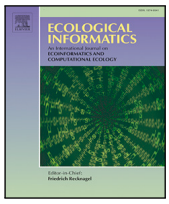
Advanced optical-based water stress detection in lettuce: a novel wavelength selection technique

Ecological Informatics 97 (2026) 103944

Contents lists available at ScienceDirect

Ecological Informatics

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



# Advanced optical-based water stress detection in lettuce: A novel wavelength selection technique

Nicola Dilillo [a,*], Gica Stefanescu Miralles [b,1], Lorenzo Comba [b], Massimo Pugliese [b,c], Maurizio Rebaudengo [a], Renato Ferrero [a]

[a] Politecnico di Torino - DAUIN, Corso Duca degli Abruzzi 24, Turin, 10129, IT, Italy
[b] Università degli Studi di Torino - DISAFA, Largo Paolo Braccini 2, Grugliasco, 10095, IT, Italy
[c] Università degli Studi di Torino - AGROINNOVA, Largo Paolo Braccini 2, Grugliasco, 10095, IT, Italy

## ARTICLE INFO

Dataset link: https://github.com/nicoladililllo/advancing_spectral_based_lettuce_classification

*Keywords:*
Vis–NIR–SWIR spectroscopy
Wavelength selection
Lettuce classification
Spectral preprocessing
Feature selection algorithms

## ABSTRACT

Water stress is a critical factor limiting agricultural productivity, particularly in controlled environments such as greenhouses, where precision irrigation is essential. Accurate and timely detection of crop water status is vital to prevent yield losses and enhance resource use efficiency. Visible–near infrared spectral reflectance data can be acquired through non-destructive leaf-clip measurements and used to monitoring water stress in lettuce-grown plants. However, given the high dimensionality of hyperspectral data, which often includes redundant or non-informative variables that can lead to collinearity, several wavelength selection algorithms have been proposed to enhance classification performance. In this study, a new method for spectral features selection, called *iGA-BOSS*, is developed, specifically conceived for water stress detection. Its effectiveness was proven on lettuce plants under different water stress conditions induced through controlled irrigation deficits relative to optimal watering regimes. Spectral reflectance measurements were collected in vivo at two growth stages and analyzed using advanced machine learning techniques. For comparison, the reflectance data were processed using the most relevant algorithm available in the literature, and the *iGA-BOSS* achieved the best performance, with an accuracy of 0.95 and an F1 score of 0.95. The selected features were used to train Partial Least Squares Discriminant Analysis (PLS-DA) for binary classification of water stress. Notably, the proposed method outperforms models trained on the complete wavelength set and competing algorithms, offering the best trade-off between classification accuracy, spectral compactness, and computational efficiency. These results demonstrate the effectiveness of combining robust preprocessing, targeted wavelength selection, and supervised classification models for monitoring spectral water stress. The identified key wavelengths may inform the development of tailored multispectral sensors, enabling scalable and efficient irrigation management in precision agriculture.

## 1. Introduction

Environmental stresses are among the primary causes of crop losses in agricultural systems, affecting both yield and quality (de la Pena and Hughes, 2007; Junaid and Gokce, 2024). Among these, water stress is particularly detrimental to crop productivity and can lead to significant yield reductions (Francini and Sebastiani, 2019). It is one of the major limiting factors for horticultural crops, especially in regions like the Mediterranean basin, where water availability is often suboptimal (Toscano et al., 2023). The alterations in physiological and biochemical parameters due to drought stress in horticulture, such as reduced leaf water content and increased membrane damage (Yavuz et al., 2021), can lead to yield reductions of up to 50%, as observed on lettuce (Lactuca sativa L.) (Galieni et al., 2015). In other horticultural crops, such as sweet basil (Ocimum basilicum L.), water stress has been shown to inhibit plant growth, decrease chlorophyll content, and cause substantial crop losses (Al-Huqail et al., 2020).

To develop resilient crop systems, it is thus crucial to detect the occurrence of water stress in advance, before shriveling signs are visible, by monitoring the crop status in an extensive and continuous way. Remote and proximal sensing approaches are key to implementing prompt and timely assessments. Applied in vegetable crops, optical sensing has

* Corresponding author.
*E-mail address:* nicola.dilillo@polito.it (N. Dilillo).
*URL:* https://nicoladililllo.github.io/ (N. Dilillo).
[1] Co-first author.

https://doi.org/10.1016/j.ecoinf.2026.103944
Received 12 March 2026; Received in revised form 16 July 2026; Accepted 17 July 2026
Available online 24 July 2026

demonstrated its potential to monitor plant stress response (Maimaitiyiming et al., 2017). Evaluating induced stresses by studying plant/leaf spectral reflectance signatures enables a broad range of data analysis techniques (Shimada et al., 2012). In addition, the non-destructive nature of this technique is advantageous with respect to traditional phenotyping methods, which are labor-intensive, time-consuming, and destructive in most cases (Kizil et al., 2012).

The potential of remote sensing techniques to detect water stress in horticulture is ascertained: Duarte-Carvajalino et al. (2021) showed that machine learning (ML) algorithms can determine influential regions in the visible (VIS) and near infrared (NIR) wavelengths to estimate water stress levels in potato crops. Also in potato, Romero et al. (2017) used a leaf-clip spectroradiometer to measure leaf reflectance in crops induced by an irrigation deficit. Correlation between spectral information and physiological parameters, such as foliar area or total water content, was found. Caturegli et al. (2020) utilized a handheld spectroradiometer to explore the effect of water stress on the spectral reflectance of bermudagrass, deriving specific vegetation indices (VIs) that could distinguish water absorption features. In tomato plants, the combination of hyperspectral imagery and partial least squares discriminant analysis (PLS-DA) allowed for the detection of abiotic drought stress at very high accuracies (Susič et al., 2018).

Considering leafy species, particularly lettuce varieties, few studies have been performed to investigate the correlation between spectral reflectance alteration and the occurrence of water stress, and they mainly focused on a small portion of the spectral range, between wavebands from 325 to 1075 nm. This is the case of Osco et al. (2019), where the hyperspectral response of water-stressed lettuce was evaluated using a handheld spectroradiometer. The obtained data were processed by a Neural Network (NN) tool, which could separate stressed and non-stressed groups with high accuracy at the end of the experiment. However, an early-stage detection of the stress was not confirmed. A similar work presented by Zhao et al. (2022) collected hyperspectral images of potted lettuce plants in outdoor conditions, and by means of Competitive Adaptive Reweighted Sampling (*CARS*) and PLS allowed a prediction of water content with a correlation coefficient of 85%. The uncontrolled outdoor growing conditions of the crops induced additional variability in the experimental campaign, which negatively affected the results. Kizil et al. (2012) presented a similar study, in which spectral response in lettuce plants under different water stress levels was measured with a handheld spectroradiometer, and an NN was used to predict the crop yield. Even if the results were promising (with $R^2$ values up to 0.92), they could not be linked to stress detection in-vivo. Kumar et al. (2021) also used hyperspectral imaging to distinguish water-stressed lettuce plants from normal irrigated ones. A set of VIs derived from the spectral data and an NN algorithm allowed a classification accuracy of 89.8% . All these preliminary and promising works adopted hyperspectral instrumentation that acquired wavelength bands only up to 950 nm. At the same time, it is known that bands in the NIR and short-wave infrared (SWIR) regions also enclose potentially valuable information for water stress detection. Up to date, only Gao et al. (2021) considered the NIR and SWIR regions in their analysis, aimed at detecting water content in lettuce through a combination of RGB images, canopy temperature, and spectral reflectance. However, results showed a correlation coefficient above 0.8 between the selected features and water content only when the models analyzed were based on multisensory inputs. Thus, additional rigorous investigations regarding the specific effects induced by water stress on spectral reflectance in lettuce leaves appear to be necessary.

Since the full Vis–NIR–SWIR spectrum contains thousands of wavelengths, many of them could become redundant to a given classification task. Using the entire spectrum may reduce model performance due to overfitting and increased computational burden (Radhesyam et al., 2024). As a result, selecting a compact and informative subset of wavelengths can become a crucial step in developing efficient and accurate classification or regression models.

Recent studies in food and agricultural spectroscopy confirm the importance of wavelength selection and feature extraction for improving NIR-based models. Xu et al. (2025) proposed GSA-iPLS and GSA-iSVM for selecting characteristic spectral intervals in rice protein prediction, showing that optimized interval selection can reduce redundancy while maintaining high regression accuracy. Dong et al. (2024) combined 1D-CNN-SVM classification with BiPLS-IBPSO and CARS-IBPSO wavelength selection for detecting illegal additives in wheat flour, demonstrating the benefit of coupling chemometrics with intelligent optimization. Similarly, Wei et al. (2025) compared CARS, SPA, CNN, LSTM, and CNN–LSTM for NIR-based beer quality assessment and authenticity verification, showing that advanced feature extraction can improve both prediction accuracy and model compactness. Although these works focus on food-quality applications, they support the use of compact spectral representations for robust hyperspectral modeling in plant-stress detection.

Indeed, several studies have demonstrated that reducing the number of wavelengths can enhance the performance of classification and regression models in hyperspectral imaging tasks. Specifically, both the *CARS* algorithm and its Calibrated variant (*CCARS*) were employed to reduce the number of wavelengths used to train a PLS-DA model for the automatic detection of lettuce health status (Dilillo et al., 2025). This research not only showed that it is possible to significantly decrease the number of features without compromising accuracy, but also emphasized the importance of feature selection in improving the model's generalizability. The authors performed a permutation test to validate these findings and analyzed learning curves, confirming that the model did not suffer from overfitting or underfitting. Similarly, Peraza-Alemán et al. (2025), used hyperspectral imaging to predict reducing sugar content in various potato genotypes. Variable selection methods such as *CARS* and interval PLS (iPLS) were used to reduce data complexity while preserving model accuracy. The models were further validated at the pixel level, ensuring spatial robustness, and visual maps were generated to effectively display the distribution of reducing sugars on the surface of potato slices. In the review presented by Wang et al. (2025), several feature selection methods have been identified as especially effective in handling NIR spectral data. In particular, models such as *CARS*, Genetic Algorithm-integrated Interval Partial Least Squares (*GA-iPLS*), and Bootstrapping Soft Shrinkage (*BOSS*) have shown promising results across multiple agricultural applications. While *GA-iPLS* and *BOSS* have individually shown promise for variable selection in spectral analysis, their combination remains underexplored. This study investigates a hybrid *GA-iPLS* + *BOSS* strategy, called *iGA-BOSS*, which leverages evolutionary selection with ensemble refinement to improve classification accuracy and model robustness.

Building upon this foundation, the present study presents an enhanced method for water stress detection on lettuce crop, based on non-destructive optical sensing within the spectral range of 400–2500 nm. The proposed workflow combines wavelength selection algorithms with robust preprocessing techniques, namely Savitzky–Golay smoothing (SG), first-derivative Savitzky–Golay preprocessing (SG1), Standard Normal Variate (SNV), and Multiplicative Scatter Correction (MSC), and evaluates multiple classification models, including Partial Least Squares Discriminant Analysis (PLS-DA), Support Vector Machine (SVM), Extreme Learning Machine (ELM), Extreme Gradient Boosting (XGBoost), and Random Forest (RF).

The remainder of this paper is organized as follows. Section 2 describes the experimental setup, the data acquisition procedure, and the preprocessing steps. Details on the stress detection methods based on classification models, the required data partitioning strategies, and the training procedures are also included in this Section. Section 3 presents the experimental outcomes, including model performance metrics, wavelength selection results, and comparative analyses. A critical interpretation of the findings, with particular emphasis on the significance of the selected features and the reliability of the classification pipelines, is also discussed. Finally, Section 4 concludes the study by summarizing the key contributions and suggesting future research directions for practical deployment and scalability.

**Table 1**
Summary of the spectral dataset characteristics used in this study.

| Attribute | Value |
|---|---|
| Total samples | 216 |
| Number of spectral wavelengths | 2101 |
| Spectral range (nm) | 400–2500 |
| Sampling interval (nm) | 1 |
| Number of classes | 2 |
| Class labels | Control ($C_i$)<br>Water Stressed ($W_i$) |
| Cycle ($i$) | 1<br>2<br>3 |

## 2. Materials and methods

### 2.1. Greenhouse experimental setup

Lettuce (*Lactuca sativa* L.) plantlets were produced at the Centre Agroinnova of the University of Turin (Grugliasco), in greenhouse at an average daily temperature of 23–24 °C, with daily watering. The cultivar selected was 'lattuga gentile' (Four, Italy). Seeds were sown into 60-holes trays filled with peat substrate. Two-week-old plantlets were then transplanted in plastic pots (2 liters), on a substrate composed of a mix of peat and perlite (50:50), moved in a greenhouse and cultivated in a closed-loop soilless cultivation system with solution recirculating for each treatment. The environmental conditions in the greenhouse were controlled during the experimental period, resulting in an average temperature of 22–25 °C, 80% relative humidity (RH), and a 14-h light/10-h dark photoperiod. The concentrations of nutrients in the nutrient solution, expressed as mM, were: 11.24 mM NO3-, 4.8 mM NH4+, 0.75 mM KH2PO4, 12.2 mM K+, 0.75 mM K2SO4, 3.1 mM CaO, 2 mM MgO, 0.012 mM Fe chelate EDTA, 2 mM SO32-, 0.2 mM B, 0.001 mM Mo, 0.15 mM Zn chelate EDTA, 0.05 mM Cu chelate EDTA, 0.25 mM Mn chelate EDTA. The EC value for the nutrient solution was 1500 $\mu S/cm$, and the average pH value was 7.06. Part of the plants were subjected to water stress (W) and compared to full (100%) irrigation control (C), corresponding to 150 ml of water/pot, provided three times per day. Water stress was caused by a reduction of 20% irrigation water in cycle 1, of 40% in cycle 2, and of 60% in cycle 3. Each cropping cycle lasted 50–60 days. Treatments started the same day the plants were transplanted. One plant was present in each pot, and 15 pots were considered for each treatment. The cultivation period was equivalent to the duration of the treatment application.

### 2.2. Data acquisition

Leaf reflectance variability was analyzed as an indicator of plant response to artificially induced water stress. To this end, leaf spectral data were collected using an RS-5400 UV–visible–infrared spectroradiometer[2] fitted with a leaf clip accessory. Spectral measurements were recorded through the DARWin™ SP software[3] operating on a portable field computer. Each acquisition produced a reflectance spectrum covering the 350–2500 nm wavelength range with a spectral resolution of 1 nm.

Due to excessive noise affecting the initial portion of the spectrum, the first 50 wavelength bands were excluded from further analysis, yielding 2101 reflectance values per spectrum. During each experimental run, 18 lettuce plants from both the control group and the water-stressed treatment were selected. Spectral measurements were collected at two phenological stages of crop development, when plants had achieved approximately 50% and 70% of their final head size, respectively.

All measurements were conducted in-vivo within the greenhouse, with leaves remaining attached to the plants throughout the process. Reflectance data were consistently recorded from an area of approximately 1 $cm^2$ located in the central region of the adaxial leaf surface. Each measurement session took place around midday and lasted approximately three hours. The final dataset comprised 216 individual spectral signatures. Each spectrum was assigned a categorical label indicating whether the corresponding plant was subjected to water stress ($W_i$) or maintained under control conditions ($C_i$), where $i$ described the cycle number. A summary of the dataset's structure and characteristics is provided in Table 1.

### 2.3. Data preprocessing

The spectral data preprocessing involved three main steps: noise filtering, data normalization, and outlier removal.

The noise filter, based on the Savitzky–Golay (SG) algorithm (Savitzky and Golay, 1964), was applied to all spectra both without derivatives (SG) and with the first derivative (SG1), to reduce high-frequency noise while preserving the shape and position of spectral features. Exploiting the first derivative, instead, is used to highlight subtle variations in the spectra, enhance the resolution of overlapping peaks, and emphasize changes in slope that are often masked in the raw signal.

Next, the spectra were normalized, considering two different techniques: SNV (Barnes et al., 1989) and MSC (Martens and Naes, 1989). SNV corrects for both additive and multiplicative scatter effects in spectroscopic data by forcing each spectrum to have zero mean and unit variance, thus improving comparability across samples and enhancing classification robustness. MSC corrects multiplicative and additive effects caused by light scattering or baseline shifts by aligning all spectra to a common baseline (in this work, that was the mean spectrum of the dataset).

To enhance the robustness of subsequent modeling steps, an outlier detection procedure was performed using PCA in combination with the Mahalanobis distance (Jolliffe and Cadima, 2016; Ghorbani, 2019). This method allows for identifying and removing samples that exhibit unusual patterns not aligned with the dataset's dominant variance structure of the dataset (Mahalanobis, 1936; Johnson and Wichern, 2002).

After projecting the data onto the PCA space, each observation $\mathbf{x}_i$ was represented by its corresponding score vector $\mathbf{t}_i$. The squared Mahalanobis distance of each sample from the multivariate mean was then computed as:

$$D_i^2 = (\mathbf{t}_i - \boldsymbol{\mu})^\top \mathbf{S}^{-1} (\mathbf{t}_i - \boldsymbol{\mu}) \tag{1}$$

where:

- $A$, equal to 10 in this case, is the number of retained principal components,
- $\mathbf{t}_i \in \mathbb{R}^A$ is the score vector of the $i$th sample in the latent space,
- $\boldsymbol{\mu} \in \mathbb{R}^A$ is the mean of the score vectors (typically zero if the data are centered),
- $\mathbf{S} \in \mathbb{R}^{A \times A}$ is the sample covariance matrix of the scores.

Under the assumption of multivariate normality, the squared Mahalanobis distances $D_i^2$ follow a Chi-square distribution with $A$ degrees of freedom. Therefore, a statistically grounded threshold for outlier detection was defined as:

$$D^2_{\text{threshold}} = \chi^2_{1-\alpha,A} = \chi^2_{0.95,10} = 18.31 \tag{2}$$

Where $\chi^2_{0.95,10}$ represents the critical value of the Chi-squared distribution with 10 degrees of freedom at the 0.95 confidence level. This

[2] https://www.alphaomega-electronics.com/en/spectrometers-spectroradiometers/4550-rs-5400-high-resolution-high-sensitivity-uv-vis-nir-spectroradiometer.html

[3] https://spectralevolution.com/product/darwin-sp-data-acquisition-software/

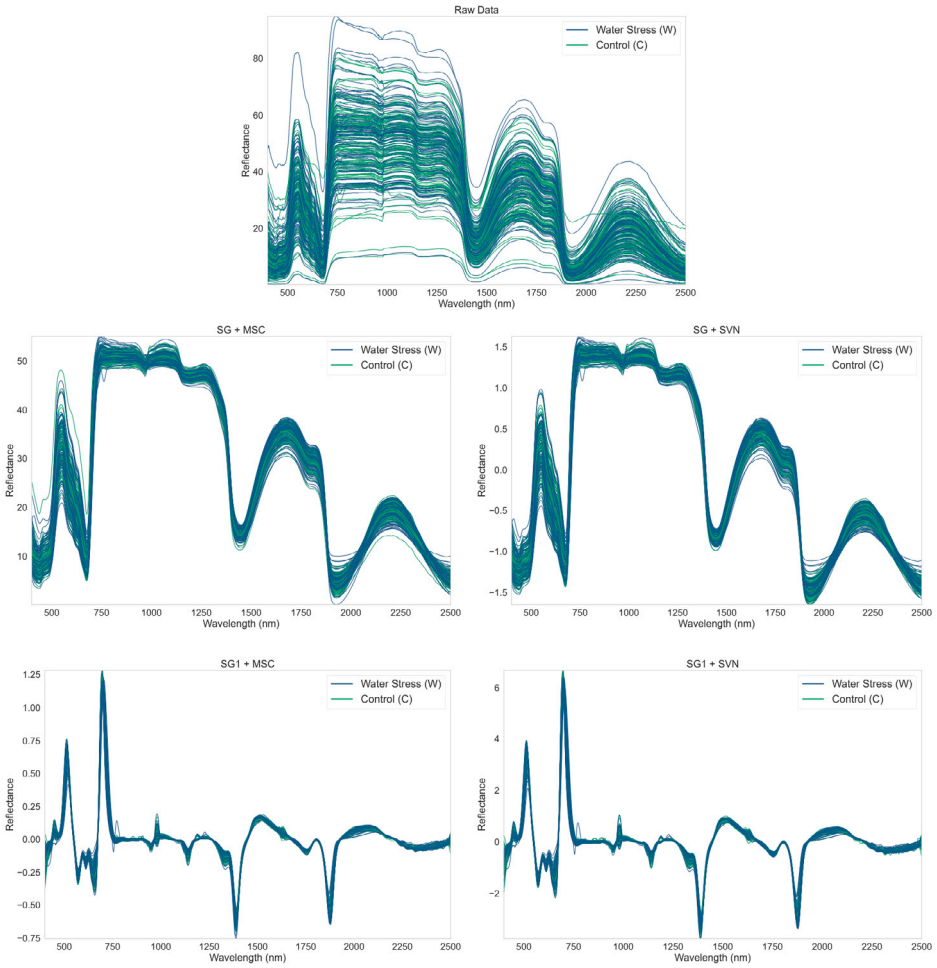



**Fig. 1.** Raw spectral signals before and after outlier removal, applying SG filtering with MSC and SNV normalization.

value remains unchanged regardless of the normalization technique applied. Observations for which $D_i^2 > 18.31$ were classified as outliers and excluded from further analysis (Garrett, 1989).

The Mahalanobis distance was computed in the latent space obtained via PCA after two different normalization strategies: combined SG and SG1 with MSC and SNV. As shown in Fig. 2, both methods have the same Chi-squared threshold, as also reported previously. Observations exceeding this threshold were identified as outliers and systematically excluded from further analysis to ensure the reliability and consistency of downstream modeling.

Fig. 1 displays the original spectral signals without any preprocessing and the final cleaned spectral signals following SG and SG1 smoothing combined with SNV and MSC normalization.

Note that the outcome of the cleaned spectral dataset and the outlier removal phase vary depending on the normalization pipeline applied. Table 2 summarizes the resulting samples after these operations, highlighting the balance between the $W_i$ and $C_i$ classes.

### 2.4. Machine learning models

In this study, five machine learning models were evaluated for automatic spectral signature classification: Partial Least Squares Discriminant Analysis (PLS-DA), Support Vector Machine (SVM), Extreme Learning Machine (ELM), Extreme Gradient Boosting (XGBoost), and Random Forest (RF). These models were selected to compare a chemometric linear classifier, a margin-based kernel method, a neural-network-based classifier, and two tree-based ensemble approaches.

**Table 2**
Train/test sample distribution by preprocessing pipeline, class (Control = $C_i$, Water Stressed = $W_i$, where $i$ indicates the cycle number).

| Preprocessing | Outlier | Class | Training samples | Testing samples |
|---|---|---|---|---|
| SG + MSC | 18 | $C_1$ | 22 | 10 |
| | | $C_2$ | 22 | 9 |
| | | $C_3$ | 21 | 10 |
| | | $W_1$ | 24 | 10 |
| | | $W_2$ | 25 | 11 |
| | | $W_3$ | 24 | 10 |
| SG + SNV | 20 | $C_1$ | 22 | 10 |
| | | $C_2$ | 22 | 9 |
| | | $C_3$ | 22 | 9 |
| | | $W_1$ | 24 | 10 |
| | | $W_2$ | 24 | 11 |
| | | $W_3$ | 23 | 10 |
| SG1 + MSC | 24 | $C_1$ | 22 | 9 |
| | | $C_2$ | 20 | 9 |
| | | $C_3$ | 21 | 9 |
| | | $W_1$ | 24 | 10 |
| | | $W_2$ | 24 | 11 |
| | | $W_3$ | 23 | 10 |
| SG1 + SNV | 23 | $C_1$ | 22 | 9 |
| | | $C_2$ | 20 | 9 |
| | | $C_3$ | 22 | 9 |
| | | $W_1$ | 24 | 10 |
| | | $W_2$ | 24 | 11 |
| | | $W_3$ | 23 | 10 |

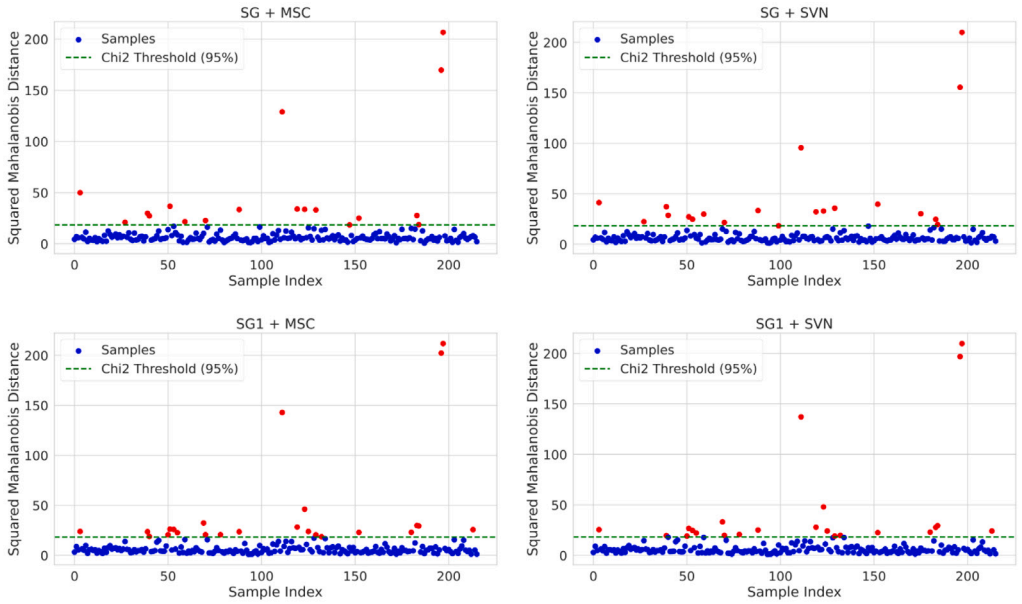



**Fig. 2.** Mahalanobis distance computed in the PCA-transformed space for SG and SG1 combined with MSC and SNV preprocessing strategies.

PLS-DA, a supervised technique that integrates elements of PLS regression with discriminant analysis, has been designed to classify and predict categorical outcomes based on multivariate data (Lee et al., 2018). It achieves dimensionality reduction while maximizing the variance between predefined classes. PLS-DA offers a computationally efficient and interpretable framework by emphasizing latent variables (LVs) that best separate these classes, particularly effective when spectral variation is primarily linear (Cozzolino, 2018; Yin et al., 2017; Shi et al., 2023).

SVMs have demonstrated strong performance in handling complex and high-dimensional spectral data, particularly due to their ability to model nonlinear relationships (Devos et al., 2014). Their robustness and adaptability make them well-suited for applications such as seed quality assessment and varietal discrimination, with several studies reporting high classification accuracies (Zhang et al., 2018).

ELM was included as a fast neural-network-based classifier (Huang et al., 2006). It consists of a single-hidden-layer feedforward neural network in which the hidden-layer parameters are randomly assigned, while only the output weights are learned analytically. This structure allows rapid model training while retaining the ability to model nonlinear relationships in the spectral data.

XGBoost was considered a gradient-boosted decision-tree ensemble (Chen and Guestrin, 2016). It builds sequential trees, where each new tree attempts to correct the errors of the previous ones. This makes XGBoost suitable for capturing nonlinear relationships and interactions among spectral variables.

Random Forest was used as a bagging-based ensemble classifier composed of multiple decision trees (Breiman, 2001). Each tree is trained on a bootstrap sample of the training data, and random subsets of variables are considered at each split. This strategy reduces overfitting and provides a robust nonlinear baseline for classification.

After data normalization and outlier removal, the datasets were divided into training and test sets at a 70:30 ratio. The sample set was divided using random stratified sampling to preserve the class distribution in both sets and ensure a balanced division. A 5-fold cross-validation scheme was used within the training set to optimize model parameters. For PLS-DA, the optimization included the number of latent variables and, where applicable, wavelength selection using *CARS*, *GA-iPLS*, *BOSS*, and the proposed *iGA-BOSS* method. SVM, ELM, XGBoost, and Random Forest were instead evaluated as full-spectrum classifiers under the same preprocessing combinations. Fig. 3 shows and summarizes the pipeline used to train and test all models and the different wavelength selection algorithms under study.

### 2.4.1. Permutation test

For all trained and tested models, permutation tests were included in the evaluation phase. A permutation test is a non-parametric statistical method used to determine whether a model's performance significantly differs from what would be expected under the null hypothesis, namely, the scenario where there is no real distinction between the classes. Instead of relying on a single performance indicator, the permutation test thoroughly evaluates classification metrics by randomly reassigning class labels and recalculating the model's performance for each permutation. The class labels are randomly shuffled only within the training set, while the original ordering is preserved for the test set. Class labels are shuffled for the training data and kept the same for the testing ones (Good, 2005; Hastie et al., 2009). This process is repeated many times to create a distribution of performance metrics based on the null hypothesis, which serves as a solid basis for assessing the statistical significance of the analyzed model (Foody, 2010, 2024).

The core idea of this method is that a reliable classification model should show decreased performance when trained or evaluated on datasets with randomly permuted labels. By comparing the performance metric obtained from the original data to the distribution created through permutations, one can calculate the *p*-value, a crucial indicator in this testing framework (Sedgwick, 2014). The *p*-value quantifies the probability of observing a performance metric as extreme as, or more extreme than, that of the original data, assuming the null hypothesis is true (Foody, 2009). It is calculated by determining the proportion of permuted instances where the resulting metric equals or surpasses the metric from the original dataset.

A *p*-value below the conventional threshold of 0.05 suggests that the observed performance is unlikely to have occurred under the null hypothesis. This implies that the model likely captures genuine patterns in the data rather than random variations. Conversely, a *p*-value greater than 0.05 indicates that the model's observed performance is statistically indistinguishable from what could be expected by chance, as reflected by its similarity to performance on datasets with randomized labels (Westerhuis et al., 2008).

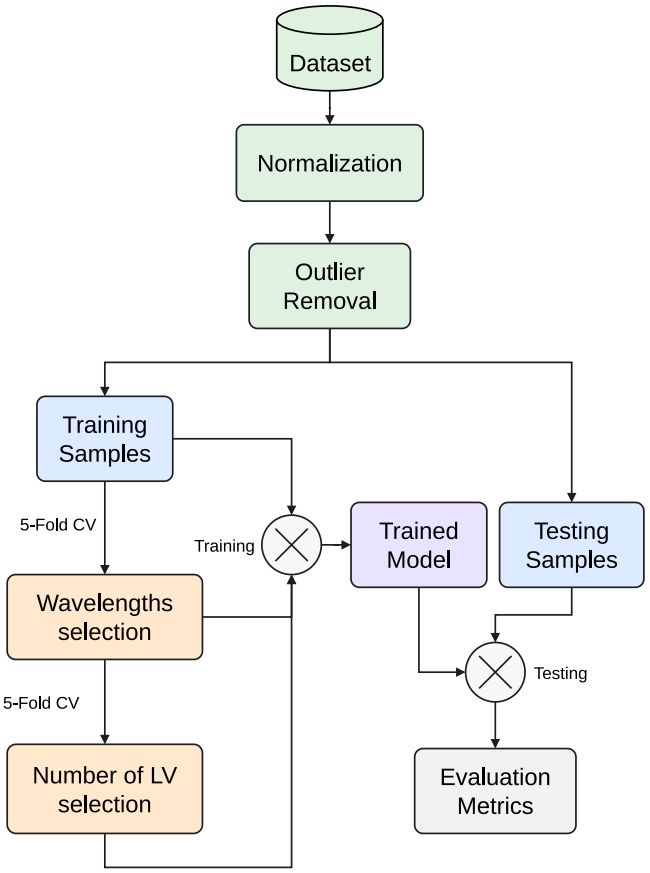



**Fig. 3.** Data division, training, and testing workflow.

### 2.4.2. Performance metrics

Model performance was assessed on the independent test set using accuracy, precision, recall, F1 score, and specificity. These indicators are commonly used for classification model evaluation and are derived from the confusion matrix (Sokolova and Lapalme, 2009; Tharwat, 2020). Accuracy measures the overall proportion of correctly classified samples. Precision indicates the proportion of samples predicted as positive that were correctly classified, whereas recall, also referred to as sensitivity, measures the proportion of actual positive samples that were correctly identified. Specificity measures the proportion of actual negative samples that were correctly classified. The F1 score represents the harmonic mean of precision and recall and provides a balanced measure when both false positives and false negatives are relevant.

The metrics are defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

$$\text{Recall (Sensitivity)} = \frac{TP}{TP + FN} \tag{5}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{6}$$

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

### 2.5. Wavelength selection algorithms

Continuous spectral data often suffer from multicollinearity, where adjacent wavelengths exhibit high correlations (Xiaobo et al., 2010b). This redundancy can adversely affect the performance of regression and classification models. Selecting a subset of informative wavelengths through feature selection techniques helps mitigate multicollinearity, enhancing model accuracy and interpretability. Moreover, focusing on key wavelengths can inform the development of more cost-effective and efficient measurement instruments (Ma et al., 2024; Cheng et al., 2022).

In this work, three different algorithms were considered to perform the wavelength selection task:

- **Competitive Adaptive Reweighted Sampling (*CARS*)** (Li et al., 2009), which is a variable selection method consisting of an iterative selection of subsets of variables by simulating a survival-of-the-fittest process. Each iteration uses Monte Carlo sampling and regression coefficients to evaluate the importance of variables, gradually eliminating the least informative ones. It efficiently reduces redundant variables by dynamically updating weights and progressively focusing on the most informative spectral regions.
- **Genetic Algorithm with interval PLS (*GA-iPLS*)** (Mitchell, 1998; Nørgaard et al., 2000; Xiaobo et al., 2010a). The spectrum is divided into intervals, and GA is used to search for the best combination of intervals to improve model performance. It can explore a large search space for optimal variable intervals, though computationally more intensive than CARS.
- **Bootstrapping Soft Shrinkage (*BOSS*)** (Deng et al., 2016), which uses bootstrap sampling to generate different subsets of variables and applies soft shrinkage to control their contribution. It maintains diversity in variable subsets while gradually focusing on those contributing most to prediction accuracy. It balances exploration and exploitation by combining bootstrap-based diversity with soft selection pressure, often leading to robust and stable models.

These algorithms were selected based on evidence from a recent study (Wang et al., 2025), which found that although several methods perform well in this context, *BOSS* and *GA-iPLS* yield the most favorable results. Nonetheless, *CARS* continues to be one of the most widely adopted and frequently cited techniques in the literature (Rogers et al., 2023).

Since the present work addresses a classification task, wavelength selection was performed by maximizing cross-validation accuracy (Accuracy-CV) rather than minimizing the Root Mean Square Error of Cross-Validation (RMSECV), which is commonly used in regression-oriented wavelength selection studies. Therefore, candidate wavelength subsets were selected according to their classification performance.

The main parameter settings used for the wavelength selection algorithms are reported in Table 3. For *CARS*, 500 independent runs were performed, each consisting of 100 Monte Carlo sampling iterations, with 80% of the calibration samples used at each iteration. For all wavelength selection methods, candidate wavelength subsets were evaluated using PLS-DA models with 5-fold cross-validation and a maximum of 10 latent variables. The classification threshold was set to 0.5, and model performance was optimized according to cross-validation accuracy.

### 2.6. Proposed approach: iGA-BOSS

Studies employing *GA-iPLS* have shown that this method is effective for wavelength selection in high-dimensional spectroscopic datasets, mainly because it explores combinations of spectral intervals through a global search strategy. However, *GA-iPLS* may retain a relatively large number of variables compared with other wavelength selection techniques. As shown in Table 4, based on the results reported by

**Table 3**
Main parameter settings used for the wavelength selection algorithms.

| Algorithm | Parameter | Value |
|---|---|---|
| *CARS* | Number of runs | 500 |
| *CARS* | Monte Carlo iterations per run | 100 |
| *CARS* | Monte Carlo sample fraction | 0.8 |
| *CARS* | Starting run index | 0 |
| *GA-iPLS* | Number of spectral intervals | 10 |
| *GA-iPLS* | Population size | 200 |
| *GA-iPLS* | Number of generations | 100 |
| *GA-iPLS* | Crossover probability | 0.6 |
| *GA-iPLS* | Mutation probability | 0.1 |
| *BOSS* | Number of bootstrap runs | 1000 |
| *BOSS* | Number of outer iterations | 100 |
| PLS-DA evaluation | Cross-validation folds | 5 |
| PLS-DA evaluation | Max number of latent variables | 10 |
| PLS-DA evaluation | Classification threshold | 0.5 |
| PLS-DA evaluation | Optimization criterion | Cross-validation accuracy |

Wang et al. (2025), *GA-iPLS* selected up to 377 variables in the *tobacco* dataset, with RMSEP (Root Mean Square Error of Prediction) values ranging from 0.0086 to 2.8201, achieving high prediction accuracy across different chemical and agricultural datasets.

In contrast, the *BOSS* algorithm has been identified as a competitive alternative because it is able to select a smaller number of informative wavelengths while maintaining comparable predictive performance. For instance, *BOSS* selected as few as 7 variables for the *wheat protein* dataset and achieved RMSEP values comparable to those obtained with *GA-iPLS* in most cases. However, this compact wavelength selection can come at the cost of high computational time. In the same benchmark, *BOSS* required up to 431 min for a single dataset due to its iterative bootstrap resampling and variable shrinkage steps.

To combine the strengths of these two methods while mitigating their limitations, we propose a hybrid wavelength selection algorithm named *iGA-BOSS*. The proposed approach follows a two-stage pipeline. First, *GA-iPLS* is applied to the full spectral matrix to perform a coarse but effective dimensionality reduction, retaining informative spectral intervals while discarding redundant or uninformative ones. Then, the reduced feature space is processed by *BOSS*, which performs a finer wavelength-level selection more efficiently because it operates on a reduced input space. The proposed *iGA-BOSS* method can therefore be considered a hybrid wrapper–embedded variable-selection strategy. In the first stage, *iGA-iPLS* acts as a wrapper method, selecting spectral intervals based on the calibration model's predictive performance. In the second stage, *BOSS* acts as an embedded/model-based selection method, because individual wavelengths are further selected through the modeling process, using bootstrap sampling and model-derived variable weights. Thus, the final selected subset is not determined only by interval elimination, but by a subsequent wavelength-level refinement performed by *BOSS*.

The complete workflow of the proposed *iGA-BOSS* wavelength selection strategy is summarized in Algorithm 1. The same parameter settings reported in Table 3 were used for the corresponding *GA-iPLS*, *BOSS*, and PLS-DA evaluation stages.

This hybrid strategy leverages the global interval-search capability of *GA-iPLS* and the sparsity-promoting refinement of *BOSS*. As a result, the proposed *iGA-BOSS* framework provides a practical balance between selection quality, model accuracy, and runtime efficiency. Although the present work focuses on a classification task, the computational-time comparison remains relevant for other modeling scenarios, including regression, because the main computational burden of both *GA-iPLS* and *BOSS* depends primarily on the number of candidate wavelengths and repeated model-fitting operations.

**Algorithm 1** Proposed *iGA-BOSS* wavelength selection strategy

**Require:** Spectral matrix $\mathbf{X}$, class vector $\mathbf{y}$, number of spectral intervals $K$
**Require:** GA-iPLS parameters: population size $N_p$, number of generations $G$, crossover probability $p_c$, mutation probability $p_m$
**Require:** BOSS parameters: number of bootstrap runs $B$, number of outer iterations $T$, maximum number of latent variables $A_{\max}$
**Ensure:** Final selected wavelength subset $\mathcal{W}^*$
Divide the full wavelength range into $K$ contiguous spectral intervals
Initialize a population of chromosomes, where each chromosome encodes a candidate subset of spectral intervals
**for** $g = 1$ to $G$ **do**
  **for** each chromosome $c$ in the population **do**
    Decode $c$ into the corresponding interval subset $\mathcal{I}_c$
    Extract the wavelengths belonging to $\mathcal{I}_c$
    Train a PLS-DA model using the selected wavelengths
    Evaluate the fitness of $c$ using cross-validation accuracy
  **end for**
  Select parent chromosomes according to their fitness
  Apply crossover with probability $p_c$
  Apply mutation with probability $p_m$
  Generate the population for the next generation
**end for**
Select the best interval subset $\mathcal{I}^*$ obtained by *GA-iPLS*
Define the reduced wavelength set $\mathcal{W}_{GA}$ as all wavelengths belonging to $\mathcal{I}^*$
Apply *BOSS* to the reduced spectral matrix $\mathbf{X}_{\mathcal{W}_{GA}}$
**for** $t = 1$ to $T$ **do**
  **for** $b = 1$ to $B$ **do**
    Generate bootstrap samples from $\mathbf{X}_{\mathcal{W}_{GA}}$ and $\mathbf{y}$
    Train weighted PLS-DA sub-models using candidate wavelength subsets
  **end for**
  Update wavelength weights according to the stability of the PLS-DA model coefficients
  Remove wavelengths with low weights
**end for**
Select the wavelength subset that maximizes cross-validation accuracy
Return the final selected wavelength subset $\mathcal{W}^*$

### 2.7. Computational environment

All experiments were conducted on a workstation with the following specifications:

- *CPU*: Intel Xeon E3-1245 v5 @ 3.50 GHz (4 cores/8 threads, Hyper-Threading enabled)
- *Memory*: 32 GB DDR4 RAM installed
- *Operating System*: Ubuntu 18.04.6 LTS

**Table 4**
Comparison of BOSS and GA-iPLS for selected variables, RMSEP, and computation time across various datasets, as reported in Wang et al. (2025).

| Dataset | Algorithm | Vars | RMSEP | Time (min.) |
|---|---|---|---|---|
| Beer | BOSS | 24 | 0.1535 | 9.87 |
| | GA-iPLS | 227 | 0.1286 | 6.70 |
| Corn oil | BOSS | 23 | 0.0180 | 18.27 |
| | GA-iPLS | 140 | 0.0086 | 72.06 |
| Wheat protein | BOSS | 7 | 0.1908 | 26.19 |
| | GA-iPLS | 239 | 0.2573 | 21.07 |
| Tobacco | BOSS | 24 | 0.0402 | 431.56 |
| | GA-iPLS | 377 | 0.0277 | 208.70 |
| Soy moisture | BOSS | 14 | 0.7397 | 8.12 |
| | GA-iPLS | 18 | 0.9848 | 3.99 |
| Diesel fuels | BOSS | 16 | 2.9979 | 110.19 |
| | GA-iPLS | 88 | 2.9067 | 22.68 |
| Tablets | BOSS | 19 | 0.3252 | 102.99 |
| | GA-iPLS | 98 | 0.3228 | 17.72 |
| Shootout-3 | BOSS | 15 | 2.8190 | 186.37 |
| | GA-iPLS | 182 | 2.8201 | 128.58 |

The source code developed for this work, written in Python, is open-source and freely available.[4] The repository provides implementations of the CARS, GA-iPLS, BOSS, and iGA-BOSS variable selection algorithms for classification tasks, with cross-validated classification accuracy used as the optimization criterion.

## 3. Results and discussion

All results are summarized in Tables 5 and 6. Computational time is reported only for configurations involving wavelength selection, since full-spectrum classifiers were trained directly on the preprocessed spectra without an additional wavelength-selection step. For *iGA-BOSS*, the wavelength subsets obtained from *GA-iPLS* under the corresponding preprocessing pipeline were used as input to the subsequent *BOSS* refinement stage, as listed in Appendix B.

### 3.1. General results analysis

Tables 5 and 6 report the test set classification performance obtained with the SG and SG1 preprocessing pipelines, respectively. In both tables, PLS-DA, SVM, ELM, XGBoost, and Random Forest are classification models trained using the full spectrum, composed of 2101 wavelengths. The remaining rows correspond to PLS-DA models in which the input wavelengths were selected using CARS, GA-iPLS, BOSS, and the proposed iGA-BOSS method.

Overall, wavelength selection substantially improved classification performance compared with the full-spectrum baseline models, while drastically reducing the number of input wavelengths. The performance of the full-spectrum classifiers varied considerably across preprocessing pipelines, confirming that both preprocessing and model choice strongly influence the classification of spectral signatures.

Among the full-spectrum classifiers, PLS-DA and SVM provided the strongest baseline performances. PLS-DA achieved accuracies ranging from 0.73 to 0.85, whereas SVM ranged from 0.69 to 0.83. PLS-DA outperformed SVM in three of the four preprocessing pipelines, namely SG+SNV, SG1+MSC, and SG1+SNV, confirming its suitability for high-dimensional and highly collinear spectroscopic data. SVM achieved its best result under SG+MSC preprocessing, reaching an accuracy of 0.83.

The additional nonlinear classifiers did not consistently improve the results when applied directly to the full spectral matrix. Random Forest achieved accuracies ranging from 0.59 to 0.76, with its best result obtained under SG1+MSC preprocessing. XGBoost showed limited performance, with accuracies between 0.55 and 0.66, suggesting that the high dimensionality and collinearity of the full spectral matrix limited its effectiveness in this dataset. ELM showed intermediate but inconsistent behavior, with accuracies ranging from 0.66 to 0.75, without outperforming the strongest full-spectrum baselines.

In contrast, the wavelength-selection approaches generally improved classification performance while greatly reducing the number of wavelengths. The proposed *iGA-BOSS* achieved the highest overall accuracy in most cases. Under the SG1+MSC pipeline, it reached an accuracy, precision, and F1-score of 0.95 using only 20 selected wavelengths, representing the best performance observed across all methods and preprocessing strategies. With SG+MSC preprocessing, *iGA-BOSS* achieved 0.92 accuracy and 0.93 F1-score using 25 wavelengths, again outperforming all competing algorithms within the same preprocessing pipeline.

Under the SG1+SNV pipeline, *iGA-BOSS* matched the performance of *BOSS*, with both methods reaching an accuracy of 0.93 and an F1-score of 0.94. However, *BOSS* achieved this result using only 19 wavelengths, whereas *iGA-BOSS* selected 64 wavelengths. Despite this larger subset, *iGA-BOSS* required less computational time than *BOSS*. Under SG+SNV preprocessing, *BOSS* achieved the best result, with 0.92 accuracy and 0.92 F1-score using only 18 wavelengths, while *iGA-BOSS* obtained 0.90 accuracy and 0.90 F1-score using 38 wavelengths. These results indicate that the proposed hybrid strategy provides competitive or superior classification performance across preprocessing pipelines, although BOSS remains particularly strong under SNV normalization.

Importantly, *iGA-BOSS* achieved its results using compact wavelength subsets ranging from 20 to 64 wavelengths. These subsets were slightly larger than those selected by *BOSS*, which retained only 18–19 wavelengths, but were generally smaller than or comparable to those selected by *CARS*, which retained 25–90 wavelengths. Moreover, *iGA-BOSS* showed lower computational time than *BOSS*, requiring on average approximately 1 h 54 min per run compared with approximately 3 h 12 min for *BOSS*. Although *CARS* required a similar or slightly longer runtime than *iGA-BOSS*, its classification accuracy was generally lower, ranging from 0.78 to 0.83.

When wavelength selection was applied, preprocessing had a clear effect on model performance. *BOSS* consistently benefited from SNV normalization: under SG preprocessing, accuracy increased from 0.88 with MSC to 0.92 with SNV, while under SG1 preprocessing it increased from 0.91 with MSC to 0.93 with SNV. *GA-iPLS* showed smaller changes, remaining at 0.88 under SG preprocessing and increasing from 0.90 to 0.91 under SG1 preprocessing. In contrast, *iGA-BOSS* performed best with MSC preprocessing, reaching 0.92 accuracy under SG+MSC and 0.95 under SG1+MSC, but slightly decreasing with SNV to 0.90 and 0.93, respectively.

In absolute terms, the best wavelength-selection method within each preprocessing pipeline improved the strongest full-spectrum baseline classifier by +0.09 for SG+MSC, from 0.83 with SVM to 0.92 with *iGA-BOSS*; +0.07 for SG+SNV, from 0.85 with PLS-DA to 0.92 with *BOSS*; +0.14 for SG1+MSC, from 0.81 with PLS-DA to 0.95 with *iGA-BOSS*; and +0.10 for SG1+SNV, from 0.83 with PLS-DA to 0.93 with both *BOSS* and *iGA-BOSS*. Therefore, the best or co-best performance was obtained by *iGA-BOSS* in three of the four preprocessing configurations.

A visual comparison of accuracies across preprocessing pipelines is provided in Fig. 5, highlighting the robustness and efficiency of *iGA-BOSS* relative to the competing methods.

The methods also differed substantially in computational efficiency and subset size. *GA-iPLS* was the fastest wavelength-selection method, requiring approximately 12 min per run, but it selected very large wavelength subsets, ranging from 588 to 757 wavelengths. Among compact selectors, *BOSS* was the most parsimonious, selecting only 18–19 wavelengths, and achieved the highest accuracy-per-wavelength value, reaching 0.05085 under SG+SNV. However, it was also the

[4] https://github.com/nicoladililo/advancing_spectral_based_lettuce_classification

**Table 5**
Test set classification performance using SG preprocessing combined with MSC and SNV. For feature-selection approaches, the model column reports the classifier on the first line and the wavelength-selection algorithm on the second line.

| Preproc. | Model (Algo.) | #Wav. | #LV | Acc. | Rec. | Prec. | F1 | Spec. | Acc./ Wav. | Comput. time |
|---|---|---|---|---|---|---|---|---|---|---|
| MSC | PLS-DA | 2101 | 7 | 0.73 | 0.81 | 0.71 | 0.76 | 0.66 | 0.00035 | – |
| | SVM | 2101 | – | 0.83 | 0.87 | 0.82 | 0.84 | 0.79 | 0.00040 | – |
| | ELM | 2101 | – | 0.73 | 0.81 | 0.71 | 0.76 | 0.66 | 0.00035 | – |
| | XGB | 2101 | – | 0.55 | 0.58 | 0.56 | 0.57 | 0.52 | 0.00026 | – |
| | RF | 2101 | – | 0.67 | 0.77 | 0.65 | 0.71 | 0.55 | 0.00032 | – |
| | PLS-DA (CARS) | 25 | 6 | 0.83 | 0.90 | 0.80 | 0.85 | 0.76 | 0.03333 | 2:40:43 |
| | PLS-DA (GA-iPLS) | 609 | 10 | 0.88 | 0.94 | 0.85 | 0.89 | 0.83 | 0.00145 | 0:12:07 |
| | PLS-DA (BOSS) | 18 | 10 | 0.88 | 0.90 | 0.88 | 0.89 | 0.86 | 0.04907 | 3:14:08 |
| | PLS-DA (iGA-BOSS) | 25 | 10 | 0.92 | 1.00 | 0.86 | 0.93 | 0.83 | 0.03667 | 1:52:53 |
| SNV | PLS-DA | 2101 | 9 | 0.85 | 0.81 | 0.89 | 0.85 | 0.89 | 0.00040 | – |
| | SVM | 2101 | – | 0.75 | 0.81 | 0.74 | 0.77 | 0.68 | 0.00035 | – |
| | ELM | 2101 | – | 0.75 | 0.84 | 0.72 | 0.78 | 0.64 | 0.00035 | – |
| | XGB | 2101 | – | 0.61 | 0.55 | 0.65 | 0.60 | 0.68 | 0.00029 | – |
| | RF | 2101 | – | 0.59 | 0.61 | 0.61 | 0.61 | 0.57 | 0.00028 | – |
| | PLS-DA (CARS) | 48 | 10 | 0.81 | 0.81 | 0.83 | 0.82 | 0.82 | 0.01695 | 2:37:54 |
| | PLS-DA (GA-iPLS) | 588 | 8 | 0.88 | 0.87 | 0.90 | 0.89 | 0.89 | 0.00150 | 0:11:48 |
| | PLS-DA (BOSS) | 18 | 10 | 0.92 | 0.90 | 0.93 | 0.92 | 0.93 | 0.05085 | 3:12:51 |
| | PLS-DA (iGA-BOSS) | 38 | 10 | 0.90 | 0.90 | 0.90 | 0.90 | 0.89 | 0.02364 | 1:49:19 |

**Table 6**
Test set classification performance using SG1 preprocessing combined with MSC and SNV. For feature-selection approaches, the model column reports the classifier on the first line and the wavelength-selection algorithm on the second line.

| Preproc. | Model (Algo.) | #LV | Acc. | Rec. | Prec. | F1 | Spec. | Acc./Wav. | Comput. time | |
|---|---|---|---|---|---|---|---|---|---|---|
| MSC | PLS-DA | 2101 | 8 | 0.81 | 0.87 | 0.79 | 0.83 | 0.74 | 0.00039 | – |
| | SVM | 2101 | – | 0.69 | 0.74 | 0.70 | 0.72 | 0.63 | 0.00033 | – |
| | ELM | 2101 | – | 0.66 | 0.71 | 0.67 | 0.69 | 0.59 | 0.00031 | – |
| | XGB | 2101 | – | 0.64 | 0.61 | 0.68 | 0.64 | 0.67 | 0.00030 | – |
| | RF | 2101 | – | 0.76 | 0.74 | 0.79 | 0.77 | 0.78 | 0.00036 | – |
| | PLS-DA (CARS) | 90 | 7 | 0.83 | 0.90 | 0.80 | 0.85 | 0.74 | 0.00920 | 2:34:30 |
| | PLS-DA (GA-iPLS) | 757 | 8 | 0.90 | 0.94 | 0.88 | 0.91 | 0.85 | 0.00118 | 0:12:21 |
| | PLS-DA (BOSS) | 19 | 9 | 0.91 | 0.97 | 0.88 | 0.92 | 0.85 | 0.04809 | 3:11:52 |
| | PLS-DA (iGA-BOSS) | 20 | 10 | 0.95 | 0.94 | 0.97 | 0.95 | 0.96 | 0.04741 | 2:01:37 |
| SNV | PLS-DA | 2101 | 9 | 0.83 | 0.84 | 0.84 | 0.84 | 0.81 | 0.00039 | – |
| | SVM | 2101 | – | 0.81 | 0.84 | 0.81 | 0.83 | 0.78 | 0.00039 | – |
| | ELM | 2101 | – | 0.71 | 0.71 | 0.73 | 0.72 | 0.70 | 0.00034 | – |
| | XGB | 2101 | – | 0.66 | 0.61 | 0.70 | 0.66 | 0.70 | 0.00031 | – |
| | RF | 2101 | – | 0.71 | 0.65 | 0.77 | 0.70 | 0.78 | 0.00034 | – |
| | PLS-DA (CARS) | 37 | 8 | 0.78 | 0.77 | 0.80 | 0.79 | 0.78 | 0.02097 | 2:34:14 |
| | PLS-DA (GA-iPLS) | 631 | 6 | 0.91 | 0.90 | 0.93 | 0.92 | 0.93 | 0.00145 | 0:12:07 |
| | PLS-DA (BOSS) | 19 | 6 | 0.93 | 0.94 | 0.94 | 0.94 | 0.93 | 0.04900 | 3:11:11 |
| | PLS-DA (iGA-BOSS) | 64 | 6 | 0.93 | 0.94 | 0.94 | 0.94 | 0.93 | 0.01455 | 1:51:30 |

slowest method, requiring approximately 3 h 12 min per run. *iGA-BOSS* provided the most balanced profile, achieving best or near-best accuracies with compact subsets of 20–64 wavelengths and runtimes between 1 h 49 min and 2 h 02 min. This corresponds to an approximate 41% reduction in computational time compared with *BOSS*, while maintaining comparable or superior classification performance in most preprocessing configurations.

Although *BOSS* achieved the best accuracy-per-wavelength overall, *iGA-BOSS* remained highly competitive, for example reaching 0.04741 under SG1+MSC while simultaneously delivering the highest absolute accuracy observed in the study. These findings indicate that the proposed hybrid method provides a better practical compromise between classification accuracy, spectral compactness, and computational cost than applying *BOSS* directly to the full spectral space.

These findings align with the observations discussed in Section 2.6, where the results reported by Wang et al. (2025) are considered in relation to Table 4. As observed in that study, *GA-iPLS* selects substantially larger wavelength subsets than *BOSS* but requires less computational

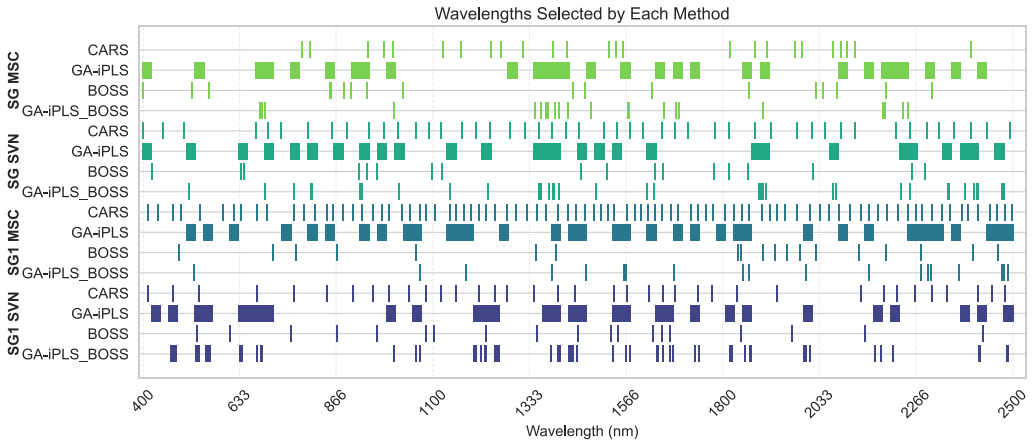



**Fig. 4.** Selected wavelengths by different feature selection methods applied to all spectral preprocessing pipelines. Vertical lines mark the positions of the selected wavelengths across the spectral range.

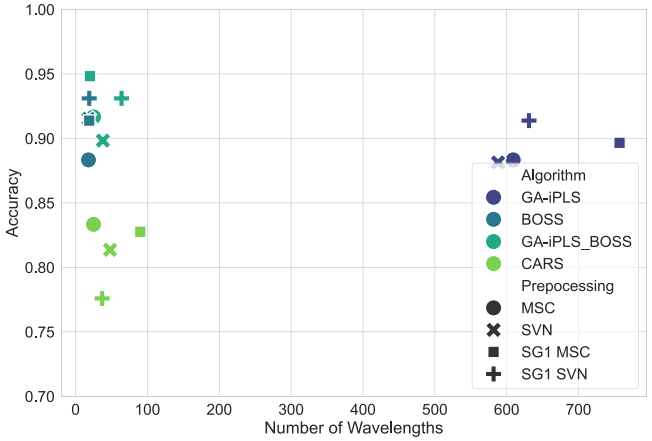



**Fig. 5.** Comparison of classification accuracy obtained on the test set across all preprocessing pipelines, full-spectrum classifiers, and wavelength-selection algorithms.

time. The main difference in the present classification context is that *BOSS* generally outperformed *GA-iPLS* in terms of classification accuracy, whereas in the regression-oriented benchmark the two methods showed more comparable predictive performance.

Finally, permutation tests were performed to assess the statistical robustness of all model configurations. As reported in Tables A.7 and A.8 in the Appendix, almost all p-values were below 0.05, confirming that the observed classification performances were statistically significant in nearly all cases. The only exceptions were Random Forest under SG+SNV and SG1+SNV preprocessing for recall, where p-values of 0.28 and 0.30 were obtained, respectively, and the RF accuracy, precision, and F1-score under SG+SNV, which also did not reach statistical significance.

### 3.2. Confusion–matrix analysis

A qualitative reading of the confusion matrices in Figs. 6, 7, 8, and 9 confirms the trends observed in Tables 5 and 6. Overall, wavelength selection reduced the number of misclassified samples compared with the full-spectrum classifiers and produced more balanced confusion matrices across the two classes. The full-spectrum baselines, namely PLS-DA, SVM, ELM, XGBoost, and Random Forest, showed preprocessing-dependent behavior, whereas the wavelength-selection methods, particularly *BOSS* and *iGA-BOSS*, generally provided lower false-positive and false-negative counts.

Under *SG+MSC* preprocessing (Fig. 6), the proposed *iGA-BOSS* achieved the best overall performance, with an accuracy of 0.92. Its confusion matrix showed no false negatives, with FN = 0, indicating maximal sensitivity for the W class. This result was obtained with only five false positives, corresponding to the misclassification of five C samples as W. Compared with the full-spectrum baselines, the improvement is clear: PLS-DA produced 10 false positives and 6 false negatives, SVM produced 6 false positives and 4 false negatives, while Random Forest produced 13 false positives and 7 false negatives. XGBoost showed weak discrimination, with 14 false positives and 13 false negatives.

Under *SG+SNV* preprocessing (Fig. 7), *BOSS* was the strongest performer, reaching an accuracy of 0.92. Its confusion matrix showed only two false positives and three false negatives, corresponding to high precision and specificity. The proposed *iGA-BOSS* produced a very similar profile, with three false positives and three false negatives. In contrast, the full-spectrum classifiers showed a higher number of errors: PLS-DA produced three false positives and six false negatives, SVM produced nine false positives and six false negatives, ELM produced eight false positives and ten false negatives, XGBoost produced nine false positives and fourteen false negatives, and Random Forest produced twelve false positives and twelve false negatives. These results indicate that SNV particularly benefited *BOSS*, helping reduce false alarms while maintaining low false-negative rates.

With *SG1+MSC* preprocessing (Fig. 8), *iGA-BOSS* delivered the best result of the entire study, reaching an accuracy of 0.95. Its confusion matrix contained only one false positive and two false negatives, showing the most favorable balance between sensitivity and specificity. This result was better than the full-spectrum classifiers: PLS-DA produced seven false positives and four false negatives, SVM produced ten false positives and eight false negatives, ELM produced eight false positives and four false negatives, XGBoost produced nine false positives and twelve false negatives, and Random Forest produced six false positives and eight false negatives. Among the wavelength-selection methods, *BOSS* also performed strongly, with four false positives and one false negative, while *GA-iPLS* produced four false positives and two false negatives. However, *iGA-BOSS* achieved the lowest false-positive count under this preprocessing pipeline, explaining its superior overall performance.

Finally, under *SG1+SNV* preprocessing (Fig. 9), *BOSS* and *iGA-BOSS* converged to identical confusion matrices, both producing two false positives and two false negatives. This explains their identical accuracy, F1-score, precision, recall, and specificity values. *GA-iPLS* also performed well, with two false positives and three false negatives. By comparison, the full-spectrum baselines produced more errors: PLS-DA and SVM each produced five false negatives, while XGBoost and Random Forest produced twelve and eleven false negatives, respectively. ELM again showed an extreme class bias, classifying almost all

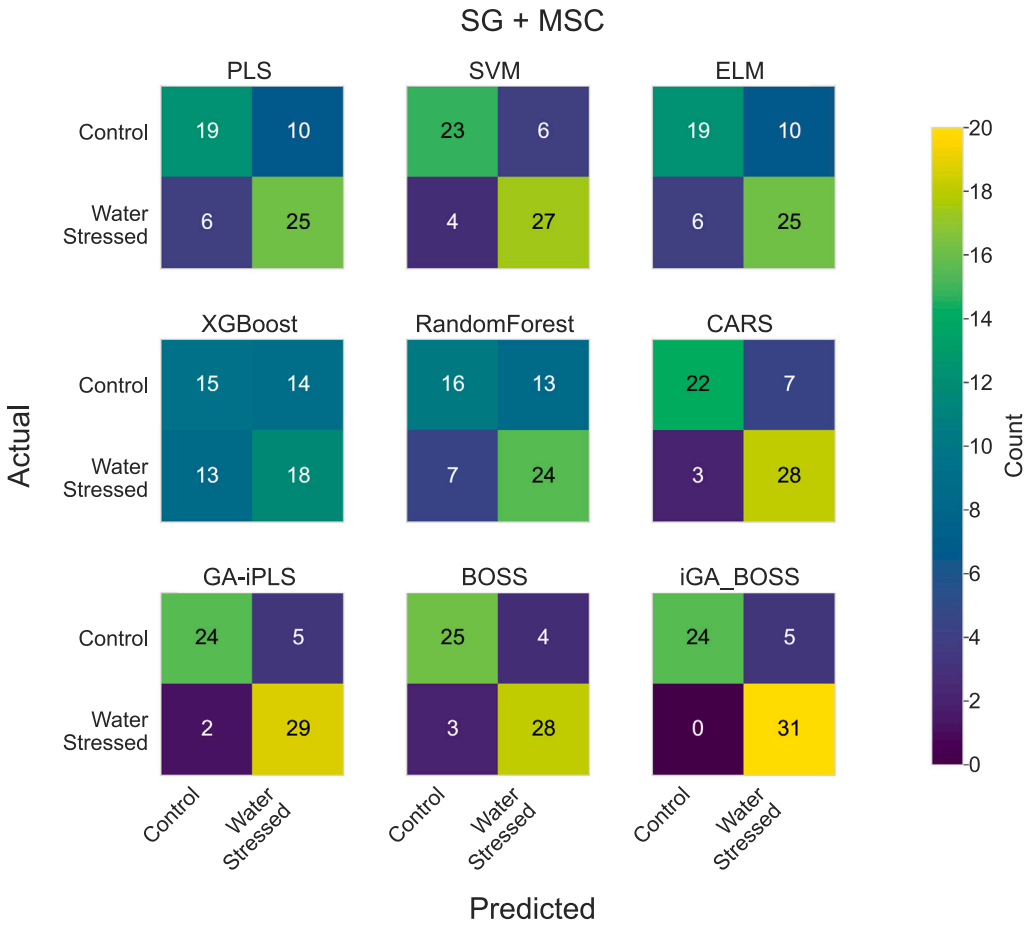



**Fig. 6.** Confusion matrices obtained under SG+MSC preprocessing for the full-spectrum classifiers and wavelength-selection-based PLS-DA models. The matrices report the number of C and W lettuce leaves correctly and incorrectly classified on the test set.

samples as C, with no false positives but thirty false negatives. This behavior confirms that ELM was highly sensitive to preprocessing and unreliable for this dataset when applied directly to the full spectral matrix.

These confusion-matrix patterns clarify how preprocessing shaped each method's error profile. For *BOSS*, SNV generally reduced false positives and improved specificity, especially under SG+SNV and SG1+SNV. For *iGA-BOSS*, MSC was more favorable, particularly in combination with SG1, where the method achieved the best overall balance between false positives and false negatives. *CARS* generally improved over the full-spectrum baselines but remained below the top-performing selectors, mainly because of residual false positives and false negatives. Among the full-spectrum classifiers, PLS-DA and SVM were the most reliable, while ELM, XGBoost, and Random Forest did not consistently improve discrimination. Overall, the confusion matrices confirm that the best classification behavior was obtained when PLS-DA was combined with wavelength selection, especially through the proposed *iGA-BOSS* approach.

### 3.3. Key spectral regions identified through wavelength selection

Fig. 4 displays the wavelengths selected by the different feature selection methods. In most cases, wavelengths were selected individually across several spectral regions, whereas *GA-iPLS* retained contiguous spectral intervals due to its interval-based selection mechanism. Although wavelengths across the full spectral range were selected in most cases, some regions were consistently identified by multiple algorithms. This was observed around 970–980 nm, 1400–1450 nm, near 1560 nm, between 1900 and 2000 nm, and in the spectral region immediately before 2500 nm. These results agree with previous studies that reported strong water absorption features at 960, 1440 and 1950 nm in leaf reflectance spectra (Roman and Ursu, 2016; Cui et al., 2024; Peñuelas et al., 1993). Indeed, the VIS region of a leaf spectrum is mainly associated with chlorophyll absorption (Gitelson et al., 2003; Hunt et al., 2011), while the NIR region contains information related to leaf internal structure and morphology (Buitrago et al., 2018; Michèle R. Slaton, 2001). For the proposed *iGA-BOSS* framework, a closer examination of the wavelengths selected by it reveals that a substantial proportion of them are located within, or adjacent to, well-established water-sensitive spectral regions. Additional selected wavelengths were found in neighboring SWIR regions, where reflectance variations are known to be influenced by changes in plant moisture content and drought-induced physiological responses (Danson and Bowyer, 2004; Arevalo-Ramirez et al., 2020). Overall, the consistency between the selected wavelengths and the known spectral behavior of water-stressed vegetation suggests that the proposed *iGA-BOSS* workflow effectively captures relevant information related to plant water content, particularly within the NIR and SWIR domains.

### 3.4. Advantages, limitations, and future perspectives

The results of this study show that the proposed *iGA-BOSS* strategy provides a favorable compromise between classification accuracy, wavelength compactness, and computational efficiency. The best configuration, obtained with SG1+MSC preprocessing, reached 0.95 accuracy and 0.95 F1-score using only 20 selected wavelengths. Although

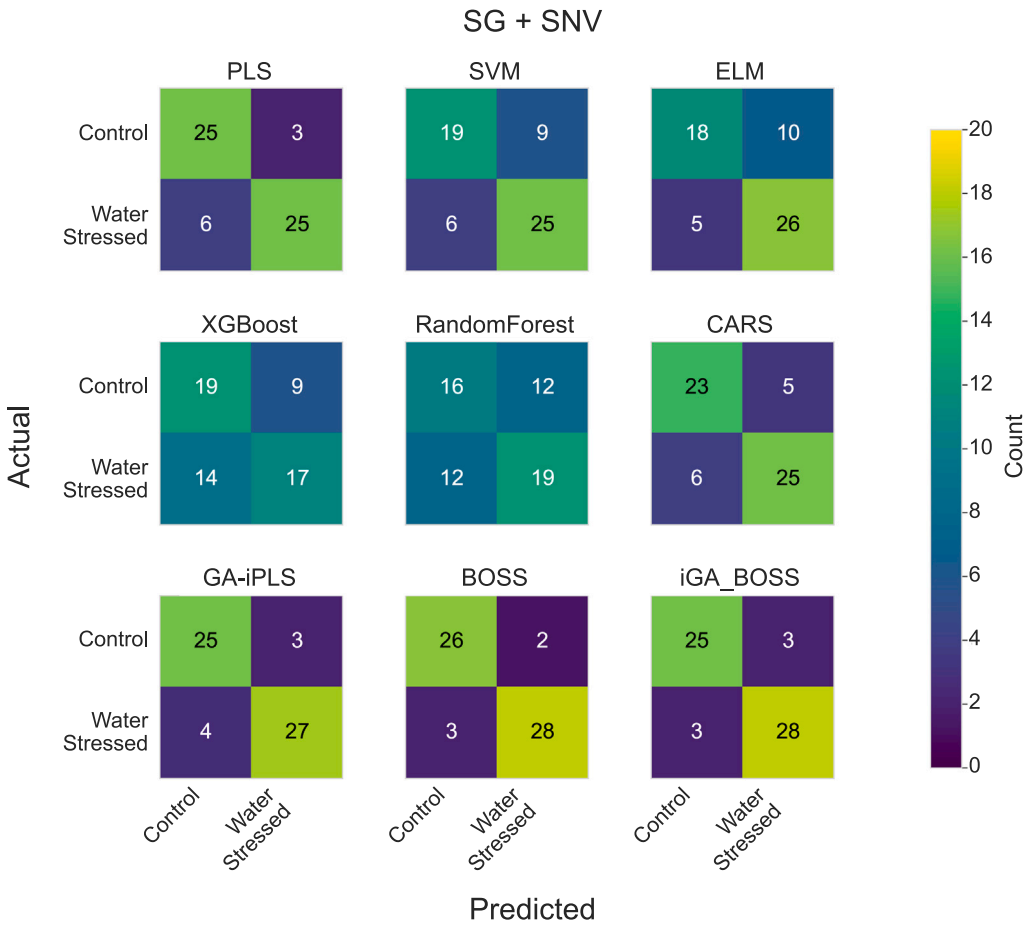



**Fig. 7.** Confusion matrices obtained under SG+SNV preprocessing for the full-spectrum classifiers and wavelength-selection-based PLS-DA models. The matrices report prediction results for C and W classes on the test set.

this performance is high, especially considering the strong dimensionality reduction from 2101 wavelengths to 20 wavelengths, further improvements may be possible by exploring alternative feature-extraction strategies.

The main advantage of the proposed approach is its interpretability. Unlike black-box feature extraction methods, wavelength selection explicitly identifies spectral bands associated with the discrimination between control and water-stressed lettuce samples. This is relevant for agronomic interpretation and for the future development of simplified optical sensing systems based on a limited number of wavelengths. In addition, the combination of *GA-iPLS* and *BOSS* reduces the computational burden of applying *BOSS* directly to the full spectral space, while preserving compact and informative wavelength subsets.

However, the proposed method also has limitations. First, the approach remains dependent on the quality of the preprocessing pipeline and on the stability of the wavelength-selection procedure. Second, because the first stage is based on interval selection, potentially informative wavelengths located in discarded intervals may be removed before the *BOSS* refinement stage. Third, although 0.95 accuracy represents the best result obtained in this study, the classification task remains binary, and future work should investigate whether more advanced feature-extraction approaches can further improve discrimination performance.

Recent studies have shown that deep learning can be used as an automatic feature-extraction strategy for spectroscopic data, such as CNN, LSTM, CNN–LSTM, or attention-based architectures. For example, CNN–LSTM-based feature extraction has been used with near-infrared spectroscopy for beer quality and authenticity assessment, achieving strong regression performance and 100% classification accuracy when combined with PLS-DA (Wei et al., 2025). Deep learning approaches can automatically learn nonlinear spectral representations and may therefore capture complex patterns that are not fully described by manually selected wavelengths.

Based on these considerations, future work should compare the proposed wavelength-selection framework with deep learning-based feature extraction methods. Such a comparison would clarify whether automatically learned spectral features can outperform compact wavelength subsets selected by chemometric methods. Nevertheless, PLS-DA remains useful in this context as a interpretable and computationally efficient reference model. Future research may investigate hybrid approaches in which deep learning is used for nonlinear feature extraction, while PLS-DA is retained for interpretable classification and comparison with wavelength-selection-based models.

Another limitation of the present study is that the different water-stress levels were merged into a binary classification problem, distinguishing only between control and stressed samples. This formulation was adopted to evaluate the ability of the proposed workflow to detect the presence of water stress in a robust and operationally simple way. However, merging the 20%, 40%, and 60% water-reduction treatments into a single stressed class inevitably removes information related to stress severity. Future work will therefore extend the proposed framework to a multi-class classification problem, in which control plants and different water-stress levels are modeled as separate classes. This will allow the evaluation of whether the selected wavelengths and

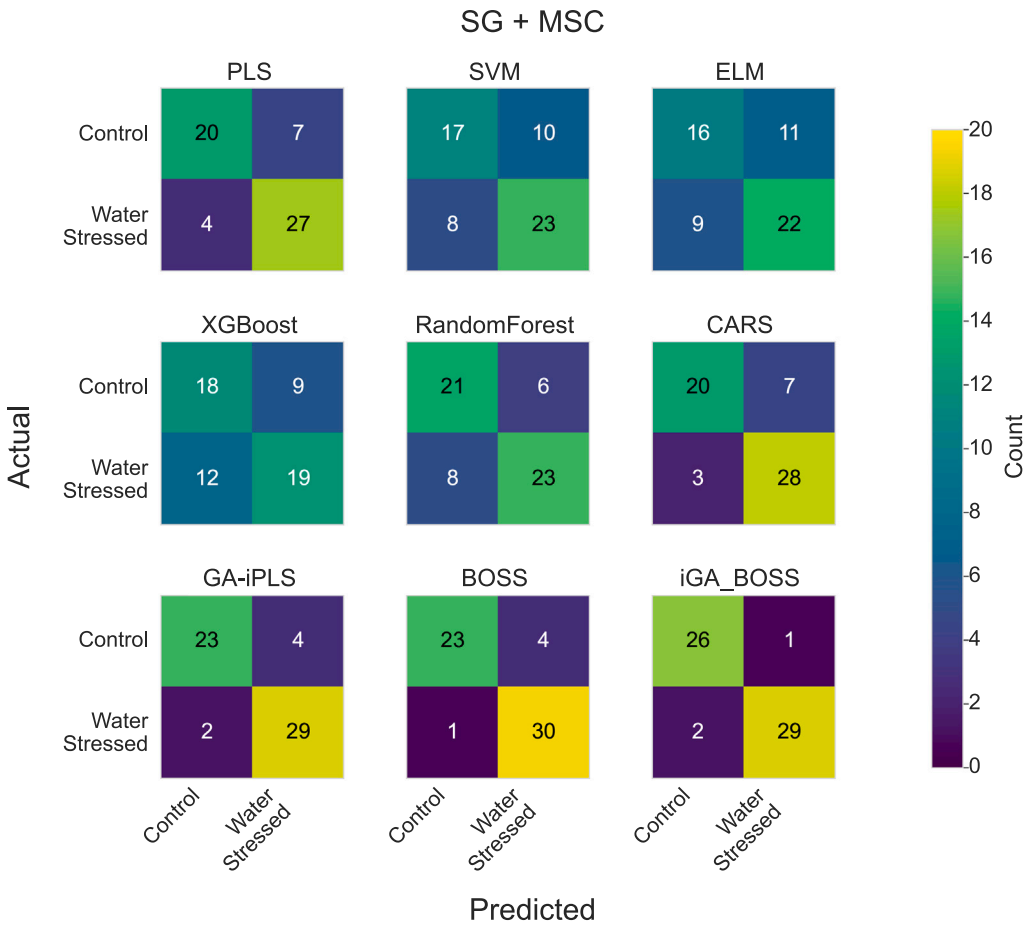



**Fig. 8.** Confusion matrices obtained under SG1+MSC preprocessing for the full-spectrum classifiers and wavelength-selection-based PLS-DA models. The matrices report the number of C and W lettuce leaves correctly and incorrectly classified on the test set.

classification models can discriminate not only the presence of water stress, but also its intensity.

## 4. Conclusion

This study presented a comprehensive evaluation of wavelength selection methods combined with classical classifiers for accurately assessing lettuce physiological status under induced water stress using spectral signatures. The results demonstrate that integrating preprocessing techniques with advanced feature selection algorithms can significantly enhance classification performance while reducing model complexity.

Among the tested combinations, the integration of SG1 smoothing and MSC normalization with the *iGA-BOSS* selection method, followed by PLS-DA classification, yielded the highest accuracy (0.95) and F1-score (0.95). Permutation tests confirmed the statistical significance of all models, with p-values consistently below 0.05, reinforcing the reliability of the results.

Beyond overall performance, the accuracy analysis per selected wavelength shows that the proposed *iGA-BOSS* method achieves the best performance in most cases, while also providing an excellent trade-off between classification accuracy and spectral compactness.

Wavelength selection analysis identified several recurring spectral regions, particularly around 1450 nm and 1950 nm, as informative across multiple algorithms. These bands correspond to absorption features related to water content in the leaf, underscoring their relevance for crop monitoring applications. *GA-iPLS*-based methods were further distinguished by their ability to select contiguous wavelength intervals, reflecting robust underlying spectral signatures.

The proposed methodology demonstrates the potential for building interpretable, computationally efficient, and high-performing models suitable for implementation in resource-constrained environments, such as onboard processing for agricultural robotics or satellite platforms.

Although the obtained results are promising, the present study was conducted using a single lettuce cultivar within a single greenhouse environment under controlled experimental conditions. Therefore, the generalization capability of the proposed models across different cultivars, greenhouse configurations, seasons, or environmental conditions cannot yet be fully assessed. Variations in genotype, illumination, temperature, humidity, and cultivation practices may affect spectral responses and consequently model performance.

Future works will include multi-cultivar and multi-environment datasets to evaluate the robustness, transferability, and scalability of the proposed methodology under more heterogeneous agricultural scenarios. In addition, the current binary classification framework will be extended to a multi-class classification problem in order to distinguish among control, 20%, 40%, and 60% water-reduction treatments. This will allow the assessment of whether the proposed wavelength-selection strategy can capture progressive spectral changes associated with increasing stress severity. Finally, the proposed *iGA-BOSS* approach will be compared with deep learning-based feature extraction methods, such as CNN-, LSTM-, and CNN–LSTM-based architectures, to evaluate whether automatically learned spectral representations can further improve classification performance while maintaining interpretability through comparison with selected wavelength regions.

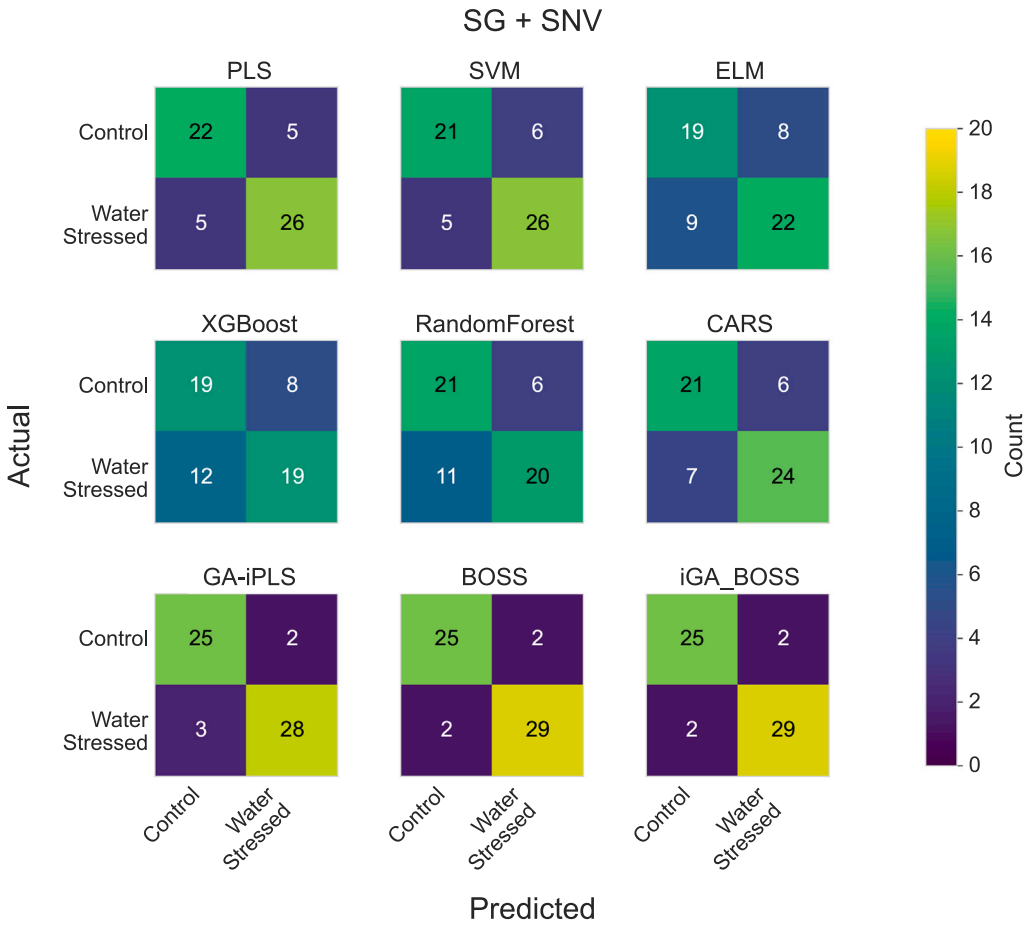



**Fig. 9.** Confusion matrices obtained under SG1+SNV preprocessing for the full-spectrum classifiers and wavelength-selection-based PLS-DA models. The matrices report prediction results for C and W classes on the test set.

## CRediT authorship contribution statement

**Nicola Dilillo:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Gica Stefanescu Miralles:** Writing – review & editing, Writing – original draft, Visualization, Validation, Resources, Data curation. **Lorenzo Comba:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Conceptualization. **Massimo Pugliese:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision. **Maurizio Rebaudengo:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision. **Renato Ferrero:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration, Methodology, Formal analysis.

## 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.

## Acknowledgments

This study was carried out within the Agritech National Research Center and received funding from the European Union Next-GenerationEU (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. Permutation test

See Tables A.7 and A.8.

## Appendix B. Selected wavelengths

See Tables B.9–B.12.

## Appendix C. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ecoinf.2026.103944.

## Data availability

The data supporting the findings of this study are available in the supplementary material. The source code used for the analyses is publicly available in the following GitHub repository: https://github.com/nicoladilillo/advancing_spectral_based_lettuce_classification.

**Table A.7**
Permutation test p-values associated with each model configuration using SG preprocessing. Values below 0.05 indicate statistically significant performance improvements compared to chance. For feature-selection approaches, the model column reports the classifier on the first line and the wavelength-selection algorithm on the second line.

| Preprocessing | Model (Algo.) | Accuracy | Recall | Precision | F1 |
|---|---|---|---|---|---|
| MSC | PLS-DA | 0.01 | 0.03 | 0.01 | 0.00 |
| | ELM | 0.00 | 0.00 | 0.00 | 0.00 |
| | XGB | 0.01 | 0.03 | 0.02 | 0.01 |
| | RF | 0.01 | 0.02 | 0.01 | 0.00 |
| | SVM | 0.01 | 0.03 | 0.02 | 0.01 |
| | PLS-DA (CARS) | 0.00 | 0.00 | 0.00 | 0.00 |
| | PLS-DA (BOSS) | 0.00 | 0.00 | 0.00 | 0.00 |
| | PLS-DA (GA-iPLS) | 0.00 | 0.00 | 0.00 | 0.00 |
| | PLS-DA (iGA-BOSS) | 0.00 | 0.00 | 0.00 | 0.00 |
| SNV | PLS-DA | 0.00 | 0.00 | 0.00 | 0.00 |
| | ELM | 0.00 | 0.00 | 0.01 | 0.00 |
| | XGB | 0.00 | 0.01 | 0.00 | 0.00 |
| | RF | 0.10 | 0.28 | 0.10 | 0.15 |
| | SVM | 0.00 | 0.01 | 0.00 | 0.00 |
| | PLS-DA (CARS) | 0.00 | 0.00 | 0.00 | 0.00 |
| | PLS-DA (BOSS) | 0.00 | 0.00 | 0.00 | 0.00 |
| | PLS-DA (GA-iPLS) | 0.00 | 0.00 | 0.00 | 0.00 |
| | PLS-DA (iGA-BOSS) | 0.00 | 0.00 | 0.00 | 0.00 |

**Table A.8**
Permutation test p-values associated with each model configuration using SG1 preprocessing. Values below 0.05 indicate statistically significant performance improvements compared to chance. For feature-selection approaches, the model column reports the classifier on the first line and the wavelength-selection algorithm on the second line.

| Preprocessing | Model (Algo.) | Accuracy | Recall | Precision | F1 |
|---|---|---|---|---|---|
| MSC | PLS-DA | 0.00 | 0.00 | 0.00 | 0.00 |
| | ELM | 0.02 | 0.03 | 0.05 | 0.01 |
| | XGB | 0.00 | 0.00 | 0.00 | 0.00 |
| | RF | 0.00 | 0.07 | 0.00 | 0.00 |
| | SVM | 0.00 | 0.00 | 0.00 | 0.00 |
| | PLS-DA (CARS) | 0.00 | 0.00 | 0.00 | 0.00 |
| | PLS-DA (BOSS) | 0.00 | 0.00 | 0.00 | 0.00 |
| | PLS-DA (GA-iPLS) | 0.00 | 0.00 | 0.00 | 0.00 |
| | PLS-DA (iGA-BOSS) | 0.00 | 0.00 | 0.00 | 0.00 |
| SNV | PLS-DA | 0.00 | 0.00 | 0.00 | 0.00 |
| | ELM | 0.01 | 0.02 | 0.01 | 0.01 |
| | XGB | 0.00 | 0.00 | 0.00 | 0.00 |
| | RF | 0.01 | 0.30 | 0.00 | 0.02 |
| | SVM | 0.00 | 0.00 | 0.00 | 0.00 |
| | PLS-DA (CARS) | 0.00 | 0.02 | 0.00 | 0.00 |
| | PLS-DA (BOSS) | 0.00 | 0.00 | 0.00 | 0.00 |
| | PLS-DA (GA-iPLS) | 0.00 | 0.00 | 0.00 | 0.00 |
| | PLS-DA (iGA-BOSS) | 0.00 | 0.00 | 0.00 | 0.00 |

**Table B.9**
Wavelengths selected by each feature selection method under SG + MSC preprocessing. Ranges indicate contiguous spectral intervals; comma-separated values indicate individual wavelengths.

| Algorithm | Wavelengths |
|---|---|
| CARS | 783, 804, 943, 981, 1003, 1125, 1168, 1239, 1264, 1318, 1390, 1423, 1524, 1541, 1558, 1817, 1878, 1907, 1973, 1991, 2067, 2085, 2100, 2120, 2399 |
| GA-iPLS | 400–420, 526–546, 673–714, 757–777, 841–861, 904–945, 988–1008, 1282–1302, 1345–1428, 1471–1491, 1555–1575, 1639–1659, 1681–1701, 1723–1743, 1849–1869, 1891–1911, 2080–2100, 2143–2163, 2185–2247, 2290–2310, 2353–2373, 2416–2436 |
| BOSS | 401, 518, 560, 852, 854, 884, 901, 940, 1027, 1438, 1467, 1630, 1863, 2026, 2041, 2076, 2194, 2304 |
| GA-iPLS + BOSS | 683, 688, 695, 1007, 1347, 1361, 1372, 1378, 1394–1395, 1405, 1426, 1482, 1570, 1574, 1657–1658, 1687, 1695, 1897, 2187, 2190, 2192, 2235, 2247 |

**Table B.10**
Wavelengths selected by each feature selection method under SG + SNV preprocessing.

| Algorithm | Wavelengths |
|---|---|
| CARS | 401, 447, 499, 672, 702, 733, 798, 857, 893, 945, 981, 1016, 1060, 1090, 1120, 1171, 1203, 1237, 1285, 1322, 1357, 1387, 1421, 1452, 1512, 1544, 1576, 1607, 1654, 1684, 1715, 1782, 1816, 1878, 1917, 1980, 2016, 2046, 2085, 2118, 2219, 2252, 2292, 2323, 2358, 2398, 2437, 2494 |
| GA-iPLS | 400–420, 505–525, 631–651, 694–714, 757–777, 799–819, 862–882, 925–945, 967–987, 1009–1029, 1135–1155, 1219–1239, 1345–1407, 1450–1470, 1492–1512, 1534–1554, 1618–1638, 1870–1911, 2059–2079, 2227–2268, 2332–2352, 2374–2415, 2458–2478 |
| BOSS | 422, 637, 644, 922, 941, 964, 1098, 1123, 1458, 1520, 1636, 1655, 1779, 1814, 1860, 2018, 2257, 2287 |
| GA-iPLS + BOSS | 511, 695, 764, 805, 807, 925, 928, 1018, 1141, 1234, 1357, 1362, 1380–1381, 1389, 1393, 1404, 1494, 1618, 1633, 1887, 1890, 1892, 1897, 1904, 2066, 2074, 2229, 2251, 2343, 2346, 2384, 2407, 2413, 2415, 2475, 2477–2478 |

**Table B.11**
Wavelengths selected by each feature selection method under SG1 + MSC preprocessing.

| Algorithm | Wavelengths |
|---|---|
| CARS | 413, 436, 473, 492, 538, 594, 620, 637, 675, 699, 764, 789, 816, 843, 859, 883, 907, 929, 954, 978, 995, 1025, 1045, 1068, 1083, 1106, 1141, 1156, 1174, 1193, 1211, 1227, 1250, 1279, 1301, 1327, 1350, 1372, 1402, 1426, 1445, 1467, 1488, 1505, 1523, 1538, 1569, 1588, 1603, 1620, 1636, 1654, 1674, 1692, 1722, 1739, 1759, 1775, 1804, 1827, 1845, 1863, 1895, 1914, 1930, 1951, 1979, 2011, 2035, 2057, 2081, 2108, 2133, 2153, 2175, 2197, 2223, 2248, 2266, 2281, 2305, 2330, 2349, 2378, 2393, 2416, 2446, 2461, 2482, 2499 |
| GA-iPLS | 505–525, 547–567, 610–630, 736–756, 799–819, 841–861, 925–945, 967–987, 1030–1071, 1135–1197, 1261–1281, 1387–1407, 1429–1470, 1534–1575, 1618–1638, 1681–1701, 1723–1743, 1786–1806, 1828–1869, 1996–2016, 2080–2100, 2143–2163, 2248–2331, 2353–2373, 2437–2500 |
| BOSS | 486, 714, 769, 868, 1058, 1349, 1397, 1836, 1843, 1897, 1926, 1955, 1987, 2024, 2129, 2193, 2279, 2403, 2464 |
| GA-iPLS + BOSS | 522, 1068–1069, 1180, 1388, 1469, 1561, 1565, 1683, 1849, 1863, 2001, 2154, 2279, 2295, 2302, 2371, 2475, 2478, 2488 |

**Table B.12**
Wavelengths selected by each feature selection method under SG1 + SNV preprocessing.

| Algorithm | Wavelengths |
|---|---|
| CARS | 413, 472, 536, 675, 764, 844, 907, 956, 995, 1042, 1084, 1119, 1155, 1212, 1249, 1279, 1345, 1401, 1444, 1536, 1568, 1621, 1655, 1685, 1739, 1774, 1834, 1931, 2133, 2188, 2221, 2265, 2304, 2342, 2415, 2449, 2482 |
| GA-iPLS | 421–441, 463–483, 526–567, 631–714, 988–1008, 1051–1071, 1198–1260, 1366–1407, 1429–1470, 1534–1575, 1639–1680, 1723–1743, 1807–1827, 1849–1869, 1996–2016, 2164–2184, 2206–2226, 2374–2394, 2416–2436, 2479–2500 |
| BOSS | 531, 610, 757, 869, 964, 1083, 1103, 1229, 1351, 1449, 1529, 1546, 1632, 1653, 1673, 1843, 1967, 2142, 2428 |
| GA-iPLS + BOSS | 468, 471, 475, 479, 528–529, 535, 551, 557, 559, 562, 633, 637–638, 675, 684, 686, 688, 1006, 1059, 1069, 1199–1200, 1204, 1216, 1226, 1228, 1249, 1254, 1260, 1386, 1403, 1406, 1429–1430, 1435, 1438, 1448, 1535, 1565, 1575, 1642–1643, 1656, 1673, 1680, 1732, 1742, 1818, 1821, 1854, 1865, 1867, 1997, 2000, 2010, 2168, 2181, 2183, 2210, 2419, 2422, 2485, 2488 |

## References


Al-Huqail, A., El-Dakak, R.M., Sanad, M.N., Badr, R.H., Ibrahim, M.M., Soliman, D., et al., 2020. Effects of climate temperature and water stress on plant growth and accumulation of antioxidant compounds in sweet basil (Ocimum basilicum L.) leafy vegetable. Scientifica 12, http://dx.doi.org/10.1155/2020/3808909.

Arevalo-Ramirez, T., Villacrés, J., Fuentes, A., Reszka, P., Auat Cheein, F.A., 2020. Moisture content estimation of pinus radiata and eucalyptus globulus from reconstructed leaf reflectance in the SWIR region. Biosyst. Eng. 193, 187–205. http://dx.doi.org/10.1016/j.biosystemseng.2020.03.004.

Barnes, R.J., Dhanoa, M.S., Lister, S.J., 1989. Standard normal variate transformation and detrending of near infrared diffuse reflectance spectra. Appl. Spectrosc. 43 (5), 772–777.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32. http://dx.doi.org/10.1023/A:1010933404324.

Buitrago, M.F., Groen, T.A., Hecker, C.A., Skidmore, A.K., 2018. Spectroscopic determination of leaf traits using infrared spectra. Int. J. Appl. Earth Obs. Geoinf. 69, 237–250. http://dx.doi.org/10.1016/j.jag.2017.11.014.

Caturegli, L., Matteoli, S., Gaetani, M., et al., 2020. Effects of water stress on spectral reflectance of bermudagrass. Sci Rep 10 (15055), http://dx.doi.org/10.1038/s41598-020-72006-6.

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785–794.

Cheng, J., Sun, J., Yao, K., Xu, M., Cao, Y., 2022. A variable selection method based on mutual information and variance inflation factor. Spectrochim. Acta Part A: Mol. Biomol. Spectrosc. 268, 120652. http://dx.doi.org/10.1016/j.saa.2021.120652.

Cozzolino, D., 2018. An overview of the application of near infrared spectroscopy for non-destructive quality assessment of cereal grains. Appl. Spectrosc. Rev. 53 (8), 667–687. http://dx.doi.org/10.1080/05704928.2018.1425214.

Cui, J., Sawut, M., Ailijiang, N., Manlike, A., Hu, X., 2024. Estimation of leaf water content of a fruit tree by in situ vis-nir spectroscopy using multiple machine learning methods in southern xinjiang, China. Agronomy 14 (8), http://dx.doi.org/10.3390/agronomy14081664.

Danson, F., Bowyer, P., 2004. Estimating live fuel moisture content from remotely sensed reflectance. Remote Sens. Environ. 92 (3), 309–321. http://dx.doi.org/10.1016/j.rse.2004.03.017.

Deng, B.-C., Yun, Y.-H., Cao, D.-S., Yin, Y.-L., Wang, W.-T., Lu, H.-M., Luo, Q.-Y., Liang, Y.-Z., 2016. A bootstrapping soft shrinkage approach for variable selection in chemical modeling. Anal. Chim. Acta 908, 63–74. http://dx.doi.org/10.1016/j.aca.2016.01.001.

Devos, O., Downey, G., Duponchel, L., 2014. 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, 124–130. http://dx.doi.org/10.1016/j.foodchem.2013.10.020.

Dilillo, N., Sanna, A., Belcore, E., Smith, K., Piras, M., Montrucchio, B., Ferrero, R., 2025. Enhancing lettuce classification: Optimizing spectral wavelength selection via CCARS and PLS-DA. Smart Agric. Technol. 11, 100962. http://dx.doi.org/10.1016/j.atech.2025.100962.

Dong, X., Dong, Y., Liu, J., Wang, C., Bao, C., Wang, N., Zhao, X., Chen, Z., 2024. Identification and quantitative detection of illegal additives in wheat flour based on near-infrared spectroscopy combined with chemometrics. Spectrochim. Acta Part A: Mol. Biomol. Spectrosc. 323, 124938. http://dx.doi.org/10.1016/j.saa.2024.124938, URL https://www.sciencedirect.com/science/article/pii/S1386142524011041.

Duarte-Carvajalino, J.M., Silva-Arero, E.A., Góez-Vinasco, G.A., Torres-Delgado, L.M., Ocampo-Paez, O.D., Castaño-Marín, A.M., 2021. Estimation of water stress in potato plants using hyperspectral imagery and machine learning algorithms. Horticulturae 7 (7), http://dx.doi.org/10.3390/horticulturae7070176.

Foody, G.M., 2009. Classification accuracy comparison: Hypothesis tests and the use of confidence intervals in evaluations of difference, equivalence and non-inferiority. Remote Sens. Environ. 113 (8), 1658–1663.

Foody, G.M., 2010. Assessing the accuracy of land cover change with imperfect ground reference data. Remote Sens. Environ. 114 (10), 2271–2285. http://dx.doi.org/10.1016/j.rse.2010.05.003.

Foody, G.M., 2024. Ground truth in classification accuracy assessment: Myth and reality. Geomatics 4 (1), 81–90. http://dx.doi.org/10.3390/geomatics4010005.

Francini, A., Sebastiani, L., 2019. Abiotic stress effects on performance of horticultural crops. Horticulturae 5 (4), http://dx.doi.org/10.3390/horticulturae5040067.

Galieni, A., Di Mattia, C., De Gregorio, M., Speca, S., Mastrocola, D., Pisante, M., Stagnari, F., 2015. Effects of nutrient deficiency and abiotic environmental stresses on yield, phenolic compounds and antiradical activity in lettuce (Lactuca sativa L.). Sci. Hort. 187, 93–101. http://dx.doi.org/10.1016/j.scienta.2015.02.036.

Gao, H., Mao, H., Zhang, X., et al., 2021. Fusion of three optical sensors for nondestructive detection of water content in lettuce canopies. J. Appl. Spectrosc 88 (194–202), http://dx.doi.org/10.1007/s10812-021-01158-8.

Garrett, R.G., 1989. The chi-square plot: a tool for multivariate outlier recognition. J. Geochem. Explor. 32 (1), 319–341. http://dx.doi.org/10.1016/0375-6742(89)90071-X.

Ghorbani, H., 2019. Mahalanobis distance and its application for detecting multivariate outliers. Facta Univ. Ser.: Math. Informatics 583–595.

Gitelson, A.A., Gritz, Y., Merzlyak, M.N., 2003. Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. J. Plant Physiol. 160 (3), 271–282. http://dx.doi.org/10.1078/0176-1617-00887.

Good, P., 2005. Permutation, Parametric and Bootstrap Tests of Hypotheses. Springer.

Hastie, T., Tibshirani, R., Friedman, J.H., Friedman, J.H., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. vol. 2, Springer.

Huang, G.-B., Zhu, Q.-Y., Siew, C.-K., 2006. Extreme learning machine: Theory and applications. Neurocomputing 70 (1–3), 489–501. http://dx.doi.org/10.1016/j.neucom.2005.12.126.

Hunt, J., Daughtry, C.S.T., Eitel, J.U.H., Long, D.S., 2011. Remote sensing leaf chlorophyll content using a visible band index. Agron. J. 103 (4), 1090–1099. http://dx.doi.org/10.2134/agronj2010.0395.

Johnson, R.A., Wichern, D.W., 2002. Applied Multivariate Statistical Analysis, fifth ed. Prentice Hall, Upper Saddle River, NJ.

Jolliffe, I.T., Cadima, J., 2016. Principal component analysis: A review and recent developments. Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci. 374 (2065), 20150202. http://dx.doi.org/10.1098/rsta.2015.0202.

Junaid, M., Gokce, A., 2024. Global agricultural losses and their causes. Bull. Biological Allied Sci. Res. 9 (1), 66. http://dx.doi.org/10.54112/bbasr.v2024i1.66.

Kizil, U., Genc, L., Melis, S., M.M., D., 2012. Lettuce (lactuca sativa l.) yield prediction under water stress using artificial neural network (ANN) model and vegetation indices. Zemdirbyste 99, 409–418.

Kumar, P., Eriksen, R.L., Simko, I., Mou, B., 2021. Molecular mapping of water-stress responsive genomic loci in lettuce (Lactuca spp.) using kinetics chlorophyll fluorescence, hyperspectral imaging and machine learning. Front. Genet. Volume 12 - 2021, http://dx.doi.org/10.3389/fgene.2021.634554.

Lee, L.C., Liong, C.-Y., Jemain, A.A., 2018. 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), 3526–3539.

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

Ma, X.-H., Chen, Z.-G., Liu, J.-M., 2024. Wavelength selection method for near-infrared spectroscopy based on max-relevance min-redundancy. Spectrochim. Acta Part A: Mol. Biomol. Spectrosc. 310, 123933. http://dx.doi.org/10.1016/j.saa.2024.123933.

Mahalanobis, P.C., 1936. On the generalized distance in statistics. Proc. Natl. Inst. Sci. (Calcutta) 2 (1), 49–55.

Maimaitiyiming, M., Ghulam, A., Bozzolo, A., Wilkins, J.L., Kwasniewski, M.T., 2017. Early detection of plant physiological responses to different levels of water stress using reflectance spectroscopy. Remote. Sens. 9 (7), http://dx.doi.org/10.3390/rs9070745.

Martens, H., Naes, T., 1989. Multivariate Calibration. John Wiley & Sons.

Michèle R. Slaton, W.K.S., 2001. Estimating near-infrared leaf reflectance from leaf structural characteristics. Am. J. Bot. 88 (2), 278–284. http://dx.doi.org/10.2307/2657019.

Mitchell, M., 1998. An Introduction to Genetic Algorithms. MIT Press.

Nørgaard, L., Saudland, A., Wagner, J., Nielsen, J.P., Munck, L., Engelsen, S.B., 2000. Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy. Appl. Spectrosc. http://dx.doi.org/10.1366/0003702001949500.

Osco, L.P., Ramos, A.P.M., Moriya, É.A.S., Bavaresco, L.G., Lima, B.C.d., Estrabis, N., Pereira, D.R., Creste, J.E., Júnior, J.M., Gonçalves, W.N., Imai, N.N., Li, J., Liesenberg, V., Araújo, F.F.d., 2019. Modeling hyperspectral response of water-stress induced lettuce plants using artificial neural networks. Remote. Sens. 11 (23), http://dx.doi.org/10.3390/rs11232797.

de la Pena, R., Hughes, J., 2007. Improving vegetable productivity in a variable and changing climate. J. SAT Agric. Res. 4.

Peñuelas, J., Filella, I., Serrano, L., Savé, R., 1993. The reflectance at the 950–970 nm region as an indicator of plant water status. Int. J. Remote Sens. 14 (10), 1887–1905. http://dx.doi.org/10.1080/01431169308954010.

Peraza-Alemán, C.M., Arazuri, S., Jarén, C., Ruiz de Galarreta, J.I., Barandalla, L., López-Maestresalas, A., 2025. Predicting the spatial distribution of reducing sugars using near-infrared hyperspectral imaging and chemometrics: a study in multiple potato genotypes. Comput. Electron. Agric. 235, 110323. http://dx.doi.org/10.1016/j.compag.2025.110323.

Radhesyam, V., Phaneendra, K., Prabukumar, M., L. Agilandeeswari, S., 2024. Strategies for dimensionality reduction in hyperspectral remote sensing: A comprehensive overview. Egypt. J. Remote. Sens. Space Sci. 27 (1), 82–92. http://dx.doi.org/10.1016/j.ejrs.2024.01.005.

Rogers, M., Blanc-Talon, J., Urschler, M., Delmas, P., 2023. Wavelength and texture feature selection for hyperspectral imaging: a systematic literature review. J. Food Meas. Charact. 17 (6), 6039–6064. http://dx.doi.org/10.1007/s11694-023-02044-x.

Roman, A., Ursu, T., 2016. Multispectral satellite imagery and airborne laser scanning techniques for the detection of archaeological vegetation marks. Landsc. Archaeol. North. Front. Rom. Emp. At Porolissum: An Interdiscip. Res. Proj. 141–152.

Romero, A.P., Alarcón, A., Valbuena, R.I., Galeano, C.H., 2017. Physiological assessment of water stress in potato using spectral information. Front. Plant Sci. Volume 8 - 2017, http://dx.doi.org/10.3389/fpls.2017.01608.

Savitzky, A., Golay, M.J.E., 1964. Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. 36, 1627–1639.

Sedgwick, P., 2014. Understanding P values. BMJ 349.

Shi, S., Feng, J., Yang, L., Xing, J., Pan, G., Tang, J., Wang, J., Liu, J., Cao, C., Jiang, Y., 2023. 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, 122343. http://dx.doi.org/10.1016/j.saa.2023.122343.

Shimada, S., Funatsuka, E., Ooda, M., 2012. Developing the monitoring method for plant water stress using spectral reflectance measurement.

Sokolova, M., Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Inf. Process. Manage. 45 (4), 427–437. http://dx.doi.org/10.1016/j.ipm.2009.03.002.

Susič, N., Žibrat, U., Širca, S., Strajnar, P., Razinger, J., Knapič, M., Vončina, A., Urek, G., Gerič Stare, B., 2018. Discrimination between abiotic and biotic drought stress in tomatoes using hyperspectral imaging. Sensors Actuators B: Chem. 273, 842–852. http://dx.doi.org/10.1016/j.snb.2018.06.121.

Tharwat, A., 2020. Classification assessment methods. Appl. Comput. Inform. 17 (1), 168–192. http://dx.doi.org/10.1016/j.aci.2018.08.003.

Toscano, S., Franzoni, G., Álvarez, S., 2023. Drought stress in horticultural plants. Horticulturae 9 (1), http://dx.doi.org/10.3390/horticulturae9010007.

Wang, T., Zheng, Y., Xu, L., Yun, Y.-H., 2025. Comprehensive comparison on different wavelength selection methods using several near-infrared spectral datasets with different dimensionalities. Spectrochim. Acta Part A: Mol. Biomol. Spectrosc. 331, 125767. http://dx.doi.org/10.1016/j.saa.2025.125767.

Wei, Y., Liu, J., Xi, G., Lu, Y., 2025. Quantitative detection of key parameters and authenticity verification for beer using near-infrared spectroscopy. Foods 14 (22), http://dx.doi.org/10.3390/foods14223936, URL https://www.mdpi.com/2304-8158/14/22/3936.

Westerhuis, J.A., Hoefsloot, H.C., Smit, S., Vis, D.J., Smilde, A.K., van Velzen, E.J., van Duijnhoven, J.P., van Dorsten, F.A., 2008. Assessment of PLSDA cross validation. Metabolomics 4, 81–89.

Xiaobo, Z., Jiewen, Z., Hanpin, M., Jiyong, S., Xiaopin, Y., Yanxiao, L., 2010a. Genetic algorithm interval partial least squares regression combined successive projections algorithm for variable selection in near-infrared quantitative analysis of pigment in cucumber leaves. Appl. Spectrosc. 64 (7), 786–794. http://dx.doi.org/10.1366/000370210791666246, (Accessed 2025-04-03).

Xiaobo, Z., Jiewen, Z., Povey, M.J., Holmes, M., Hanpin, M., 2010b. Variables selection methods in near-infrared spectroscopy. Anal. Chim. Acta 667 (1), 14–32. http://dx.doi.org/10.1016/j.aca.2010.03.048.

Xu, Y., Dong, Y., Liu, J., Wang, C., Li, Z., 2025. Combination of near infrared spectroscopy with characteristic interval selection for rapid detection of rice protein content. J. Food Comp. Anal. 137, 106995. http://dx.doi.org/10.1016/j.jfca.2024.106995, URL https://www.sciencedirect.com/science/article/pii/S0889157524010299.

Yavuz, N., Seymen, M., Kal, Ü., 2021. Impacts of water stress and harvest time on physio-biochemical characteristics of lettuce. Int. J. Agric. Nat. Sci. 14, 61–77.

Yin, W., Zhang, C., Zhu, H., Zhao, Y., He, Y., 2017. Application of near-infrared hyperspectral imaging to discriminate different geographical origins of Chinese wolfberries. PLoS ONE 12 (7), e0180534. http://dx.doi.org/10.1371/journal.pone.0180534.

Zhang, J., Feng, X., Liu, X., He, Y., 2018. Identification of hybrid okra seeds based on near-infrared hyperspectral imaging technology. Appl. Sci. 8 (10), 1793. http://dx.doi.org/10.3390/app8101793.

Zhao, J., Li, H., Chen, C., Pang, Y., Zhu, X., 2022. Detection of water content in lettuce canopies based on hyperspectral imaging technology under outdoor conditions. Agriculture 12 (11), http://dx.doi.org/10.3390/agriculture12111796.