POLITECNICO DI TORINO
Repository ISTITUZIONALE

Enhancing lettuce classification: Optimizing spectral wavelength selection via CCARS and PLS-DA

Original
Enhancing lettuce classification: Optimizing spectral wavelength selection via CCARS and PLS-DA / Dilillo, N., Sanna, A., Belcore, E., Smith, K., Piras, M., Montrucchio, B., Ferrero, R.. - In: SMART AGRICULTURAL TECHNOLOGY. - ISSN 2772-3755. - 11:(2025). [10.1016/j.atech.2025.100962]

Availability:
This version is available at: 11583/3000931 since: 2025-06-16T14:38:13Z

Publisher:
Elsevier

Published
DOI:10.1016/j.atech.2025.100962

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

19 November 2026

Smart Agricultural Technology 11 (2025) 100962

Contents lists available at ScienceDirect

Smart Agricultural Technology

journal homepage: www.journals.elsevier.com/smart-agricultural-technology

Enhancing lettuce classification: Optimizing spectral wavelength selection via CCARS and PLS-DA

Nicola Dilillo [a,*], Andrea Sanna [a], Elena Belcore [b], Kyra Smith [b], Marco Piras [b], Bartolomeo Montrucchio [a], Renato Ferrero [a]

[a] Politecnico di Torino - DAUIN, Corso Duca degli Abruzzi 24, Turin, 10129, IT, Italy
[b] Politecnico di Torino - DIATI, Corso Duca degli Abruzzi 24, Turin, 10129, IT, Italy

ARTICLE INFO

Dataset link: Enhancing Lettuce Classification: Optimizing Spectral Wavelength Selection via CCARS and PLS-DA (Original data)

Keywords:
Wavelength selection
Vis-NIR spectroscopy
PLS-DA
Leaf analysis
Chemometrics
CARS

ABSTRACT

Spectroscopy is a valuable tool for analyzing the inside of plants. In this field, plant health is evaluated through light analysis, specifically by examining wavelengths beyond the visible spectrum, making it essential to select the most representative wavelength. The Competitive Adaptive Reweighted Sampling (CARS) algorithm has been applied efficiently in the literature to select the best variables in several applications, including agricultural monitoring, nutrient analysis, and chemometrics. This study presents the Calibrated CARS (CCARS) algorithm, an extension of CARS, alongside the Partial Least Square Discriminant Analysis (PLS-DA) model. The algorithm is developed to identify critical informative wavelengths of a spectral dataset of lettuce to facilitate the creation of streamlined and efficient models for lettuce health classification. While effective with spectral data, the PLS-DA models tend to overfit, and to address this problem a rigorous systematic evaluation approach is employed. Permutation tests are conducted to verify the model's robustness, while learning curve analyses ensure the model's capacity to generalize data. With this comprehensive evaluation method, confidence in the robustness of the PLS-DA models is instilled, ensuring model stability, which is achieved thanks to the CCARS algorithm instead of the original version. The results demonstrate that using CCARS with 3 or 4 PLS components and only 30 or 19 selected wavelengths reduces the number of variables by 97%, without sacrificing accuracy, and with a statistically significant robust model.

1. Introduction

One of the primary challenges in precision agriculture is accurately quantifying plant vigor to detect health changes. According to the International Society of Precision Agriculture (ISPA), "Precision Agriculture is a management strategy that gathers, processes and analyzes temporal, spatial and individual plant and animal data and combines it with other information to support management decisions according to estimated variability for improved resource use efficiency, productivity, quality, profitability and sustainability of agricultural production".[1] This allows for targeted and timely interventions, reducing the need for extended preventive measures. Indeed, precision agriculture seeks to maximize yield production while minimizing resource inputs, making it an essential approach to addressing global food security and environmental sustainability [1].

The response of vegetation to visible and near-infrared light (Vis-NIR) has been widely utilized in agriculture applications, ranging from satellite imagery for field mapping and soil characterization to the creation of prescription maps at the field level using Uncrewed Aerial Systems (UASs) [2,3], and even to non-destructive measurements for defining the organoleptic characteristics of agricultural products [4–6]. Most of these applications are grounded in the fact that plant species may exhibit slightly different responses in the infrared region influenced by the composition of the foliar pigment and the concentration and the characteristics of anatomical plant cells [7]. Indeed, photosynthetic pigments are characterized by strong absorption in the red region of the electromagnetic spectrum and weak absorption in the NIR. This is due to chlorophyll, which, within chloroplasts, triggers the photosynthetic process. Even though all wavelengths between 400 and 700 nm affect the photosynthetic proteins, specifically, two protein complexes are known

* Corresponding author.
E-mail address: nicola.dilillo@polito.it (N. Dilillo).
[1] https://www.ispag.org/about/definition.

https://doi.org/10.1016/j.atech.2025.100962
Received 27 January 2025; Received in revised form 24 March 2025; Accepted 18 April 2025
Available online 24 April 2025
2772-3755/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

for the highest absorption peaks at specific wavelengths: photosystem (PS) II absorbs optimally around 680 nm, while PSI absorbs optimally around 700 nm, with peak absorption shifting to different wavelengths depending on several variables, including the type of organism and environmental conditions [7–9]. The most significant absorption peaks and reflectance levels are identified and utilized to design *multispectral* and *hyperspectral* imaging sensors, incorporating filters for the most relevant wavelengths [10].

Multispectral data employ a limited number of broad wavelength bands —typically between three and ten— while hyperspectral ones capture data in a much larger number of narrow, contiguous bands, often numbering in the hundreds. This higher spectral resolution in hyperspectral systems enables a more detailed analysis of material properties and subtle differences in composition [11,12], allowing for a more precise differentiation of vegetation traits, improving the ability to discriminate between different species, physiological stress conditions, and biochemical variation [13]. However, this wealth of information has some limitations, including a generally lower spatial resolution than multispectral sensors, as increasing spectral resolution often reduces the ability to capture fine details at the ground level. A specific challenge of hyperspectral imaging is the need for accurate coregistration of spectral bands, as any misalignment between images acquired at different wavelengths can compromise data analysis and classification [14]. Hyperspectral sensing generates significantly larger data volumes compared to multispectral ones, increasing the requirements for storage and data transmission [11]. This greater data load results in longer processing times and the need for advanced computational infrastructure for image analysis and interpretation. In contrast, multispectral data are more manageable and allow for faster analysis, making them suitable for operational systems with limited resources [13]. However, the limited number of bands may not be sufficient to capture the characteristic features of vegetation or specific phenomena.

Point spectroscopy offers a compelling alternative to multispectral and hyperspectral imaging by gathering spectral data from a narrowly defined area. This focused method delivers superior spectral resolution and sensitivity compared to hyperspectral approaches, enabling the detection of subtle features and compositional variations. Additionally, its targeted nature enhances the signal-to-noise ratio and simplifies calibration, making it exceptionally well-suited for precise chemical identification and on-site measurements [15,16]. In contrast to imaging sensors that depend on natural electromagnetic signals, spectrometers measure spectral responses at a single point using an artificial light source with a consistent, controlled emission. This controlled setup facilitates accurate spectral characterization [17].

In this context, accurately identifying significant spectral bands and their variations under stress conditions is crucial for precision agriculture, where effective classification enables targeted and timely interventions proportional to the stress level. Although band selection in classification models reduces noise and overfitting while optimizing computational efficiency, an inadequate selection may lead to the loss of critical information, potentially compromising both classification accuracy and trait characterization of the sample [18–21].

Determining critical informative wavelengths using spectrometers poses challenges, as spectral data often contains redundant or collinear information. This requires selection algorithms, such as *Competitive Adaptive Reweighted Sampling (CARS)* [22], which exploits robust classification models for handling such complexities, such as *Partial Least Squares Discriminant Analysis (PLS-DA*). Originating from the foundational principles of PLS regression [23], PLS-DA has evolved into a robust classification technique. Moreover, PLS-DA has underscored the effectiveness of integrating spectroscopy with chemometric methods in agriculture [24]. The CARS method has shown notable success in identifying the most informative wavelengths of spectral signals across various fields, including agriculture and general spectroscopy. Some notable examples are briefly reviewed in the following.

Xiong et al. [25] explored CARS for wavelength selection in Vis-NIR spectroscopy to estimate potassium concentrations in lettuce. Their study demonstrated that integrating CARS with PLS models enhanced predictive capabilities, confirming its utility in agricultural monitoring and nutrient analysis. Similarly, Zhang et al. [26] employed CARS to improve the predictive performance of Vis-NIR spectroscopy for assessing potassium concentrations in lettuce. These findings collectively highlight the effectiveness of CARS in refining spectroscopic models for agricultural applications. Jiang et al. [27] compared CARS with a stability-enhanced variant, stability CARS (SCARS), for identifying key wavelengths in Fourier Transform NIR spectral data—spectral measurements that capture the molecular vibration signatures corresponding to chemical compositions in samples. SCARS improved performance by reducing the number of selected variables while maintaining high model accuracy, underscoring the importance of precise wavelength selection in applications like fermentation monitoring. Yuan et al. [28] applied CARS and the Successive Projections Algorithm (SPA) to select characteristic wavelengths for discriminating abdominal tissues in rabbits. Their results highlighted the significance of wavelength optimization in enhancing spectral discrimination in biological applications. Wang et al. [29] combined CARS with SPA to optimize wavelength selection for predicting moisture content in maize seeds. Their findings emphasized the importance of selecting optimal wavelengths to improve model accuracy and stability. Integrating CARS and SPA significantly reduced redundancy in spectral data, enhancing calibration and validation processes.

Following the results of these researches, CARS has been chosen in the current work as the wavelength selection algorithm to improve spectroscopic analysis, employing the PLS-DA model for lettuce plant stress classification, after the selection of the best features. Still, it was crucial to ensure the reliability and practical usability of the developed classification model. The primary aim of this study, beyond demonstrating the efficiency of the algorithm, was to conduct an in-depth evaluation of the model to assess the potential for overfitting that sometimes can be caused by the PLS-DA model. The novelties of the current paper are:

- a new version of CARS, to reduce the number of wavelengths, called *Calibrated CARS (CCARS)*;
- adopting a systematic evaluation approach to assess the robustness of PLS-DA models, after feature selection, which is missing in state-of-the-art CARS-related works, increasing data generalization and ensuring model robustness and stability.

The remainder of the article is organized as follows. Section 2 introduces the concepts of PLS-DA and CARS, instead Section 3 describes the data set generation process, while Section 4 describes the CCARS method, and the tools employed in the study. The procedures for obtaining the findings are detailed in Section 5. Section 6 provides insights and interpretations of the results, and Section 7 draws some conclusions.

2. Background

In spectral data, many wavelengths can suffer from redundant or collinear information, impairing model performance. By selecting a subset of wavelengths that deliver the most relevant information, models can become less complex, more interpretable, and achieve improved predictive performance. CARS, and the modified version CCARS, have been selected as the wavelength selection algorithm, which exploits the PLS-DA model for the classification. Additionally, preprocessing steps play a crucial role in enhancing data quality before further analysis. These three steps —classification, wavelength selection, and preprocessing— are detailed in the following.

2.1. Classification

Classification is a key focus of this research, with chemometric approaches such as *Support Vector Machines (SVM)* and PLS-DA commonly proposed in the literature.

SVMs are highly effective for handling complex, high-dimensional spectral data by modeling nonlinear relationships using kernel functions [30]. Their robustness makes them well-suited for tasks such as seed quality assessment and varietal discrimination, as demonstrated by their high classification accuracy in several studies [31].

PLS-DA is a supervised statistical technique for PLS regression elements with discriminant analysis to classify and predict outcomes based on multiple variables [32]. It reduces dimensionality while maximizing the variance explained by predefined groups. It provides an interpretable framework through dimensionality reduction, emphasizing latent variables that maximize class separation. This method is computationally efficient and excels in scenarios where the spectral differences are predominantly linear [33–35]. Together, these complementary techniques enhance the reliability and precision of plant classification based on spectroscopic signals. In this study, the PLS-DA model was employed because it plays a central role in the CARS algorithm.

While PLS-DA offers several advantages, it is particularly prone to overfitting, especially when applied to high-dimensional datasets [36–38]. To mitigate this risk and ensure reliable results, it is essential to carefully tune model parameters and employ rigorous validation techniques, practices that contribute to the overall statistical robustness of the analysis [36]. Moreover, issues such as improper selection of latent components, suboptimal distribution of training samples, and the use of non-independent test sets can further exacerbate overfitting and compromise model performance [39]. Nevertheless, several studies have reported successful outcomes using PLS-DA, demonstrating its potential when appropriately applied [40,41].

2.2. Wavelength selection

CARS is a wavelength selection algorithm based on the evolutionary principle of *survival of the fittest* [22]. In this approach, wavelengths are assessed using PLS-DA. The least useful wavelengths, or variables, are eliminated, while the more relevant ones are preserved for subsequent iterations.

The whole process is summarized in Fig. 1: the algorithm consists of R iterations to improve precision, and each iteration is structured as described in the following steps.

- *Observation Random Sampling.* In the beginning, to obtain a significant selection of variables and to ensure that the selected wavelengths are less dependent on the training set, 80% of the observations are randomly sampled for the following steps.
- *Variable Evaluation.* The chosen observations and wavelengths selected from the previous iterations (at the beginning, all wavelengths are chosen) are used to construct the PLS-DA model. Once the model is developed, the regression coefficients β_i, associated with each wavelength, are used to calculate the weight w_i associated, as shown in (1). The algorithm discards the associated variable when a weight is set to zero.

$$w_i = \frac{|\beta_i|}{\sum_{i=1}^{p} |\beta_i|}, \qquad \text{for} \quad i = 1, \dots, p. \tag{1}$$

- *Number of Variable Selection.* The percentage r_i of survived variables decreases exponentially at each iteration, following the function $r_i = \alpha e^{-ki}$ based on the following parameters.

$$\alpha = \left(\frac{p}{2}\right)^{\frac{1}{N-1}}, \qquad k = \frac{\ln\left(\frac{p}{2}\right)}{N-1} \tag{2}$$

The number r_i gives the upper bound of the maximum variables to keep, which are selected according to the weight w_i associated, calculated before. It can be observed in Fig. 2.B that the exponential function decreases very rapidly in the initial iterations, performing a fast selection. The reduction is slower in later iterations, resulting in a refined selection. The function's coefficients, α and l described in (2), are determined using two conditions: at the first iteration, all variables must be used (i.e., $r_0 = 1$), and only two variables must remain at the last one (i.e., $r_N = \frac{2}{N}$).

- *Adaptive Reweighted Sampling (ARS).* The remaining wavelengths are resampled according to their respective weights w_i. Variables with higher weights have a greater probability of being selected.
- *Final Accuracy.* After selecting the variables, the PLS-DA model is built using all observations along with the currently chosen wavelengths. Then, the model's accuracy is assessed using a dedicated test set. Based on the identified wavelengths, the next iteration of the analysis will begin.

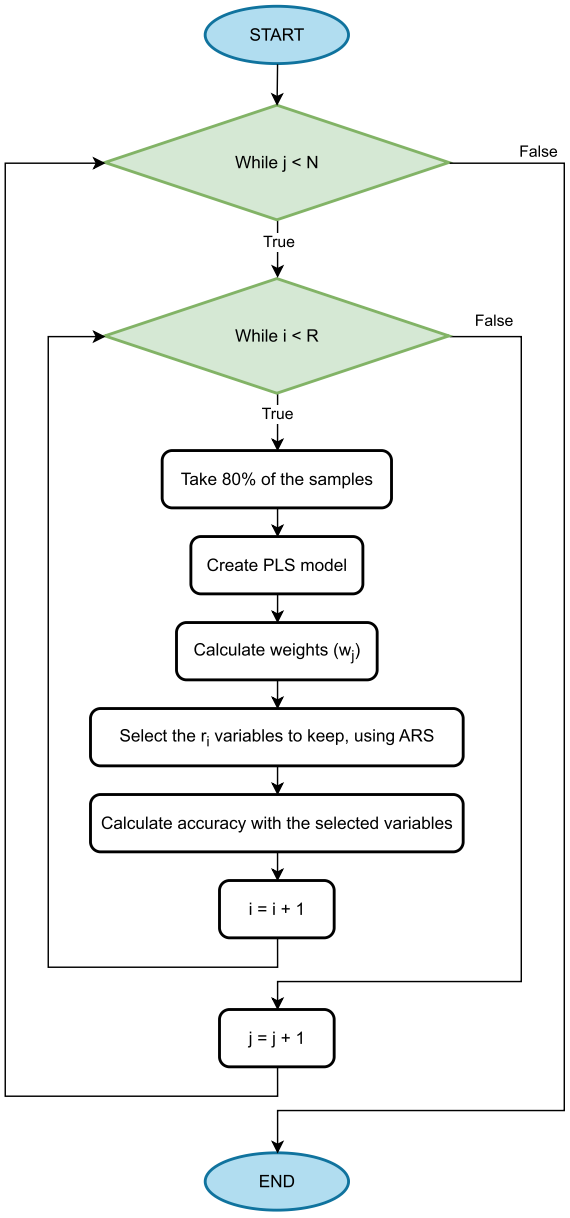


Fig. 1. CARS flowchart.

Fig. 2.A exemplifies how accuracy evolves based on the selection of variables during a single CARS run. Fig. 2.B shows the variables selected at each iteration. These graphs, which belong to the same run, demon-

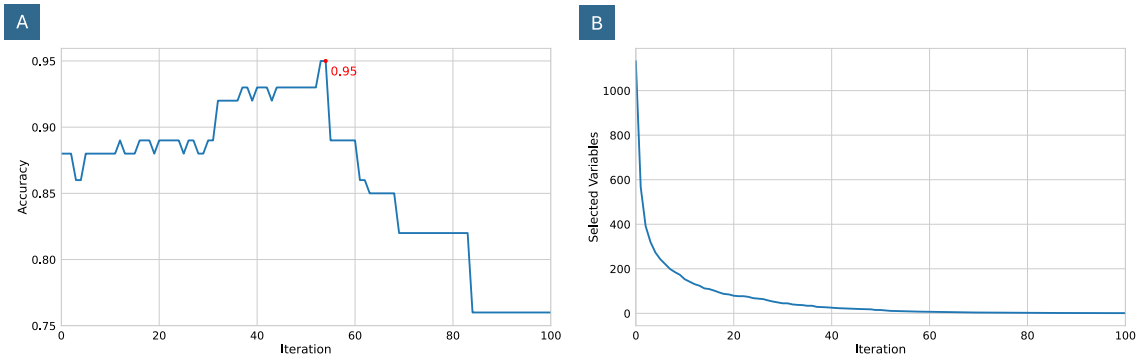


Fig. 2. Figure A represents the accuracy over different iterations in the same run. The red dot represents the best point that has been chosen for this specific run. Figure B shows the number of variables selected during the same run.

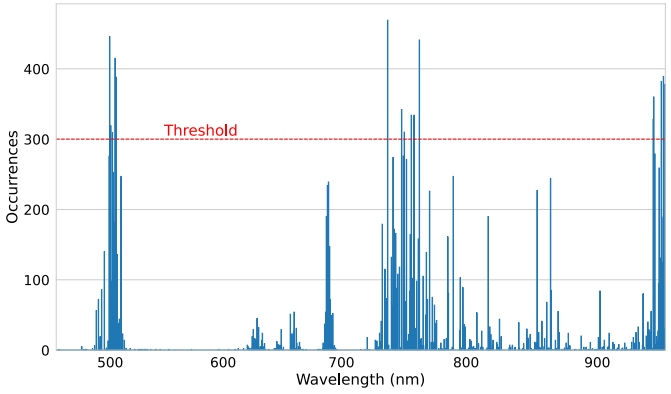


Fig. 3. Occurrence of wavelengths extraction from a whole cycle of CARS runs.

strate that optimal accuracy is not achieved at the initial stages when the maximum number of variables is included. Instead, it is reached at an intermediate phase when the most relevant variables are effectively identified.

The entire algorithm that performs R iterations is repeated for N runs, and for each of them, the iteration point that achieves the highest accuracy with the fewest wavelengths is identified and selected. The wavelengths selected at these optimal points are aggregated together and, for simplicity and better understanding, are plotted into a frequency histogram, as shown in Fig. 3. Wavelengths with the highest frequencies are highlighted as the most utilized, underscoring their importance in achieving the best model performance. After the selection of these highlight wavelengths, the PLS-DA is computed and evaluated based only on these selected features.

2.3. Preprocessing

Standard Normal Variate (SNV) is a statistical normalization technique used in the preprocessing of spectral data, especially in the fields of chemometrics, remote sensing, and other applications involving spectral signals. It corrects scattering effects and variability caused by factors such as particle size or sample thickness, which can influence signal capture [42].

SNV standardizes the spectra by centering their mean around zero and setting their standard deviation to one. If m_i represents the mean of all 1133 wavelengths for sample i and s_i represents their standard deviation, the standardization is achieved as:

$$x_{ik} = \frac{x_{ik} - m_i}{s_i} \tag{3}$$

This process eliminates systematic variations in the data intercept caused by scattering and removes irrelevant amplitude variations, enhancing the identification of meaningful variations in the analysis.

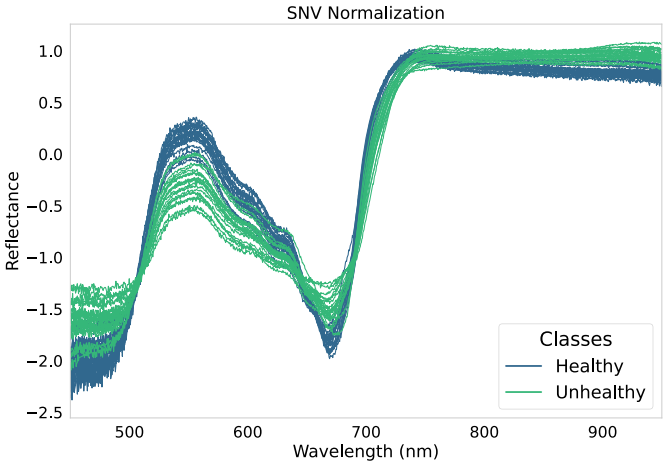


Fig. 4. SNV normalized lettuce spectral signals of Healty and Unhealty plants.

3. Materials

The Ocean HDX-XR spectrometer acquired spectral signatures characterized by a spectral resolution between 200 nm and 1100 nm. The system consists of the spectrometer connected via optical fiber to a tungsten-halogen lamp emitting between 450 nm and 950 nm on one side and a sampler with a specific holder on the other. The system is connected to a PC to control the tool and data storage. Acquisitions involved turning on the instrument and lamp, waiting for lamp stabilization, followed by black calibration (removing the light source) and white calibration using the system's reflectance calibrator. The acquisition was done in reflection mode with a 45-degree pen angle to avoid specular reflection.

The data was collected on 22 lettuce plants aged 28 days and grown in a controlled hydroponic indoor environment [43]. In this study, water stress was induced by withholding water for five consecutive days. During this period, each plant was measured multiple times daily—ranging from three to ten measurements per plant—with readings taken on different leaves and at various points on the same leaf. Importantly, the individual spectral signatures were preserved without averaging, thereby maintaining the inherent variability of each measurement. A total of 745 spectral signatures were collected, each characterized by 1133 unique features corresponding to wavelengths from 450.404 to 949.698 nm. An expert then manually labeled each sample as either Healthy or Unhealthy, dividing the dataset into two distinct classes as detailed in Table 1. The dataset used to validate the findings and to support the development of the analysis methods is available online.[2]

[2] Dataset available at https://zenodo.org/records/15019866.

Table 1
Class occurrence.

Class	Occurrence
Healthy	384
Unhealthy	361

To enhance the quality and reliability of the data, outlier samples were meticulously identified and manually removed. In addition, all the signals were normalized using SNV, the only preprocessing applied, as showed in Fig. 4. This step was crucial to minimize noise and potential biases, ensuring that the subsequent analysis remains accurate and representative of the underlying patterns.

4. Methods

This Section describes the *Calibrated CARS (CCARS)* algorithm, which introduces the following novelties with respect to the classic CARS:

1. a division of the dataset into subsets to be employed at different stages of the algorithm
2. a new criterion for the PLS-DA threshold according to the ROC curve analysis;
3. an enhanced statistical significance of the model, driven by an additional permutation test analysis;
4. a selection of a model with a better generalization capability by means of a learning curve evaluation.

4.1. Dataset division

In the CARS approach, wavelengths are selected using the entire dataset, and then the PLS-DA model is built on that same data. In contrast, the CCARS method selects wavelengths exclusively from the calibration dataset and computes the final PLS-DA model on an independent dataset using the chosen features. In detail, the dataset is divided into two distinct groups:

- *calibration*: a dedicated subset of the dataset was processed to identify the optimal wavelengths,
- *final model development*: the wavelengths selected during calibration were employed to construct the final PLS-DA classification model using a separate subset of the dataset.

Both groups are then partitioned into training and testing sets. In this study, an 80-20 split is applied to both subsets, with a 5-fold cross-validation procedure, as described in Section 2.2.
The strategy of separating feature selection from model validation is also used in synthetic data to enhance the robustness and validity of predictive models. This approach ensures that the data used to generate synthetic datasets for training are not reused in the final testing phase, thereby preventing overfitting and improving model generalization [44].

4.2. PLS-DA threshold

A critical step in transitioning PLS regression to PLS-DA is determining the appropriate threshold to classify observations into distinct groups. To make discrete group assignments, a threshold must be established. The default threshold is often set to 0.5 for binary classifications; however, this may not always yield optimal results, especially in imbalanced datasets or when the distributions of the classes overlap significantly.

In CCARS, ROC curve analysis is employed to identify the optimal threshold. By plotting the true positive rate against the false positive rate across various threshold values, the ROC curve provides a visual and quantitative means to determine a threshold that maximizes the model's sensitivity and specificity. The point on the curve closest to the top-left corner, representing perfect classification, is chosen as the optimal threshold.

4.3. Permutation test

A permutation test is a statistical method that determines whether a model's performance metrics differ significantly from what would be expected under the null hypothesis—that is when there is no true difference between the classes. Instead of relying on a single summary statistic, this method rigorously compares classification metrics by randomly shuffling the class labels and recalculating the performance measures. Repeating this process multiple times generates a distribution of performance metrics under the null hypothesis, that according to the studies from G. M. Foody, provides a robust basis for significance testing [45,46].

The core idea of the permutation test is that a stable classification model should perform poorly on data with permuted labels. By comparing the statistic derived from the actual data with the null distribution, the *p-value* is calculated. This value, which represents the likelihood of observing such a statistic under the null hypothesis, is a crucial measure in the permutation test [47]. It is determined by counting the times the statistics from the permuted data equals or exceeds the original statistics, providing a measure of statistical significance for the model result.

P-values lower than 0.05 indicate that the statistics derived from the original data are unlikely to originate from the distribution of statistics generated using permuted data, suggesting that the model captures meaningful patterns rather than random noise. P-values greater than 0.05 suggest that the model's performance on the actual labels closely mirrors its performance on permuted labels [36].

4.4. Learning curve

Learning curve is a visual tool that shows how a machine learning model's performance changes as training data increases. In the context of PLS-DA, it helps determine how well the model generalizes to new, unseen data. A typical learning curve plot includes *Training Error* line, which represents the model's error (or loss) on the training dataset, and the *Validation Error*, which instead depicts the model's error on an independent validation dataset not used for training [48].

In this study a 5-fold cross-validation has been employed. The average accuracy and the variability in performance across the folds are used to analyze the learning curve and efficiently evaluate the model [49]. If the variance is low, this means that the model performs consistently across different data splits, indicating good generalization. While, if it is high, this can mean the model's performance depends heavily on the training or validation subsets, indicating overfitting, insufficient data, or an overly complex model. This problem can be resolved using more data, simplifying the model, or adding more regularization.

Comparing training and validation curves of mean accuracy, instead, is crucial for diagnosing underfitting, overfitting, and overall model quality [50]. A good fit occurs when the training and validation curves converge at high accuracy. Underfitting appears when both curves are low and close together, suggesting that the model is too simple to capture the data's underlying patterns [48,51]. Overfitting is evident when the training curve remains high while the validation lags behind, indicating that the model is overly complex and fails to generalize [52,53]. An early plateau in the validation curve hints at data limitations, while large fluctuations or sudden drops imply high variance.

By carefully examining these learning curves, it is possible to understand whether the model is learning and generalizing appropriately, gain critical insight into whether it is overfitting or underfitting, and adopt the correct strategy to overcome the problem [48].

5. Results

In this study, multiple parameter combinations were tested across various models. Initially, CARS was applied to identify the optimal wavelength combinations, using the classical or the calibrated version, which were then used to design several PLS-DA classification models. All of these models were first analyzed by looking at the confusion matrix, then evaluated for their performance using the appropriate metrics described in Section 5.1, and in the end, they were assessed for their quality through permutation testing and learning curve analysis.

The Python code used for implementing CARS and CCARS, along with the complete analysis for evaluating the results, is freely available online to enhance transparency and reproducibility.[3]

5.1. Metrics

In classification models, different types of predictions determine their performance. A True Positive (TP) occurs when the model correctly identifies a positive case, while a True Negative (TN) means the model has accurately recognized a negative instance. On the other hand, errors can arise in two ways: a False Positive (FP) happens when the model mistakenly classifies a negative case as positive, whereas a False Negative (FN) occurs when a positive case goes undetected and is incorrectly labeled as negative.

In evaluating the performance of a classification model, several metrics provide insights into different aspects of its accuracy, precision, and overall effectiveness.

Accuracy measures the proportion of correctly predicted instances out of the total instances. It is useful when classes are balanced but can be misleading with imbalanced classes.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

Precision indicates how many of the instances predicted as positive are actually positive. It is important in scenarios where false positives are costly.

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

Recall, or sensitivity, measures the proportion of actual positives that were correctly identified. It is essential in cases where missing positive cases (false negatives) are costly.

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

F1 Score is the harmonic mean of precision and recall, balancing both metrics. It is useful when there is an uneven class distribution and when both false positives and false negatives are essential.

$$F1 = 2\frac{Precision \times Recall}{Precision + Recall} \tag{7}$$

Q^2 is a statistical measure that evaluates the predictive accuracy of a regression model. It indicates how well the model's predicted values approximate the actual observed data. In this formula, y_i represents the observed values, $\hat{y}_i$ represents the predicted values from the model, and $\bar{y}_i$ represents the mean of the observed values.

$$Q^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2} \tag{8}$$

AUC (Area Under the ROC Curve) measures the area under the ROC curve, which plots the trade-off between the recall and the False Positive Rate (FPR), which is defined in (9), across different thresholds. A higher AUC indicates better model separability between classes, with values close to 1.0 representing excellent classification performance.

$$FPR = \frac{FP}{FP + TN} \tag{9}$$

5.2. Wavelengths selection

In the CARS and CCARS algorithms, each run produces a set of optimal wavelength combinations chosen based on the highest precision and the fewest wavelengths. Some wavelengths were selected repeatedly in different runs, as illustrated in Fig. 3. A selective threshold was applied to determine which wavelengths were retained based on the frequency of occurrence for each wavelength. Only wavelengths that meet or exceed this threshold were selected in the classification model. After selecting the initial set of features, an additional manual analysis was conducted to refine the selection by removing highly near wavelengths.

Tables 3, 4 and 5, in the appendix, present these results, where a threshold was applied to the frequency graph, as in Fig. 3, to select specific wavelengths. When the field is marked as 'Hand-picking,' similar wavelengths were consolidated by removing the most similar ones in order to reach the best accuracy value.

5.3. Model performance

Based on the specific wavelengths that were identified as significant in previous steps, a PLS-DA model was generated for each combination of these wavelengths. The overall classification results are depicted from the correlation matrix in Fig. 7 and Fig. 8, while performances have been summarized in Table 2, which compares CARS and CCARS. In the row with 1133 wavelengths, no feature selection was performed—the simple PLS-DA model was built using all features. With CARS, which used 745 samples, every sample was employed for both wavelength selection and final model computation. In contrast, CCARS used only 50% of the samples (372) for feature selection, and the final model was then computed on the remaining ones.

In contrast, the metric results of the same model subjected to a permutation test have been summarized in Table 6. The dataset division to compute this curve depends on the best fold division that used the fold for validation, which gives the highest accuracy. In this case, the other folds were used as training sets, while the remaining was used as the validation set.

Results from Table 2 indicate good values, but some specific combinations exhibit unusually high performances. Given the tendency of PLS-DA models to overfit, a more in-depth analysis is essential to address this potential issue [54]. This analysis is detailed in Section 6.6, and it is based on a combination of the tabulated obtained p-values and the learning curves.

6. Discussions

6.1. Plot

Although the score plot should not be used to infer class separation, it might reveal structure (e.g., subgroups) within a class. Since the model is not forced to show this difference, this is not a result of overfitting, and thus, such information could be inferred from the PLS plot.

Fig. 5.A and Fig. 5.B show the plot of all PLS components graphs obtained using respectfully all wavelengths and only 30 selected ones, employing 3 PLS components and the CCARS method. In both cases, it is possible to see a good separation between the two classes.

Instead, Fig. 6.A and Fig. 6.B show the plot of all PLS components graphs when only 2 components and CARS are used. They were obtained using respectively all wavelengths and only 10 selected ones. In this case, the separation is less sharp when all wavelengths are used, while it is more highlighted when only the selected ones are chosen.

6.2. Wavelenghts selection

Analyzing Tables 3, 4 and 5, the number of PLS components significantly influences the selection of wavelengths. Utilizing 2 components

[3] https://github.com/nicoladilillo/CARS_PLSDA_wavelenghts_selection.

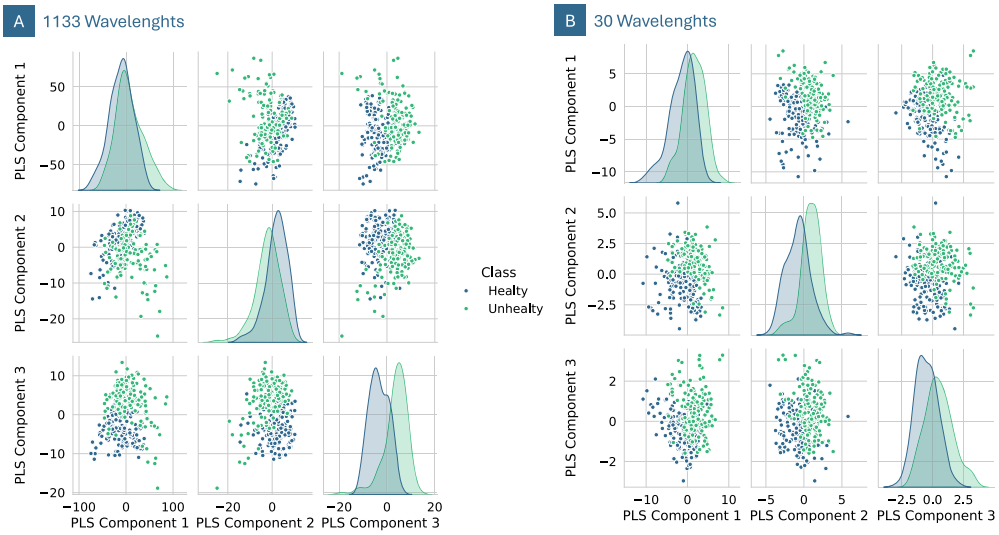


Fig. 5. PLS components graph using 3 components, CCARS method, with all wavelengths (A) and only 30 of them (B).

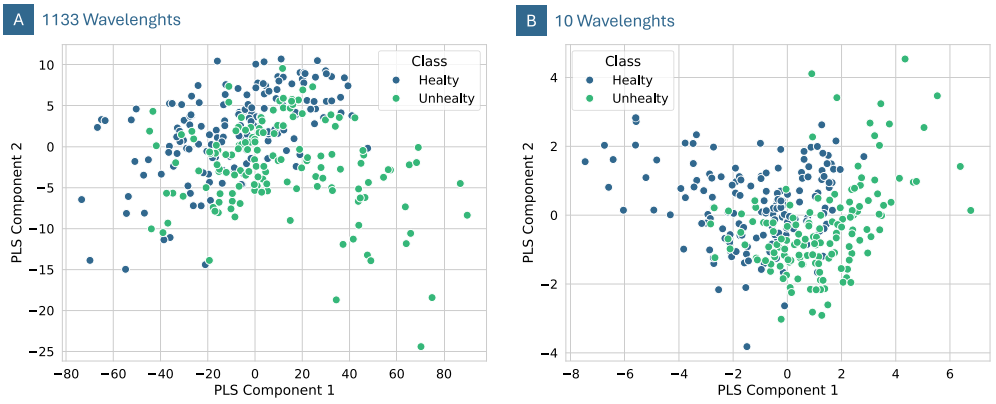


Fig. 6. PLS components graph using 2 components, CARS, and all wavelengths (A) and also 10 of them (B).

generally yields a smaller set of wavelengths, varying from 6 to 12. In contrast, increasing to 3 components allows for a broader selection, up to 30 wavelengths, instead of 4 components ranging from 19 to 34. More features could benefit complex analytical tasks where detailed differentiation is crucial.

Several wavelengths, such as 498.821, 504.19, 689.301, and 941.537, recur across various model configurations. Their selection across different settings emphasizes their potential as critical indicators in the spectral analysis classification. This observation confirms the importance of these wavelengths and suggests that they should be prioritized in future spectroscopic studies.

6.3. Confusion matrix

The analysis of both Fig. 7 and Fig. 8 underscores that the choice and range of wavelengths are pivotal in spectral classification. Deliberate selection of specific spectral subsets not only refines the level of detail captured in the spectral profiles but also plays a crucial role in balancing model complexity with classification accuracy. In Fig. 7, confusion matrices were generated using the CARS method on a dataset comprising 77 healthy and 72 unhealthy samples. In contrast, Fig. 8 depicts confusion matrices derived from the CCARS method, which utilized a reduced dataset for testing 38 healthy and 36 unhealthy samples. The smaller sample size in the CCARS method resulted from dividing the whole dataset into two halves—one for calibration and one for the final

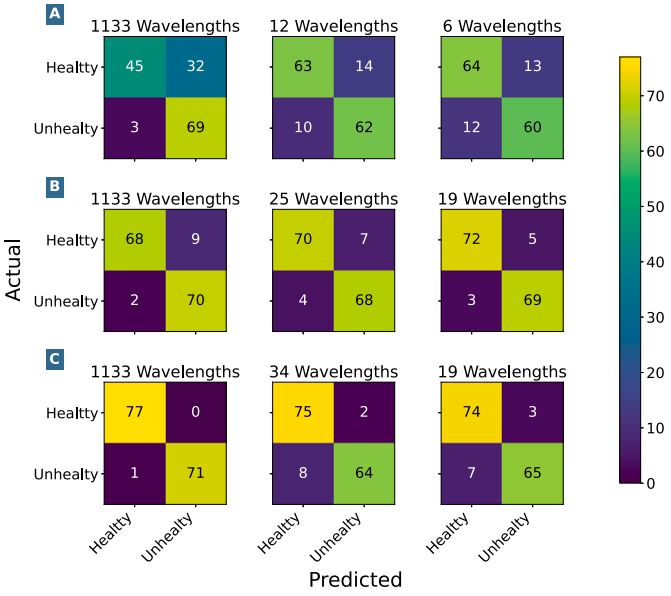


Fig. 7. These are the correlation matrix of CARS where (A) use 2 PLS components, (B) 3 and (C) 4.

Table 2
CARS and CCARS results.

	PLS Comp.	No. of Wave.	Acc.	Recall	Prec.	F1	Q^2	AUC	Cutoff
CARS	2	1133	0.77	**0.96**	0.68	0.80	0.27	0.83	0.37
		12	0.84	0.86	0.82	0.84	0.33	**0.87**	0.49
		6	0.83	0.83	0.82	0.83	0.34	**0.87**	0.49
	3	1133	0.93	**0.97**	0.89	0.93	**0.60**	0.97	0.44
		25	0.93	0.94	0.91	0.93	0.50	0.95	0.51
		19	**0.95**	0.96	0.93	**0.95**	0.55	0.96	0.51
	4	1133	**0.95**	0.92	**0.97**	0.94	0.55	**0.98**	0.57
		30	**0.95**	0.92	**0.97**	0.94	**0.60**	0.97	0.52
		20	0.91	**0.97**	0.85	0.91	0.57	0.96	0.40
CCARS	2	1133	0.76	0.81	0.72	0.76	0.25	0.81	0.43
		10	**0.85**	0.92	0.80	**0.86**	0.33	**0.87**	0.53
		6	**0.85**	0.94	0.79	**0.86**	**0.36**	**0.87**	0.54
	3	1133	0.99	0.99	1.00	0.99	0.78	1.00	0.51
		34	**0.93**	0.89	**0.97**	**0.93**	0.58	**0.97**	0.54
		19	**0.93**	0.90	0.96	**0.93**	0.59	**0.97**	0.55
	4	1133	0.96	1.00	0.92	0.96	0.67	0.99	0.46
		19	**0.93**	**0.92**	0.94	**0.93**	**0.60**	0.96	0.49

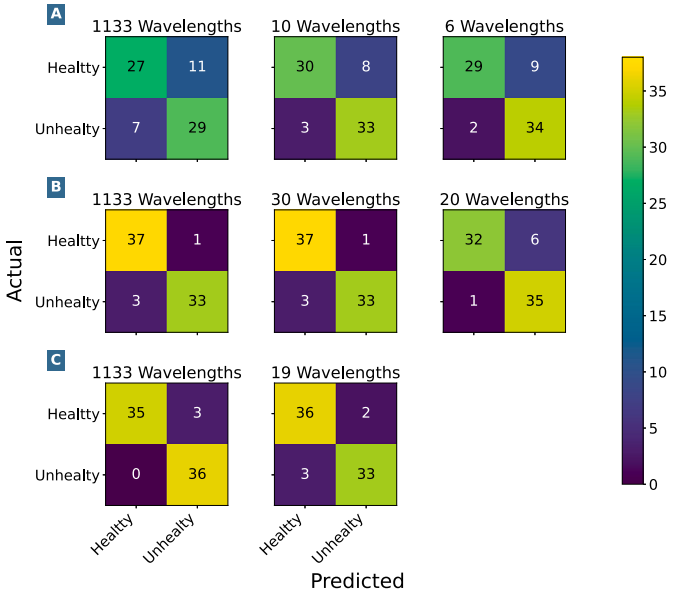


Fig. 8. These are the correlation matrix of CCARS where (A) use 2 PLS components, (B) 3 and (C) 4.

model. When only 2 PLS components were used (labeled A), both methods exhibited a significantly higher number of mispredictions compared to models employing 3 (B) or 4 (C) PLS components. In particular, when all 1133 wavelengths were used, a noticeable number of unhealthy signals were classified as healthy. In CARS, Fig. 7, the model with 3 PLS components misclassified more samples as unhealthy, while the model with 4 components showed a tendency to misclassify healthy samples as unhealthy. Notably, with 4 PLS components, the model with 1133 wavelengths mispredicted only one value, suggesting an almost perfect model that could indicate overfitting. A similar trend is observed in the CCARS method, Fig. 8; when all wavelengths and 30 are used, with 3 PLS components model, it tended to misclassify samples as unhealthy, while with 20 it misclassified much more healthy as unhealthy. For the model with 4 PLS components, when all wavelengths are used, it misclassifies just healthy samples as unhealthy, while using 19 wavelengths, the number samples wrongly classified is almost the same.

6.4. Model

In Table 2, the model's performance in multiple metrics has been analyzed while adjusting the number of components of PLS-DA and applying the calibration approach described previously.

Models with 3 and 4 PLS components typically outperform those with only 2 components, indicating that additional components help increase the evaluation metrics of the models. In some cases, CCARS affects performance, mainly when only 2 components are used. However, when using 3 or 4 components, the identical highest scores are achieved using both calibrated and original algorithms.

It is evident how, in most cases, the results are better when only a subset of all wavelengths are selected. For instance, the model utilizing 12 wavelengths, the CARS, with 2 PLS components, scores notably higher in accuracy, recall, and AUC compared to using all wavelengths under the same conditions. The only case in which the performances are better using all wavelengths is obtained using 4 PLS components and original CARS. In this case, accuracy, recall, and precision are almost perfect (1.0). However, this model has a high probability of overfitting, as shown in Section 6.6.

It can be noted that when using only 2 PLS components, CCARS makes the difference. In this case, the highest recall (0.94) and Q^2 (0.36) are obtained with calibration and employing only 6 wavelengths. Always with CCARS, only 10 wavelengths return the best precision (0.80), while the best F1 (0.86) and accuracy (0.85) are obtained when both 6 and 10 wavelengths are selected. The best value of AUC is obtained for all combinations that don't use wavelengths.

For what concerns results with 3 PLS components, it is possible to see that several combinations have the same highest accuracy (0.95), even if the best one is with the lowest number of wavelengths, selected using the classical CARS, which is also the one with the highest F1 score (0.95). The high recall value, instead, is associated with the first and the last combination, even if, in Section 6.6, the latter one will be analyzed to demonstrate that it is not a testable result. The best precision (0.97) belongs to a combination that exploits CCARS with 1133 and 30 wavelengths; the latter also has the best Q^2 (0.60), while the first has the best AUC (0.98).

From the results with 4 PLS components, it becomes evident that including all wavelengths leads to overfitting. For this reason, they will not be considered in subsequent analysis. This behavior can be deducted from the obtained results metrics, which, in most cases, tend to perfect values (e.g., 0.99 or 1.00) and analyses done in section 6.6. The best recall (0.92) and Q^2 (0.60) values are associated with the final combi-

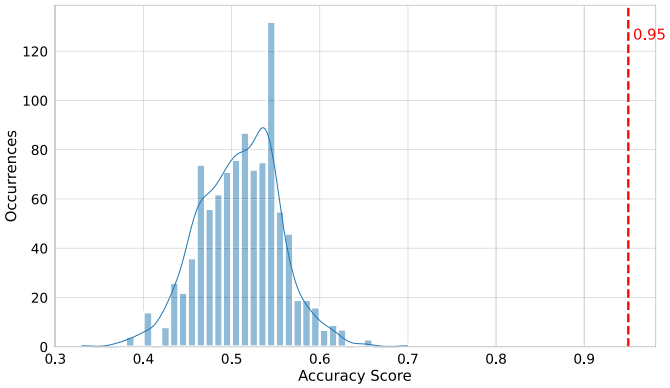


Fig. 9. Permutation test Accuracy graph performed using 3 PLS components and CCARS.

nation that used CCARS. In contrast, the combinations found using CARS achieve the highest AUC (0.97), while the model using 34 wavelengths delivers the highest precision (0.97).

With different PLS components, in all the combinations, it is possible to see how the cutoff values selected for performing classification have results that range from 0.37 to 0.57, making the selection of these parameters a pivotal step of this analysis for model selection.

At the end of this analysis, the use of CARS resulted in a reduction in measured wavelengths of about **97%** without loss in performance metrics. From these results, there is no evident difference between the values of the CARS and CCARS algorithms. However, further analyses are needed to confirm which is the best model.

6.5. Permutation test

An example of the permutation test result is represented in Fig. 9. Table 6, in the Appendix A, lists the p-values obtained from the permutation tests applied to the models developed using the wavelengths selected by the CARS or CCARS algorithm in the context of PLS-DA. These p-values indicate the statistical significance of each model's predictive ability. They are almost all null, except for the recall of 3 combinations.

The relatively higher recall p-values (0.91 and 0.80) observed in the two PLS component models, using all wavelengths, strongly suggest that these models are not statistically significant for what concerns the recall performance metric. Also, in three PLS components, when CCARS is used, and 20 wavelengths are employed, the return of a p-value of recall (0.06), which is higher than 0.05, suggests a model not statistically meaningful. On the other hand, using four PLS components suggests that it does not imply any such problem, highlighting the absence of issues.

Lower values of the PLS components may lead to higher recall p-values, but this problem seems to decrease when more components are used.

6.6. Overfitting analysis

The graphs in Fig. 10 reveal that using only two PLS components with 1133 wavelengths and CARS results in overfitting, as evidenced by a high training accuracy mean and a significantly lower validation accuracy, with substantial variances (recall: 0.96, precision: 0.68). Reducing the wavelength count to 12 or 6 narrows this performance gap, although variance remains high; the corresponding learning curves suggest improved metric balance (accuracy: 0.84 and 0.83; recall: 0.86 and 0.83; AUC: 0.87). CCARS methods with 1133 wavelengths still indicate overfitting but with reduced variance. In contrast, using 10 or 6 wavelengths mitigates overfitting, aligning training and validation accuracies more closely and achieving robust metrics (accuracy: 0.85; recall: 0.92, 0.94; AUC: 0.87), albeit with slight residual variance.

For configurations with three PLS components, as shown in Fig. 11, the learning curves for 1133 wavelengths and CARS display overfitting, with a notable discrepancy between training and validation accuracy despite balanced metrics such as accuracy (0.93), recall (0.97), and AUC (0.97). Conversely, using 25 and 19 wavelengths results in strong generalization and no signs of overfitting, although high variance persists in validation across all graphs. With CCARS, variance issues are resolved. The learning curves for 1133 wavelengths exhibit a smaller gap between training and validation accuracies towards the end of the curve, while those with 30 or 20 wavelengths will show a consistent closeness throughout. Nonetheless, while the 20-wavelength model demonstrates good generalization, a permutation test yields a p-value of 0.06 for recall, indicating a possible concern despite high recall (0.97) and relatively lower precision (0.85).

With four PLS components as depicted in Fig. 12, the learning curves for 1133 wavelengths and CARS display nearly perfect mean training accuracy. However, a significant gap in validation accuracy aligns with metrics such as accuracy (0.99), precision (1.00), and AUC (1.00), suggesting a tendency toward memorization over-generalization. Reducing the number of wavelengths to 34 or 19 results in less perfect training accuracies, though the gap remains unchanged, and the metrics are balanced (accuracy: 0.93; AUC: 0.97). Variance remains high across all graphs regardless of the wavelength count. Implementing CCARS method separation significantly reduces variance. With 1133 wavelengths, the learning curves show perfect training accuracy with no discrepancy with validation accuracy. Using 19 wavelengths under dataset separation leads to strong generalization, evidenced by converging learning curves where training and validation accuracies overlap perfectly and balanced metrics (accuracy: 0.93; recall: 0.92; AUC: 0.96).

This analysis highlights several important factors to consider when developing effective models that can generalize well.

- *Optimal PLS Component Selection.* It is crucial to choose the right number of PLS components to prevent overfitting and ensure the model performs well during classification tasks.
- *Wavelength Selection.* Reducing the number of wavelengths is important to generalize the data, but it is still important to understand if the subset selected is statistically significant.
- *Calibration Method.* Employing CCARS can mitigate overfitting by narrowing the gap between training and validation accuracies and reducing variance.

Additionally, the integration of permutation tests provides valuable insights into the model's performance, mainly when using PLS-DA, which is prone to overfitting. These tests are instrumental in assessing model robustness and refining its predictive capabilities.

6.7. Best results

As shown in Fig. 13, utilizing only 2 PLS components results in lower accuracy when a limited number of selected wavelengths is used. In contrast, when 3 or 4 PLS components were employed, a significantly more extensive set of features is needed, also increasing the overall accuracy.

When CCARS was used instead of the original version, the model's accuracy increases a bit. However, as discussed in Section 6.6, it is crucial to evaluate the robustness of the model with PLS-DA, and in this regard, CCARS performs better than the classical one. This aspect of the model is another important feature for selecting the best candidate.

The optimal configuration, considering the accuracy metric, is achieved using 3 PLS components and 30 wavelengths. The configuration with 4 components and 19 features is the best if fewer wavelengths are preferred. Alternatively, using 2 components with either 10 or 6 variables can be a viable option. However, there is not a best result, but a *Pareto Point* from which it will be possible to choose the best model according to needs.

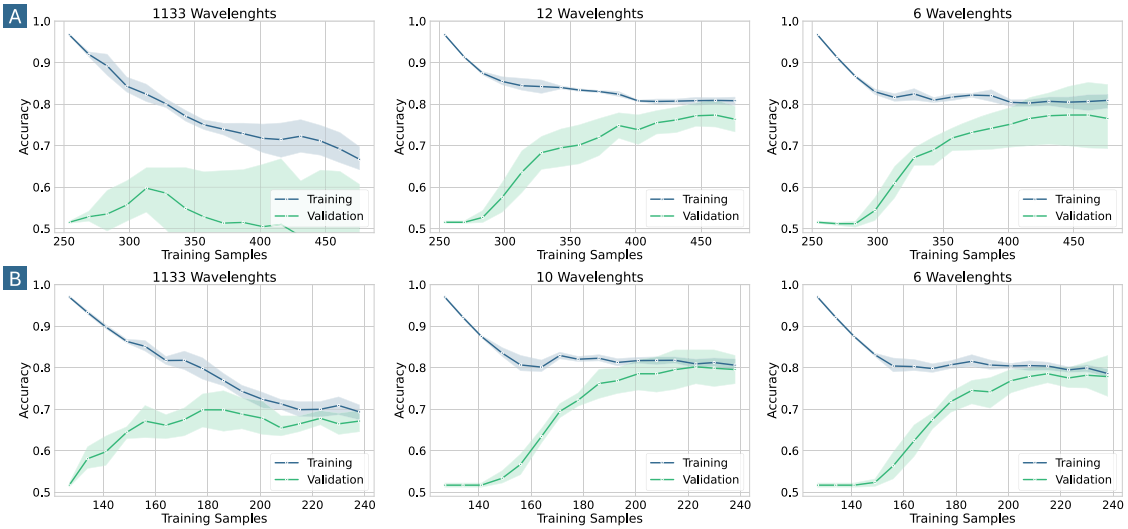


Fig. 10. Learning curve with 2 PLS components: figure A employs classical CARS, while figure B employs CCARS.

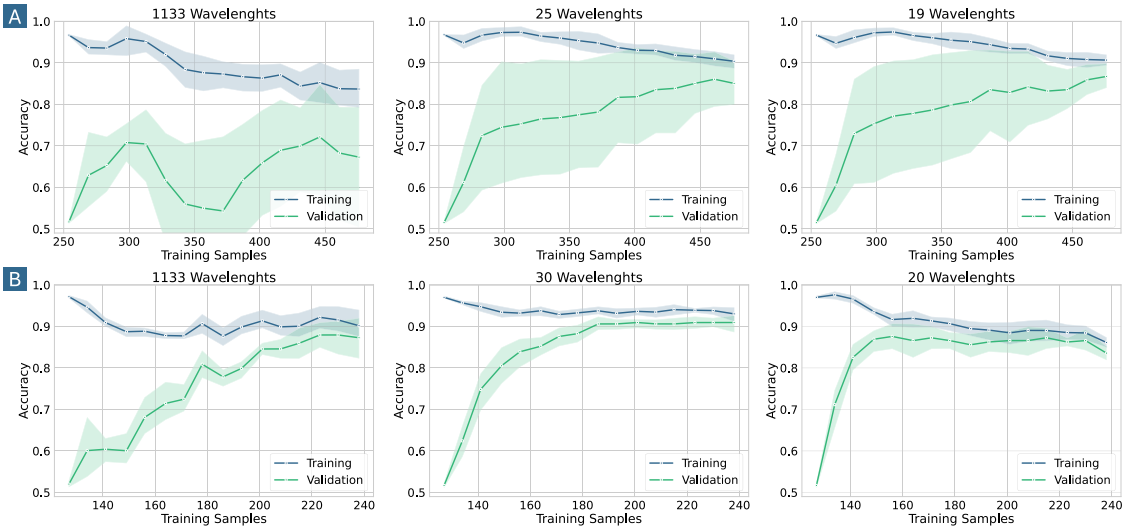


Fig. 11. Learning curve with 3 PLS components: Figure A employs classical CARS, while figure B employs CCARS.

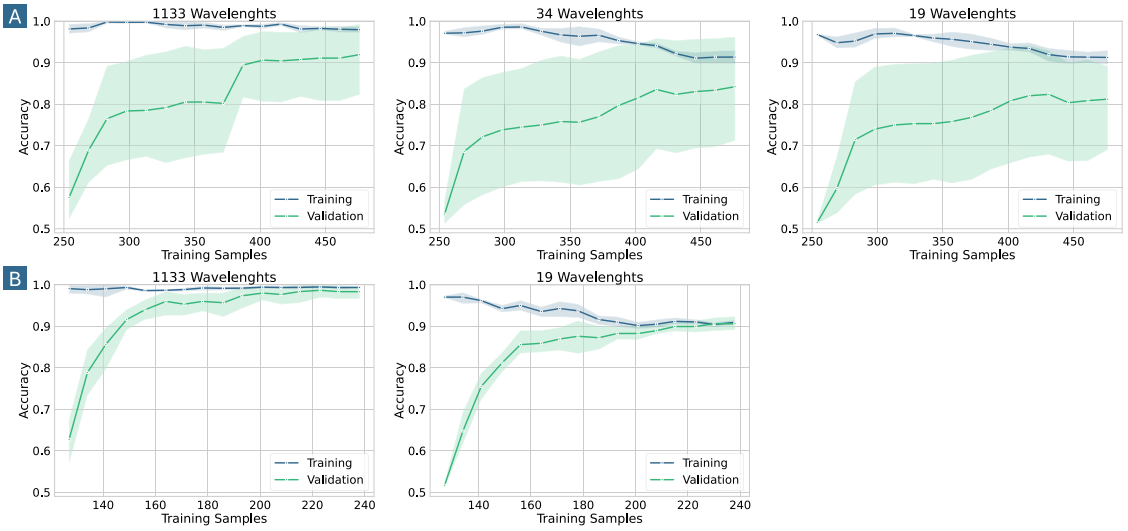


Fig. 12. Learning curve with 4 PLS components: Figure A employs classical CARS, while figure B employs CCARS.

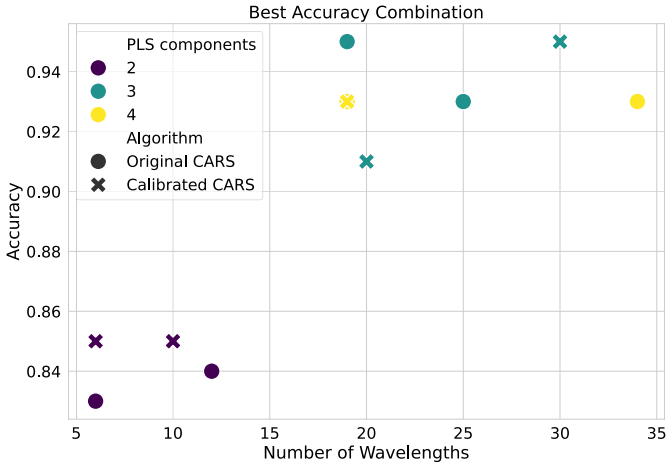


Fig. 13. Best accuracy results plot.

7. Conclusion

This research has presented the CCARS algorithm to identify critical informative wavelengths for constructing robust models. Indeed, using a subset of carefully selected variables, that account for 97% of the overall, simplifies the PLS-DA model without losing accuracy and improves its ability to generalize new data, making it more efficient and less prone to overfitting. The CCARS method proposed in this work is suitable for avoiding data generalization problems that can lead to overfitting, which can arise when PLS-DA models are employed.

The learning curve analysis provides essential insights about the model generalization, while the permutation tests confirm that these models are statistically significant. Both evaluations were missing in related work and are used to assess the robustness and effectiveness of the models. The methodological evaluation approach presented in this research provides a balance between predictive accuracy and model data generalization, which is rarely discussed when the classic CARS algorithm and PLS-DA models are used. Overall, the analyses strongly support using the CCARS algorithm and PLS-DA in constructing classification models for spectral data analysis in agriculture.

In the present study, a fixed number of PLS-DA components was maintained to isolate and evaluate the performance of the proposed algorithms under consistent conditions. Future work will include an analysis of the error associated with varying the number of components, thereby optimizing component selection and further enhancing model transparency. In addition, several algorithms for preprocessing and wavelength selection will be explored to enhance results.

CRediT authorship contribution statement

Nicola Dilillo: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Data curation, Conceptualization. **Andrea Sanna:** Writing – original draft, Software, Methodology, Formal analysis, Data curation. **Elena Belcore:** Writing – review & editing, Writing – original draft, Validation, Project administration, Investigation, Data curation. **Kyra Smith:** Writing – review & editing, Writing – original draft, Project administration, Data curation. **Marco Piras:** Writing – review & editing, Resources, Project administration. **Bartolomeo Montrucchio:** Writing – review & editing, Validation, Supervision, Data curation, Conceptualization. **Renato Ferrero:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

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

Acknowledgements

This study was carried out within the Agritech National Research Center and received funding from the European Union NextGenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4 – D.D. 1032 17/06/2022, CN00000022). The study is also part of the project NODES, which has received funding from the MUR—M4C2 1.5 of PNRR with the grant agreement no. ECS00000036. This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

Appendix A. Tables

Table 3
CARS and CCARS Wavelengths selection results using 2 PLS components.

	No. of Wave.	Threshold	Wavelengths
CARS	12	300	498.821, 499.31, 502.727, 504.19, 506.139, 689.301, 763.223, 941.537, 942.315, 946.982, 948.534, 949.698
	6	Hand-picking	498.821, 506.139, 689.301, 763.223, 941.537, 949.698
CCARS	10	300	499.31, 502.727, 504.19, 751.169, 763.223, 941.537, 942.315, 946.982, 948.534, 949.698
	6	Hand-picking	499.31, 504.19, 751.169, 763.223, 941.537, 949.698

Table 4
CARS and CCARS Wavelengths selection results using 3 PLS components.

	No. of Wave.	Threshold	Wavelengths
CARS	25	390	498.821, 499.31, 500.287, 502.727, 504.19, 505.164, 506.139, 689.301, 737.734, 742.078, 749.008, 750.305, 752.896, 756.775, 758.927, 763.223, 771.36, 790.071, 941.147, 941.537, 942.315, 945.428, 946.982, 948.534, 949.698
	19	Hand-picking	498.821, 499.31, 500.287, 502.727, 504.19, 505.164, 506.139, 689.301, 737.734, 742.078, 749.008, 750.305, 752.896, 756.775, 758.927, 790.071, 941.147, 941.537, 946.982
CARS	30	220	498.821, 499.31, 500.287, 501.751, 502.727, 504.19, 505.164, 509.545, 688.41, 689.301, 737.734, 742.078, 749.008, 750.305, 751.169, 752.896, 756.775, 758.927, 763.223, 771.36, 790.071, 854.949, 865.138, 941.147, 941.537, 942.315, 945.428, 946.982, 948.534, 949.698
	20	Hand-picking	498.821, 502.727, 504.19, 509.545, 688.41, 737.734, 742.078, 749.008, 752.896, 756.775, 758.927, 763.223, 771.36, 790.071, 854.949, 865.138, 941.147, 942.315, 945.428, 949.698

Table 5
CARS and CCARS Wavelengths selection results using 4 PLS components.

	No. of Wave.	Threshold	Wavelengths
CARS	34	450	498.821, 499.31, 500.775, 504.19, 505.164, 737.734, 742.078, 742.946, 747.278, 749.008, 750.305, 752.896, 755.914, 756.775, 758.927, 762.365, 763.223, 766.225, 768.795, 771.36, 790.071, 817.378, 854.949, 865.138, 941.147, 941.537, 942.315, 945.428, 946.205, 946.982, 947.37, 947.758, 948.534, 949.698
	19	Hand-picking	498.821, 499.31, 500.287, 502.727, 504.19, 505.164, 506.139, 689.301, 737.734, 742.078, 749.008, 750.305, 752.896, 756.775, 758.927, 790.071, 941.147, 941.537, 946.982
CCARS	19	400	499.31, 500.775, 504.19, 505.164, 737.734, 742.078, 749.008, 750.305, 751.169, 752.896, 756.775, 758.927, 763.223, 854.949, 941.147, 941.537, 946.982, 948.534, 949.698

Table 6
P-values obtained for each combination and each metric.

	PLS Comp.	No. of Wave.	Acc.	Recall	Prec.	F1	Q^2	AUC
CARS	2	1133	0.00	**0.91**	0.00	0.00	0.00	0.00
		12	0.00	0.00	0.00	0.00	0.00	0.00
		6	0.00	0.00	0.00	0.00	0.00	0.00
	3	1133	0.00	0.00	0.00	0.00	0.00	0.00
		25	0.00	0.00	0.00	0.00	0.00	0.00
		19	0.00	0.00	0.00	0.00	0.00	0.00
	4	1133	0.00	0.00	0.00	0.00	0.00	0.00
		35	0.00	0.00	0.00	0.00	0.00	0.00
		19	0.00	0.00	0.00	0.00	0.00	0.00
CCARS	2	1133	0.00	**0.80**	0.00	0.00	0.00	0.00
		10	0.00	0.00	0.00	0.00	0.00	0.00
		6	0.00	0.00	0.00	0.00	0.00	0.00
	3	1133	0.00	0.00	0.00	0.00	0.00	0.00
		30	0.00	0.00	0.00	0.00	0.00	0.00
		20	0.00	**0.06**	0.00	0.00	0.00	0.00
	4	1133	0.00	0.00	0.00	0.00	0.00	0.00
		19	0.00	0.00	0.00	0.00	0.00	0.00

Data availability

Data are published in an online available dataset at the following:

Enhancing Lettuce Classification: Optimizing Spectral Wavelength Selection via CCARS and PLS-DA (Original data) (Zenodo)

References

[1] U. Garlando, S. Calvo, M. Barezzi, A. Sanginario, P.M. Ros, D. Demarchi, Ask the plants directly: understanding plant needs using electrical impedance measurements, Comput. Electron. Agric. 193 (2022) 106707, https://doi.org/10.1016/j.compag.2022.106707.

[2] E. Belcore, S. Angeli, E. Colucci, M.A. Musci, I. Aicardi, Precision agriculture workflow, from data collection to data management using FOSS tools: an application in Northern Italy vineyard, ISPRS Int. J. Geo-Inf. 10 (4) (2021) 236, https://doi.org/10.3390/ijgi10040236.

[3] I. Blekanov, A. Molin, D. Zhang, E. Mitrofanov, O. Mitrofanova, Y. Li, Monitoring of grain crops nitrogen status from UAV multispectral images coupled with deep learning approaches, Comput. Electron. Agric. 212 (2023) 108047, https://doi.org/10.1016/j.compag.2023.108047.

[4] S. Ditcharoen, P. Sirisomboon, K. Saengprachatanarug, A. Phuphaphud, R. Rittiron, A. Terdwongworakul, C. Malai, C. Saenphon, L. Panduangnate, J. Posom, Improving the non-destructive maturity classification model for durian fruit using near-infrared spectroscopy, Artif. Intell. Agric. 7 (2023) 35–43, https://doi.org/10.1016/j.aiia.2023.02.002.

[5] X. Tian, J. Li, S. Yi, G. Jin, X. Qiu, Y. Li, Nondestructive determining the soluble solids content of citrus using near infrared transmittance technology combined with the variable selection algorithm, Artif. Intell. Agric. 4 (2020) 48–57, https://doi.org/10.1016/j.aiia.2020.05.001.

[6] A. Raghavendra, D.S. Guru, M.K. Rao, Mango internal defect detection based on optimal wavelength selection method using NIR spectroscopy, Artif. Intell. Agric. 5 (2021) 43–51, https://doi.org/10.1016/j.aiia.2021.01.005.

[7] H. Mattila, S. Khorobrykh, M. Hakala-Yatkin, V. Havurinne, I. Kuusisto, T. Antal, T. Tyystjärvi, E. Tyystjärvi, Action spectrum of the redox state of the plastoquinone pool defines its function in plant acclimation, Plant J. 104 (4) (2020) 1088–1104, https://doi.org/10.1111/tpj.14983.

[8] R. Blankenship, Molecular Mechanisms of Photosynthesis, John Wiley & Sons, 2021.

[9] S. Hogewoning, E. Wientjes, P. Douwstra, G. Trouwborst, W. van Ieperen, R. Croce, J. Harbinson, Photosynthetic quantum yield dynamics: from photosystems to leaves, Plant Cell 24 (5) (2012) 1921–1935, https://doi.org/10.1105/tpc.112.097972.

[10] H. Chandra, R.R. Nidamanuri, Object-based spectral library for knowledge-transfer-based crop detection in drone-based hyperspectral imagery, Precis. Agric. 26 (1) (2024) 6, https://doi.org/10.1007/s11119-024-10203-3.

[11] M. Govender, K. Chetty, H. Bulcock, A review of hyperspectral remote sensing and its application in vegetation and water resource studies, Water SA 33 (2) (2007) 145–151.

[12] N. Yokoya, C. Grohnfeldt, J. Chanussot, Hyperspectral and multispectral data fusion: a comparative review of the recent literature, IEEE Geosci. Remote Sens. Mag. 5 (2) (2017) 29–56.

[13] B. Lu, P.D. Dao, J. Liu, Y. He, J. Shang, Recent advances of hyperspectral imaging technology and applications in agriculture, Remote Sens. 12 (16) (2020) 2659, https://doi.org/10.3390/rs12162659.

[14] R.S. Haynes, A. Lucieer, D. Turner, E. Cimoli, Co-registration of multi-modal UAS pushbroom imaging spectroscopy and RGB imagery using optical flow, Drones 9 (2) (2025) 132, https://doi.org/10.3390/drones9020132.

[15] D.L. Pavia, G.M. Lampman, G.S. Kriz, J.R. Vyvyan, et al., Introduction to Spectroscopy, Cengage Learning Stamford, CT, 2015.

[16] J.M. Hollas, Modern Spectroscopy, John Wiley & Sons, 2004.

[17] C.J. Perera, C. Premachandra, H. Kawanaka, Comparison of light weight hyperspectral camera spectral signatures with field spectral signatures for agricultural applications, in: IEEE Int. Conference on Consumer Electronics (ICCE), 2023, pp. 1–3.

[18] B. Guo, R.I. Damper, S.R. Gunn, J.D.B. Nelson, Improving hyperspectral band selection by constructing an estimated reference map, J. Appl. Remote Sens. 8 (1) (2014) 083692, https://doi.org/10.1117/1.JRS.8.083692.

[19] X. Jia, Q. Du, Improving hyperspectral band selection by constructing an estimated reference map, J. Appl. Remote Sens. 8 (1) (2014) 083692, https://doi.org/10.1117/1.JRS.8.083692.

[20] S. Li, J. Qiu, X. Yang, H. Liu, D. Wan, Y. Zhu, A novel approach to hyperspectral band selection based on spectral shape similarity analysis and fast branch and bound search, Eng. Appl. Artif. Intell. 27 (2014) 241–250.

[21] X. Jia, Q. Du, Improving hyperspectral band selection by constructing an estimated reference map, J. Appl. Remote Sens. 8 (1) (2014) 083692, https://doi.org/10.1117/1.JRS.8.083692.

[22] H. Li, Y. Liang, Q. Xu, D. Cao, Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration, Anal. Chim. Acta 648 (1) (2009) 77–84, https://doi.org/10.1016/j.aca.2009.06.046.

[23] S. Wold, M. Sjöström, L. Eriksson, Pls-regression: a basic tool of chemometrics, Chemom. Intell. Lab. Syst. 58 (2) (2001) 109–130.

[24] C. Kumaravelu, A. Gopal, A review on the applications of near-infrared spectrometer and chemometrics for the agro-food processing industries, in: 2015 IEEE Technological Innovation in ICT for Agriculture and Rural Development (TIAR), 2015, pp. 8–12.

[25] Y. Xiong, S. Ohashi, K. Nakano, W. Jiang, K. Takizawa, K. Iijima, P. Maniwara, Application of the radial basis function neural networks to improve the nondestructive Vis/NIR spectrophotometric analysis of potassium in fresh lettuces, J. Food Eng. 298 (2021) 110417, https://doi.org/10.1016/j.jfoodeng.2020.110417.

[26] Z. Zhang, Y. Pu, Z. Wei, H. Liu, D. Zhang, B. Zhang, Z. Zhang, J. Zhao, J. Hu, Combination of interactance and transmittance modes of Vis/NIR spectroscopy improved the performance of PLS-DA model for moldy apple core, Infrared Phys. Technol. 126 (2022) 104366, https://doi.org/10.1016/j.infrared.2022.104366.

[27] H. Jiang, H. Zhang, Q. Chen, C. Mei, G. Liu, Identification of solid state fermentation degree with FT-NIR spectroscopy: comparison of wavelength variable selection methods of CARS and SCARS, Spectrochim. Acta, Part A, Mol. Biomol. Spectrosc. 149 (2015) 1–7, https://doi.org/10.1016/j.saa.2015.04.024.

[28] H. Yuan, C. Liu, H. Wang, L. Wang, L. Dai, Pls-da and vis-nir spectroscopy based discrimination of abdominal tissues of female rabbits, Spectrochim. Acta, Part A, Mol. Biomol. Spectrosc. 271 (2022) 120887, https://doi.org/10.1016/j.saa.2022.120887.

[29] Z. Wang, S. Fan, J. Wu, C. Zhang, F. Xu, X. Yang, J. Li, Application of long-wave near infrared hyperspectral imaging for determination of moisture content of single

maize seed, Spectrochim. Acta, Part A, Mol. Biomol. Spectrosc. 254 (2021) 119666, https://doi.org/10.1016/j.saa.2021.119666.
[30] O. Devos, G. Downey, L. Duponchel, Simultaneous data pre-processing and svm classification model selection based on a parallel genetic algorithm applied to spectroscopic data of olive oils, Food Chem. 148 (2014) 124–130, https://doi.org/10.1016/j.foodchem.2013.10.020.
[31] J. Zhang, X. Feng, X. Liu, Y. He, Identification of hybrid okra seeds based on near-infrared hyperspectral imaging technology, Appl. Sci. 8 (10) (2018) 1793, https://doi.org/10.3390/app8101793.
[32] L.C. Lee, C.-Y. Liong, A.A. Jemain, Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice strategies and knowledge gaps, Analyst 143 (15) (2018) 3526–3539.
[33] D. Cozzolino, An overview of the application of near infrared spectroscopy for non-destructive quality assessment of cereal grains, Appl. Spectrosc. Rev. 53 (8) (2018) 667–687, https://doi.org/10.1080/05704928.2018.1425214.
[34] W. Yin, C. Zhang, H. Zhu, Y. Zhao, Y. He, Application of near-infrared hyperspectral imaging to discriminate different geographical origins of Chinese wolfberries, PLoS ONE 12 (7) (2017) e0180534, https://doi.org/10.1371/journal.pone.0180534.
[35] S. Shi, J. Feng, L. Yang, J. Xing, G. Pan, J. Tang, J. Wang, J. Liu, C. Cao, Y. Jiang, Combination of NIR spectroscopy and algorithms for rapid differentiation between one-year and two-year stored rice, Spectrochim. Acta, Part A, Mol. Biomol. Spectrosc. 291 (2023) 122343, https://doi.org/10.1016/j.saa.2023.122343.
[36] J.A. Westerhuis, H.C. Hoefsloot, S. Smit, D.J. Vis, A.K. Smilde, E.J. van Velzen, J.P. van Duijnhoven, F.A. van Dorsten, Assessment of plsda cross validation, Metabolomics 4 (2008) 81–89.
[37] D. Ruiz-Perez, H. Guan, P. Madhivanan, K. Mathee, G. Narasimhan, So you think you can pls-da?, BMC Bioinform. 21 (1) (2020) 2, https://doi.org/10.1186/s12859-019-3310-7.
[38] K. Liu, D. Tian, H. Wang, G. Yang, Rapid classification of plastics by laser-induced breakdown spectroscopy (libs) coupled with partial least squares discrimination analysis based on variable importance (vi-pls-da), Anal. Methods 11 (9) (2019) 1174–1179.
[39] E. Szymańska, E. Saccenti, A. Smilde, J. Westerhuis, Double-check: validation of diagnostic statistics for pls-da models in metabolomics studies, Metabolomics 8 (1) (2012) 3–16, https://doi.org/10.1007/s11306-011-0388-6.
[40] M. Yulia, D. Suhandy, Authentication of Organic Lampung Robusta Ground Roasted Coffee by UV-Visible Spectroscopy and PLS-DA Method, Journal of Physics: Conference Series, vol. 1341, IOP Publishing, 2019, p. 022006.
[41] S. Chevallier, Collaborators, application of partial least squares discriminant analysis to authentication problems in the food industry, Food Chem. 95 (2) (2006) 380–389, https://doi.org/10.1016/j.foodchem.2005.07.025.
[42] R. Gautam, S. Vanga, F. Ariese, S. Umapathy, Review of multidimensional data processing approaches for Raman and infrared spectroscopy, EPJ Tech. Instrum. 2 (1) (2015) 8, https://doi.org/10.1140/epjti/s40485-015-0018-6.
[43] N. Grasso, B. Fasciolo, A.M.M. Awouda, G. Bruno, A smart aeroponic chamber: structure and architecture for an efficient production and resource management, in: N. Kumar (Ed.), Hydroponics: The Future of Sustainable Farming, Springer US, New York, NY, 2024, pp. 353–380.
[44] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, A review of feature selection methods on synthetic data, Knowl. Inf. Syst. 34 (2013) 483–519, https://doi.org/10.1007/s10115-012-0487-8.
[45] G.M. Foody, Assessing the accuracy of land cover change with imperfect ground reference data, Remote Sens. Environ. 114 (10) (2010) 2271–2285, https://doi.org/10.1016/j.rse.2010.05.003.
[46] G.M. Foody, Ground truth in classification accuracy assessment: myth and reality, Geomatics 4 (1) (2024) 81–90, https://doi.org/10.3390/geomatics4010005.
[47] G.M. Foody, Classification accuracy comparison: hypothesis tests and the use of confidence intervals in evaluations of difference, equivalence and non-inferiority, Remote Sens. Environ. 113 (8) (2009) 1658–1663.
[48] T. Hastie, R. Tibshirani, J.H. Friedman, J.H. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, vol. 2, Springer, 2009.
[49] F. Mohr, J.N. van Rijn, Fast and informative model selection using learning curve cross-validation, IEEE Trans. Pattern Anal. Mach. Intell. 45 (8) (2023) 9669–9680, https://doi.org/10.1109/TPAMI.2023.3251957.
[50] C. Perlich, Learning Curves in Machine Learning, Springer US, Boston, MA, 2010, pp. 577–580.
[51] I. Goodfellow, Y. Bengio, A. Courville, Y. Bengio, Deep Learning, vol. 1, MIT Press, Cambridge, 2016.
[52] S. Geman, E. Bienenstock, R. Doursat, Neural networks and the bias/variance dilemma, Neural Comput. 4 (1) (1992) 1–58, https://doi.org/10.1162/neco.1992.4.1.1.
[53] C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, Understanding deep learning (still) requires rethinking generalization, Commun. ACM 64 (3) (2021) 107–115.
[54] R. Rodríguez-Pérez, L. Fernández, S. Marco, Overoptimism in cross-validation when using partial least squares-discriminant analysis for omics data: a systematic study, Anal. Bioanal. Chem. 410 (23) (2018) 5981–5992, https://doi.org/10.1007/s00216-018-1217-1.