Characterizing Cross-Crop Stink Bug Spectral Signatures from Hyperspectral Data

João de Deus Ferreira e Silva[a†], André Oliveira Françani[a†], Jianglong Yan[b], Ednaldo José Ferreira[c], Lúcio André de Castro Jorge[c], Liang Zhao[a].

[a] Faculty of Philosophy, Sciences and Letters at Ribeirão Preto (FFCLRP), University of São Paulo, Ribeirão Preto, 14040-901, SP, Brazil.

[b] Institute of Mathematics and Computer Science (ICMC), University of São Paulo, São Carlos, 13566-590, SP, Brazil.

[c] National Research Center for Development of Agricultural Instrumentation, Brazilian Agricultural Research Agency, São Carlos, 13561-206, SP, Brazil.

[†] These authors contributed equally to this work.

A paper from the Proceedings of the
17th International Conference on Precision Agriculture and 11th Brazilian Congress on Precision and Digital Agriculture
13-15 July 2026
Porto Alegre, Brazil

Abstract.

Effective crop protection in agricultural production systems requires the ability to detect pest-induced stress in a timely and reliable manner. In large-scale farming systems, stink bugs attack multiple crop species, making cross-crop pest detection a critical capability for scalable monitoring solutions. Rather than developing crop-specific models that require retraining for each species, identifying crop-independent spectral signatures of stink bug infestation enables transferable detection across different crops and growing conditions. Accurately characterizing pest-specific spectral responses from hyperspectral data supports early interventions, improving productivity and reducing losses. However, hyperspectral sensors generate high-dimensional and often redundant data, which is costly to acquire and process. These challenges make it difficult to isolate spectral patterns that are biologically meaningful and robust across crops. Therefore, a key research challenge lies in extracting pest-driven spectral signatures that generalize beyond individual crops, capturing stink bug damage while remaining invariant to crop-specific reflectance properties. By focusing on cross-crop characterization, this work advances the development of transferable and scalable pest monitoring systems. In this work, we propose a methodology to characterize the spectral signature of the stink bug using a reduced set of informative wavelengths rather than the full spectrum. Reflectance spectra were collected from soybean and corn leaves over six days, with plants cultivated under greenhouse conditions and stink bugs reared in a controlled laboratory environment before infestation. Prior to analysis, we filtered the raw spectra *to remove atmospheric water absorption bands and extreme regions affected by calibration instability and low signal-to-noise ratio. The proposed methodology comprises extensive experiments across multiple classification models and preprocessing techniques to extract robust signatures independent of model selection. In addition to the raw spectra, we evaluated first- and second-order derivatives to reveal subtle spectral features not evident in the original spectra. To estimate a reduced spectrum independent of both model and crop, we evaluated all combinations of crop type, preprocessing strategy, and classification model. For each experiment, SHAP importance scores were computed for every wavelength. The results were then grouped by crop, and average importance scores were calculated. To preserve crop-specific yet complementary information, we normalized and fused the averaged importance scores from both crops, ensuring that relevant bands from one crop were not discarded due to lower relevance in the other. Experimental results demonstrate that stink bug infestation can be detected using only 62 selected wavelengths, as opposed to the original 1,559 bands, with an area under the receiver operating characteristic curve (AUROC) of 0.84, Precision of 0.78, Recall of 0.76, and F1-score of 0.77 for the soybean, and AUROC of 0.82, Precision of 0.73, Recall of 0.70 and F1-score of 0.70 for the corn crop. We further constructed a performance decay curve to identify the optimal trade-off between classification performance and the number of selected wavelengths, defining the reduced set as the pest spectral signature. The results also confirm that the proposed importance fusion strategy is essential for transferring relevant spectral features across crops. Finally, the methodology is robust to limited data and independent of specific models and crops, highlighting its potential for broader application in pest detection and spectral signature characterization in precision agriculture.*

Keywords.
Spectral signature, Hyperspectral data, Pest detection, Feature extraction, Machine learning.

1. Introduction

Stink bugs (Hemiptera: Pentatomidae) are among the most economically important insect pests affecting agricultural production worldwide, causing substantial yield and quality losses in several crops, particularly soybean and maize. In Brazil, species such as *Dichelops melacanthus*, *Euschistus heros*, and *Nezara viridula* have become increasingly relevant due to the expansion of no-tillage systems and crop succession, which provide favorable conditions for population persistence throughout the year (Sosa-Gómez et al., 2020). Feeding by these insects damages vegetative and reproductive tissues, compromises plant physiological processes, reduces grain filling, and negatively affects final productivity. Consequently, the development of rapid and reliable approaches for early pest detection has become a strategic priority for sustainable crop protection and precision agriculture (Reisig et al., 2023; Mishra et al., 2025).

Conventional pest monitoring is still predominantly based on manual scouting and visual inspection performed by trained personnel. Although these methods constitute the foundation of integrated pest management programs, they are labor-intensive, time-consuming, and often incapable of providing continuous spatial information over large production areas (Nansen et al., 2021). Moreover, visible symptoms generally appear only after physiological damage has already progressed, reducing the opportunity for timely intervention. Recent advances in precision agriculture have therefore stimulated the integration of proximal and remote sensing technologies with artificial intelligence to enable automated, scalable, and non-destructive monitoring systems capable of identifying pest-induced stress before irreversible damage occurs (Alsadik et al., 2024; Saran et al., 2025).

Among the available sensing technologies, hyperspectral sensing has emerged as one of the most promising approaches for characterizing plant physiological status because it simultaneously captures detailed spectral information across hundreds or even thousands of contiguous narrow bands distributed throughout the visible (VIS), near-infrared (NIR), and shortwave infrared (SWIR) regions. Unlike multispectral systems, which are limited to a reduced number of predefined bands, hyperspectral sensors provide continuous spectral signatures that reflect variations in pigments, internal leaf structure, water content, and biochemical composition. These properties make hyperspectral data particularly suitable for detecting subtle physiological alterations caused by insect attacks before external symptoms become visually evident (Ferreira et al., 2024; Cheng et al., 2025; Hong et al., 2026).

Recent investigations have demonstrated that hyperspectral data combined with machine learning algorithms can successfully identify multiple sources of crop stress, including diseases, nutritional deficiencies, drought conditions, and insect infestations. Ensemble methods, support vector machines, artificial neural networks, and deep learning architectures have consistently achieved high predictive performance by exploiting the rich spectral information contained in hyperspectral measurements (García-Vera et al., 2024; Ualiyeva et al., 2025). Nevertheless, the practical application of these approaches remains challenging because hyperspectral datasets are intrinsically high-dimensional and contain substantial redundancy among neighboring wavelengths, increasing computational costs and reducing model interpretability. Efficient dimensionality reduction and informative wavelength selection therefore represent fundamental steps toward operational deployment in agricultural monitoring systems (Guerri et al., 2023).

The problem becomes even more complex when the objective is to develop generalized pest detection systems that can be transferred across different crop species. Most existing studies train and validate predictive models independently for each crop, causing selected spectral features to reflect crop-specific characteristics rather than physiological responses directly associated with insect feeding (Behmann et al., 2023). Consequently, wavelength subsets obtained for one host species frequently fail to generalize to another, limiting the scalability of hyperspectral monitoring systems. From both biological and engineering perspectives, identifying crop-independent spectral signatures associated with pest infestation represents an important open challenge whose solution could facilitate the future development of dedicated multispectral sensors and transferable decision-support tools for precision agriculture (Pfordt et al., 2025; Ualiyeva et al., 2025).

Recent advances in Explainable Artificial Intelligence (XAI) provide new opportunities to address this challenge by improving the interpretation of complex machine learning models. In particular, SHapley Additive exPlanations (SHAP) quantify the individual contribution of each input variable to the final prediction, allowing wavelength importance to be estimated in a mathematically consistent manner (Molnar, 2022). Beyond model interpretation, SHAP has increasingly been employed for feature ranking and dimensionality reduction in remote sensing applications, enabling the identification of compact and biologically meaningful spectral subsets while reducing dependence on specific algorithms or preprocessing strategies. Such explainability-driven methodologies contribute not only to prediction accuracy but also to scientific understanding of the physiological mechanisms underlying spectral responses (Ualiyeva et al., 2025).

Despite these recent developments, there is still limited evidence regarding the extraction of transferable hyperspectral signatures capable of characterizing stink bug infestation across multiple host crops. Most available studies focus either on individual species or on crop-specific classification frameworks, leaving unexplored the possibility of combining spectral importance information from different crops into a unified representation. A generalized spectral signature could substantially reduce sensor complexity, facilitate wavelength optimization, improve model robustness, and support the design of practical monitoring systems suitable for different agricultural environments.

Therefore, the objective of this study was to develop and evaluate a crop-independent methodology for characterizing stink bug spectral signatures using hyperspectral data acquired from soybean and maize plants. The proposed framework integrates raw reflectance spectra and derivative representations with Random Forest and Extreme Gradient Boosting classifiers, while SHAP values are employed to estimate wavelength relevance across multiple experimental configurations. Subsequently, crop-specific importance profiles are normalized and fused to identify a compact subset of informative wavelengths capable of preserving predictive performance while providing a transferable spectral signature for stink bug detection, contributing to the development of scalable, interpretable, and sensor-oriented solutions for precision agriculture.

2. Materials and Methods

2.1 Experimental Setup

The data collection experiments were carried out in collaboration with Embrapa Genetic Resources and Biotechnology, Brasília, DF, Brazil. Hyperspectral measurements were collected from soybean (*Glycine max*) and corn (*Zea mays*) plants grown under greenhouse conditions. The plants were maintained under controlled conditions before infestation to support regular development and reduce unwanted variation among samples.

The pest evaluated in this study was the stink bug (*Dichelops melacanthus*), an insect of economic importance in Brazilian agriculture (Sosa-Gómez et al. 2020). The insects were reared under laboratory conditions and later introduced into selected groups of plants to induce infestation. Plants grown under the same conditions but without insects were used as control samples.

The experiment was monitored over six days. During this period, hyperspectral measurements were collected from both control and infested plants, allowing the analysis of spectral changes associated with stink bug damage. Spectral data were acquired under natural illumination using an ASD FieldSpec 3 portable spectroradiometer (Analytical Spectral Devices Inc., Boulder, USA). Before the measurements, the instrument was calibrated with a white reference panel. The sensor was positioned approximately 50 cm above the plant canopy using an 8° field-of-view lens. The equipment records spectra in the 350–2460 nm range, covering parts of the ultraviolet (UV), visible (VIS), near-infrared (NIR), and shortwave infrared (SWIR) regions.

Reflectance values $\rho(\lambda)$ at wavelength λ were computed from the ratio between the radiance measured from the plant surface and the radiance measured from the white reference panel:

$$\rho(\lambda) = \kappa \frac{\phi_T(\lambda)}{\phi_R(\lambda)}, \tag{1}$$

where $\phi_T(\lambda)$ denotes the radiance reflected by the target plant surface, $\phi_R(\lambda)$ represents the radiance measured from the reference white panel, and κ is a calibration factor.

2.2 Hyperspectral Dataset

The dataset used in this work contains 1,965 hyperspectral reflectance samples acquired from soybean and corn plants under control and stink bug infestation conditions. Each sample corresponds to one hyperspectral signature associated with a crop species, acquisition day, and class label. Table 1 presents the distribution of samples used in this study. The dataset includes

959 control samples and 1,006 stink bug-damaged samples. Soybean contributed with 1,141 samples, while corn has 824 samples.

Table 1. Distribution of hyperspectral samples according to crop and plant condition (control or damaged).

Crop	Control	Stink Bug	Total
Soybean	485	656	1,141
Corn	474	350	824
Total	959	1,006	1,965

In this work, stink bug detection was formulated as a binary classification problem, where control samples were compared against stink bug-infested samples. The experiments were conducted separately for soybean and corn, and the resulting information was later combined to characterize a crop-independent spectral signature of stink bug infestation.

The hyperspectral signatures provide information related to biochemical and structural properties of leaves across the UV, VIS, NIR, and SWIR spectral regions (Blackburn 2007; Homolová et al. 2013). Since stink bug feeding can affect pigment concentration, tissue structure, water-related responses, and plant metabolism, these measurements provide useful information for detecting pest-induced stress and identifying informative spectral regions (Nansen et al. 2021).

2.3 Hyperspectral Preprocessing

After reflectance computation, the hyperspectral spectra were preprocessed to remove unreliable wavelength regions before feature extraction and model training. Although the sensor acquired data within the 350–2460 nm range, only wavelengths between 380 and 2250 nm were retained. This step removed extreme spectral regions more affected by calibration instability and lower signal-to-noise ratio.

In addition, two spectral intervals were excluded due to strong atmospheric water absorption effects: 1330–1440 nm and 1800–2000 nm. These regions are commonly associated with reduced reliability in hyperspectral reflectance measurements and may introduce noise unrelated to plant physiological responses. After applying the wavelength range filtering and removing the water absorption bands, 1,559 valid wavelengths were retained for the experiments. Fig. 1 illustrates the resulting hyperspectral signal after preprocessing.

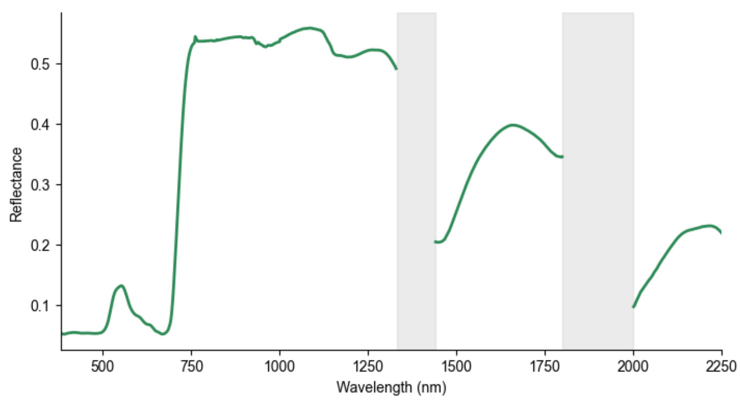


Fig. 1. Average hyperspectral reflectance spectrum for soybean samples after preprocessing. Gray shaded regions indicate wavelength intervals removed due to atmospheric water absorption effects.

2.3 Feature Representations

As feature representations used in this work, three spectral representations were evaluated: raw reflectance, first-order derivative, and second-order derivative. The raw representation corresponds to the preprocessed reflectance spectrum without additional transformation, preserving the original spectral response measured from the leaves.

Derivative-based representations were also considered to highlight local changes along the spectrum. The first derivative describes how the reflectance changes with wavelength, while the second derivative represents the change in this rate, emphasizing the curvature of the spectral signal. These representations can be written as

$$\mathbf{y}_{\mathrm{D1}} = \frac{d\mathbf{x}}{d\lambda}, \quad \mathbf{y}_{\mathrm{D2}} = \frac{d^2\mathbf{x}}{d\lambda^2}, \tag{2}$$

where $\mathbf{x}$ is the reflectance spectrum, λ is the wavelength, and $\mathbf{y}_{\mathrm{D1}}$ and $\mathbf{y}_{\mathrm{D2}}$ correspond to the first and second-order derivatives, respectively.

Using derivative representations can help reduce the effect of absolute reflectance differences and make subtle spectral variations more visible. This is relevant for pest detection since stink bug damage may produce small changes in the spectral curve that are not easily observed in the raw signal alone. In this work, the first and second derivatives were estimated using the Savitzky-Golay filter (Savitzky and Golay 1964), as implemented in the SciPy library (Virtanen et al. 2020).

2.4 Classification Models

During the evaluation step, two tree-based classifiers were used to evaluate the spectral representations considered in this work: Random Forest (RF) and Extreme Gradient Boosting (XGBoost). These methods were chosen since they are well established for supervised classification problems and are commonly applied without extensive manual feature design.

The Random Forest (RF) (Breiman 2001) combines several decision trees trained on different bootstrap samples of the training data. At each tree split, only a random subset of features is considered, which helps increase diversity among trees. The final class prediction is obtained from the aggregation of the individual tree outputs. Due to this ensemble structure, RF is robust to noisy variables and less prone to overfitting than a single decision tree, although complex problems may require a larger number of trees.

XGBoost (Chen and Guestrin 2016) was also evaluated as a second classifier. In contrast to RF, XGBoost builds trees sequentially, where each new tree is optimized to reduce the errors made by the previous ensemble. The method includes regularization mechanisms and is widely used for structured data classification. However, its performance can be more affected by hyperparameter selection when compared with RF.

In this work, RF and XGBoost were trained independently for each feature representation, including raw reflectance, first-derivative spectra, and second-derivative spectra. The use of two different classifiers aimed to make the analysis less dependent on a single model. Therefore, wavelength relevance patterns that were consistent across both classifiers were considered more reliable for the characterization of stink bug spectral signatures.

As for the implementation details, both models were implemented using standard Python machine learning libraries. The RF classifier was implemented with *scikit-learn* using 150 trees and a maximum depth of 6. The XGBoost classifier was implemented using the XGBoost Python library, with 150 estimators, maximum depth equal to 4, subsampling ratio of 0.8, and binary logistic objective function.

2.5 Evaluation Protocol

Stink bug detection was treated as a binary classification problem for each crop. In this setup, stink bug-infested samples were assigned to the positive class, while control samples were assigned to the negative class. The same evaluation procedure was applied to all feature representations considered in this work, namely raw reflectance, first derivative, and second derivative.

For each experiment, the data were divided using random sampling to preserve the proportion of control and infested samples in each subset. First, 30% of the samples were separated as the test set. From the remaining 70%, 10% was used as a validation set for hyperparameter tuning and wavelength subset selection. Therefore, the final data partition corresponded to 63% for training, 7% for validation, and 30% for testing. The random seed was fixed in all experiments to ensure reproducibility, and the same split was used across all experiments.

The model performance was evaluated using precision, recall, $F1$-score, and area under the receiver operating characteristic curve (AUROC) (Bishop 2006). These metrics were computed from the confusion matrix, using true positives (T_P), true negatives (T_N), false positives (F_P), and false negatives (F_N). Precision (P) measures the proportion of samples predicted as infested that were correctly classified, while recall (R) measures the proportion of infested samples correctly detected:

$$P = \frac{T_P}{T_P + F_P}, \quad R = \frac{T_P}{T_P + F_N}. \tag{3}$$

The $F1$-score combines precision and recall through their harmonic mean, that is:

$$F1 = 2\,\frac{P \cdot R}{P + R} = \frac{2T_P}{2T_P + F_P + F_N}. \tag{4}$$

It is important to mention that reported $F1$ metric corresponds to Macro-$F1$, and both terms are used interchangeably throughout the text. Macro-$F1$ is obtained by computing the $F1$-score separately for each class and then averaging the class-wise scores with equal weight.

AUROC was also used as a threshold-independent metric to measure the ability of the classifier to separate control and infested samples. It is obtained from the ROC curve, which relates the true positive rate (T_{PR}) and the false positive rate (F_{PR}) over different decision thresholds:

$$T_{PR} = \frac{T_P}{T_P + F_N}, \quad T_{PR} = \frac{F_P}{F_P + T_N}. \tag{5}$$

AUROC values close to 1 indicate strong class separation, while values close to 0.5 indicate performance similar to a random classification.

2.5 SHAP Importance and Aggregation

Estimating the contribution of each wavelength to stink bug detection is an important step in the spectral signature characterization. To do that, the SHAP values (Shapley Additive exPlanations) (Lundberg and Lee 2017) were computed as explainable method for each trained classifier. SHAP assigns an additive contribution value to each input feature for a given prediction, allowing the influence of each wavelength on the model output to be quantified. In this work, SHAP was used not only for model interpretation, but also as a wavelength relevance measure for spectral signature extraction.

For each experimental configuration, defined by crop, spectral representation (raw, first and second derivatives), and classification model, SHAP values were computed for all retained wavelengths. Since the objective was to obtain a global relevance score per wavelength, the absolute SHAP values were averaged across samples. Thus, for a given experiment p, the importance of wavelength λ was computed as

$$I_{\mathrm{p}}(\lambda) = \frac{1}{N_{\mathrm{p}}} \sum_{i=1}^{N_{\mathrm{p}}} \left|\phi_{i,\lambda}^{\mathrm{p}}\right|, \tag{6}$$

where $\phi_{i,\lambda}^{\mathrm{p}}$ is the SHAP value assigned to wavelength λ for sample i, and N_{p} is the number of samples used for the SHAP analysis in experiment p. This procedure produced one wavelength-importance curve for each combination of crop, spectral representation, and classifier.

The final crop-specific importance profile was obtained by averaging the SHAP-based importance curves across all experiments associated with the same crop (per-crop aggregation). In this work, the experiments included raw reflectance, first-derivative, and second-derivative spectra, combined with Random Forest and XGBoost classifiers. This averaging step reduced the dependence of the wavelength ranking on a specific model or preprocessing strategy.

2.6 Cross-Crop Fusion Module and Spectral Subset Selection

After computing the average SHAP importance curves for each crop (soybean and corn), a fusion module was applied to obtain a single cross-crop importance profile. The goal of this step was to identify wavelengths that remain informative across crop-dependent spectral variability, while avoiding a wavelength selection biased toward only one crop.

First, each crop-specific importance curve was independently normalized using min–max normalization. Then, the normalized soybean ($I_{\text{soybean}}(\lambda)$) and corn ($I_{\text{corn}}(\lambda)$) importance curves were then summed wavelength by wavelength:

$$I_{\text{fused}}(\lambda) = \text{min-max}\left(I_{\text{soybean}}(\lambda) + I_{\text{corn}}(\lambda)\right), \tag{7}$$

where $\text{min-max}(\cdot)$ is the min-max normalization to the interval $[0, 1]$ and $I_{\text{fused}}(\lambda)$ is the final fused importance curve.

This fused curve was used as the final wavelength relevance profile. To obtain compact spectral subsets, different thresholds were applied to $I_{\text{fused}}(\lambda)$. For each threshold thr, the selected wavelength set was defined as

$$\mathcal{S}_{thr} = \{\lambda: I_{\text{fused}}(\lambda) > thr\}. \tag{8}$$

Each selected subset $\mathcal{S}_{thr}$ was then used to train and evaluate new classification models. This enabled the construction of a performance decay analysis, relating the number of selected wavelengths to the classification performance. The final spectral signature was chosen from this trade-off between compactness and predictive performance.

2.7 Spectral Signature Extraction Pipeline

The complete workflow of the employed spectral signature extraction pipeline is depicted in Fig. 2. The main modules of the pipeline, including spectral preprocessing, feature representations (derivatives), training classification models, SHAP importance estimation, cross-crop fusion, and threshold-based subset selection, were described in the previous subsections.

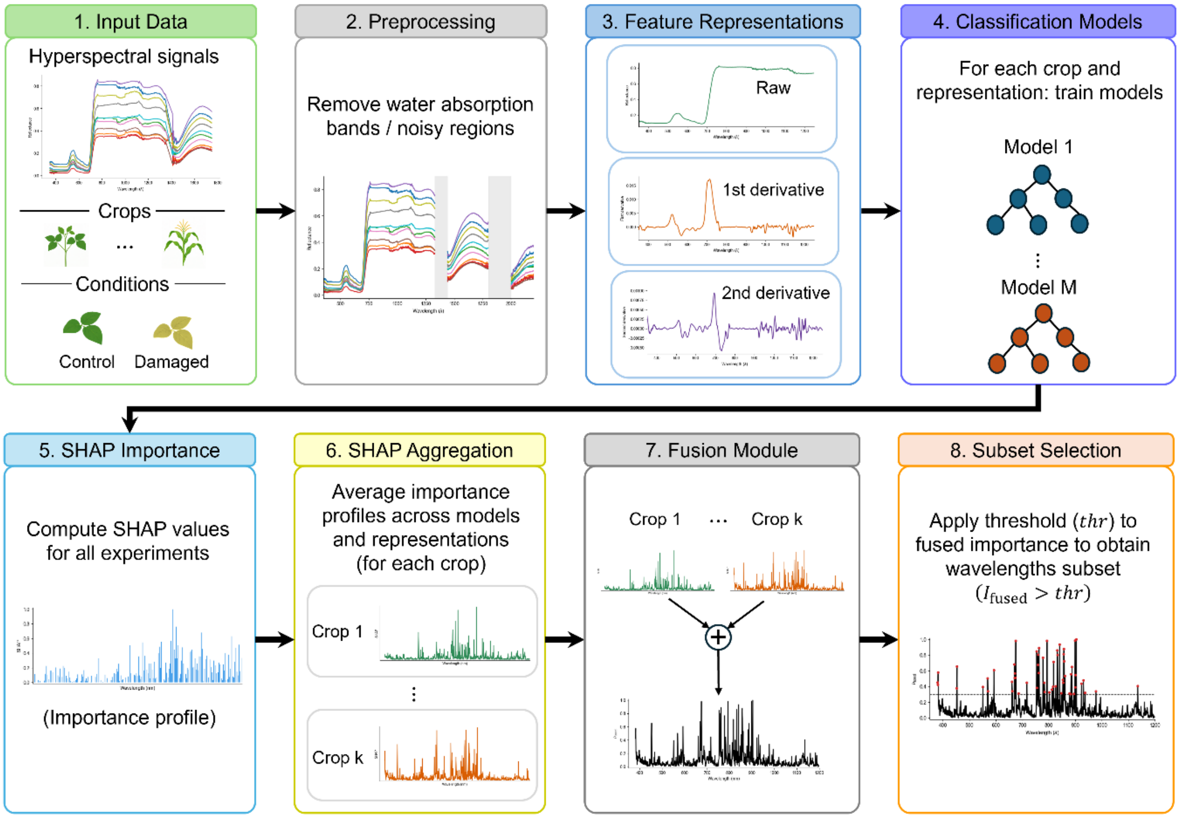


Fig. 2. Overview of the proposed pipeline for extracting cross-crop spectral signatures. SHAP importance profiles are computed for each experiment, aggregated by crop, fused across crops, and thresholded to select the final wavelength subset.

The workflow starts from hyperspectral signals acquired from multiple crops and experimental conditions. After removing unreliable spectral regions, different spectral representations are generated and used to train binary classifiers for the target condition. For each experiment, defined by a crop, spectral representation, and classifier, SHAP importance values are computed for all wavelengths. These importance profiles are averaged per crop and then combined by the fusion module to produce a single cross-crop relevance curve. Finally, thresholding is applied to this fused curve to obtain reduced wavelength subsets, which are re-evaluated to select the final compact spectral signature.

In this work, the pipeline illustrated in Fig. 2 was instantiated for stink bug detection using soybean and corn spectra, raw reflectance, first derivative, second derivative, and two classification models: Random Forest and XGBoost.

3. Experimental Results and Discussion

The experimental results were organized to evaluate the proposed pipeline in terms of cross-crop spectral importance, threshold-based wavelength reduction, fusion strategy ablation, and characterization of the final reduced spectral signature.

3.1 Cross-Crop Spectral Importance Analysis

To characterize stink bug-induced spectral responses independently of a specific classifier or crop, SHAP importance values were computed for all experiments combining the evaluated classifiers and spectral representations, namely raw reflectance, first-order derivative, and second-order derivative. For each crop, the SHAP importance values were averaged across models and preprocessing strategies, resulting in one mean importance profile for soybean and one mean importance profile for corn. The crop-specific importance curves were then fused according to Eq. (7) to obtain a single cross-crop relevance profile.

Fig. 3 presents the resulting average importance profiles for soybean, corn, and the final fused importance curve. The soybean and corn curves exhibit partially overlapping but non-identical relevance patterns, indicating that some spectral regions are consistently informative across both crops, while others contribute more strongly in a crop-dependent manner. This behavior reinforces the importance of combining information from multiple crops during wavelength selection, since a crop-specific ranking may fail to retain wavelengths that are highly relevant in another crop.

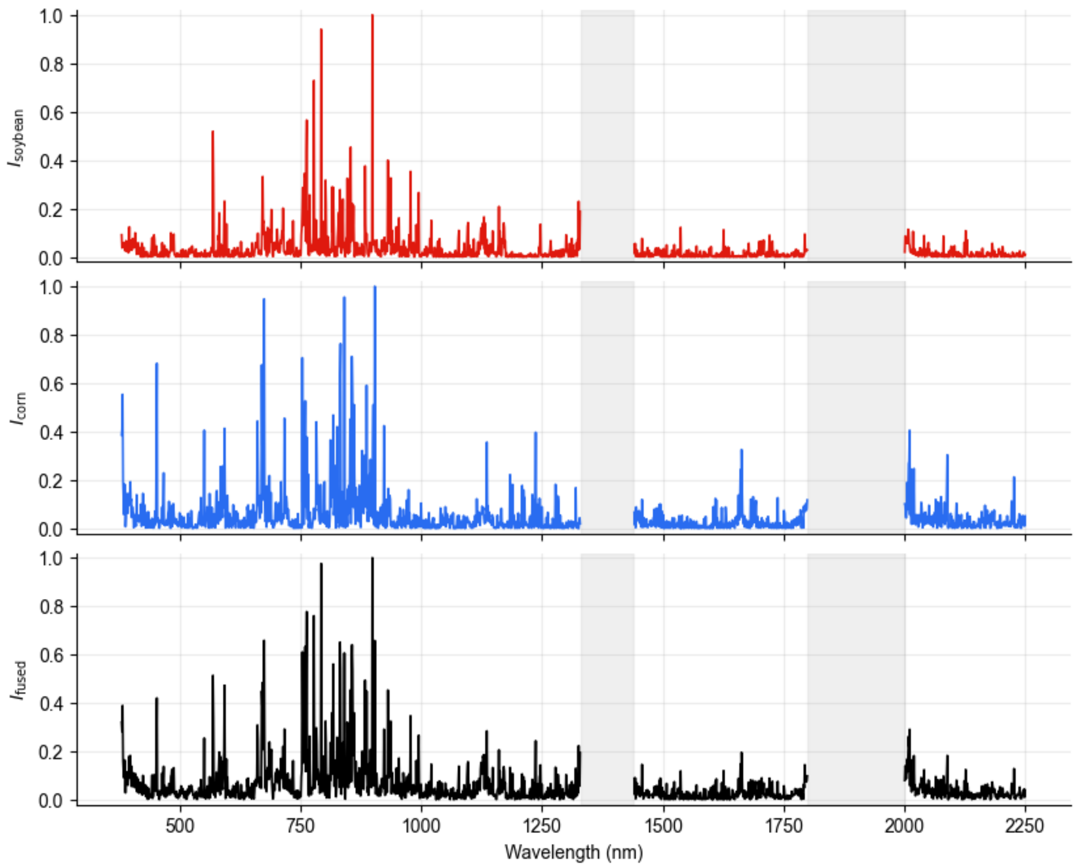


Fig. 3. Normalized SHAP importance profiles for soybean (I_{soybean}), corn (I_{corn}), and the fused cross-crop importance curve (I_{fused}). Gray regions indicate excluded spectral intervals.

The fused importance curve summarizes spectral regions that remain informative after combining the crop-specific relevance profiles. Before fusion, the soybean and corn importance curves were independently normalized, ensuring that both crops contributed on a comparable scale. The normalized curves were then summed, allowing wavelengths relevant to either soybean or corn to contribute to the final cross-crop importance score. This strategy reduces the influence of crop-dependent magnitude differences and provides a suitable basis for defining a crop-independent spectral signature of stink bug infestation.

From a methodological perspective, these results support the use of multiple spectral representations and classification models during importance estimation. Since the importance profiles were averaged across raw, first-derivative, and second-derivative spectra, as well as across Random Forest and XGBoost, the resulting fused signature is less dependent on a specific preprocessing strategy or model configuration. This increases the robustness of the selected wavelengths and strengthens their interpretation as pest-related rather than model-specific spectral features.

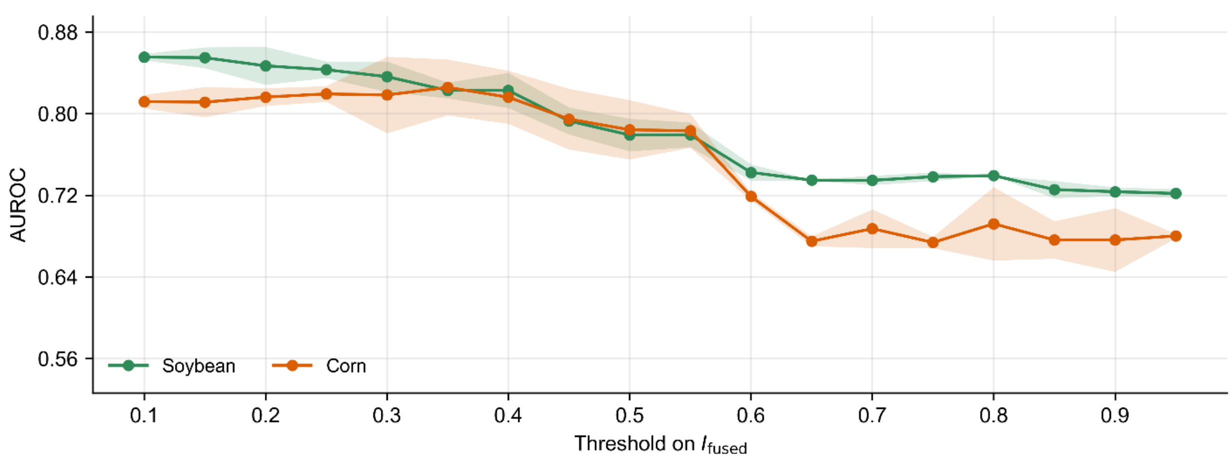
3.2 Threshold-Based wavelength reduction using fused importance

After obtaining the fused importance curve, a threshold-based selection procedure was applied to identify compact wavelength subsets. For different threshold values, only wavelengths with I_{fused} above the threshold were retained, and new classification experiments were performed using the resulting reduced spectral subset (the selected wavelengths with highest final importance). Table 2 summarizes the number of wavelengths retained for each threshold value. This procedure enabled the construction of a performance decay curve relating classification performance to the number of selected wavelengths.

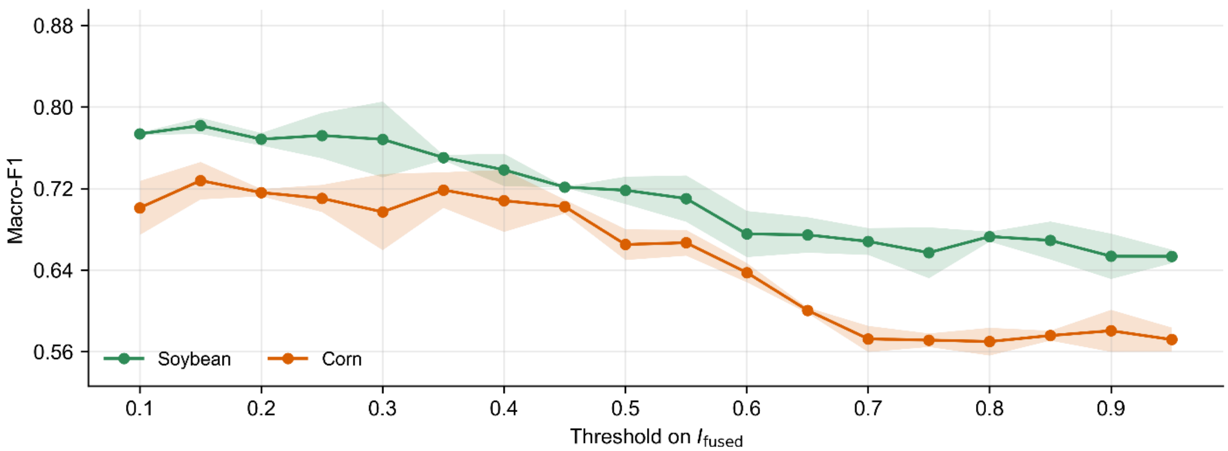
Table 2. Number of selected wavelengths retained at each threshold of the fused importance curve. The selected operating point used in the paper is highlighted in bold.

Threshold	#Selected Wavelengths
0.10	329
0.15	176
0.20	115
0.25	76
0.30	**62**
0.35	49
0.40	41
0.45	33
0.50	29
0.55	24
0.60	20
0.65	18
0.70	15
0.75	13
0.80	12
0.85	8
0.90	5
0.95	4

Fig. 4 shows the performance-drop analysis obtained by thresholding the fused importance curve. Specifically, Fig. 4(a) presents the AUROC values, while Fig. 4(b) shows the Macro-F1 scores obtained for soybean and corn using the wavelength subsets selected at each threshold. As expected, increasing the threshold progressively reduced the number of retained wavelengths, yielding a more compact spectral representation. Importantly, the results indicate that classification performance remained relatively stable over a range of threshold values, showing that a large portion of the original hyperspectral spectrum is redundant for stink bug detection. Only under more aggressive thresholding did performance begin to decline more noticeably, suggesting that relevant spectral information was being removed.



(a) Performance decay for AUROC



(b) Performance decay for Macro-F1.

Fig. 4. Performance decay obtained by thresholding the fused importance curve (I_{fused}). Curves show soybean and corn performance using the wavelength subsets selected at each threshold.

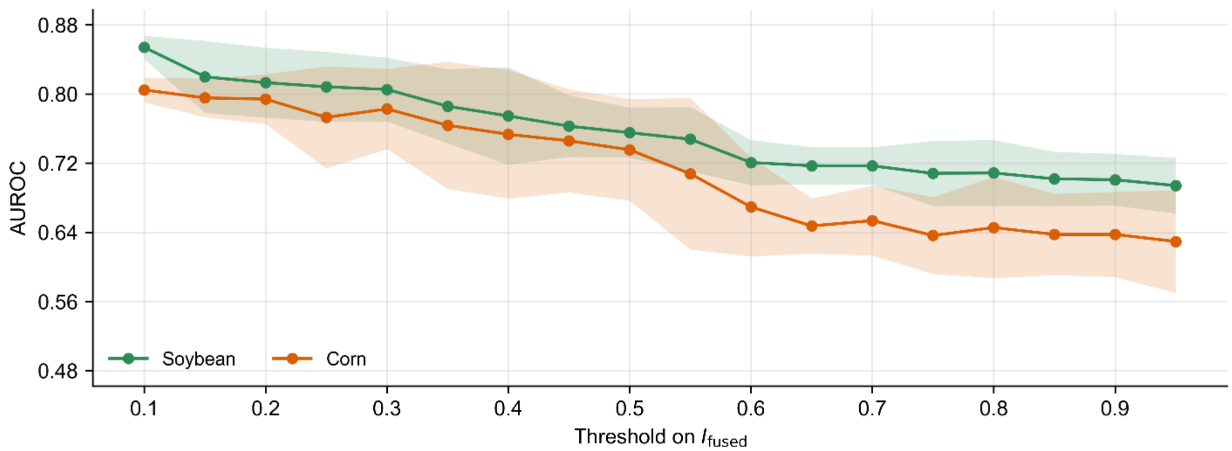
This behavior demonstrates that the proposed methodology can identify a reduced set of wavelengths while preserving most of the predictive information contained in the original hyperspectral signal. Based on the trade-off between compactness and classification performance, the selected operating point corresponded to 62 wavelengths. Using this reduced spectral signature, stink bug detection achieved an AUROC of approximately 0.84 and a Macro-$F1$ score of approximately 0.77, confirming that the selected wavelengths preserve strong discrimination performance despite the substantial dimensionality reduction.

These findings highlight that the objective of the proposed methodology is not only dimensionality reduction, but the extraction of a compact and informative spectral signature associated with stink bug infestation. The threshold analysis provides a principled way to define this signature by balancing spectral compactness and predictive performance.

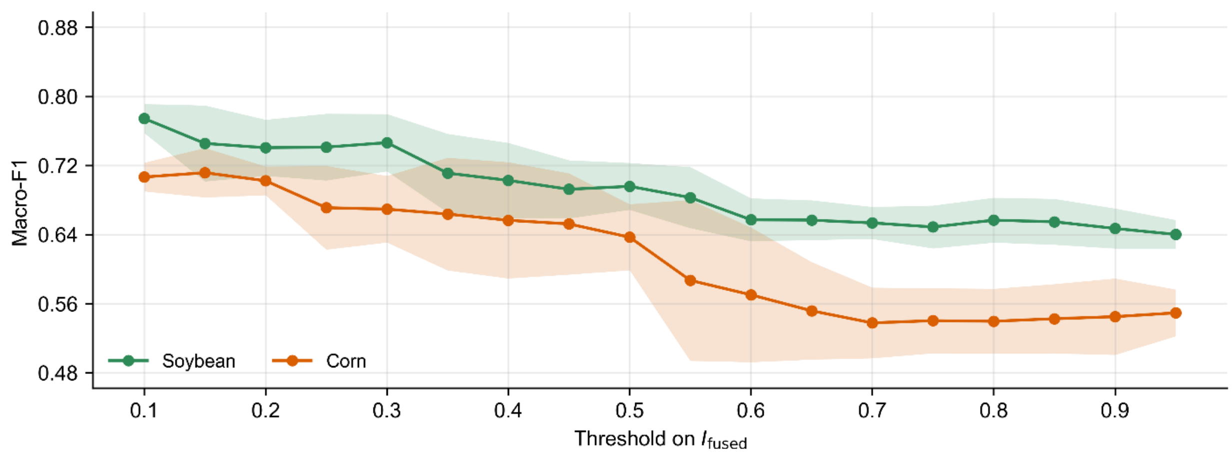
3.3 Ablation Study: Fused Importance Versus Soybean-Only Importance

To evaluate the contribution of the cross-crop fusion strategy, an ablation study was performed by repeating the threshold-based wavelength selection using only the soybean importance curve (I_{soybean}) instead of the fused importance curve (I_{fused}). The selected wavelength subsets were then evaluated for both soybean and corn using the same experimental protocol.

Fig. 5 shows the performance decay obtained when thresholding was applied to I_{soybean}. Fig. 5(a) presents AUROC, and Fig. 5(b) presents Macro-F1. Compared with the fused strategy in Fig. 4, soybean-only thresholding showed a decreasing trend from the beginning of the curve and lower average performance, especially for corn. At the selected threshold of 0.30, corresponding to 62 wavelengths, the fused strategy improved AUROC from 0.81 to 0.84 for soybean and from 0.78 to 0.82 for corn. Macro-F1 also increased from 0.75 to 0.77 for soybean and from 0.67 to 0.70 for corn.



(a) Performance decay for AUROC



(b) Performance decay for Macro-F1.

Fig. 5. Performance decay obtained by thresholding the fused importance curve ($I_{soybean}$). Curves show soybean and corn performance using the wavelength subsets selected at each threshold.

These results indicate that selecting wavelengths from a single crop may bias the reduced signature toward crop-specific spectral patterns, limiting transferability. In contrast, the fused strategy incorporates relevance information from both soybean and corn, producing a more stable performance curve and a more robust cross-crop spectral signature for stink bug characterization.

3.4 Discussion of the proposed spectral signature

Experimental results indicate that stink bug infestation can be characterized using a compact subset of wavelengths identified through the proposed pipeline. At the selected threshold (0.30), the final spectral signature contained 62 wavelengths. These wavelengths were distributed across the visible, RE/NIR, and SWIR regions, rather than being concentrated in a single narrow interval. Table 3 summarizes the selected wavelengths as spectral groups.

Table 3. Selected wavelength groups obtained at the selected threshold 0.30, corresponding to 62 wavelengths.

Spectral region	Selected wavelength groups (nm)	Number of wavelengths
Visible	380–382, 452–453, 551, 569–570, 593, 661–675, 686	18
RE / NIR	717, 754–763, 778–861, 878–904, 923–937, 978	37
SWIR	1136, 1237, 1663, 2009–2011	7

The selected signature was mainly concentrated in the visible and RE/NIR regions, with additional wavelengths in the SWIR domain. This distribution suggests that stink bug-induced stress is reflected by multiple spectral mechanisms, including changes related to pigment absorption, leaf structure, and water-sensitive spectral responses. Therefore, the proposed signature should not be interpreted as a single isolated band, but as a compact set of complementary discrete wavelengths that jointly contribute to pest discrimination.

Considering the quantitative analysis, the reduced signature preserved discrimination performance, achieving an AUROC of 0.84 and Macro-F1 of 0.77 for soybean, and an AUROC of 0.82 and Macro-F1 of 0.70 for corn. Table 4 summarizes the final performance on the test set obtained with the 62 selected wavelengths, also with Precision and Recall metrics.

Table 4. Performance obtained on the test set with the selected 62-wavelength signature at threshold 0.30.

Crop	AUROC	Precision	Recall	Macro-F1
Soybean	0.84	0.78	0.76	0.77
Corn	0.82	0.73	0.70	0.70

An important aspect of the proposed methodology is its robustness. By combining raw, first-derivative, and second-derivative spectra, the method captures both baseline reflectance behavior (from the raw signal) and local spectral variations (from the derivatives). By averaging SHAP values across Random Forest and XGBoost, it reduces dependence on a single classifier. Finally, by normalizing and fusing the crop-specific importance curves, it emphasizes wavelengths that remain informative across soybean and corn, which is essential for deriving a transferable spectral signature.

With this approach, instead of relying on the full hyperspectral spectrum, the method identifies a compact subset of wavelengths that preserves relevant information for stink bug detection. This can support the future design of multispectral sensors or targeted acquisition protocols for pest monitoring across different crops.

4. Conclusion

This work presented a crop- and model-independent pipeline to characterize stink bug spectral signatures from hyperspectral leaf data. The method combined raw reflectance, first-derivative, and second-derivative spectra with Random Forest and XGBoost classifiers, using SHAP importance values to identify wavelengths that remain informative across soybean and corn.

The proposed fusion strategy and threshold-based strategy selected a compact signature of 62 wavelengths while preserving strong detection performance, with AUROC of 0.84 and Macro-F1 of 0.77 for soybean, and AUROC of 0.82 and Macro-F1 of 0.70 for corn. The selected wavelengths were distributed across visible, NIR, and SWIR regions, indicating that stink bug-induced stress is represented by complementary spectral responses. The ablation study showed that fused importance produced better transferability than soybean-only selection, confirming the relevance of cross-crop fusion.

Finally, the results demonstrate that stink bug infestation can be characterized using a reduced and interpretable spectral signature. Future work should validate the pipeline on additional crops, pests, field conditions, classifiers, and spectral transformations, as well as investigate the use of the selected wavelengths for reduced-band sensor design.

Acknowledgments

This work was supported in part by the São Paulo Research Foundation (FAPESP) under Grants #2019/07665-4, #2024/15430-5, #2025/10840-3, and #2025/08971-2, by the Brazilian National Council for Scientific and Technological Development (CNPq) under Grants #406417/2022-9, # 403307/2025-2, and #306788/2023-3, and by the National Natural Science Foundation of China (NSFC) under Grant #W2412112.

References

Alsadik, B., et al. (2024). Hyperspectral remote sensing for crop monitoring and precision agriculture: Recent advances and future perspectives. *Remote Sensing, 16*(23), 4371.

Behmann, J., Steinrücken, J., & Plümer, L. (2023). Hyperspectral imaging for plant disease and stress detection: Recent advances and future challenges. *Precision Agriculture*.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer.

Blackburn, G. A. (2007). Hyperspectral remote sensing of plant pigments. *Journal of Experimental Botany, 58*(4), 855–867.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

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). New York, NY, USA: ACM.

Ferreira, M. P., et al. (2024). Hyperspectral sensing and machine learning for early crop stress detection. *Frontiers in Plant Science*.

García-Vera, D., et al. (2024). Machine learning and hyperspectral feature selection in agricultural applications. *Sustainability, 16*, 6064.

Guerri, G., et al. (2023). Explainable machine learning for hyperspectral remote sensing: A review. *Remote Sensing*.

Homolová, L., Malenovský, Z., Clevers, J. G. P. W., García-Santos, G., & Schaepman, M. E. (2013). Review of optical-based remote sensing for plant trait mapping. *Ecological Complexity, 15*, 1–16.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*.

Mahlein, A. K. (2022). Plant disease detection by imaging sensors: Parallels and specific demands for precision agriculture and plant phenotyping. *Plant Disease*.

Molnar, C. (2022). *Interpretable Machine Learning* (2nd ed.). Lulu Press.

Nansen, C., Murdock, M., Purington, R., & Marshall, S. (2021). Early infestations by arthropod pests induce unique changes in plant compositional traits and leaf reflectance. *Pest Management Science, 77*(11), 5158–5169.

Savitzky, A., & Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least-squares procedures. *Analytical Chemistry, 36*(8), 1627–1639.

Sosa-Gómez, D. R., Corrêa-Ferreira, B. S., Kraemer, B., Pasini, A., Husch, P. E., Vieira, C. E. D., & Lopes, I. O. N. (2020). Prevalence, damage, management and insecticide resistance of stink bug populations (Hemiptera: Pentatomidae) in commodity crops. *Agricultural and Forest Entomology, 22*(2), 99–118.

Ualiyeva, A., et al. (2025). Explainable artificial intelligence for hyperspectral feature selection in plant phenotyping. *Biology, 14*(12), 1715.

Verrelst, J., et al. (2021). Optical remote sensing and machine learning in vegetation monitoring: Current status and perspectives. *ISPRS Journal of Photogrammetry and Remote Sensing*.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods, 17*(3), 261–272.

Wang, H., et al. (2023). Feature selection strategies for hyperspectral data analysis in precision agriculture. *Computers and Electronics in Agriculture*.

Zhao, L., et al. (2024). Explainable artificial intelligence for precision agriculture: Integrating machine learning and hyperspectral sensing. *Remote Sensing*.