João de Deus Ferreira e Silva[a]; André Oliveira Françani[a]; Ednaldo José Ferreira[b], Lúcio André de Castro[b], Joaquim Cezar Felipe[a], 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]National Research Center for Development of Agricultural Instrumentation, Brazilian Agricultural Research Agency, São Carlos, 13561-206, SP, Brazil.

A stability-driven framework for robust plant spectral signature identification

ABSTRACT - Accurate identification of agricultural crops through spectral signatures poses a critical challenge for large-scale phytosanitary monitoring. This work proposes a stability-based framework for the robust identification of plant spectral signatures, applied to the discrimination of soybean (Glycine max) in comparison of maize (Zea mays) and cotton (Gossypium hirsutum) under conditions of biotic damage caused by the insect pest Spodoptera frugiperda and stink bugs. Spectral data were collected with an ASD FieldSpec 3 portable spectroradiometer (350–2500 nm) over eight consecutive days, totaling 1,436 spectral samples from control and damaged plants. After pre-processing with first- and second-order derivatives, machine learning algorithms — Random Forest (RF) and Extreme Gradient Boosting (XGBoost) — were applied to classify the species. The results demonstrated high predictive performance, with AUC of 0.99 for the control data and between 0.97 and 0.99 for the damaged plant data. The sequential reclassification framework allowed the identification of 63 wavelengths representative of the stable spectral signature of soybean, concentrated in the regions of UV (380–400 nm), visible, NIR (870–900 nm and 1300–1450 nm) and SWIR (1700–2100 nm). The results indicate that the proposed methodology is effective for the definition of robust spectral signatures, with potential application in remote sensing systems and field-scale pest monitoring.

Keywords: *remote sensing; machine learning; soybean; spectral signature; Spodoptera frugiperda; reflectance spectroscopy.*

1. Introduction

Soybean (Glycine max (L.) Merrill) has been demonstrating its important role in the production of grains in Brazilian agriculture, sustaining both domestic food security and international trade (Vargas-Almendra et al., 2024). In the last decade, Brazil has consolidated its position as the world's largest producer of soybeans, with a record production of 169 million tons in the 2024/25 harvest, cultivated on about 47.4 million hectares (USDA, 2024; Santos et al., 2023; Crusiol et al., 2023). This oilseed represents almost 60% of global soybean exports, constituting a pillar of the national economy and a crucial source of foreign exchange, which makes soybean an engine of agronomic innovation and an object of growing scientific interest (Falcioni et al., 2024).

In this context, the phytosanitary protection of crops is crucial for maintaining productivity. Among the main biotic agents that threaten agricultural production, pest insects stand out, especially the fall armyworm Spodoptera frugiperda (J.E. Smith, 1797) (Lepidoptera: Noctuidae), which attacks a wide range of host crops — including corn, cotton, and soybeans — and is responsible for significant production losses worldwide (Barros et al., 2010). Stink bugs (Hemiptera: Pentatomidae) are also relevant pests, causing damage by direct feeding to developing grains, with direct effects on soybean quality and yield (Bi et al., 1997).

Efficient and early monitoring of infestations is, therefore, an indispensable step in integrated pest management. Traditional phytosanitary evaluation techniques, based on direct visual inspection, demand a high level of specialized labor and are subject to the subjectivity of the observer, making them unfeasible for the scale of large Brazilian crops. In this scenario, hyperspectral remote sensing emerges as an alternative with great potential, since plants stressed by herbivory modify their biochemical and anatomical leaf patterns, which is reflected in detectable changes in the spectral reflectance profile (Eisenring et al., 2019; Prabhakar et al., 2020).

Reflectance spectroscopy in the range 350–2500 nm (VIS-SWIR) has been widely used to discriminate plant species, identify abiotic and biotic stresses, and monitor plant biochemical attributes, such as chlorophyll, leaf water, and phenolic compounds (Clark & Roush, 1984; Buddenbamm & Steffens, 2012; Martinelli et al., 2015). The use of first- and second-order spectral derivatives is a well-established procedure to amplify absorption characteristics and remove baseline trends, increasing the resolution of the relevant spectral peaks (Rinnan et al., 2009).

Associated with spectroscopy, machine learning algorithms — such as Random Forest (RF) and Extreme Gradient Boosting (XGBoost) — have demonstrated excellent performance in crop classification and detection of pest damage (Osco et al., 2019a; Zhang et al., 2019; Singh et al., 2020; Osco et al., 2020). The ability of these models to handle high-dimensional data, such as reflectance spectra with hundreds of wavelengths, and to automatically extract the most informative variables makes them especially suitable for constructing robust spectral signatures (Martins et al., 2017).

Few studies address the issue of spectral signature stability over time, which is critical for the implementation of continuous monitoring systems. The definition of a minimum and stable set of wavelengths that allows for high-performance discrimination of species of interest — regardless of the state of damage — represents a gap in the literature.

In view of the above, the present work proposes a stability-based framework for the robust identification of plant spectral signatures. The method combines the acquisition of spectral data in multiple days, derivative pre-processing, RF and XGBoost classification and a sequential reclassification procedure with progressive reduction of wavelengths, in order to determine the minimum set of spectral signals that sustain high predictive performance. The study was conducted involving soybean, corn and cotton crops, exposed or not to damage by Spodoptera frugiperda and stink bugs, in order to evaluate both the discrimination between species and the robustness of the spectral signature in the face of adverse biotic conditions. In view of the above, this work aims to find the spectral signature of soybean against corn and cotton crops.

2 Material and methods

2.1 Location of the study area.

The experiments were conducted in an outdoor area, inside a greenhouse, in association with Embrapa - Genetic Resources and Biotechnology in Brasília, DF, Brazil.

2.2 Cultivation of soybean crop.

The soybean plants were cultivated for approximately two weeks after seed emergence and had fully expanded leaves. After this period, the experiment was conducted for approximately 8 days, during which time the spectral behavior of the plants was measured.

2.3 Cultivation of corn crops.

Corn seeds were obtained from the Embrapa Maize and Sorghum Germplasm Bank, in Sete Lagoas, MG, Brazil (19°27'57 " S and 44° 14'48 " W) and germinated on moistened paper. After 4 days, the seeds were transplanted into pots with a mixture of soil and organic substrate (in the proportion of 1:1 m/m) and kept in a greenhouse (photoperiod of 14 h). The plants used in the experiments were cultivated for 9 to 10 days after emergence and had three fully expanded leaves.

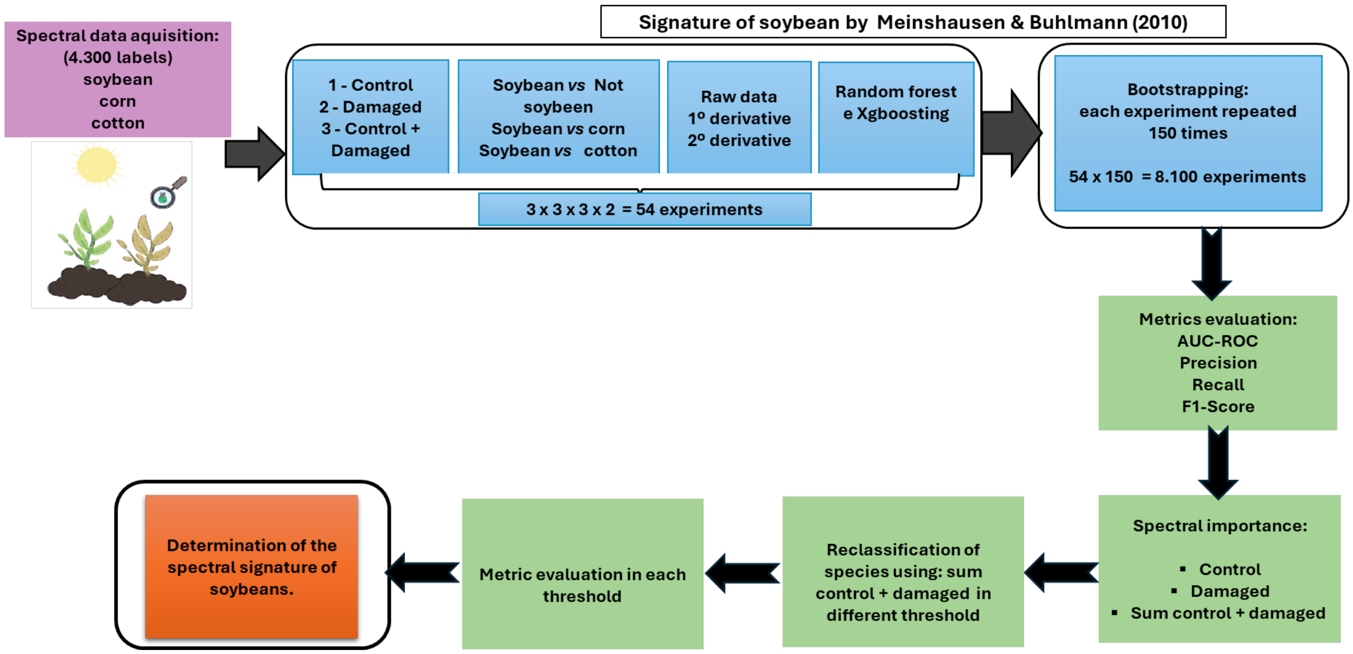
2.4 Cultivation of cotton crops.

Cotton plants (var. Delta Opal) used were 6 weeks old, in vegetative stage (with up to 6 expanded leaves and about 30 cm in height).

2.5 Insect production

Specimens of Spodoptera frugiperda were kept in separate climatized rooms at 27 ± 1 °C, with 65 ± 10% relative humidity and a photoperiod of 14 h. The larvae of S. frugiperda were obtained from a laboratory colony maintained at Embrapa Genetic Resources and Biotechnology, in Brasília, DF, Brazil. The larvae were raised in plastic containers with an artificial diet based on beans (Phaseolus vulgaris). Second instar larvae (Schmidt et al., 2009) were used in the experiments and fasted for 24 h before the experiment.

The stink bugs were raised in 8 L plastic containers, fed with soybean seeds (cv. Conquista), sunflower seeds (Helianthus annuus), raw peanuts (Arachis hypogaea), fresh green beans (Phaseolus vulgaris) and water. The food was renewed twice a week. To provide substrate for oviposition and shelter, a 15 cm² nylon screen was placed inside the container. The stink bugs were kept in a room with a controlled environment, photoperiod of 14 hours of light and 10 hours of darkness (L14:D10), temperature of 26 ± 0.3 °C and relative humidity of 70 ± 10%.



2.6 Spectral Data Acquisition and Organization

For the spectral reading, insects were removed from the plants and measurements of damaged and undamaged plants (healthy/control) were performed, with direct exposure of the plants to sunlight. Spectral measurements were conducted with a high-resolution ASD FieldSpec 3 portable spectroradiometer (Analytical Spectral Devices Inc., Boulder, USA). Calibration of the instrument was performed using a white Spectralon reference panel, immediately prior to the first measurement. The equipment was positioned 25 cm high in relation to the sheet, and an 8-degree lens was used to capture the readings. The procedure was repeated periodically during the measurement process. This equipment operates in the spectral range of 350–2500 nm (VIS-SWIR) with spectral resolution of 1.4 nm between 350 and 1000 nm and 2 nm between 1000 and 2500 nm. The device was connected to a portable microcomputer, where the data was stored after being obtained.

Two spectral regions were removed from the spectral dataset due to absorption by water vapor in the atmosphere: 1350–1410 nm and 1820–1940 nm, and 2460–2500 nm (Jensen, 2014). In addition to these regions mentioned above, the final and initial regions between 350 – 380 and 2250 – 2500 were removed, which are characterized by high noise from the spectroradiometer signal. Spectral measurements were taken between 9 a.m. and 3 p.m. local time. This range is recommended for the collection of spectral measurements considering a more regular solar irradiation in plants (Jensen, 2014). The experiment was monitored for eight consecutive days, and spectral measurements were extracted according to each day of analysis. A total of 1436 spectral samples were collected almost uniformly over the eight days.

2.7 Data processing and machine learning evaluation.

2.7.1 Application of 1st and 2nd derivative.

The conversion of spectra into first- and second-order derivatives aims to accentuate the absorption characteristics contained in the spectra. Derivatives also remove additive and multiplicative effects in spectra (Rinnan et al. 2009). Calculating the first-order derivative removes the trend from the spectrum, while calculating the second-order derivative also removes a linear trend.

2.7.2 Machine Learning: Xgboost and Random Forest

The classification of agricultural crop species was obtained by two machine learning algorithms: random forest and xgboosting. The *Random Forest* (RF) Algorithm is an algorithm widely used for applications in the field of agricultural sciences. The RF model is a set of decision trees organized as a set of structured trees and can be used for regression and classification (Breiman, 2001). The advantage of RF when compared to other traditional algorithms is that it is based on several rules, where each one is constructed using different sizes of the original dataset (Breiman, 2001). The parameters used for the random forest were: bootstrap: 150 repartitions, max_depth: 10, n_estimators: 300.

Extreme Gradient Boosting (XGBoost) (model 2) is a hybrid algorithm of the tree model, applied to regression and classification predictions. It is an *ensemble* model that works based on multi-tree prediction, which is highly effective and widely used in various fields of agricultural sciences for solving regression or classification problems (Chen and Guestrin, 2016). Its scalable work mechanism is stage-based with increases in the number of trees. A differential aspect of the model is its ability to handle a database with a large number of variables and observations. The parameters used in the model were: bootstrap:150 repartitions, max_depth:6, n_estimators: 300.

The database was standardized to present null mean and unit variance, transforming all variables to the same order of magnitude. To carry out the classification, the database was divided into training and testing. Thus, 70% of the data was left for model training and 30% for testing, using different parameters for each algorithm.

At the end of the modeling, some metrics were obtained to evaluate the level of precision and accuracy of the classification: ROC curve metrics, precision, Recall, F1 score.

$$P = \frac{TP}{TP + FP} \tag{1}$$

$$R = \frac{TP}{TP + FN} \tag{2}$$

$$F.Score = 2 \times \frac{P \times R}{P + R} \tag{3}$$

The accuracy metric relates to the number of samples returned by the model as belonging to a class, divided by the total number of samples in that class. The recall consists of the relevant samples obtained by the classification of the model, divided by the total number of existing samples. The F-score metric is the harmonic mean of the precision and accuracy values and is useful to indicate the overall performance of the model.

3 Results

3.1 Metrics obtained in the modeling of the spectral signature of species.

The classification of species for each group of data, considering the different models and pre-processing, were evaluated using the classification metrics (Table 1). Considering the values of the metrics evaluated by the database (control and damaged), it is possible to observe that the values that AUC did not change, remaining at 0.99 in all experiments performed for this dataset. However, the AUC evaluated in the Damaged database showed a slight range of variation, with values ranging from 0.97 to 0.99.

For the Precision metric, the data with the label Control presented the range ranging from 0.94 – 0.96, while the data with the label Damaged presented values between 0.91 – 0.95. For the recall, the metrics followed a similar trend, in which the best performance was presented for the control data, ranging from 0.93 – 0.96, for the Damaged the range was from 0.91 - 0.94. For the F1 score, the control data ranged from 0.94 to 0.96 and the Damaged data ranged from 0.91 to 0.94.

Table 1. General plant species classification metrics (soybean, corn, and cotton)

Label data	Model	Pre-processing data	AUC	Precision	Recall	F1 score
Control	Random forest	Raw	0.99	0.94	0.94	0.94
		first derivative	0.99	0.96	0.95	0.95
		second derivative	0.99	0.95	0.93	0.94
	XGBoost	Raw	0.99	0.94	0.94	0.94
		first derivative	0.99	0.96	0.96	0.96
		second derivative	0.99	0.96	0.95	0.96
Damaged	Random forest	Raw	0.97	0.91	0.91	0.91
		first derivative	0.98	0.94	0.92	0.91
		second derivative	0.98	0.94	0.91	0.93
	XGBoost	Raw	0.97	0.91	0.91	0.94
		first derivative	0.98	0.94	0.94	0.92
		second derivative	0.99	0.95	0.94	0.94

3.2 - Importance of spectral signals for species classification.

After obtaining excellent classification metrics for both the control and Damaged data, the importance curves of each sign were plotted for the correct classification of the species (Figure 1). First, each curve was analyzed individually, after this step the two individual curves (control and damaged) were added to obtain a representative curve of the two plant situations and obtain the spectral signature.

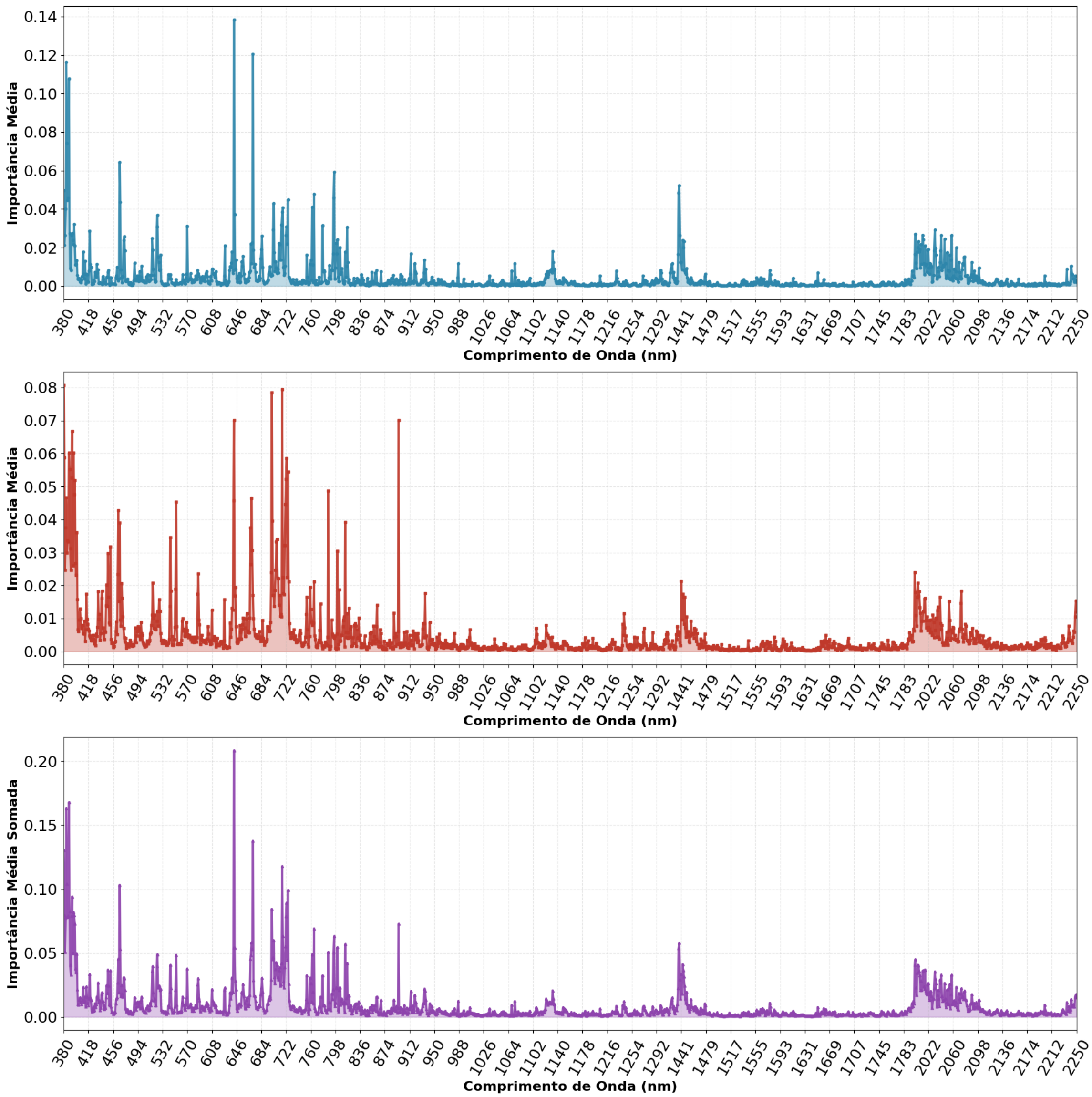


Figure 1. Curves of spectral importance for the classification of agricultural crop species.

In the analysis of the spectral curve added to the two situations (control and Damaged), we can conclude that there were regions with greater spectral importance than the others. With this, we can specify some regions marked as important in the graph. The first region that presented important spectral signals was the ultraviolet region between 380 and 400 nm. After this region, several spectral signals in the visible range were shown to be important. In the NIR region we can identify three specific ones that differentiate them from the others in terms of spectral importance, they are: 870 to 900 nm, 1300 to 1450 and at the end of the spectral curve encompassing the regions 1700 to 2100.

3.3 Definition of limits for the spectral signature.

In order to find the wavelengths that best specify the signature of the soybean crop in relation to the other crops (corn and cotton), a sequential flow of species reclassification was performed, selecting different amounts of spectral signals by their level of importance (Figure 2A), with this, it was possible to determine the quantity and identify each spectral signal for each established limit (2B). It can be noted that there was a great variation in the number of selected wavelengths, with the first limit represented by 778 wavelengths and the last limit represented by the two most important wavelengths.

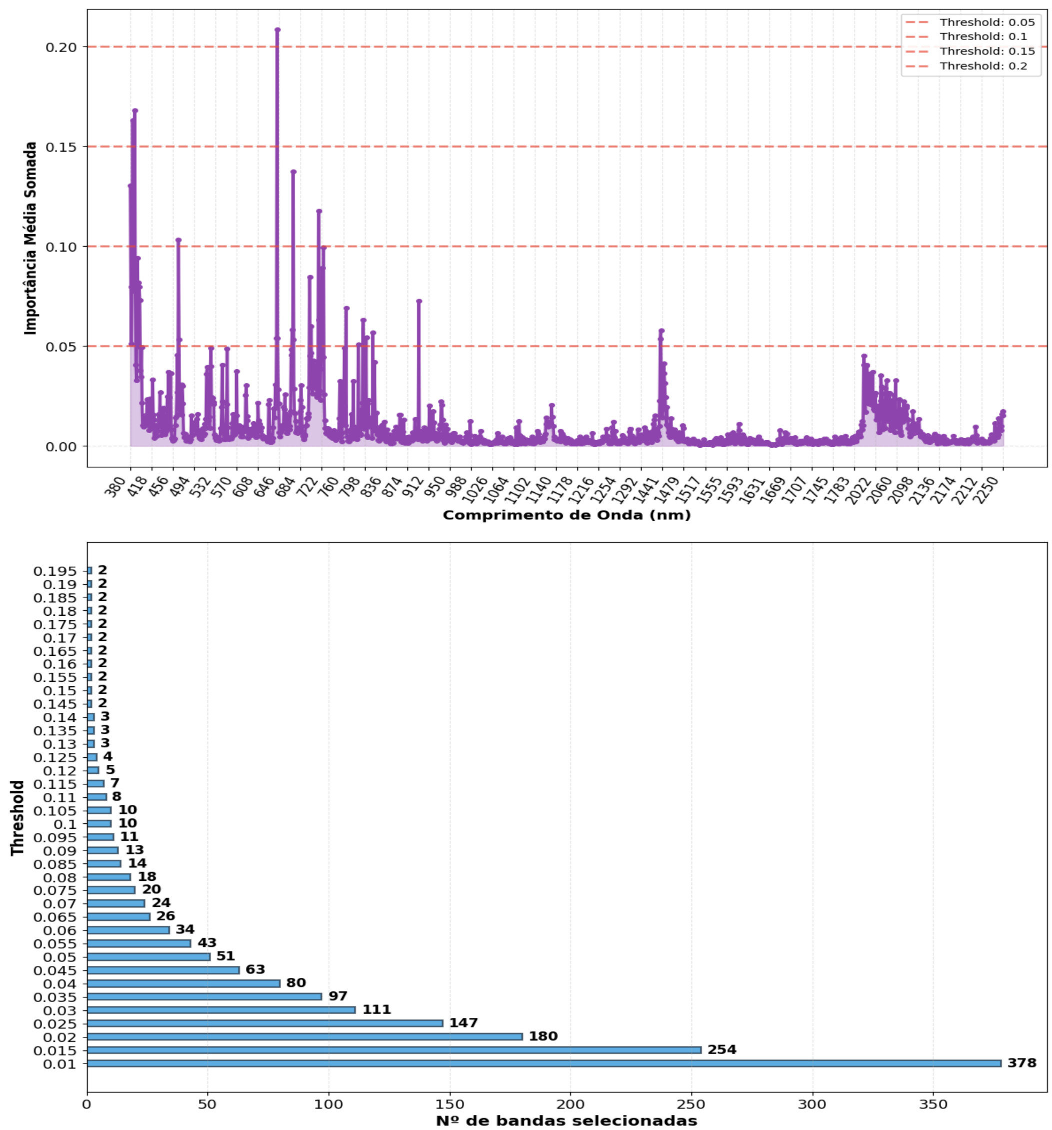


Figure 2. Established limits and amount of curve at each limit.

3.4 Reclassification of species and definition of signature.

By reclassifying the species to the determined limits, it is possible to identify the behavior of the metrics of the specified limits (Figure 3). The limit identified as spectral signature was the limit of 0.045, in which it has a total of 63 spectral signals associated with this limit establishment. We can see that despite a significant reduction in spectral signals, the classification metrics were still at excellent classification levels, thus determining the limit specified as the spectral signature of the target crop, which is soybean (Table 2).

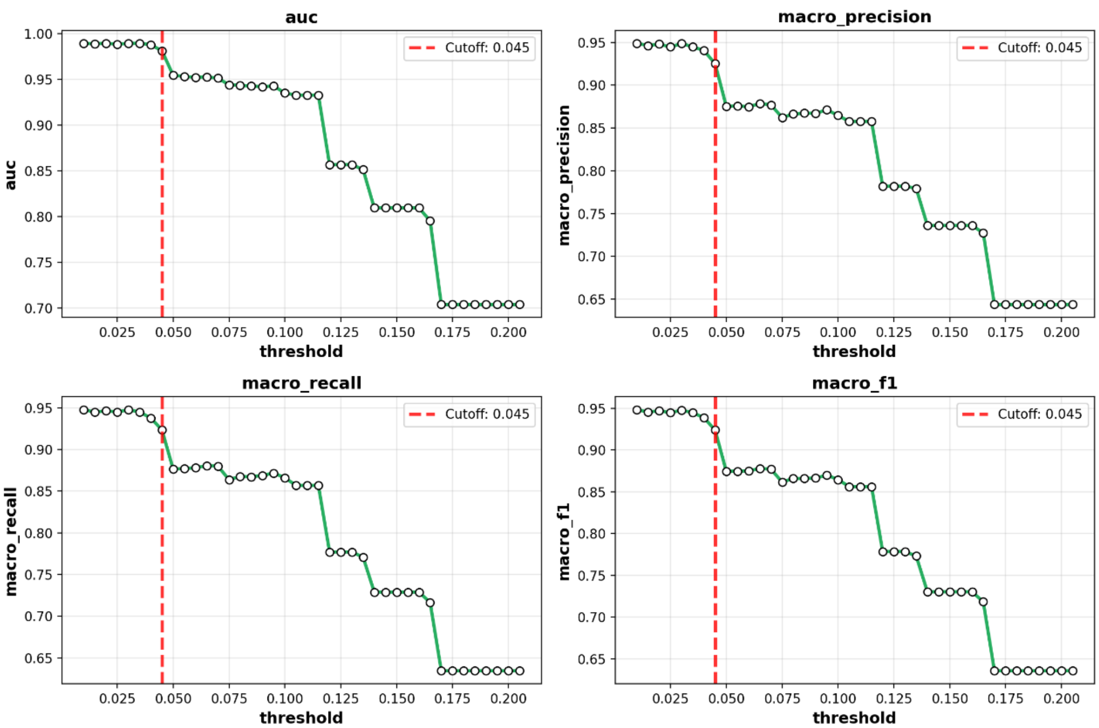


Figure 3. Metrics of the specified thresholds for spectral signature identification.

Table 2. Wavelengths identified as spectral signature.

Ranges	Wavelengths
380 - 400	351 352 354 355 356 358 359 361 362 363 364 365 366 367 371 372 373 374 376 377 378 381 383 384 385 386 388 389 392 393 395 396 397 400
401 - 600	400 464 524
601 - 800	641 642 643 671 700 703 711 716 717 721 722 723 725 765 787 796
801 – 1000	801 813 895
1000 - 2500	1346 1983 1986 1990 1996 2313 2336 2344

4 Discussion

The results obtained in this study demonstrate that the proposed framework is able to identify robust spectral signatures for the discrimination of soybean crops against corn and cotton, even under conditions of biotic damage caused by S. frugiperda and stink bugs. The high predictive performance of the RF and XGBoost models, with AUC of up to 0.99 for both control and damaged data, is in line with previous studies that indicate the efficiency of these algorithms in classifying crops based on spectral data (Osco et al., 2019a; Zhang et al., 2019; Singh et al., 2020).

Maintaining high metrics even in damaged plant data is especially relevant from an applied point of view. Studies such as that of Prabhakar et al. (2020) and Buchaillot et al. (2020) show that insect damage substantially modifies the leaf reflectance profile, altering concentrations of photosynthetic pigments, water content, and the structural integrity of the mesophyll. The ability of the models to maintain discriminative power in this scenario suggests that the spectral regions identified as important are biologically robust and less susceptible to variation induced by biotic stress.

The spectral regions identified as most informative—UV (380–400 nm), visible, NIR (870–900 nm, and 1300–1450 nm), and SWIR (1700–2100 nm)—have a biochemical and physiological basis that is well documented in the literature. The ultraviolet and visible regions are associated with absorption by leaf pigments, such as chlorophylls a and b and carotenoids (Clark & Roush, 1984). The NIR regions, in turn, strongly reflect the internal structure of the leaf mesophyll and the scattering of light between cells, characteristics that differ between species with distinct leaf anatomies (Marin et al., 2019). The SWIR range (1700–2100 nm) is sensitive to lignin, cellulose, and water content in the leaf, compounds ranging from grasses (maize), herbaceous dicots (soybeans), and plants with a high concentration of fiber in the stem (cotton) (Buddenbamm & Steffens, 2012; Martinelli et al., 2015).

The sequential reclassification approach with progressive reduction of wavelengths introduced in this work constitutes a relevant methodological contribution. The proposed framework anchors the selection of wavelengths in the stability of predictive performance over multiple days of evaluation. This ensures that the selected signals are representative not only of a particular phenological or environmental condition, but of persistent spectral behavior over time — a key characteristic for continuous field monitoring applications.

The definition of 63 wavelengths as a stable spectral signature of soybean, maintaining classification metrics at excellent levels (AUC ≥ 0.97), represents a reduction of more than 90% from the original number of spectral variables. This compression is decisive for the feasibility of application in multispectral sensors embedded in drones and satellites, where the number of spectral bands is limited. Similar results of efficient dimensionality reduction with preservation of accuracy were reported by Martins et al. (2017) and Osco et al. (2020) in different contexts of crop classification by remote sensing.

It is important to note that the analyses were conducted under controlled greenhouse conditions, which, although allowing for greater experimental control, may not fully reflect the complexity of field conditions, where factors such as variation in solar irradiance, illumination angle, and the presence of other plant species may introduce additional spectral noise. Future studies should evaluate the transferability of the identified spectral signature to field conditions and different stages of crop development, as well as its applicability to orbital and sub-orbital remote sensing systems.

Conclusions

The stability framework proposed in this work proved to be effective for the identification of robust spectral signatures of cultivated plants. The Random Forest and XGBoost machine learning algorithms, applied to spectral data with preprocessing by first- and second-order derivatives, showed excellent performance in discriminating between soybean, corn, and cotton, both in healthy and insect-damaged plant conditions (AUC ≥ 0.97 in all scenarios).

The sequential reclassification procedure allowed the number of wavelengths to be reduced to just 63 while maintaining classification metrics at excellent levels. These 63 wavelengths, concentrated in the regions of UV (380–400 nm), visible, NIR (870–900 nm and 1300–1450 nm) and SWIR (1700–2100 nm), constitute the stable and robust spectral signature of the soybean crop compared to corn and cotton.

The inclusion of data from plants damaged by S. frugiperda and stink bugs in the spectral signature selection process ensured the robustness of the set of wavelengths identified under biotic stress conditions, which is fundamental for practical applications in phytosanitary monitoring. The results obtained open promising perspectives for the development of dedicated multispectral sensors, calibrated from the bands identified in this study, with potential application in drone and satellite platforms for monitoring large cultivated areas.

5 references

Araújo MLS, et al. The Relationship between Climate, Agriculture and Land Cover in Matopiba, Brazil (1985–2020). *Sustainability*. 2024;16(7):2670. doi:10.3390/su16072670.

Barros EM, Torres JB, Ruberson JR, Oliveira MD. Development of *Spodoptera frugiperda* on different hosts. *Entomologia Experimentalis et Applicata*. 2010;137(3):237–245.

Bi JL, et al. Physiological effects of insect herbivory on plants. *Journal of Chemical Ecology*. 1997;23(2):293–307.

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

Buddenbamm H, Steffens M. Spectral analysis of soil and vegetation. *Remote Sensing of Environment*. 2012;124:1–10.

Buchaillot ML, et al. Remote sensing for insect damage assessment. *Precision Agriculture*. 2020;21:843–859.

Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM; 2016. p. 785–794. doi:10.1145/2939672.2939785.

Clark RN, Roush TL. Reflectance spectroscopy: Quantitative analysis techniques. *Journal of Geophysical Research*. 1984;89(B7):6329–6340.

Crusiol LGT, Furlanetto RH, Nanni MR, et al. Hyperspectral Data for Early Identification and Classification of Potassium Deficiency in Soybean Plants (*Glycine max* (L.) Merrill). *Remote Sensing*. 2024;16(11):1900. doi:10.3390/rs16111900.

Eisenring M, et al. Leaf biochemical responses under biotic stress. *Plant Physiology*. 2019;181:1234–1245.

Falcioni R, et al. Chemometric Analysis for the Prediction of Biochemical Compounds in Leaves Using UV-VIS-NIR-SWIR Hyperspectroscopy. *Plants*. 2023;12(19):3424. doi:10.3390/plants12193424.

Jensen JR. *Remote sensing of the environment: An earth resource perspective*. 2nd ed. Upper Saddle River: Pearson Prentice Hall; 2014.

Marin MA, et al. Spectral discrimination of crop stresses using hyperspectral data. *Agricultural Remote Sensing Journal*. 2019;15(4):78–92.

Martinelli T, et al. Spectral characteristics of diseased plants. *International Journal of Applied Earth Observation and Geoinformation*. 2015;35:142–150.

Martins J, et al. Hyperspectral sensing in agriculture. *Sensors*. 2017;17(5):1121–1138. doi:10.3390/s17051121.

Osco LP, et al. Hyperspectral imaging and machine learning for crop monitoring. *Computers and Electronics in Agriculture*. 2019a;162:482–491. doi:10.1016/j.compag.2019.04.003.

Osco LP, et al. Machine learning and remote sensing integration for precision agriculture. *ISPRS Journal of Photogrammetry and Remote Sensing*. 2020;168:208–224. doi:10.1016/j.isprsjprs.2020.08.014.

Prabhakar M, et al. Detection of insect stress in crops through remote sensing. *Journal of Experimental Botany*. 2020;71(8):2313–2325. doi:10.1093/jxb/eraa022.

Rinnan Å, van den Berg F, Engelsen SB. Review of normalization and derivative pre-treatment in spectroscopy. *TrAC – Trends in Analytical Chemistry*. 2009;28(10):1201–1222. doi:10.1016/j.trac.2009.07.007.

Santos WF, et al. Brazilian soybean production and land use dynamics. *Agricultural Systems*. 2023;210:103705.

Singh A, et al. Hyperspectral reflectance for pest detection in crops. *Precision Agriculture*. 2020;21:451–471. doi:10.1007/s11119-019-09678-4.

USDA – United States Department of Agriculture. World Agricultural Supply and Demand Estimates (WASDE). Washington: USDA; February 10, 2026. Disponível em: https://www.usda.gov/oce/commodity/wasde/wasde0226.pdf. (Produção de soja no Brasil 2025/26: 180 milhões de toneladas – recorde, elevada em 2 milhões de toneladas em relação ao relatório anterior, com área colhida projetada em 49,4 milhões de hectares.)

Vargas-Almendra A, Ruiz-Medrano R, Núñez-Muñoz LA, Ramírez-Pool JA, Calderón-Pérez B, Xoconostle-Cázares B. Advances in Soybean Genetic Improvement. *Plants (Basel)*. 2024;13(21):3073. doi:10.3390/plants13213073.

Zhang J, et al. Machine learning for detection of crop pest damage using spectral features. *Remote Sensing*. 2019;11(21):2499. doi:10.3390/rs11212499.