Non-destructive classification of Peruvian medicinal plants using NIR hyperspectral imaging (1300–1700 nm) and machine learning

Clasificación no destructiva de plantas medicinales peruanas mediante imágenes hiperespectrales NIR (1300–1700 nm) y aprendizaje automático

André Rodríguez León[1*], Jimy Oblitas Cruz[1], Jhonsson Luis Quevedo-Olaya[1]

Highlights

- NIR-HSI (1300–1700 nm) combined with machine learning enabled non-destructive discrimination of Peruvian medicinal plants.
- Supervised algorithms achieved accurate multiclass classification even among species with high spectral similarity.
- The methodology is applicable to quality control, authentication, and adulteration prevention in phytotherapeutic products

Innovaciencia ISSN: 2346-075X
E- ISSN: 2346-075X
Innovaciencia 2026; 14(1):e5680
http://dx.doi.org/10.15649/2346075X.5680

ORIGINAL RESEARCH

How to cite this article:

Rodríguez A, Oblitas J, Quevedo-Olaya JL, Non-destructive classification of Peruvian medicinal plants using NIR hyperspectral imaging (1300–1700 nm) and machine learning, 2026; 14 (1): e5680
http://dx.doi.org/10.15649/2346075X.5680

Received: 11 September 2025
Accepted: 27 March 2026
Published: 20 April 2026

Keywords:

NIR-HSI, machine learning, medicinal plants, supervised classification, quality control.

Palabras clave:

NIR-HSI, aprendizaje automático, plantas medicinales, clasificación supervisada, control de calidad

ABSTRACT

Introduction. The identification of native medicinal plants often relies on destructive, time-consuming, or expert-dependent methods. This study proposes the use of near-infrared hyperspectral imaging (NIR-HSI) and machine learning as alternative, non-invasive tools for the rapid discrimination of three medicinal species from northern Peru: *Peperomia inaequalifolia* (congona), *Alternanthera* sp. (lancetilla), and *Teline monspessulana* (retama). **Objectives.** To evaluate the performance of multiclass classification models applied to preprocessed NIR-HSI spectra, aiming to develop a reliable system for plant identification and quality control. **Materials and Methods.** A total of 1467 spectra were collected using a NIR-HSI camera in the range of 1300–1700 nm. Spectral data were preprocessed using Savitzky–Golay smoothing and standard normal variate (SNV). Seven machine learning classifiers were trained and evaluated through stratified 5-fold cross-validation, including Random Forest, Gradient Boosting, XGBoost, and Ridge Classifier. **Results.** Random Forest achieved the highest performance (accuracy = 0.9959), with a ROC-AUC of 1.00. The remaining models yielded mean accuracies ranging from 0.9720 to 0.9945, with ROC-AUC values close to 1.00, indicating strong discriminative capability. **Conclusions.** The combination of NIR-HSI and supervised learning models enables highly accurate classification of medicinal plants. This approach shows potential for traceability, quality assurance, and ethnobotanical validation, particularly in decentralized or resource-limited settings.

RESUMEN

Introducción. La identificación de plantas medicinales nativas suele depender de métodos destructivos, lentos o altamente dependientes de experiencia especializada. Este estudio propone el uso de imágenes hiperespectrales en el infrarrojo cercano (NIR-HSI) y aprendizaje automático como herramientas alternativas, no invasivas, para la discriminación rápida de tres especies medicinales del norte del Perú: *Peperomia inaequalifolia* (congona), *Alternanthera* sp. (lancetilla) y *Teline monspessulana* (retama). **Objetivos.** Evaluar el desempeño de modelos de clasificación multiclase aplicados a espectros NIR-HSI preprocesados, con el fin de desarrollar un sistema confiable para la identificación vegetal y el control de calidad. **Materiales y Métodos.** Se recolectaron un total de 1467 espectros mediante una cámara NIR-HSI en el rango de 1300–1700 nm. Los datos espectrales fueron preprocesados utilizando suavizado Savitzky–Golay y normalización por variación normal estándar (SNV). Se entrenaron y evaluaron siete clasificadores de aprendizaje automático mediante validación cruzada estratificada de 5 pliegues, incluyendo Random Forest, Gradient Boosting, XGBoost y Ridge Classifier. **Resultados.** Random Forest alcanzó el mejor desempeño (accuracy = 0.9959), con un ROC-AUC de 1.00. Los demás modelos presentaron valores de exactitud entre 0.9720 y 0.9945, con valores de ROC-AUC cercanos a 1.00, lo que indica una alta capacidad discriminativa. **Conclusiones.** La combinación de NIR-HSI y modelos supervisados de aprendizaje automático permite una clasificación altamente precisa de plantas medicinales. Este enfoque muestra potencial para aplicaciones en trazabilidad, aseguramiento de la calidad y validación etnobotánica, particularmente en contextos descentralizados o con recursos limitados.

CC BY-NC 4.0

1 Universidad Nacional de Cajamarca, Cajamarca, Perú *Corresponding Author: arodriguezl@unc.edu.pe.

Open access

INTRODUCTION

The ancestral use of medicinal plants such as congona (*Peperomia inaequalifolia*), lancetilla (*Alternanthera* sp.), and retama (*Teline monspessulana*) remains widespread in Peruvian Andean and Amazonian communities, where these species are traditionally used to treat digestive, hepatic, respiratory, and dermatological conditions[1–2]. These plants are currently marketed in dried, fresh, or powdered forms; however, their informal commercialization, often lacking adequate quality control, facilitates adulteration, substitution, and undeclared mixtures, thereby compromising consumer safety and undermining the value of traditional ethnobotanical knowledge.

Despite their medicinal relevance, these species have received limited attention using advanced analytical technologies. Congona, cultivated in inter-Andean regions, has traditionally been used as a cardiotonic, anti-inflammatory, and hepatoprotective agent and is commonly administered as an infusion or macerate to treat headaches, liver disorders, and menstrual discomfort[1]. Lancetilla, in turn, has been used in Amazonian contexts as an anti-inflammatory, febrifuge, and bronchodilator, typically prepared as a decoction for intestinal infections and respiratory diseases[2]. Retama, a plant historically associated with Inca and colonial medicine, contains bioactive alkaloids that, when carefully dosed, have been used as diuretic, wound-healing, and hepatoprotective agents[1–3].

Conventional taxonomic approaches for plant authentication are often impractical because they are time-consuming, destructive, and highly dependent on specialized expertise. In this context, near-infrared hyperspectral imaging (NIR-HSI) has emerged as a promising analytical tool, as it combines spectral and spatial information within the 900–1700 nm range, enabling the detection of chemical signatures associated with bioactive compounds even in dried or ground samples. The spectral information captured by NIR-HSI is based on the interaction of radiation with C–H, O–H, and N–H bond vibrations, whose absorption features provide indirect information related to molecular composition[4–6]. In particular, the 1300–1700 nm region is sensitive to overtones and combination bands associated with water, lipids, phenolic compounds, and alkaloids[7–9]. However, due to the high dimensionality of hyperspectral data and the overlapping nature of NIR absorption bands, direct extraction of structural information remains limited[10]. When samples exhibit similar compositions that are difficult to distinguish visually, machine learning–based classification models provide an effective strategy for discriminating species with minimal human intervention and reduced error rates.

Chemometric tools such as Principal Component Analysis (PCA) and supervised models, including Support Vector Machines, Random Forest, and XGBoost, have demonstrated high accuracy in the discrimination of plant species, even after processing or storage[11–12]. These multivariate approaches exploit spectral complexity to identify patterns that are not visually discernible, making them a reliable alternative to traditional botanical identification methods[13–14].

In this study, the potential of NIR-HSI combined with machine learning algorithms was evaluated for differentiating three Peruvian medicinal plant species—congona, lancetilla, and retama—which have been used

since pre-Inca times in traditional medicine[3,15–16]. This approach is supported by previous studies. Singh et al.[17] evaluated *Andrographis paniculata* powder in the 900–1700 nm range using Support Vector Machines and ten-fold cross-validation, classifying three quality categories with a mean accuracy of 83% after SNV normalization. In the same year, Kasemsumran et al.[18] developed PLS-DA models to distinguish *A. paniculata* samples with high and low andrographolide content, achieving 100% correct classification in both calibration and validation using second-derivative preprocessing and six PLS factors. More recently, Zhang et al.[19] discriminated five chrysanthemum tea varieties using a portable NIR system and LDA-based approaches combined with K-nearest neighbor after PCA, reaching accuracies of 87.2%, 94.4%, and 99.2%. Similarly, Li et al.[20] showed that NIR-HSI coupled with machine learning and deep learning models enabled rapid and accurate identification of medicinal species (*Gastrodia elata* Blume), achieving perfect classification rates without excessively complex preprocessing. In addition, Jayapal et al.[21] used regression-based models, including PLSR, SVM, regression trees, and PCR, to predict phenolic compounds in *Arabidopsis*, obtaining a standard error of prediction of 0.07 mg/g, although classification was not addressed.

The cultural and pharmacological relevance of these species, together with the need for reliable identification methods in industrial and decentralized settings, provides a strong rationale for proposing NIR-HSI as a non-invasive, reproducible, and portable technique suitable for monitoring plant materials in markets, processing centers, and germplasm banks[19]. Accordingly, the aim of this study was to evaluate the discriminative performance of machine learning models applied to NIR-HSI data for identifying these three species and to assess their potential for quality control, traceability, and the conservation of Peruvian plant diversity.

MATERIALS AND METHODS

Study type and design

An observational, analytical, cross-sectional study was conducted

Population, sample, and preparation

The target population consisted of medicinal plant species from the Andean region of La Libertad, Peru. The sample included three species: *Peperomia inaequalifolia* (congona), *Alternanthera* sp. (lancetilla), and *Teline monspessulana* (retama).

Spectral acquisition

As shown in (**Figure 1a**), the plants were collected by local specialists from their natural distribution areas, transported to the laboratory in paper bags, and stored without further processing, preserving their original commercial presentation. This strategy allowed the spectral analysis to reflect real conditions of use and commercialization. For hyperspectral acquisition, the leaf was selected as the target tissue, and regions of interest were defined by identifying spectrally homogeneous areas across the leaf surface (**Figure 1b**). This selection was performed using the discrimination capabilities of the Resonon software, which allows clustering

and visualization of pixels with similar spectral signatures, thereby reducing intra-leaf variability associated with structural heterogeneity (e.g., veins and edges). A total of 20 independent leaves per herb type were analyzed, from which an average of 24 spectra per sample was extracted. Hyperspectral images were acquired using a NIR RESONON PIKA IR+ system equipped with an InGaAs camera (900–1700 nm range), an optical spectrograph (5.6 nm FWHM spectral resolution), and an illumination system with two 150 W halogen lamps positioned at 45°. The camera was mounted on a motorized conveyor belt, and the acquisition parameters were adjusted to avoid saturation: focal distance, 25 cm; speed, 20 mm/s; and exposure time, 5 ms[22,23]. Raw images were corrected by reflectance normalization using white (Teflon, 99%) and black (dark capture) references according to the standard equation[24–26].

Spectral preprocessing

The corrected images were segmented using binary masks for each spectral band. Mean spectra were extracted, and the analysis was restricted to the 1300–1700 nm range because of its high density of vibrational information[27,28].

A Savitzky–Golay filter was applied (31-point window, first-order polynomial) to reduce high-frequency noise while preserving peak shape, which is essential for retaining relevant spectral features[29]. The data were then normalized using Standard Normal Variate (SNV), a technique that corrects baseline shifts and intensity variations caused by scattering effects[23].

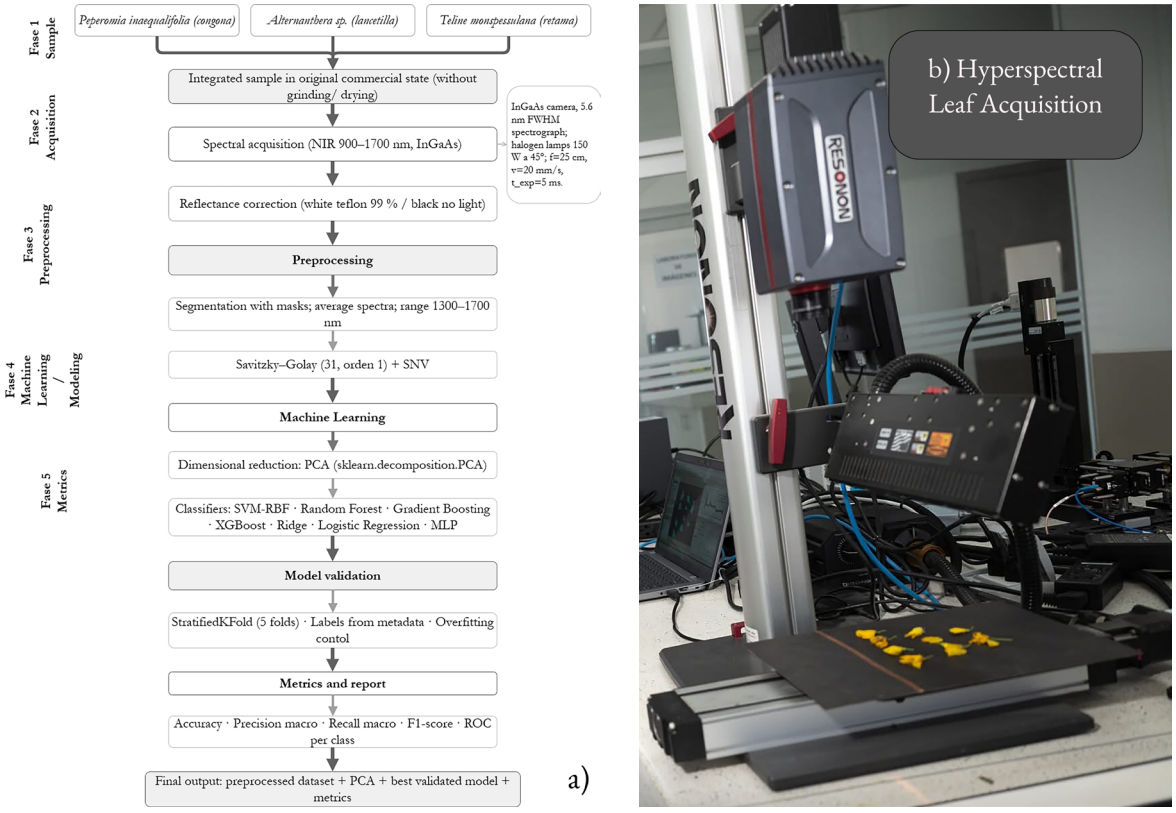

Figure 1. a) Workflow for medicinal plant classification using machine learning; **b)** Spectral acquisition by hyperspectral imaging.

Computational and statistical analysis

Machine learning

Data processing was performed in Google Colab using Python 3.10. The following libraries were used: NumPy for array handling, Matplotlib for visualization, Scikit-learn for machine learning algorithms, and XGBoost for gradient boosting models. Dimensionality reduction was performed using Principal Component Analysis (PCA) implemented through sklearn.decomposition.PCA. Seven classifiers were trained and validated: i) Support Vector Machine (RBF kernel); ii) Random Forest; iii) Gradient Boosting; iv) XGBoost; v) Ridge Classifier; vi) Logistic Regression; vii) MLPClassifier (Multilayer Perceptron).

Statistical analysis and validation

Model evaluation was performed using stratified five-fold cross-validation (StratifiedKFold) to preserve class balance. Labels were assigned based on collection metadata and numerically encoded. Performance was assessed using accuracy, macro-precision, macro-recall, F1-score, and ROC curves[30–31].

Data availability

The dataset generated during image acquisition is publicly available in Mendeley Data: Rodríguez León, André (2025), *Dataset clasificación HM*, Mendeley Data, V1, doi: 10.17632/kythyt9dt2.1.

RESULTS

Average spectral characteristics

The spectra were initially obtained in their raw form, as shown in (**Figure 2a**). Owing to sample characteristics and acquisition conditions, the raw spectra exhibited substantial noise. To prevent this noise from affecting PCA-based description, the spectra were preprocessed, as shown in (**Figure 2b**).

Analysis of the mean spectra obtained in the 1300–1700 nm range revealed distinct patterns among the three evaluated species. Five spectral regions with prominent peaks and valleys were identified. The first peak, located around 1360 nm, was common to all three species and was associated with the second overtone of the C–H bond, a region that has been related in the NIR literature to phenolic compounds and lipids[32–33]. In this region, *P. inaequalifolia* showed greater relative signal intensity. A second peak, near 1450 nm, was attributed to the first overtone of O–H stretching, commonly associated with structured water or hydroxylated compounds[34]. In this case, *T. monspessulana* exhibited a more pronounced valley. Between 1450 and 1540 nm, a common pek was observed in the three species, associated with combination bands of N–H and C–H bonds, which have frequently been related in the literature to alkaloids, tannins, or secondary amino acids. In this region, *Alternanthera* sp. showed a smoother slope, which may reflect a lower relative contribution of spectral signals associated with nitrogen-containing compounds compared with the other two species.

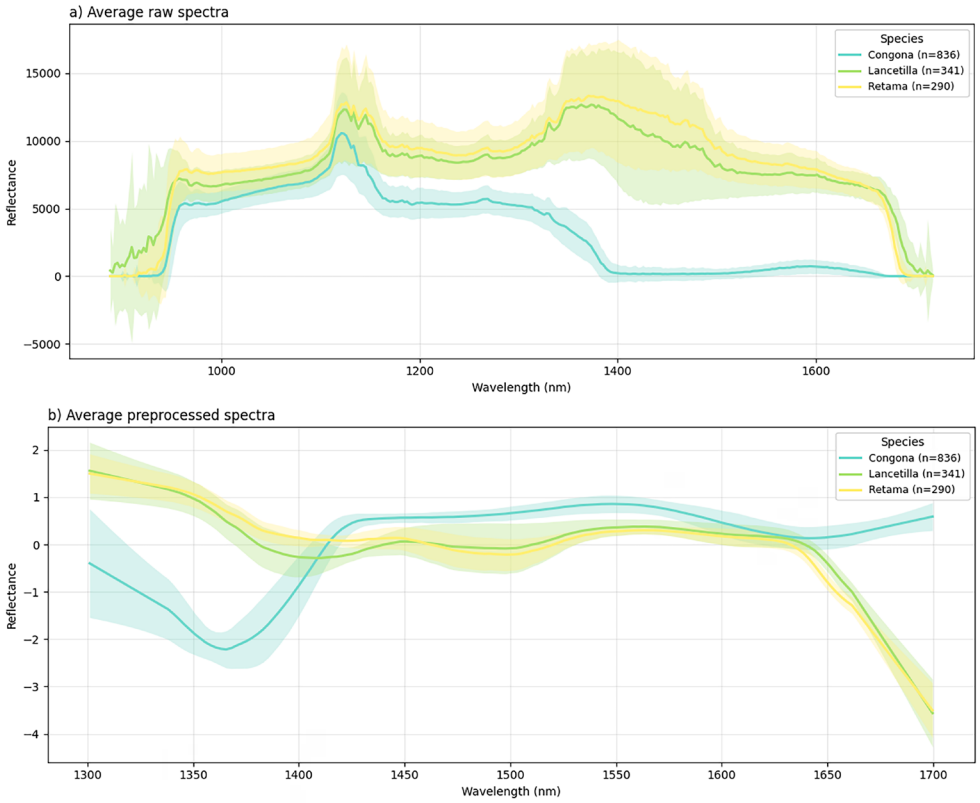

Figure 2. Relative absorbance spectra of the three studied species: *Peperomia inaequalifolia*, *Alternanthera* sp., and *Teline monspessulana*. (**a**) Raw spectra. (**b**) Cropped spectra (1300–1700 nm) preprocessed with Savitzky–Golay derivative and SNV.

The region between 1600 and 1650 nm exhibited marked differences among species. In *Teline monspessulana* and *Alternanthera* sp., greater relative absorbance was observed around 1620 nm, a band associated with CH2 and aromatic C–H stretching combinations[35]. In contrast, *P. inaequalifolia* showed a pronounced decrease in this region, which may indicate a lower relative contribution of signals associated with metabolites containing conjugated aromatic structures.

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) was applied to the preprocessed spectra (first-derivative Savitzky–Golay + SNV) to reduce dataset dimensionality and explore unsupervised grouping patterns among the studied species. In the PCA plot (**Figure 3**), the blue arrows represent the loading vectors corresponding to the 20 most influential wavelengths. These vectors indicate how each wavelength contributed to group separation in the PC1–PC2 space, highlighting key directions that reflect differences in spectral patterns among species. The first two principal components jointly explained 92.5% of the total variance (PC1: 82.7%; PC2: 9.8%), indicating that most of the relevant spectral information was concentrated in a two-dimensional space. The distribution of samples in the PC1 vs. PC2 plane showed that *Peperomia inaequalifolia* (congona) formed a distinct and well-defined cluster relative to the other two species. These samples (gray points) were located at positive PC1 values and associated with vectors at wavelengths close to 1680–1699 nm.

The congona group exhibited a crescent-shaped distribution, corresponding to the horseshoe effect in PCA, which is associated with nonlinear gradients in spectral data[36]. Although the vectors in the 1682–1699 nm range accounted for most of the variance along PC1, they did not capture the vertical dispersion observed in the dataset. In contrast, *Alternanthera* sp. (lancetilla) and *Teline monspessulana* (retama) were grouped at negative PC1 values, with vectors associated with wavelengths between 1360 and 1440 nm. These two species showed partial overlap, indicating similar spectral signatures within the analyzed range. Because PCA did not achieve complete separation among the three species, particularly between lancetilla and retama, the results obtained using supervised classification models are presented below.

Supervised classification

Structure and selection of supervised models

Seven supervised learning models were implemented to cover a range of methodological approaches suitable for complex spectral data. These included linear, nonlinear, ensemble-based, and kernel-based algorithms. The evaluated models were Random Forest (RF), Gradient Boosting (GB), Support Vector Machine with RBF kernel (SVM), Ridge Classifier (RC), XGBoost (XGB), Logistic Regression (LR), and Multilayer Perceptron (MLP). Each model was trained using spectral data numerically labeled according to species of origin. Performance was assessed using stratified five-fold cross-validation (StratifiedKFold, k = 5), a standard approach in supervised modeling of spectral data[37–38].

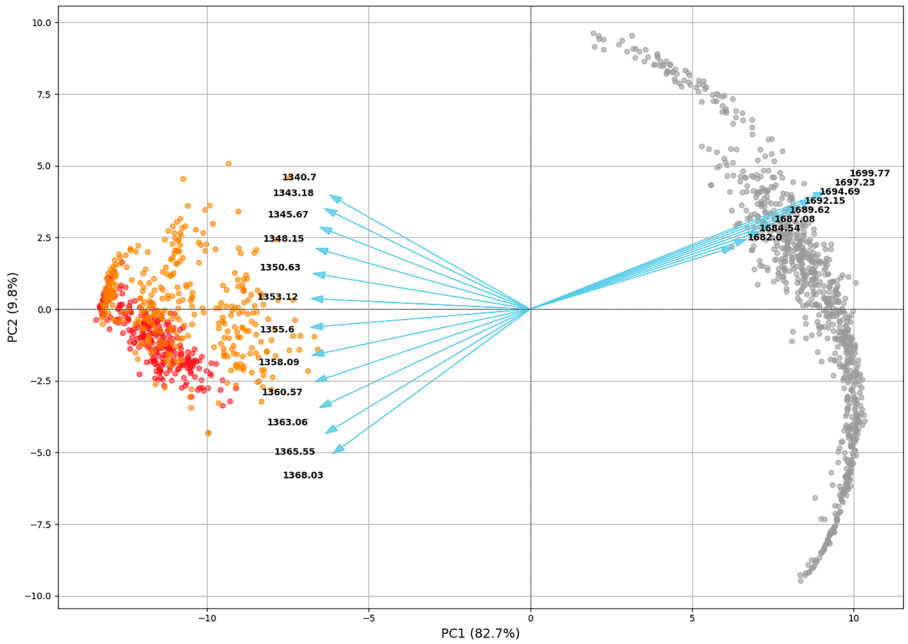

Figure 3. PCA biplot showing samples by species and the 20 most influential wavelengths. Points represent leaf samples of Retama (**red**), Lancetilla (orange), and Congona (**gray**). Blue vectors indicate the wavelengths with the greatest contribution to group separation; arrows pointing left (≈1340–1368 nm) correspond to O–H and C–H related regions, whereas arrows pointing right (≈1682–1699 nm) correspond to N–H/C=O related regions. PC1 explains 82.7% and PC2 explains 9.8% of total variance.

No hyperparameter fine-tuning was performed. Model configurations were as follows: SVM with RBF kernel (C = 1.0, gamma = scale); RF, GB, and XGB with *n_estimators* = 100; RF using the Gini criterion; GB using log loss; MLP with one hidden layer of 100 neurons and ReLU activation; and RC with default L2 regularization.

Classification performance evaluation

Comparative metrics

The quantitative evaluation of the seven supervised models was performed using standard metrics for multiclass classification, with accuracy as the primary metric in this comparison. **Table 1** summarizes the mean accuracy values obtained through stratified five-fold cross-validation. Random Forest showed the highest performance, whereas MLPClassifier yielded the lowest value.

Table 1. Mean performance metrics for each supervised model using five-fold cross-validation.

Modelo	Accuracy
XGBoost	0.9891
Random Forest	0.9959
Gradient Boosting	0.9932
MLPClassifier	0.9720
SVM (RBF)	0.9932
Logistic Regression	0.9925
Ridge Classifier	0.9945

Error analysis using confusion matrices

Confusion matrices were used to identify classification error patterns and determine which species were more frequently confused. Figure 4 presents the confusion matrices for the four models with the highest performance: Random Forest, Ridge Classifier, SVM, and Gradient Boosting.

Random Forest (**Figure 4a**) produced one classification error, assigning one lancetilla sample to retama. Ridge Classifier (**Figure 4b**) and SVM (**Figure 4c**) each produced two classification errors, distributed between lancetilla and retama. A similar pattern was observed for Gradient Boosting (**Figure 4d**), which also produced two classification errors between the same classes. In all four models, congona was correctly classified in 100% of the repetitions.

ROC curves and AUC

ROC curves were used to assess model discrimination in a multiclass setting using a one-vs-rest binarization

scheme. As shown in **Figure 5** , all models exhibited near-perfect discrimination, with ROC curves approaching the upper-left corner and AUC values close to 1.00. Minor differences between classes were observed in the numerical AUC values (0.994–1.000), particularly for Lancetilla in the Gradient Boosting model, consistent with the few observed classification errors.

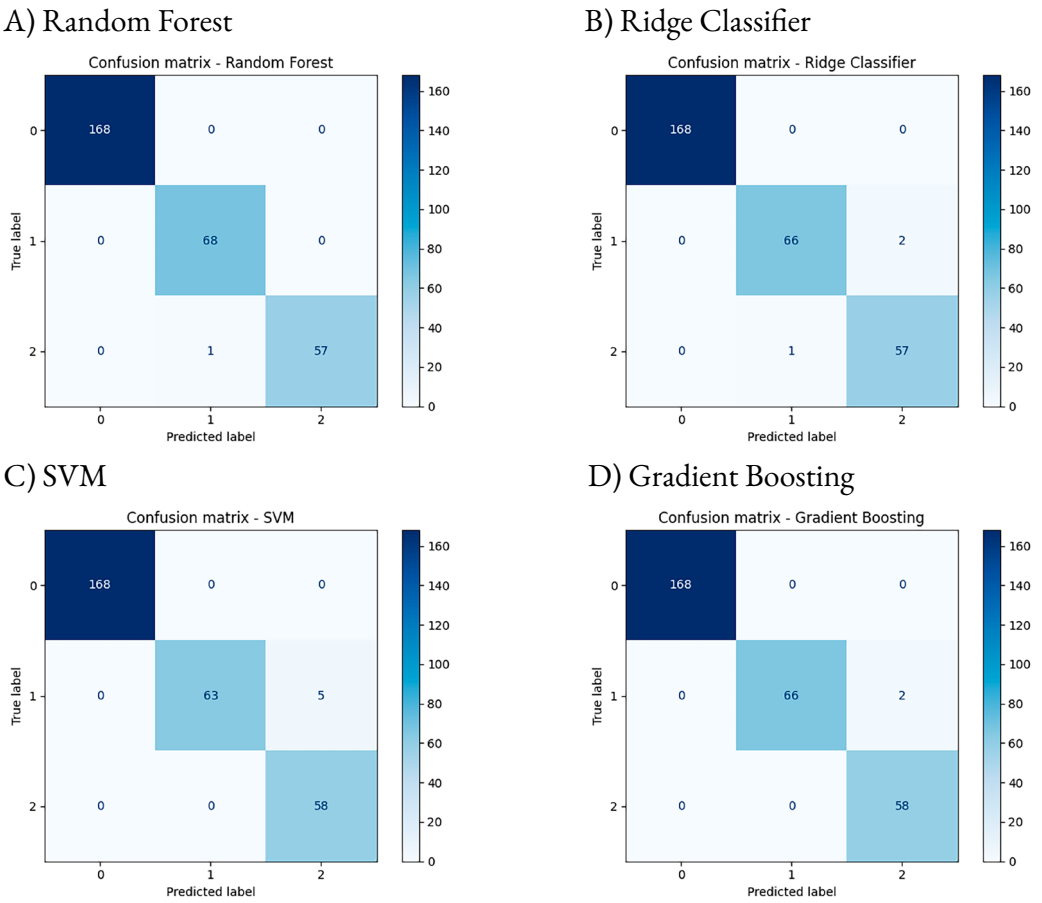

Figure 4. Confusion matrices for Random Forest (**a**), Ridge Classifier (**b**), SVM (**c**), and Gradient Boosting (**d**), expressed as the number of samples classified per class.

Comparative performance of multiclass classification models

The performance of the four selected multiclass classification models was evaluated using accuracy, macro-precision, macro-recall, macro-F1 score, and macro-ROC AUC. **Table 2** summarizes the results. Random Forest showed the strongest overall performance across most metrics, whereas Ridge Classifier and Gradient Boosting yielded closely comparable values. SVM reached the highest macro-ROC AUC, although its remaining metrics were slightly lower.

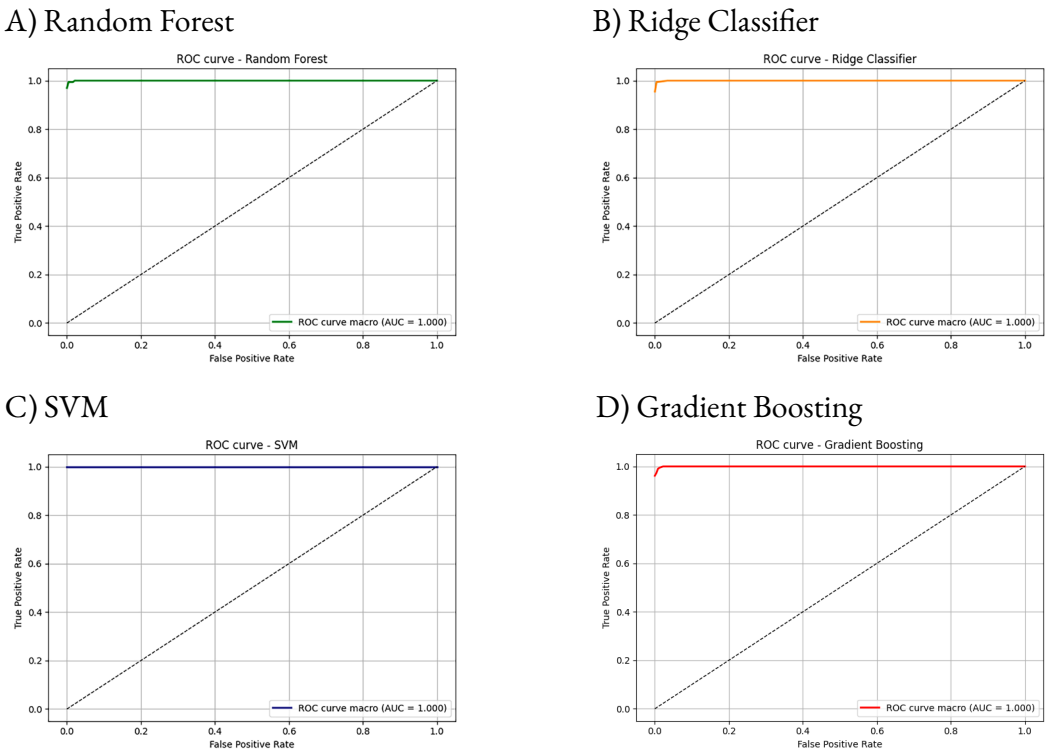

Figure 5. ROC curves for Random Forest (**a**), Ridge Classifier (**b**), SVM (**c**), and Gradient Boosting (**d**) under a one-vs-rest multiclass scheme.

Table 2. Mean performance metrics for the four selected supervised models using five-fold cross-validation.

Model	Accuracy	Macro-precision	Macro-recall	Macro-F1 score	Macro-ROC AUC
Random Forest	0.9966	0.9952	0.9943	0.9947	0.9998
Ridge Classifier	0.9932	0.9894	0.9894	0.9894	0.9996
Gradient Boosting	0.9932	0.9889	0.9902	0.9894	0.9996
SVM	0.9830	0.9735	0.9755	0.9735	1.0000

DISCUSSION

The multivariate analysis showed that PCA explained 92.5% of the total variability (PC1: 82.7%; PC2: 9.8%), although it did not achieve complete separation between retama and lancetilla. This partial overlap may be related to the similarity of their spectral signatures within the 1300–1700 nm range, particularly in regions commonly associated with O–H and C–H overtones[31,35]. These findings indicate that, although PCA is effective for exploring structural variance and identifying dominant spectral patterns, its linear nature limits class separability when species exhibit subtle or overlapping signatures. Similar limitations have been reported in hyperspectral applications, where PCA is useful for dimensionality reduction but does not necessarily maximize discrimination between classes[39–40].

In contrast, the supervised models achieved near-perfect performance. Random Forest reached 99.66% accuracy, followed by Ridge Classifier and Gradient Boosting (99.32%) and SVM (98.30%), with macro-F1 scores ranging from 0.9735 to 0.9947. These results indicate that the classifiers captured subtle and potentially nonlinear

differences among chemically related species that were not fully resolved by PCA. This behavior is consistent with previous studies demonstrating that supervised learning improves discrimination in hyperspectral datasets of plant materials and related classification tasks [22–23]. Similarly, Yang et al.[41] reported 98.95% accuracy in the classification of botanically similar materials using supervised approaches, whereas Zheng et al.[42] achieved 100% accuracy with SVM for species identification using NIR data. Liu et al.[21] also reported 97.9% accuracy using NIR-HSI combined with SVM for geographic authentication, supporting the applicability of supervised spectral models in complex classification scenarios. Together, these findings reinforce the use of NIR-HSI combined with machine learning as a reliable strategy for discriminating closely related plant species.

From an applied perspective, the proposed methodology offers advantages in terms of speed, portability, and economic feasibility. Li et al.[43] reported 100% accuracy in *Fritillaria* spp. authentication using NIR-based models, with relevant spectral information concentrated between 1400 and 1500 nm, whereas Altieri et al.[44] achieved 97.8% accuracy in ripeness classification using portable NIR devices. Compared with highly specialized VNIR-SWIR and deep learning approaches reporting accuracies above 95%[45], the present study provides a favorable balance between classification performance (>98%) and technological feasibility, allowing potential implementation in germplasm banks, nurseries, and local markets. In addition, portable NIR analysis typically requires less than two minutes and has been reported to cost approximately USD 5 per sample[43,45–46], making it an efficient alternative to conventional chromatographic methods. Chen et al.[46] further showed that combining FT-NIR with machine learning enables rapid adulteration detection and concentration prediction, supporting the usefulness of spectroscopic pattern analysis in practical authentication settings. Within this framework, the use of baseline model configurations without extensive hyperparameter optimization supports a more application-oriented evaluation, facilitating comparison across classifiers with different levels of interpretability and generalization capacity.

The proposed methodology can be adapted to low-cost, low-power platforms. The use of open-source libraries such as Scikit-learn and compatibility with compact NIR sensors based on MEMS or photodiodes enable the development of portable devices or cloud-connected mobile applications, facilitating model updating and recalibration. Commercial systems such as SCiO and NIRScan have demonstrated effectiveness in related applications, reinforcing the feasibility of portable NIR technologies as analytical tools. In particular, the DLP NIRScan Nano (Texas Instruments), based on digital micromirror device (DMD) technology, has been described as a promising compact NIR platform for portable analytical applications[33], supporting its potential use in rural or low-infrastructure settings.

Finally, a relevant limitation of this study is the absence of independent external validation, which limits the extrapolation of the models to new populations or operational conditions[26]. Additionally, batch variability and different collection contexts were not considered, factors that may influence spectral responses. However, the consistency observed across metrics such as accuracy, F1-score, and ROC curves suggests high internal model stability against within-sample variability. The ROC curves appear visually saturated due to the high classification performance, which may limit the graphical differentiation between models despite small numerical differences in AUC values. In this regard, although external validation is required to confirm

generalization, the obtained results provide a solid and reproducible foundation for future applications under real-world conditions.

CONCLUSIONS

The integration of NIR-HSI in the 1300–1700 nm range with machine learning models proved to be a precise, non-destructive, and reproducible strategy for the identification of Peruvian medicinal plant species. Spectral preprocessing using Savitzky–Golay filtering and SNV improved signal quality and enhanced the performance of supervised classifiers, enabling effective discrimination among *Peperomia inaequalifolia*, *Alternanthera* sp., and *Teline monspessulana*.

Although *Peperomia inaequalifolia* exhibited a clearly differentiated spectral profile, the separation between *Alternanthera* sp. and *Teline monspessulana* was only achieved through supervised models capable of capturing nonlinear relationships. Random Forest showed the best performance (macro-F1 score = 0.9947), followed by Gradient Boosting, Ridge Classifier, and SVM-RBF, all with metrics above 0.97, indicating the robustness of the proposed approach in the presence of morphological and phytochemical similarity.

The recurrence of relevant spectral bands (1360, 1450, 1470, 1530, 1600, 1620, 1650, and 1692–1699 nm), consistent with spectral regions commonly associated with O–H, C–H, and N–H vibrations, supports the feasibility of developing optimized sensors focused on these regions.

Overall, the proposed methodology represents a viable alternative for quality control, plant authentication, and real-time classification, with potential applications in markets, germplasm banks, and responsible biotrade. Future expansion of the species set and validation using portable devices would further strengthen its applicability in decentralized settings.

ACKNOWLEDGMENTS

The authors sincerely thank Dr. Wilson Castro for his methodological support and for providing access to the specialized equipment used in this study. His technical advice and scientific guidance were essential for the implementation and validation of the experimental procedures.

ETHICAL CONSIDERATIONS

This study did not involve human participants or vertebrate animals. Plant material was collected and handled in accordance with current Peruvian regulations on access to biological resources and biodiversity, under principles of sustainability and conservation.

FUNDING

This research was funded with the authors' own resources. No specific funding was received from public, commercial, or non-profit agencies.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

REFERENCES

1. **Valarezo E, Herrera-García M, Astudillo-Dávila P, Rosales-Demera I, Jaramillo-Fierro X, Cartuche L, et al.** Study of the chemical composition and biological activity of the essential oil from congona (*Peperomia inaequalifolia* Ruiz and Pav.). Plants (Basel). 2023;12:1504. Available from: https://doi.org/10.3390/plants12071504
2. **Ahmed A, Ghatas Y, Mohamed SM.** Evaluation the growth, some chemical constituents and landscape value of *Alternanthera dentata* plant grown under different planting methods and distances with herbaceous plants. Ann Agric Sci Moshtohor. 2024;62:57-66. Available from: https://doi.org/10.21608/assjm.2024.285941.1279
3. **Espinoza-Gavilanes R, Tuza-Roa I, Vásquez-Freytez C, Jaramillo-Loayza K, Noriega-Rivera P.** Efecto acaricida y ovicida de los aceites esenciales de *Chenopodium ambrosioides* L. y *Peperomia inaequalifolia* Ruiz & Pav. contra *Tetranychus urticae* en fresa (*Fragaria* spp.). Polibotanica. 2024;(57). Available from: https://doi.org/10.18387/polibotanica.57.14
4. **Islam M, Bijjahalli S, Fahey T, Gardi A, Sabatini R, Lamb DW.** Destructive and non-destructive measurement approaches and the application of AI models in precision agriculture: a review. Precis Agric. 2024;25:1127-80. Available from: https://doi.org/10.1007/s11119-024-10112-5
5. **Atanassova S, Petrova A, Yorgov D, Mineva R, Veleva P.** Visible and near-infrared spectroscopy for investigation of water and nitrogen stress in tomato plants. AgriEngineering. 2025;7:155. Available from: https://doi.org/10.3390/agriengineering7050155
6. **Zhao B, Zhang H, Liu X, Dong Q, Zang H.** Study of glycated human serum albumin in non-enzymatic glycation process based on MIR/NIR spectroscopy. J Mol Struct. 2025;1335:141928. Available from: https://doi.org/10.1016/j.molstruc.2025.141928
7. **Benetti RA, Arantes PR, Oliveira PBR, Santos TB, Filho LJR, Lima UC, et al.** Near infrared spectroscopy for the photodiagnosis of osteonecrosis, a future perspective: mini historical review. J Near Infrared Spectrosc. 2025;33:3-9. Available from: https://doi.org/10.1177/09670335251329601
8. **Santos YJS, Malegori C, Colnago LA, Vanin FM.** Application of infrared spectroscopy for the analysis of total phenolic compounds in fruits. Crit Rev Food Sci Nutr. 2024;64:2906-16. Available from: https://doi.org/10.1080/10408398.2022.2128036
9. **Barbinta-Patrascu ME, Bita B, Negut I.** From nature to technology: exploring the potential of plant-based materials and modified plants in biomimetics, bionics, and green innovations. Biomimetics (Basel). 2024;9:390. Available from: https://doi.org/10.3390/biomimetics9070390
10. **Beć KB, Grabska J, Huck CW.** Interpretability in near-infrared (NIR) spectroscopy: current pathways to the long-standing challenge. Trends Anal Chem. 2025;189:118254. Available from: https://doi.org/10.1016/j.trac.2025.118254

11. **Wan-Azemin A, Suryati Mohd K, Rao USM, Sasidharan S, Dharmaraj S.** Chemometric analysis of attenuated total reflection-Fourier transform infrared (ATR-FTIR) spectra for geographical authentication of *Melastoma malabathricum*. Res J Pharm Technol. 2024;17:3769-76. Available from: https://doi.org/10.52711/0974-360X.2024.00586
12. **Chang B, Li F, Hu Y, Yin H, Feng Z, Zhao L.** Application of UAV remote sensing for vegetation identification: a review and meta-analysis. Front Plant Sci. 2025;16:1452053. Available from: https://doi.org/10.3389/fpls.2025.1452053
13. **Cioanca O, Lungu II, Mita-Baciu I, Robu S, Burlec AF, Hancianu M, et al.** Extraction and purification of catechins from tea leaves: an overview of methods, advantages, and disadvantages. Separations. 2024;11:171. Available from: https://doi.org/10.3390/separations11060171
14. **Yu Y, Huang J, Wang L, Liang S.** A 1D-Inception-ResNet based global detection model for thin-skinned multifruit spectral quantitative analysis. Food Control. 2025;167:110823. Available from: https://doi.org/10.1016/j.foodcont.2024.110823
15. **Singla RK, Dhir V, Madaan R, Kumar D, Singh Bola S, Bansal M, et al.** The genus *Alternanthera*: phytochemical and ethnopharmacological perspectives. Front Pharmacol. 2022;13:769111. Available from: https://doi.org/10.3389/fphar.2022.769111
16. **Laranjeira IM, Dias ACP, Pinto-Ribeiro FL.** *Genista tridentata* phytochemical characterization and biological activities: a systematic review. Biology (Basel). 2023;12:1387. Available from: https://doi.org/10.3390/biology12111387
17. **Sing D, Banerjee S, Jana SN, Mallik R, Dastidar SG, Majumdar K, et al.** Estimation of andrographolides and gradation of *Andrographis paniculata* leaves using near infrared spectroscopy together with support vector machine. Front Pharmacol. 2021;12:629833. Available from: https://doi.org/10.3389/fphar.2021.629833
18. **Kasemsumran S, Apiwatanapiwat W, Ngowsuwan K, Jungtheerapanich S.** Rapid selection of *Andrographis paniculata* medicinal plant materials based on major bioactive using near-infrared spectroscopy. Chem Pap. 2021;75:5633-44. Available from: https://doi.org/10.1007/s11696-021-01746-0
19. **Zhang J, Wu X, He C, Wu B, Zhang S, Sun J.** Near-infrared spectroscopy combined with fuzzy improved direct linear discriminant analysis for nondestructive discrimination of chrysanthemum tea varieties. Foods. 2024;13:1439. Available from: https://doi.org/10.3390/foods13101439
20. **Li G, Li J, Liu H, Wang Y.** Rapid and accurate identification of *Gastrodia elata* Blume species based on FTIR and NIR spectroscopy combined with chemometric methods. Talanta. 2025;281:126910. Available from: https://doi.org/10.1016/j.talanta.2024.126910
21. **Jayapal PK, Joshi R, Sathasivam R, Van Nguyen B, Faqeerzada MA, Park SU, et al.** Non-destructive measurement of total phenolic compounds in *Arabidopsis* under various stress conditions. Front Plant Sci. 2022;13:982247. Available from: https://doi.org/10.3389/fpls.2022.982247
22. **Chen Z, Xue X, Wu H, Gao H, Wang G, Ni G, et al.** Visible/near-infrared hyperspectral imaging combined with machine learning for identification of ten *Dalbergia* species. Front Plant Sci. 2024;15:1413215. Available from: https://doi.org/10.3389/fpls.2024.1413215

23. **Liu X, Wu Z, Zhao Q, Yu Y, Li Z.** Using near-infrared hyperspectral imaging combined with machine learning to predict the components and the origin of *Radix Paeoniae Rubra*. Anal Methods. 2025;17:1334-44. Available from: https://doi.org/10.1039/D4AY01977F

24. **Ahmed MT, Monjur O, Kamruzzaman M.** Deep learning-based hyperspectral image reconstruction for quality assessment of agro-product. J Food Eng. 2024;382:112223. Available from: https://doi.org/10.1016/j.jfoodeng.2024.112223

25. **Du Z, You S, Cheng C, Wei S.** Automatic spectral calibration of hyperspectral images: method, dataset and benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2025. p. 28081-90. Available from: https://doi.org/10.1109/CVPR52734.2025.02615

26. **Rodríguez-León A, Oblitas J, Quevedo-Olaya JL, Vera W, Quispe-Santivañez GW, Salvador-Reyes R.** Non-destructive detection of *Elasmopalpus lignosellus* infestation in fresh asparagus using VIS-NIR hyperspectral imaging and machine learning. Foods. 2026;15:355. Available from: https://doi.org/10.3390/foods15020355

27. **Dalal N, Sáiz MJ, Caporale AG, Baldini F, Babayan SA, Adamo P.** Fishy forensics: FT-NIR and machine learning based authentication of Mediterranean anchovies (*Engraulis encrasicolus*). J Food Compos Anal. 2024;136:106847. Available from: https://doi.org/10.1016/j.jfca.2024.106847

28. **Krishnamoorthi S, Urano D.** Hyperspectral reflectance imaging and spectral component analysis techniques to reveal distinct color patterns on plant leaves. STAR Protoc. 2025;6:103854. Available from: https://doi.org/10.1016/j.xpro.2025.103854

29. **Sutliff BP, Beaucage PA, Audus DJ, Orski SV, Martin TB.** Sorting polyolefins with near-infrared spectroscopy: identification of optimal data analysis pipelines and machine learning classifiers. Digit Discov. 2024;3:2341-55. Available from: https://doi.org/10.1039/D4DD00235K

30. **Guo K, Shen Y, Gonzalez-Montiel GA, Huang Y, Zhou Y, Surve M, et al.** Artificial intelligence in spectroscopy: advancing chemistry from prediction to generation and beyond. In: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence. 2025. p. 10445-54. Available from: https://doi.org/10.24963/ijcai.2025/1160

31. **Beć KB.** A simple guide to complex world of overtone and combination bands: theoretical simulation and interpretation of NIR spectra—summary of the workshop at NIR-2021 Beijing Conference. NIR News. 2021;32:15-8. Available from: https://doi.org/10.1177/09603360211060966

32. **Türker-Kaya S, Huck C.** A review of mid-infrared and near-infrared imaging: principles, concepts and applications in plant tissue analysis. Molecules. 2017;22:168. Available from: https://doi.org/10.3390/molecules22010168

33. **Beć KB, Grabska J, Huck CW.** Principles and applications of miniaturized near-infrared (NIR) spectrometers. Chemistry. 2021;27:1514-32. Available from: https://doi.org/10.1002/chem.202002838

34. **Birenboim M, Brikenstein N, Kenigsbuch D, Shimshoni JA.** Aquaphotomics study of fresh cannabis inflorescence: near infrared spectral analysis of water matrix structures. Anal Bioanal Chem. 2025;417:747-60. Available from: https://doi.org/10.1007/s00216-024-05685-z

35. **Ma L, Peng Y, Pei Y, Zeng J, Shen H, Cao J, et al.** Systematic discovery about NIR spectral assignment from chemical structural property to natural chemical compounds. Sci Rep. 2019;9:9503. Available from: https://doi.org/10.1038/s41598-019-45945-y

36. **Shah N, Meng Q, Zou Z, Zhang X.** Systematic analysis on the horse-shoe-like effect in PCA plots of scRNA-seq data. Bioinform Adv. 2024;4:vbae109. Available from: https://doi.org/10.1093/bioadv/vbae109

37. **Cao L, Sun M, Yang Z, Jiang D, Yin D, Duan Y.** A novel Transformer-CNN approach for predicting soil properties from LUCAS Vis-NIR spectral data. Agronomy. 2024;14:1998. Available from: https://doi.org/10.3390/agronomy14091998

38. **Peng X, Yu X, Lu L, Ye X, Zhong L, Hu W, et al.** Application of handheld near infrared spectrometer in quality control of traditional Chinese medicine: rapid screening and quantitative analysis of *Lonicerae Japonicae* Flos adulteration. Spectrochim Acta A Mol Biomol Spectrosc. 2025;326:125215. Available from: https://doi.org/10.1016/j.saa.2024.125215

39. **Chen R, Liu F, Zhang C, Wang W, Yang R, Zhao Y, et al.** Trends in digital detection for the quality and safety of herbs using infrared and Raman spectroscopy. Front Plant Sci. 2023;14:1128300. Available from: https://doi.org/10.3389/fpls.2023.1128300

40. **Hajaj S, El Harti A, Pour AB, Khandouch Y, Üstüner M, Amiri MM.** Balancing hyperspectral dimensionality reduction and information preservation for machine learning-based lithological classification using EnMAP hyperspectral imagery. Remote Sens Appl Soc Environ. 2025;38:101618. Available from: https://doi.org/10.1016/j.rsase.2025.101618

41. **Yang Y, Wang S, Zhu Q, Qin Y, Zhai D, Lian F, et al.** Non-destructive geographical traceability of American ginseng using near-infrared spectroscopy combined with a novel deep learning model. J Food Compos Anal. 2024;136:106736. Available from: https://doi.org/10.1016/j.jfca.2024.106736

42. **Zheng C, Li J, Liu H, Wang Y.** Rapid and non-invasive estimation of total phenol content and species identification in dried wild edible bolete using FT-NIR spectroscopy. Arab J Chem. 2024;17:106011. Available from: https://doi.org/10.1016/j.arabjc.2024.106011

43. **Li G, Li J, Liu H, Wang Y.** Rapid and accurate identification of *Gastrodia elata* Blume species based on FTIR and NIR spectroscopy combined with chemometric methods. Talanta. 2025;281:126910. Available from: https://doi.org/10.1016/j.talanta.2024.126910

44. **Altieri G, Laveglia S, Rashvand M, Genovese F, Matera A, Mininni AN, et al.** Portable NIR spectroscopy combined with machine learning for kiwi ripeness classification: an approach to precision farming. Appl Sci (Basel). 2025;15:6233. Available from: https://doi.org/10.3390/app15116233

45. **Yuan M, Ding L, Bai R, Yang J, Zhan Z, Zhao Z, et al.** Feature-level hyperspectral data fusion with CNN modeling for non-destructive authentication of "Weilian" from different origins. Microchem J. 2025;215:114201. Available from: https://doi.org/10.1016/j.microc.2025.114201

46. **Chen Y, Li S, Jia J, Sun C, Cui E, Xu Y, et al.** FT-NIR combined with machine learning was used to rapidly detect the adulteration of *Pericarpium Citri Reticulatae* (Chenpi) and predict the adulteration concentration. Food Chem X. 2024;24:101798. Available from: https://doi.org/10.1016/j.fochx.2024.101798