIR spectra are often affected by instrumental noise, light scattering effects, and baseline variations. To mitigate these interferences and enhance chemically relevant information, an optimized preprocessing pipeline was applied prior to chemometric modeling [39]–[41]. To optimize the SNR and correct for physical disturbances in the samples, a preprocessing workflow divided into two critical phases was implemented, following a dual-pipeline architecture.

i) Individual cleaning phase: first, the raw spectra of each measurement point underwent signal normalization using SG smoothing (51-point window, second-order polynomial), SNV smoothing, and then SG and the first derivative (19-point window, second-order polynomial). To ensure sample purity, outlier detection was performed using PCA-OD with a 90% confidence level and 5 components. Scans identified as outliers were systematically removed, and only the spectra valid for representing each fruit zone were averaged.
ii) Global chemometric pipeline phase: once the dataset was consolidated, an advanced pipeline was applied. The sequence of the second pipeline is illustrated in Figure 5, where Figure 5(a) shows the raw absorbance spectra of all points and variety, Figure 5(b) shows after SNV correction, Figure 5(c) shows after MNF noise reduction, Figure 5(d) shows after the second derivative of SG, and Figure 5(e) shows the final spectra after Gaussian smoothing and spectral clipping; and consisted of the following steps:
    – SNV: as a first step, SNV was applied to correct light scattering variations and optical path length differences among samples, standardizing the scale of each spectrum.
    – MNF denoising: for deeper noise reduction than conventional smoothing methods, the MNF transform with five components was employed. This advanced technique was crucial to segregate and eliminate multivariate noise, retaining only the components with the highest SNR for subsequent stages.
    – SG smoothing and second derivative: a SG filter (17-point window, second-order polynomial) was then applied for smoothing and simultaneous calculation of the second derivative. This combination effectively removes baseline shifts and resolves overlapping peaks, amplifying subtle differences associated with chemical composition.
    – Gaussian smoothing: empirical tests showed that a final smoothing step using a low-impact Gaussian filter (σ =1.0) contributed to additional spectrum stabilization, eliminating residual high-frequency noise introduced by derivation without distorting peaks of interest.
    – Spectral trimming: finally, the spectral range was cropped to 1,070–1,627 nm to focus the analysis on the most informative region with the highest SNR for dry matter prediction in avocado.

It is acknowledged that this pipeline is more elaborate than conventional preprocessing sequences. However, during the model optimization phase, it was confirmed that omitting any of these steps particularly MNF denoising and the final Gaussian smoothing led to a quantifiable decrease in classification performance (lower AUC and MCC in cross-validation). Therefore, the complete sequence was retained to ensure maximum reliability and robustness of the final models presented in this work.

2.6. Development and evaluation of classification models

The development of the classifiers was designed under a predictive computational architecture structured around two approaches: i) a global model trained with combined data from both ‘Hass’ and ‘Fuerte’ varieties and ii) variety-specific models individually calibrated for each cultivar. This methodological contrast enabled the evaluation of whether spectral differences between cultivars justify distinct calibration strategies [21], [42], [43]. The predictive processing workflow was based on an ensemble stacking pipeline, integrating multiple algorithms as base learners and a meta-model for combination. Specifically, the base classifiers included RF, GB, XGBoost, categorical boosting (CatBoost), KNN, and decision trees (DT), while the fusion layer consisted of a linear regression meta-model, responsible for optimizing the weighted combination of individual predictions. To ensure training robustness and mitigate overfitting, a 20-fold cross-validation scheme was applied in all experiments. This statistical design ensured that each instance of the dataset was used for both training and validation, thereby reducing bias in error estimation and strengthening model generalization.

The performance of the classifiers was evaluated using a set of complementary metrics: confusion matrix, accuracy, precision, recall, F1-score, AUC, and MCC, the latter being particularly robust in multiclass classification contexts [44]. Hyperparameters for each algorithm were automatically optimized within the cross-validation framework to maximize predictive performance. In this way, the present study not only compares the effectiveness of global versus variety-specific models but also validates the relevance of ensemble stacking as an advanced methodological strategy to improve the stability, scalability, and interpretability of avocado maturity classification systems.

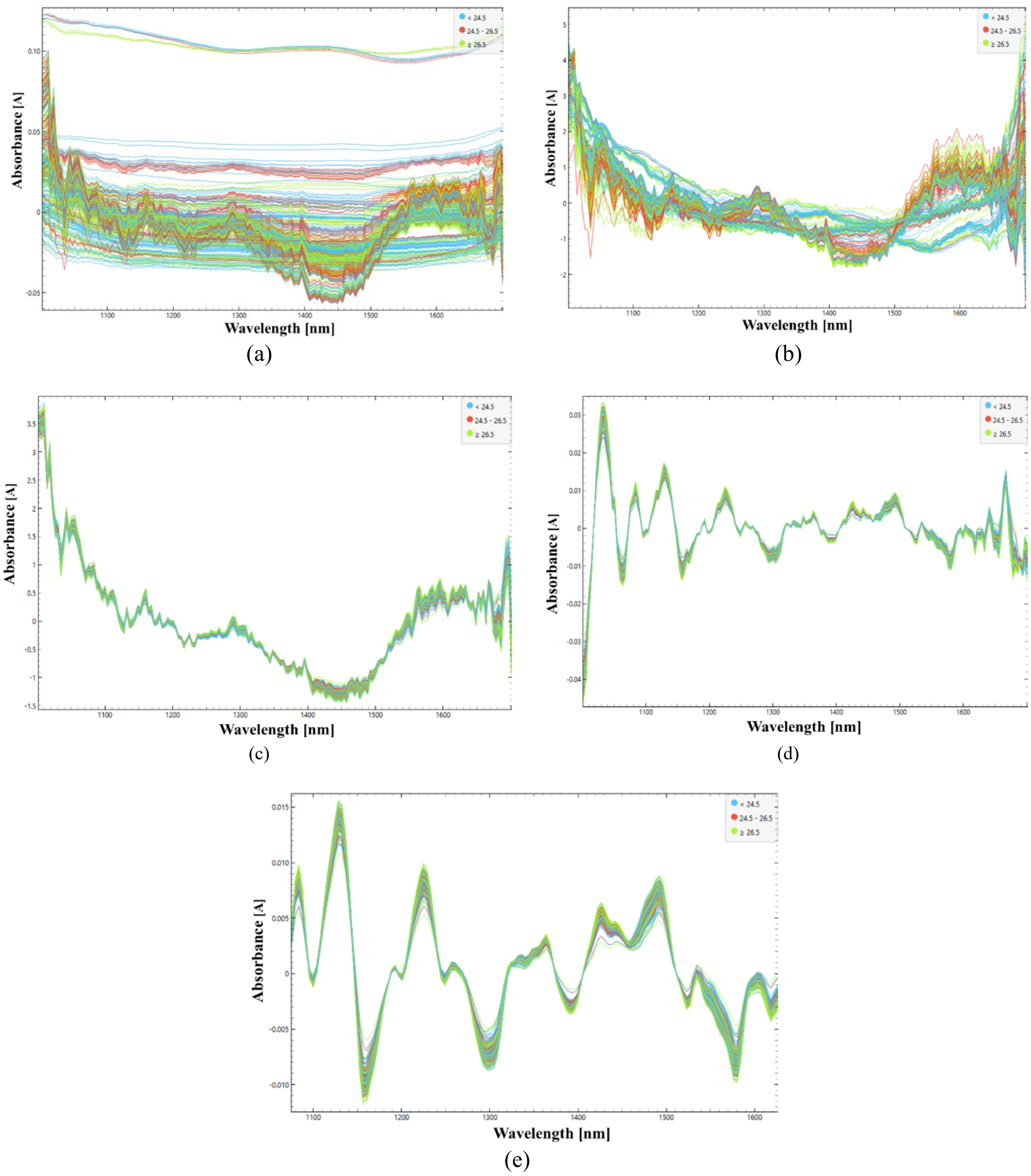


Figure 5. Visualization of the spectral preprocessing workflow for (a) raw absorbance spectra of all points and variety, (b) after SNV correction, (c) after MNF noise reduction, (d) after the second derivative of SG, and (e) final spectra after Gaussian smoothing and spectral clipping

3. RESULTS AND DISCUSSION

This section presents the key findings of the study, integrating the exploratory analysis of the spectral data with the quantitative evaluation of the classification models. Five ML algorithms were compared under two strategic approaches: a global model and variety-specific models. In addition, the impact of the ensemble stacking technique was assessed against individual models, with the aim of identifying the most robust and accurate strategy for the non-destructive classification of avocado maturity stages.

3.1. Exploratory data analysis

Figure 6 presents the average preprocessed NIR spectra for each maturity class (unripe, optimal maturity, and overripe) across the wavelength range of 1,070–1,627 nm (x-axis: wavelength in nm; y-axis: absorbance). Distinct spectral patterns are observed among the three maturity stages, indicating that NIR spectroscopy effectively captures the physicochemical changes associated with the avocado ripening process, particularly those related to dry matter accumulation and compositional shifts. To further explore the intrinsic structure of the spectral data, PCA was applied. Figure 7 illustrates the PCA score plot, where a clear separation of samples along the first principal component (PC1) is observed. PC1 is strongly correlated with the maturity stage, with Unripe samples (dry matter <24.5%) clustering predominantly on the positive side of PC1, while optimal maturity and overripe samples occupy the negative PC1 region.

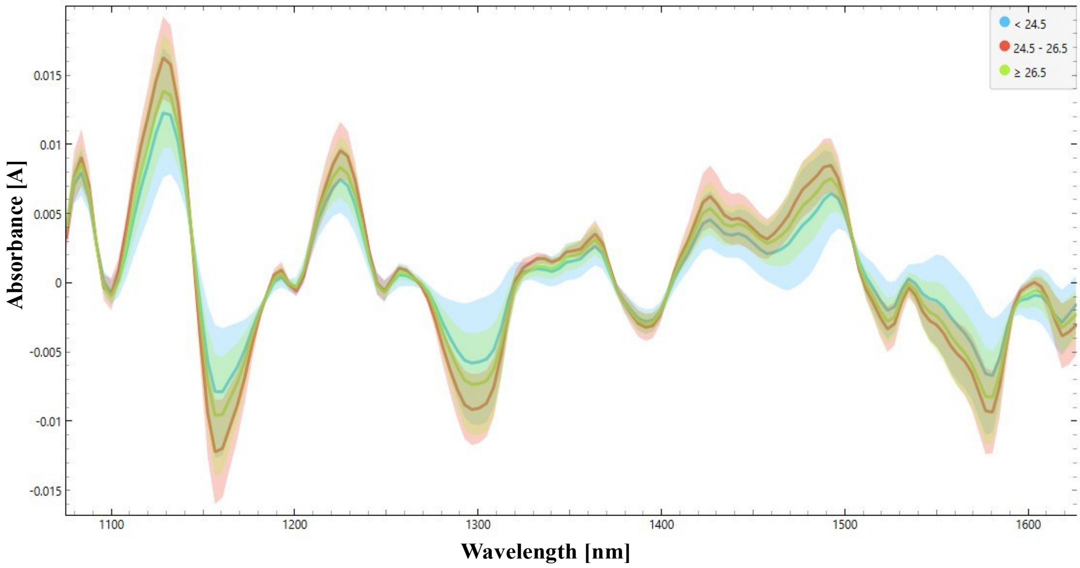


Figure 6. Average NIR spectra for each maturity class

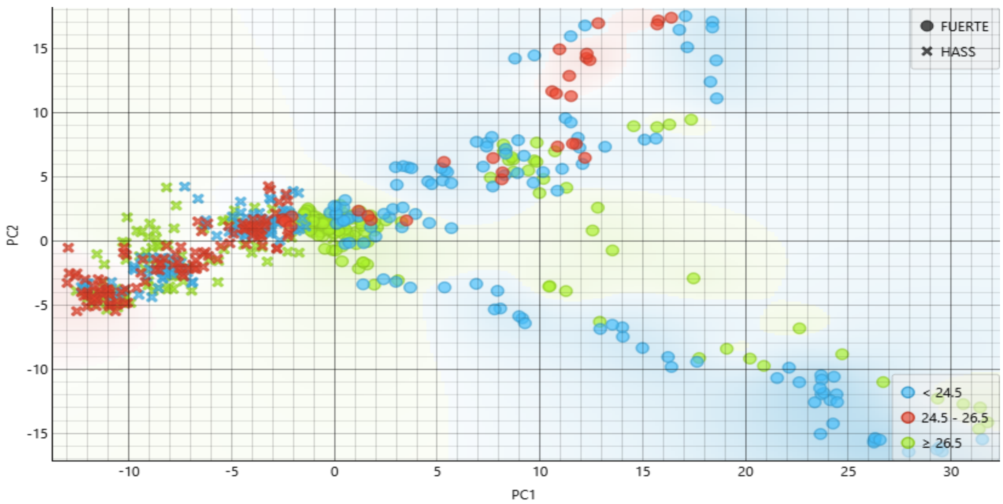


Figure 7. Principal component analysis

Beyond maturity-related separation, the PCA also reveals a partial but consistent separation between the two avocado varieties. Hass samples are mainly concentrated on one side of the spectral space, whereas Fuerte samples span a broader range across maturity levels. This varietal heterogeneity provides strong empirical evidence supporting the hypothesis that a single global model may be limited by inter-varietal spectral differences, whereas variety-specific models can better exploit cultivar-dependent spectral signatures. These findings directly justify the comparative modeling strategy adopted in this study.

3.2. Comparison of approaches: global model vs. specific models

To determine the most effective modeling approach, a systematic comparison was conducted between a global model and variety-specific models for Fuerte and Hass. For each approach, five ML algorithms were implemented, with hyperparameters optimized via cross-validation to ensure a fair and robust comparison. The quantitative results are summarized in Tables 3 to 5. For the global model (Table 3), the best performance was achieved using GB, with an MCC of 0.526 and an AUC of 0.863. While these values indicate reasonable predictive capability, they were consistently surpassed by the variety-specific models.

Table 3. Global model metrics results

Models	AUC	Accuracy	F1-score	Precision	Recall	MCC
SVM	0.805	0.676	0.676	0.677	0.676	0.505
AdaBoost	0.708	0.611	0.611	0.611	0.613	0.411
GB	0.863	0.690	0.689	0.689	0.690	0.526
RF	0.847	0.681	0.680	0.681	0.681	0.512
KNN	0.834	0.624	0.624	0.624	0.624	0.426

Table 4. Strong variety specific model metric results

Models	AUC	Accuracy	F1-score	Precision	Recall	MCC
SVM	0.915	0.789	0.786	0.788	0.789	0.633
AdaBoost	0.822	0.803	0.802	0.802	0.803	0.662
GB	0.931	0.814	0.812	0.814	0.814	0.679
RF	0.933	0.817	0.815	0.815	0.816	0.684
KNN	0.918	0.781	0.782	0.782	0.781	0.627

Table 5. Hass manifold specific model metric results

Models	AUC	Accuracy	F1-score	Precision	Recall	MCC
SVM	0.858	0.720	0.720	0.732	0.720	0.580
AdaBoost	0.733	0.641	0.641	0.641	0.641	0.459
GB	0.877	0.714	0.714	0.715	0.714	0.568
RF	0.866	0.691	0.689	0.693	0.691	0.533
KNN	0.845	0.681	0.680	0.681	0.681	0.518

For the Fuerte variety (Table 4), RF achieved the highest performance, reaching an MCC of 0.684 and an AUC of 0.933. Similarly, for the Hass variety (Table 5), the SVM classifier yielded the best balance between discrimination and stability, with an MCC of 0.633 and an AUC of 0.858. These results clearly demonstrate that variety-specific calibration strategies outperform the global approach across all evaluation metrics. Table 6 presents the optimized hyperparameters for all evaluated models. Notably, the optimal configurations differ substantially between global and variety-specific models, further reinforcing the idea that cultivar-dependent spectral characteristics require tailored modeling strategies rather than a unified global calibration. From a methodological perspective, ensemble-based algorithms particularly GB and RF consistently outperformed simpler classifiers such as AdaBoost and KNN. This trend suggests that the relationship between NIR spectral features and avocado maturity is governed by complex, non-linear interactions that are better captured by ensemble learning techniques.

3.3. Analysis of the best performance models

Following the confirmation of the superiority of variety-specific models, a detailed analysis was conducted on the best-performing classifier for each variety. For the Fuerte variety, RF demonstrated the highest robustness and discrimination capability. Figure 8 shows the confusion matrix, where a strong concentration of predictions along the main diagonal indicates high classification accuracy across maturity stages. Most misclassifications occurred between the optimal maturity and overripe classes, which is consistent with the overlap observed in the PCA (Figure 7) and reflects the continuous nature of the ripening process. The corresponding ROC curves (Figure 9) further validate this robustness, with AUC values exceeding 0.90 for all maturity classes, confirming the model's strong discriminatory power. In the case of the Hass variety, the SVM classifier achieved the best compromise between accuracy and stability. The confusion matrix presented in Figure 10 indicates high sensitivity for the optimal maturity class (80.2%), while performance for the extreme classes was more moderate: 69.8% for unripe and 64.1% for overripe.

Table 6. Model-specific and global hyperparameters

Models	Strong	Hatred	Global
SVM	Cost :10	Cost: 8.5	Cost: 1.05
	Epsilon regression: 0.10	Epsilon regression: 0.20	Epsilon regression: 0.20
	Kernel RBF g: 1.05	Kernel RBF g: 1.05	Kernel RBF g: 0.97
AdaBoost	Number of estimators: 50	Number of estimators: 50	Number of estimators: 50
	Larning rate: 0.010	Larning rate: 0.010	Larning rate: 0.010
	Loss regression: square	Loss regression: square	Loss regression: square
GB	Method: XGBoost	Method: XGBoost	Method: XGBoost
	number of trees: 200	number of trees: 200	number of trees: 100
	Learning rate: 0.30	Learning rate: 0.30	Learning rate: 0.30
RF	Number of trees: 350	Number of tree: 250	Number of tree: 100
KNN	Number of neighbors: 3	Number of neighbors: 3	Number of neighbors: 8
	Metric: Euclidean	Metric: Euclidean	Metric: Euclidean
	Weight: by distances	Weight: by distances	Weight: by distances

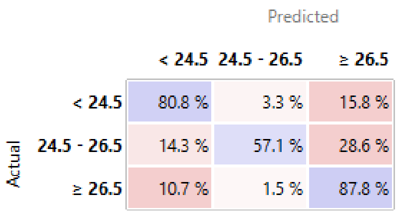


Figure 8. Strong avocado confusion matrix

The analysis of the matrix reveals high sensitivity for the 'ripe' class (24.5–26.5), with an accuracy of 80.2%. However, performance was more moderate for the extreme classes, with sensitivities of 69.8% for 'unripe' (<24.5) and 64.1% for 'overripe' (≥26.5). Notably, the main source of error for both 'unripe' and 'overripe' was their misclassification as 'ripe' (25.0% and 28.3% of the time, respectively). This suggests that, for the 'Hass' variety, the spectral characteristics of the intermediate class tend to overlap significantly with those of the early and late stages, presenting a more complex classification challenge compared to the 'Fuerte' variety.

The model demonstrates strong discrimination ability across all classes, with AUC values of 0.871 for 'unripe,' 0.898 for 'ripe,' and 0.863 for 'overripe.' These consistently high values confirm that the SVM model possesses strong predictive potential for the 'Hass' variety. The apparent discrepancy between the high AUC and the more moderate sensitivities in the extreme classes indicates that, although the model is effective at ranking probabilities, defining a single cutoff threshold for final classification is inherently challenging due to class overlap.

To contextualize the findings of this study, it is essential to compare them with the current state of the art. Previous research on NIR spectroscopy for avocado quality assessment has predominantly employed regression-based methods, particularly PLS, to predict quantitative parameters such as dry matter content. Reported R^2 coefficients typically range from 0.62 to 0.98 under controlled laboratory conditions [10], [26], [28]. Although these models have proven valuable for quantitative prediction, their application in industrial environments remains limited due to the need for frequent recalibration and the challenges associated with transferring models to dynamic, real-time classification systems. In contrast, the present study approaches the problem from a classification perspective, which is more compatible with automated sorting systems and postharvest quality management operations.

Regarding the ensemble stacking strategy, results demonstrated that combining multiple algorithms consistently outperformed individual classifiers, improving both accuracy and predictive stability. This trend aligns with findings from recent studies on tropical fruit classification using ML [30], where stacking-based frameworks achieved accuracy gains of 5–10% compared to baseline models. In our case, ensemble stacking also reduced inter-campaign variability and enhanced model robustness for underrepresented maturity categories an important factor for operational deployment in heterogeneous production contexts.

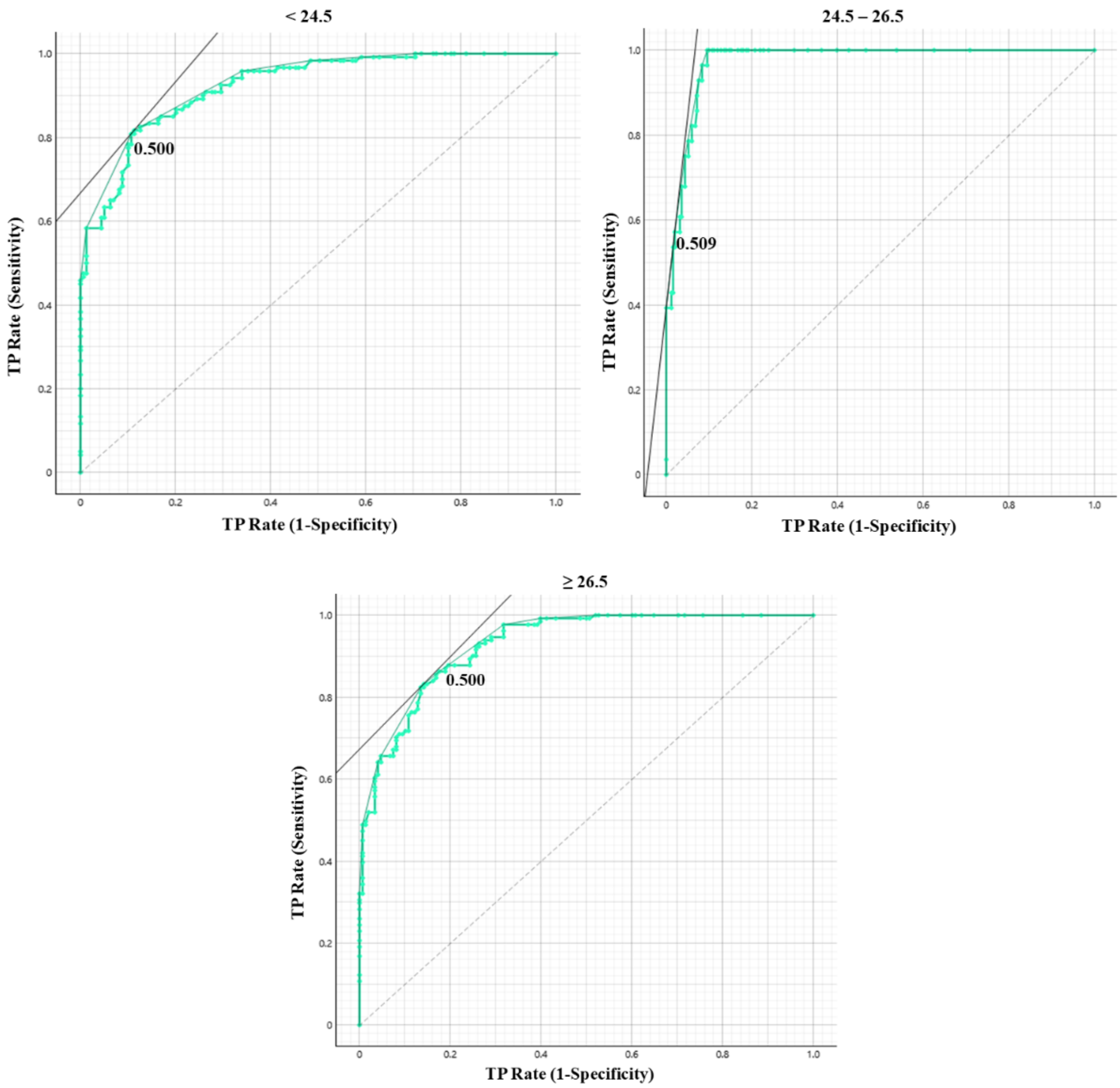


Figure 9. Strong avocado ROC curve chart for each maturity threshold

Actual \ Predicted	< 24.5	24.5 - 26.5	≥ 26.5
< 24.5	69.8 %	25.0 %	5.2 %
24.5 - 26.5	9.5 %	80.2 %	10.3 %
≥ 26.5	7.6 %	28.3 %	64.1 %

Figure 10. Hass avocado confusion matrix

When compared to previous PLS-DA models applied to avocado [45], the proposed approach achieved a more balanced and reliable classification across all categories. Whereas the referenced study reported limited performance for intermediate maturity (F1-score =0.32), the model obtained F1 values exceeding 0.78, along with high MCC and AUC scores across classes. These results indicate that advanced ML algorithms particularly when combined through stacking are more capable of managing overlapping spectral patterns and the inherent complexity of the avocado ripening process.

From a practical standpoint, the proposed framework provides a feasible alternative for real-time maturity monitoring during postharvest handling and export preparation. By offering non-destructive, rapid, and automated evaluation, it can help optimize harvest scheduling, reduce postharvest losses, and improve the uniformity of export batches. Moreover, the scalability of the ensemble approach supports integration into IoT-based quality control systems and digital traceability platforms, contributing to the advancement of precision agriculture and supply chain efficiency.

Nevertheless, certain limitations must be acknowledged. Environmental factors such as variations in ambient humidity, temperature, and lighting may influence spectral readings and introduce variability not fully captured in this study. The dataset used was limited to a single harvest season, which could affect the model's generalizability across years or geographic regions. Future work should therefore explore cross-seasonal validation, sensor calibration standardization, and the impact of different orchard management practices to strengthen model robustness and reproducibility under diverse real-world conditions. Overall, the findings validate the relevance of variety-specific modeling and ensemble learning for non-destructive maturity prediction, offering a practical and scalable framework that bridges data-driven modeling with agricultural technology innovation.

4. CONCLUSION

This study demonstrated that NIR spectroscopy combined with supervised ML provides an effective, non-destructive, and scalable framework for avocado maturity classification. The results clearly showed that variety-specific models for Hass and Fuerte consistently outperformed the global model, achieving MCC values of up to 0.667, thereby confirming that cultivar-dependent calibration enhances predictive accuracy and model stability under varietal heterogeneity. Furthermore, the validation of ensemble stacking highlighted its superiority over individual classifiers, as the integration of multiple learning algorithms resulted in more robust and reliable decision-making performance. From an applied perspective, the proposed methodology offers direct benefits for postharvest management and export logistics by enabling rapid, non-invasive quality assessment and is well suited for integration into IoT-based monitoring systems, digital traceability platforms, and cloud-based analytical infrastructures. Future research should explore broader spectral ranges, cross-seasonal validation, and longitudinal studies tracking fruit evolution throughout storage and transport, as well as the incorporation of sensor fusion and edge-computing strategies. Overall, this work establishes a solid foundation for the development of intelligent, scalable, and industry-oriented quality assessment systems that support automation, precision agriculture, and technological innovation in the global avocado supply chain.

ACKNOWLEDGMENTS

Sincerely thank to the Universidad Tecnológica del Perú for its commitment to providing the essential tools and management support.

FUNDING INFORMATION

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Christian Ovalle	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓
Jhonn Garcia Jimenez	✓		✓		✓	✓		✓	✓		✓	✓	✓	✓
Jose Briones Zuñiga	✓	✓	✓		✓	✓	✓	✓	✓		✓		✓	✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

This study did not involve human participants, patients, or identifiable personal data. Therefore, informed consent was not required.

ETHICAL APPROVAL

This research is based on the analysis of spectral data and ML models applied to agricultural products and does not involve experiments with human participants or animals. Consequently, ethical approval was not required.

DATA AVAILABILITY

Data availability is not applicable to this article, as no new publicly available datasets were created or deposited. All data supporting the findings of this study are included within the article and its references.

REFERENCES

[1] Ministry of Agrarian Development and Irrigation, "Monthly statistical bulletin 'agriculture in figures'-2024," *gob.pe*. 2025. Accessed: Apr. 7, 2025. [Online]. Available: https://www.gob.pe/institucion/midagri/informes-publicaciones/5380407-boletin-estadistico-mensual-el-agro-en-cifras-2024

[2] Agricultural News Agency of Peru, "Peruvian avocado exports will reach 630,000 tons in 2025," *agraria.pe*. Accessed: Jul. 2, 2025. [Online]. Available: https://agraria.pe/index.php/noticias/exportaciones-peruanas-de-palta-alcanzaran-las-630-000-tonel-39493

[3] M. L. Colan, "Evolution of the global and national avocado market," *cien.adexperu.org.pe*. Accessed: Apr. 7, 2025. [Online]. Available: https://cien.adexperu.org.pe/informacion-estrategica/evolucion-del-mercado-mundial-y-nacional-de-palta/

[4] M. Vasileiou *et al.*, "Digital transformation of food supply chain management using blockchain: a systematic literature review towards food safety and traceability," *Business and Information Systems Engineering*, pp. 1-28, Jun. 2025, doi: 10.1007/s12599-025-00948-0.

[5] P. Mishra, M. Paillart, L. Meesters, E. Woltering, and A. Chauhan, "Avocado dehydration negatively affects the performance of visible and near-infrared spectroscopy models for dry matter prediction," *Postharvest Biology and Technology*, vol. 183, 2022, doi: 10.1016/j.postharvbio.2021.111739.

[6] M. M. Mohsan *et al.*, "SwishFormer for robust firmness and ripeness recognition of fruits using visual tactile imagery," *Postharvest Biology and Technology*, vol. 225, 2025, doi: 10.1016/j.postharvbio.2025.113487.

[7] J. M. Godbout *et al.*, "Non-destructive identification of varieties of Hawaii-grown avocados using near-infrared spectroscopy: feasibility studies using bench-top and handheld spectrometers," *PLoS ONE*, vol. 19, no. 6, pp. 1-11, 2024, doi: 10.1371/journal.pone.0303532.

[8] P. Rodríguez, I. Soto, J. Villamizar, and A. Rebolledo, "Fatty acids and minerals as markers useful to classify Hass avocado quality: ripening patterns, internal disorders, and sensory quality," *Horticulturae*, vol. 9, no. 4, 2023, doi: 10.3390/horticulturae9040460.

[9] L. Zou, J. Yuan, X. Liu, J. Li, P. Zhang, and Z. Niu, "Burgers viscoelastic model-based variable stiffness design of compliant clamping mechanism for leafy greens harvesting," *Biosystems Engineering*, vol. 208, pp. 1–15, 2021, doi: 10.1016/j.biosystemseng.2021.05.007.

[10] M. Zhang and X. Guo, "Emerging strategies in fluorescent aptasensor toward food hazard aflatoxins detection," *Trends in Food Science and Technology*, vol. 129, pp. 621–633, 2022, doi: 10.1016/j.tifs.2022.11.013.

[11] K. B. Walsh, J. Blasco, M. Z. Sasse, and X. Sun, "Visible-NIR 'point' spectroscopy in postharvest fruit and vegetable assessment: the science behind three decades of commercial use," *Postharvest Biology and Technology*, vol. 168, 2020, doi: 10.1016/j.postharvbio.2020.111246.

[12] I. R. D.- González, C. Valero, M. A. Momin, A. Kaur, and D. C. Slaughter, "Performance evaluation of two commercially available portable spectrometers to non-invasively determine table grape and peach quality attributes," *Agronomy*, vol. 10, no. 1, 2020, doi: 10.3390/agronomy10010148.

[13] K. B. Beć, J. Grabska, and C. W. Huck, "Handheld NIR spectroscopy for real-time on-site food quality and safety monitoring," *Advances in Food and Nutrition Research*, vol. 115, pp. 293–389, 2025, doi: 10.1016/bs.afnr.2025.01.002.

[14] F. P. Lamptey, E. Teye, E. E. Abano, and C. L. Y. Amuah, "Application of handheld NIR spectrometer for simultaneous identification and quantification of quality parameters in intact mango fruits," *Smart Agricultural Technology*, vol. 6, 2023, doi: 10.1016/j.atech.2023.100357.

[15] A. Arnal *et al.*, "Rapid, non-destructive prediction of ripeness of pink lady apples by using near-infrared spectroscopic methods to monitor firmness, sugar content, juiciness, and acidity," *Sustainability*, vol. 16, no. 23, pp. 1-20, Nov. 2024, doi: 10.3390/su162310259.

[16] K. B. Beć, J. Grabska, and C. W. Huck, "Miniaturized NIR spectroscopy in food analysis and quality control: promises, challenges, and perspectives," *Foods*, vol. 11, no. 10, 2022, doi: 10.3390/foods11101465.

[17] C. Wang, X. Li, Z. Zhang, X. Luo, J. Cai, and A. Wang, "Nondestructive quality detection of characteristic fruits based on Vis/NIR spectroscopy: principles, systems, and applications," *Agriculture*, vol. 15, no. 20, 2025, doi: 10.3390/agriculture15202167.

[18] P. Mishra and R. N. Langerodi, "Partial least square regression versus domain invariant partial least square regression with application to near-infrared spectroscopy of fresh fruit," *Infrared Physics and Technology*, vol. 111, 2020, doi: 10.1016/j.infrared.2020.103547.

[19] P. Rodríguez, J. Villamizar, L. Londoño, T. Tran, and F. Davrieux, "Quantification of dry matter content in Hass avocado by near-infrared spectroscopy (NIRS) scanning different fruit zones," *Plants*, vol. 12, no. 17, Aug. 2023, doi: 10.3390/plants12173135.
[20] C. Guo, J. Zhang, W. Cai, and X. Shao, "Enhancing transferability of near-infrared spectral models for soluble solids content prediction across different fruits," *Applied Sciences*, vol. 13, no. 9, 2023, doi: 10.3390/app13095417.
[21] P. Mishra, "Developing multifruit global near-infrared model to predict dry matter based on just-in-time modeling," *Journal of Chemometrics*, vol. 38, no. 4, 2024, doi: 10.1002/cem.3540.
[22] C. V. Engeland, L. Spreutels, R. Legros, and B. Haut, "Comprehensive analysis of intermittent drying. A theoretical approach," *Food and Bioproducts Processing*, vol. 131, pp. 86–101, 2022, doi: 10.1016/j.fbp.2021.10.009.
[23] P. Mishra, J. M. Roger, F. Marini, A. Biancolillo, and D. N. Rutledge, "FRUITNIR-GUI: a graphical user interface for correcting external influences in multi-batch near infrared experiments related to fruit quality prediction," *Postharvest Biology and Technology*, vol. 175, 2021, doi: 10.1016/j.postharvbio.2020.111414.
[24] L. S. Magwaza and S. Z. Tesfay, "A review of destructive and non-destructive methods for determining avocado fruit maturity," *Food and Bioprocess Technology*, vol. 8, no. 10, pp. 1995–2011, 2015, doi: 10.1007/s11947-015-1568-y.
[25] M. Amondi, J. Ombiro, Z. Birech, and D. Oeba, "Predicting Hass avocado maturity with NIR spectroscopy for non-invasive dry matter estimation in Hass avocados," *Asian Journal of Research and Reviews in Physics*, vol. 8, no. 4, pp. 53–65, 2024, doi: 10.9734/ajr2p/2024/v8i4175.
[26] P. P. Subedi and K. B. Walsh, "Assessment of avocado fruit dry matter content using portable near infrared spectroscopy: method and instrumentation optimisation," *Postharvest Biology and Technology*, vol. 161, Mar. 2020, doi: 10.1016/j.postharvbio.2019.111078.
[27] T. Cheng *et al.*, "Near-infrared model and its robustness as affected by fruit origin for 'Dangshan' pear soluble solids content and pH measurement," *Agriculture*, vol. 12, no. 10, 2022, doi: 10.3390/agriculture12101618.
[28] J. P. Wold *et al.*, "Inter seasonal validation of non-contact NIR spectroscopy for measurement of total soluble solids in high tunnel strawberries," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 309, Mar. 2024, doi: 10.1016/j.saa.2024.123853.
[29] B. B. Wedding, C. Wright, S. Grauf, P. Gadek, and R. D. White, "The application of FT-NIRS for the detection of bruises and the prediction of rot susceptibility of 'Hass' avocado fruit," *Journal of the Science of Food and Agriculture*, vol. 99, no. 4, pp. 1880–1887, 2019, doi: 10.1002/jsfa.9383.
[30] P. Sseruwagi, E. Lehmann, P. Sigombe, G. Ddamulira, J. W. V. Casteren, and P. D. Bauw, "Characterizing avocado production systems for Ugandan exports: the need for consolidation and support for sustainable development," *Frontiers in Sustainable Food Systems*, vol. 9, 2025, doi: 10.3389/fsufs.2025.1500012.
[31] S. Akbar, F. Ali, M. Hayat, A. Ahmad, S. Khan, and S. Gul, "Prediction of antiviral peptides using transform evolutionary and SHAP analysis based descriptors by incorporation with ensemble learning strategy," *Chemometrics and Intelligent Laboratory Systems*, vol. 230, 2022, doi: 10.1016/j.chemolab.2022.104682.
[32] W. Zhang, L. C. Kasun, Q. J. Wang, Y. Zheng, and Z. Lin, "A review of machine learning for near-infrared spectroscopy," *Sensors*, vol. 22, no. 24, 2022, doi: 10.3390/s22249764.
[33] E. Aghdamifar, V. R. Sharabiani, E. Taghinezhad, A. R. Fanaei, and M. Szymanek, "Non-destructive method for identification and classification of varieties and quality of coffee beans based on soft computing models using VIS/NIR spectroscopy," *European Food Research and Technology*, vol. 249, no. 6, pp. 1599–1612, 2023, doi: 10.1007/s00217-023-04240-x.
[34] A. Rady, J. Fischer, S. Reeves, B. Logan, and N. J. Watson, "The effect of light intensity, sensor height, and spectral pre-processing methods when using NIR spectroscopy to identify different allergen- containing powdered foods," *Sensors*, vol. 20, no. 1, 2020, doi: 10.3390/s20010230.
[35] Ambimet Instrumentation, "What is the dry matter content in avocados?," *ambimet-instrumentacion.cl*. Accessed: Jun. 12, 2025. [Online]. Available: https://www.ambimet-instrumentacion.cl/que-es-el-contenido-de-materia-seca-en-paltas/
[36] L. Li, W. Huang, Z. Wang, S. Liu, X. He, and S. Fan, "Calibration transfer between developed portable Vis/NIR devices for detection of soluble solids contents in apple," *Postharvest Biology and Technology*, vol. 183, 2022, doi: 10.1016/j.postharvbio.2021.111720.
[37] X. Li *et al.*, "Combination of 1-MCP and modified atmosphere packaging (MAP) maintains banana fruit quality under high temperature storage by improving antioxidant system and cell wall structure," *Postharvest Biology and Technology*, vol. 198, 2023, doi: 10.1016/j.postharvbio.2023.112265.
[38] F. J. B. Sanchez, H. P. Espinosa, and M. A. M. Aguilar, "Development of non-destructive system for estimating avocado quality parameters," *Postharvest Biology and Technology*, vol. 212, 2024, doi: 10.1016/j.postharvbio.2024.112886.
[39] J. A. Konkol and G. Tsilomelekis, "Porchlight: an accessible and interactive aid in preprocessing of spectral data," *Journal of Chemical Education*, vol. 100, no. 3, pp. 1326–1332, 2023, doi: 10.1021/acs.jchemed.2c00812.
[40] A. A. Munawar, Hizir, C. Erika, and E. Pawelzik, "Fast and simultaneous prediction of inner quality parameters on intact mangos by near infrared spectroscopy: impact of spectra pre-processing on prediction accuracy," *Future Foods*, vol. 10, 2024, doi: 10.1016/j.fufo.2024.100463.
[41] C. E. R. López, C. H. Brenes, V. Treviño, and R. I. D. D. L. Garza, "Avocado fruit maturation and ripening: dynamics of aliphatic acetogenins and lipidomic profiles from mesocarp, idioblasts and seed," *BMC Plant Biology*, vol. 17, no. 1, 2017, doi: 10.1186/s12870-017-1103-6.
[42] T. T. H. Phan, Q. T. Vo, C. V. Bui, L. V. Nguyen, T. M. T. Pham, and T. T. Ha, "Ensemble learning techniques for improving date fruit varieties classification," in *2025 IEEE 2nd International Conference on Electronics, Communications and Intelligent Science (ECIS 2025)*, 2025, doi: 10.1109/ECIS65594.2025.11086796.
[43] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, 2020, doi: 10.1186/s12864-019-6413-7.
[44] A. M. Shimaoka, R. C. Ferreira, and A. Goldman, "The evolution of CRISP-DM for data science: methods, processes and frameworks," *SBC Reviews on Computer Science*, vol. 4, no. 1, pp. 28–43, 2024, doi: 10.5753/reviews.2024.3757.
[45] A. M. Herreros *et al.*, "Postharvest ripeness assessment of 'Hass' avocado based on development of a new ripening index and Vis-NIR spectroscopy," *Postharvest Biology and Technology*, vol. 181, Nov. 2021, doi: 10.1016/j.postharvbio.2021.111683.

BIOGRAPHIES OF AUTHORS

Christian Ovalle is an associate professor at the Faculty of Engineering of the Faculty of Engineering, Universidad Tecnológica del Perú, Lima-Peru. He has a Ph.D. in Systems Engineering with a specialization in AI. His research areas are process mining, business data analysis, and pattern recognition. He is the CEO of the 7D consultancy dedicated to the investigation of intelligent solutions. He has participated in different research projects, receiving awards from the Peruvian Ministry of Defense and the Armed Forces Army for the best general researcher, which is a technology-based company and his innovative products received national and international recognition. He has filed a number of patents and industrial designs on his innovative ideas. His research interests include data mining, AI, image/signal processing, bibliometrics, and pattern recognition. He can be contacted at email: dovalle@utp.edu.pe.

Jhonn Garcia Jimenez graduated as an industrial mechatronics technician and he is currently pursuing a degree in mechatronics engineering. Throughout academic and professional journey, he has cultivated a strong passion for research, innovation, and applied technology. His main research and development interests lie in the application of ML and AI to solve complex engineering challenges, particularly those related to automation, intelligent control systems, and predictive modeling. In addition, he has developed skills in electronics design, PCB layout, prototyping, and hardware-software integration, which allow me to address projects from both the computational and physical perspectives. He is especially motivated by projects that combine automation, robotics, and AI, aiming to create efficient, intelligent, and sustainable technological solutions. His current academic work is complemented by continuous exploration of cutting-edge tools in embedded systems, data analysis, and digital manufacturing, with the goal of contributing to the advancement of smart systems and Industry 4.0 applications. In the future, he aspires to engage in interdisciplinary research projects, collaborate with international teams, and publish in areas related to AI-driven automation and intelligent mechatronic systems. He can be contacted at email: U20204796@utp.edu.pe.

Jose Briones Zuñiga holds a master's degree in Mathematical Statistics from the National University of San Marcos and is an economist. He is a RENACYT professor and researcher at the Universidad Tecnológica del Perú, specializing in the area of AI applied to statistical and scientific problems. His work focuses on the development and application of AI techniques to solve challenges in different fields. He can be contacted at email: C19980@utp.edu.pe.