remote sensing

MDPI

Article

Machine Learning Reveals Physiologically Relevant Spectral Domains for Citrus Huanglongbing Detection

Mauro Morata Bortoloto Junior [1,*], Lucas Prado Osco [2], Lúcio André de Castro Jorge [3] and Ana Paula Marques Ramos [1]

[1] Department of Cartography, São Paulo State University (UNESP), Roberto Simonsen 305, Presidente Prudente 19060-900, SP, Brazil; marques.ramos@unesp.br
[2] Pos-Graduate Program of Enviroment and Regional Developement, Western São Paulo State University (UNESP), Raposo Tavares 572, Presidente Prudente 19067-175, SP, Brazil; lucasosco@unoeste.br
[3] Embrapa Instrumentação, Brazilian Agricultural Research Corporation, São Carlos 13560-970, SP, Brazil; lucio.jorge@embrapa.br
* Correspondence: mauro.morata@unesp.br

Highlights

What are the main findings?

- Hyperspectral reflectance enabled robust discrimination between samples from HLB-positive and HLB-negative citrus plants, as determined by plant-level PCR diagnosis, at both leaf and plant scales under leave-one-FIELD-out validation.
- Spectral domains located in the visible (400–700 nm), red-edge (700–750 nm), and shortwave infrared (1500–2500 nm) regions consistently contributed to HLB discrimination, while plant-level aggregation reduced within-class spectral variability and improved classification stability.

What are the implications of the main findings?

- Within the evaluated dataset, spectral representation was associated with larger and more consistent performance differences than class balancing, whereas hyperparameter optimization produced limited and configuration-dependent effects.
- The identified spectral domains provide testable hypotheses for subsequent evaluation at canopy, UAV, and satellite scales, although their transferability to these sensing platforms was not evaluated in the present study.

Abstract

Huanglongbing (HLB) is the most destructive disease affecting citrus production worldwide, causing severe yield losses and tree decline. Hyperspectral proximal sensing provides detailed information on leaf reflectance; however, its effectiveness for reliable HLB detection depends on the identification of consistent and discriminative spectral patterns. In this study, hyperspectral reflectance data (400–2450 nm) were combined with supervised machine learning algorithms to distinguish HLB-positive and HLB-negative citrus samples at two complementary scales: (i) leaf-level classification using individual spectral signatures and (ii) plant-level classification based on aggregated spectral responses. Model performance was evaluated under a leave-one-field-out validation framework designed to minimize spatial dependence between training and testing data. Results demonstrated consistently high classification performance, with Logistic Regression achieving accuracies above 0.94 at the leaf level and up to 1.00 at the plant level in several configurations. Dimensionality reduction using Principal Component

Academic Editors: Abid Ali and Hans-Peter Kaul

Received: 4 August 2026
Revised: 20 September 2026
Accepted: 25 September 2026
Published: 3 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Analysis (PCA) preserved most of the discriminative information, although with a slight reduction in recall under certain FIELD combinations. In contrast, Class balancing produced no observed improvement, whereas hyperparameter optimization showed limited and configuration-dependent effects in classification performance. Spectral coefficient analysis revealed that the most relevant wavelengths are concentrated in the visible (400–700 nm), red-edge (700–750 nm), near-infrared (750–1300 nm), and shortwave infrared (1500–2500 nm) regions, reflecting physiological alterations associated with pigment degradation, mesophyll disruption, and changes in water and biochemical composition. Overall, hyperspectral reflectance combined with parsimonious machine learning models enabled consistent discrimination of HLB-positive and HLB-negative samples within the evaluated orchard under validation designed to reduce spatial dependence, particularly at the plant scale.

Keywords: HLB; machine learning; spectroscopy; hyperspectral; spatial cross validation

1. Introduction

Huanglongbing (HLB) remains the most economically damaging disease affecting citrus production worldwide, causing progressive canopy deterioration, severe yield reduction, and ultimately tree death [1,2]. In countries with large-scale citrus industries, such as Brazil, accurate identification of infected trees is essential for effective disease management [3–5]. Previous studies have demonstrated the potential of machine learning for detecting plant diseases using spectral data, including applications in citrus and other crops [3–12].

The disease is associated with bacteria from the *Candidatus Liberibacter* complex and is transmitted by sap-feeding insect vectors [1,2]. Conventional diagnostic approaches, including visual scouting and laboratory-based assays such as polymerase chain reaction (PCR), present important limitations, particularly due to the heterogeneous distribution of the pathogen within the plant and the frequent absence of visible symptoms during the initial stages of infection [1,2,13].

Under these constraints, remote sensing (especially hyperspectral measurements) enables the acquisition of detailed spectral information related to plant physiological condition in a non-destructive manner [6,8,9,14–17]. This approach allows the investigation of disease-induced alterations in spectral signatures, which may support the discrimination between HLB-positive and HLB-negative plants [3–5,18–20]. Nevertheless, despite extensive research efforts, a standardized and broadly validated remote-sensing methodology for HLB detection has not yet been established, particularly regarding acquisition protocols, disease-stage variability, model transferability, and validation across orchards [9,14,16,19,21,22].

Identifying informative spectral domains is fundamental for the development of automated and non-invasive diagnostic systems and for reducing hyperspectral data dimensionality [23–25]. Evaluating spectral responses at multiple organizational and spatial scales may further strengthen this approach [11,19,21,22]. While leaf-level measurements capture localized biochemical and structural perturbations, plant- and canopy-level observations provide more integrative representations of pigment composition, internal leaf structure, water status, canopy architecture, and symptom distribution, which are more consistent with how remote-sensing data are acquired and interpreted under field conditions [26–30].

Although previous studies have demonstrated the potential of hyperspectral sensing and machine learning for HLB and other plant-disease detection applications [4,5,10–

12,18–24], comparatively little attention has been given to how leaf- and plant-level spectral representations influence model robustness when training and testing samples originate from different orchard areas. In particular, it remains unclear whether improving the biological representation of spectral information contributes more substantially to HLB classification performance than algorithmic refinements such as class balancing, hyperparameter optimization, and dimensionality reduction PCA [9,14,21,22,31].

Previous statistical analysis of this dataset demonstrated significant spectral differences between HLB-positive and HLB-negative samples at both leaf and plant levels and showed that plant-level aggregation reduced intra-plant spectral variability while preserving consistent discriminative spectral regions [32]. Building on this wavelength-wise statistical characterization, the present study investigates whether these representation-related effects are reflected in supervised machine learning performance under spatially blocked FIELD-based validation and how the observed performance patterns compare with those associated with algorithmic refinements.

Therefore, the main objective of this study was to evaluate whether the statistically demonstrated effects of spectral representation and plant-level aggregation on HLB separability are reflected in machine learning classification performance, and to compare their observed effects with those resulting from algorithmic refinements. Specifically, the study aimed to: (i) assess the consistency of HLB discrimination at leaf and plant levels; (ii) determine whether plant-level spectral aggregation reduces within-class variability while preserving disease-related information; (iii) quantify the effects of dimensionality reduction, class balancing, and hyperparameter optimization on classification performance; and (iv) identify the spectral domains most consistently associated with HLB discrimination and discuss their potential physiological relevance based on previous literature.

Based on our previous statistical evidence showing consistent HLB-related spectral separability across leaf and plant scales and reduced intra-plant variability after spectral aggregation, we hypothesized that plant-level spectral aggregation would preserve HLB-related spectral information while improving classification stability and performance under validation designed to reduce spatial dependence, whereas the effects of class balancing, hyperparameter optimization, and dimensionality reduction would depend on the evaluated data configuration.

2. Materials and Methods

The study was conducted in commercial citrus orchards located at Entre Rios Farm, São Paulo State, Brazil (Figure 1). The regional climate is classified as humid subtropical (Cwa) according to the Köppen–Geiger classification. The predominant soil type is Red Latosol, with subordinate occurrences of Red-Yellow Latosols and Quartzarenic Neosols. Sampling procedures followed the methodology described by Morata Bortoloto Junior et al. (2026) [32]. To account for the heterogeneous distribution of *Candidatus Liberibacter* spp. within the canopy, approximately twenty leaves were collected from different canopy positions and heights. All sampled leaves were collected between 10:30 a.m. and 1:30 p.m. and subsequently subjected to hyperspectral measurements using the portable field setup described below.

Molecular Detection by Real-Time PCR (qPCR). Total genomic DNA was extracted from 100–200 mg of midribs and petioles collected from adult citrus leaves using a modified cetyltrimethylammonium bromide (CTAB) protocol optimized for citrus tissues. Molecular diagnosis was performed using a real-time multiplex quantitative PCR (qPCR) assay following the Fundecitrus protocol adapted from previously established CLas detection procedures [14,33]. *Candidatus* Liberibacter asiaticus (CLas) was detected using

the HLBas/HLBr primer set and the FAM-labeled HLBp probe [14], while the citrus cytochrome oxidase (COX) gene was simultaneously amplified using COXf/COXr primers and a VIC/HEX-labeled COXp probe as an internal control for DNA quality and amplification inhibition [33]. All samples were analyzed in technical duplicate. Each qPCR plate included a positive control consisting of DNA from a CLas-infected citrus plant, a negative extraction control from a healthy plant, and a no-template control containing nuclease-free water. Samples showing a characteristic sigmoidal amplification curve with $Ct \leq 36.0$ were classified as CLas-positive, whereas samples with no amplification or $Ct \geq 40.0$ were classified as negative. Samples with $36.0 < Ct < 40.0$ or discordant technical replicates were retested using the same DNA template. Persistently equivocal samples were subjected to DNA re-extraction from the original plant material; samples remaining inconclusive after re-extraction required a new field collection.

HLB status was assigned at the plant level based on PCR analysis of pooled leaf material collected from each sampled plant. Consequently, the molecular diagnosis constituted a plant-level reference, and all hyperspectral measurements obtained from leaves belonging to the same plant inherited the corresponding HLB-positive (HLB+) or HLB-negative (HLB−) label. Individual leaves were not independently tested by PCR and are therefore not interpreted as individually PCR-confirmed samples. This labeling structure is particularly relevant at the leaf scale because pathogen distribution, symptom expression, and associated physiological alterations may vary among leaves within an HLB-positive plant.

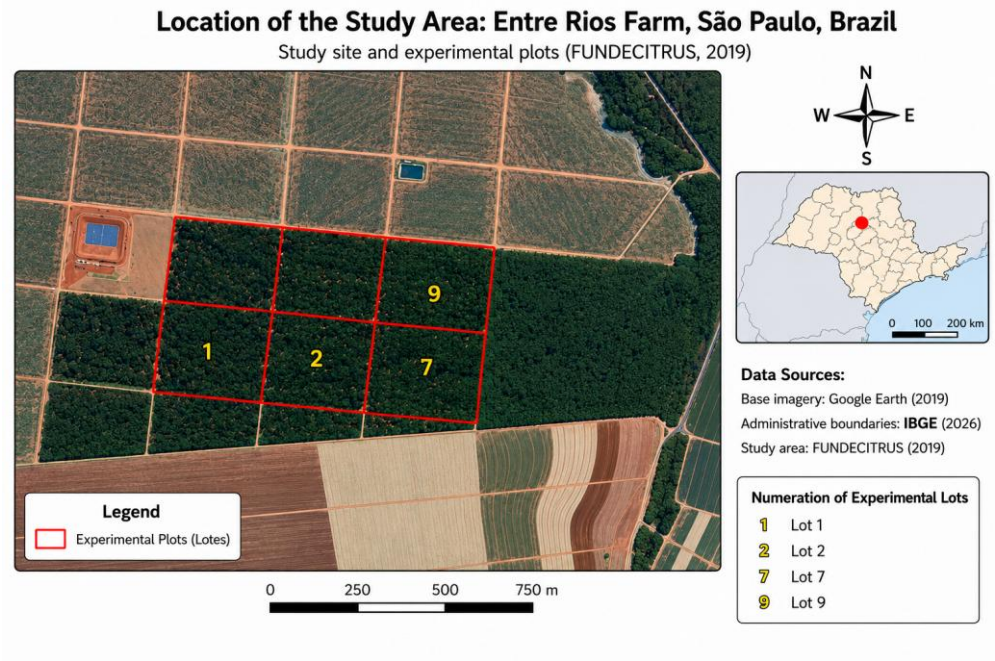


Figure 1. Location of the study area at Entre Rios Farm, São Paulo State, Brazil. The map shows the spatial distribution of the experimental FIELDs used for hyperspectral data collection within the commercial citrus orchard affected by Huanglongbing (HLB). Base imagery: Google; auxiliary cartographic data: Brazilian Institute of Geography and Statistics (IBGE); field boundaries: FUNDECITRUS.

Spectral reflectance measurements were acquired using an ASD FieldSpec® 3 spectroradiometer (Analytical Spectral Devices Inc., Boulder, CO, USA), operating over the 350–2500 nm spectral range [34]. The instrument was equipped with an 8° field-of-view (FOV) foreoptic mounted on a fixed support to maintain consistent sensor–target geometry throughout the measurements. The foreoptic was positioned approximately

0.50 m above the leaf surface and oriented perpendicular to the target, corresponding to a nadir viewing geometry. At this working distance, the 8° foreoptic provided an approximate measurement footprint of 7.8 cm in diameter.

Following collection from the trees, individual leaves were positioned in a portable measurement setup assembled in the field for hyperspectral acquisition. Measurements targeted the upper (adaxial) leaf surface. The setup consisted of the fixed ASD FieldSpec® 3 spectroradiometer and a tungsten-halogen lamp used as the artificial illumination source. The foreoptic was maintained perpendicular to the sample surface, while the illumination source was positioned at an angle of approximately 30–45° relative to the sample surface. The instrument, illumination source, and sample were maintained in a fixed configuration to provide consistent sensor–sample–illumination geometry throughout spectral acquisition. Thus, spectral measurements were obtained from detached leaves under controlled artificial illumination while still in the field.

Before spectral acquisition, the FieldSpec 3 was switched on and allowed to thermally stabilize for at least 30 min. The instrument was subsequently optimized using RS^3 acquisition software (Analytical Spectral Devices Inc., Boulder, CO, USA) and a calibrated Spectralon® diffuse reflectance reference panel positioned using the same sensor–sample–illumination geometry adopted for the leaf measurements. Instrument optimization was performed to obtain an appropriate detector signal while avoiding saturation. Following optimization, a white-reference measurement was acquired from the Spectralon panel under the same sensor distance, viewing geometry, illumination conditions, and acquisition parameters adopted for the leaf measurements. Spectral reflectance was subsequently obtained relative to the calibrated Spectralon reference.

The FieldSpec 3 covered the 350–2500 nm spectral range. Spectral measurements were provided at the instrument's native sampling intervals and internally resampled by the acquisition software to a uniform 1 nm output wavelength interval across the entire spectral range. For each leaf spectrum, approximately 10 individual scans were averaged to improve the signal-to-noise ratio and provide a representative spectral measurement. The FieldSpec 3 uses three detector arrays covering the VNIR and SWIR spectral regions. The spectral output generated by the instrument acquisition software was used in the subsequent analyses, and no additional correction was applied to the detector-overlap regions during data preprocessing.

Reflectance was obtained by normalizing the target radiance to the radiance measured from the Spectralon® white reference under the same acquisition geometry. Accordingly, relative spectral reflectance at wavelength λ was expressed as

$$R_{rel}(\lambda) = \frac{L_{target}(\lambda)}{L_{reference}(\lambda)}, \quad (1)$$

where $L_{\text{target}}(\lambda)$ is the radiance measured from the leaf and $L_{\text{reference}}(\lambda)$ is the radiance measured from the Spectralon® reference panel. When the wavelength-dependent calibrated reflectance of the reference panel is explicitly considered, target reflectance can be expressed as

$$R_{target}(\lambda) = \frac{L_{target}(\lambda)}{L_{reference}(\lambda)} R_{reference}(\lambda), \quad (2)$$

where $R_{\text{reference}}(\lambda)$ represents the calibrated spectral reflectance of the reference panel. Spectral acquisition and instrument optimization were controlled using RS^3 software.

This procedure converts radiance measurements into reflectance values, minimizing illumination effects and ensuring the comparability of spectral measurements acquired throughout the field campaign. Each spectral profile represents the average of ten individual measurements acquired per leaf. A total of 89 plants and 1912 leaves were sampled from Pera Rio grafted onto Tangerine Sunki species. The dataset was organized in spreadsheet format, including the following variables: SAMPLE, identifying each

individual leaf sample; FIELD, indicating the field where the plant was located; STREET, corresponding to the row within the orchard; PLANT, identifying the individual plant within each row; and HLB, representing the infection status of the plant.

A total of 1912 leaves were collected from 89 orange citrus plants, corresponding to an average of 21.48 ± 4.50 leaves per plant (median = 21; range = 6–42 leaves). The dataset comprised 31 HLB-positive plants (617 leaves) and 58 HLB-negative plants (1295 leaves). HLB-positive plants contributed an average of 19.90 ± 3.39 leaves per plant (median = 21; range = 9–25), whereas HLB-negative plants contributed 22.33 ± 4.80 leaves per plant (median = 22; range = 6–42).

The dataset exhibited a moderate class imbalance at both organizational levels. At the plant level, 31 of the 89 plants (34.8%) were HLB-positive and 58 (65.2%) were HLB-negative. At the leaf level, 617 of the 1912 spectra (32.3%) were obtained from HLB-positive plants and 1295 (67.7%) from HLB-negative plants. The class distribution was also relatively consistent across FIELD, although the absolute number of samples varied because of differences in the number of sampled plants and leaves. The FIELD-specific distribution of HLB-positive and HLB-negative observations is reported in Table 1.

Table 1. Distribution of HLB-positive (HLB+) and HLB-negative (HLB−) samples across the four experimental FIELDs at plant and leaf levels. Leaf-level labels were inherited from the corresponding plant-level PCR diagnosis.

FIELD	HLB− Plants	HLB+ Plants	Total Plants	HLB− Leaves	HLB+ Leaves	Total Leaves
FIELD 1	20	11	31	437	208	645
FIELD 2	19	10	29	426	205	631
FIELD 7	9	5	14	206	98	304
FIELD 9	10	5	15	226	106	332
Total	**58**	**31**	**89**	**1295**	**617**	**1912**

The original class distribution was preserved in each withheld FIELD used for final evaluation. Class balancing, when evaluated, was applied exclusively to the corresponding model-development data and did not modify the class composition of the withheld FIELD.

The hyperspectral dataset analyzed in the present study is the same dataset previously investigated using wavelength-wise statistical analyses in [32]. The previous study focused on the statistical characterization of spectral separability between HLB-positive and HLB-negative samples at leaf and plant levels. In contrast, the present study addresses a distinct analytical objective by evaluating supervised machine learning classification under spatially blocked leave-one-FIELD-out validation, examining the effects of leaf- and plant-level spectral representation and algorithmic refinements, and assessing the stability of model-associated spectral contributions across FIELD holdouts.

2.1. Data Preprocessing

The dataset was subsequently reorganized by associating each sample with a composite plant identifier (FIELD + STREET + PLANT), enabling hierarchical analysis at both leaf and plant levels. The HLB column was used to represent the phytosanitary status of each sample, and spectral bands were treated as features for statistical and computational analysis.

During preprocessing, negative reflectance values were identified primarily at the short-wavelength boundary of the retained spectral range and were replaced with zero. A total of 2650 negative values were identified across 303 spectra, occurring between 400 and 425 nm, with most occurrences concentrated between 400 and 419 nm. Despite the

higher noise level observed in this spectral region, wavelengths from 400 nm onward were intentionally retained because the lower blue-edge portion of the spectrum may contain potentially useful discriminatory information associated with HLB-related changes in leaf optical properties. Therefore, rather than removing the entire short-wavelength interval, a conservative preprocessing strategy was adopted in which physically implausible negative values were truncated to zero while preserving the corresponding wavelengths for subsequent analysis.

Spectral preprocessing was performed conservatively to avoid unnecessarily removing physiologically informative wavelengths. Measurements at the extreme ends of the original sensor range were excluded, and the analysis was restricted to 400–2450 nm. Within this interval, only the 1831–1889 nm region was excluded because it was considered insufficiently reliable for subsequent analysis. In contrast, broader wavelength regions commonly associated with strong water absorption, particularly around 1400 and 1900 nm, were intentionally retained whenever reliable measurements were available. This conservative strategy was adopted because HLB-induced physiological alterations may affect leaf water status and related biochemical and structural properties, and therefore spectral responses in water-sensitive regions could contain relevant discriminatory information. The final dataset comprised 1992 spectral wavelengths distributed between 400 and 2450 nm, with the 1831–1889 nm interval omitted.

Hyperspectral reflectance data were collected from leaves sampled across four fields (FIELD 1, 2, 7, and 9) in a commercial citrus orchard, Figure 1. Machine learning classification models [35] were designed to differentiate samples from HLB-positive and HLB-negative plants under multiple data-processing configurations.

Considering the adopted experimental framework, two complementary data structuring strategies were implemented to evaluate the ability of the algorithms to discriminate between HLB-positive and HLB-negative samples. The first strategy consisted of leaf-level classification, in which each individual leaf spectrum was treated as an observation for model training, while recognizing that leaves originating from the same plant are not statistically independent. The second strategy corresponded to plant-level classification, in which the reflectance values of all sampled leaves belonging to the same plant were averaged wavelength by wavelength, producing one representative spectrum per plant. The choice of arithmetic mean was supported by the statistical characterization of this same spectral dataset reported in our previous study [32], in which the distributional behavior of the spectral data was evaluated and plant-level aggregation showed an approximation to normality. Under this distributional behavior, the arithmetic mean provides an appropriate measure of central tendency for integrating multiple leaf observations into a representative plant-level spectrum. Moreover, mean aggregation reduces leaf-to-leaf variability while preserving the overall spectral response of each plant. Therefore, the objective of this aggregation was to characterize the central spectral response of the plant rather than to emphasize extreme observations, as would occur with maximum-based aggregation. Although alternative robust aggregation approaches, such as the median, may also be investigated, the mean was selected here based on the previous statistical characterization of the same dataset [32] and the objective of obtaining an integrated plant-level representation. For each leave-one-FIELD-out iteration, spectral standardization was performed using parameters estimated exclusively from the training data. The mean and standard deviation of each spectral variable were therefore computed from the three training FIELDs and subsequently applied, without refitting, to the withheld FIELD. This procedure prevented information from the evaluation FIELD from contributing to preprocessing [36].

Principal Component Analysis (PCA) was additionally evaluated as a dimensionality-reduction strategy at both leaf and plant levels. For the leaf-level

representation, individual leaf reflectance spectra constituted the observational units, whereas, for the plant-level representation, the reflectance spectra of the sampled leaves belonging to each plant were averaged wavelength by wavelength to obtain a single representative spectrum per plant. To prevent information from the evaluation FIELD from contributing to dimensionality reduction, PCA was performed independently within each leave-one-FIELD-out iteration. For each leave-one-FIELD-out iteration, spectral standardization was performed using parameters estimated exclusively from the training data. The mean and standard deviation of each spectral variable were computed from the three training FIELDs and subsequently applied, without refitting, to the withheld FIELD. This procedure prevented information from the evaluation FIELD from contributing to preprocessing. PCA was then fitted exclusively on the standardized model-development data, retaining the minimum number of principal components required to explain at least 95% of the cumulative variance. The fitted PCA transformation was subsequently applied, without refitting, to the withheld FIELD. Consequently, neither the standardization parameters nor the PCA loadings were estimated using observations from the evaluation FIELD. The resulting principal-component scores replaced the original wavelength predictors and were used as input features for the corresponding machine learning classification experiments.

Scatter-correction and baseline-removal techniques, such as Standard Normal Variate (SNV) and spectral derivatives, were evaluated during preliminary exploratory analyses. However, as they did not demonstrate a measurable improvement in overall classification metrics under the cross-FIELD evaluation design, they were omitted from the final processing pipeline to preserve model parsimony.

To reduce spatial dependence between model-development and evaluation samples, a spatially blocked leave-one-FIELD-out approach was adopted [37]. The hierarchical structure of the dataset was explicitly considered throughout model development. Individual leaves were nested within plants, and plants were nested within FIELD. For leaf-level analyses, all spectra originating from the same plant were maintained within the same partition during internal cross-validation by using the composite plant identifier (FIELD + STREET + PLANT) as the grouping variable. Thus, leaves from a given plant were not simultaneously included in the training and validation subsets of the same internal fold. At the outer validation level, all samples from the withheld FIELD were completely excluded from model fitting, preprocessing, internal cross-validation, and hyperparameter optimization and were used only for final FIELD-wise evaluation. This hierarchical partitioning strategy was adopted to prevent information sharing among repeated spectral observations from the same biological unit and to reduce spatial dependence between training and evaluation data. For plant-level analyses, each plant constituted a single independent observation represented by the mean spectrum of its sampled leaves; therefore, no repeated leaf-level observations from the same plant were present after aggregation. Under this configuration, three FIELDs were used to construct the training dataset, while the remaining FIELD was reserved exclusively for testing. Class imbalance was addressed in selected experiments using the Synthetic Minority Over-sampling Technique (SMOTE) [38]. SMOTE generates synthetic samples of the minority class by interpolating between neighboring observations in the feature space, rather than simply duplicating existing minority-class samples. Within each model-development partition, SMOTE was applied exclusively to the training data to increase the representation of HLB-positive samples relative to the majority HLB-negative class. The withheld FIELD was not resampled and retained its original class distribution, ensuring that model evaluation reflected the naturally imbalanced data. Balanced and unbalanced configurations were evaluated separately to assess the influence of oversampling on classification performance. For leaf-level analyses, resampling was restricted to the

corresponding model-development partitions and was performed after the definition of plant-grouped training and validation subsets, thereby preserving the separation of observations originating from the same plant.

Additionally, hyperparameter optimization was performed using Optuna [39] exclusively within the model-development data of each leave-one-FIELD-out iteration. The withheld FIELD was not accessed during optimization. For each optimization experiment, 100 trials were evaluated using a Tree-structured Parzen Estimator (TPE) sampler with a random seed of 42. Candidate configurations were assessed through 10-fold StratifiedGroupKFold internal cross-validation (n_splits = 10, shuffle = True, random_state = 42). For leaf-level models, the composite plant identifier (FIELD + STREET + PLANT) was used as the grouping variable, ensuring that all spectra originating from the same plant remained within the same internal fold while preserving the class distribution as far as permitted by the group structure. The optimization objective was the mean weighted F1-score across the internal validation folds. No early-stopping criterion was applied; each optimization study was terminated after the predefined 100 trials. After optimization, the selected hyperparameter configuration was fitted using the complete model-development dataset and subsequently evaluated on the withheld FIELD. The hyperparameters evaluated, their search ranges, and sampling distributions are reported in Table 2.

Table 2. Hyperparameter search spaces used for Optuna-based model optimization.

Classifier	Hyperparameter	Search Space/Values	Sampling Type
Logistic Regression	C	10^{-5}–10^{2}	Log-uniform
	solver	liblinear; saga	Categorical
Support Vector Machine	C	10^{-5}–10^{2}	Log-uniform
	kernel	linear; poly; rbf; sigmoid	Categorical
	gamma	scale; auto	Categorical
K-Nearest Neighbors	n_neighbors	1–30	Integer
	weights	uniform; distance	Categorical
	p	1–2	Integer
Naïve Bayes	—	Default GaussianNB configuration	Not optimized
Random Forest	n_estimators	10–50	Integer
	max_depth	2–32	Integer
	min_samples_split	2–20	Integer
	min_samples_leaf	1–10	Integer
Gradient Boosting	n_estimators	50–100	Integer
	learning_rate	10^{-3}–0.1	Log-uniform
	max_depth	2–8	Integer
Decision Tree	max_depth	2–32	Integer
	min_samples_split	2–20	Integer
	min_samples_leaf	1–10	Integer
LightGBM	n_estimators	50–100	Integer
	learning_rate	10^{-3}–0.1	Log-uniform
	num_leaves	2–64	Integer
	max_depth	−1–15	Integer
XGBoost	n_estimators	50–100	Integer
	learning_rate	10^{-3}–0.1	Log-uniform
	max_depth	2–10	Integer
	subsample	0.6–1.0	Continuous, log scale
	colsample_bytree	0.6–1.0	Continuous, log scale

Note: Optuna optimization used a Tree-structured Parzen Estimator (TPE) sampler with random seed = 42. Each study comprised 100 trials across the joint classifier–hyperparameter search space

https://doi.org/10.3390/rs18213782

and maximized the mean weighted F1-score obtained during internal cross-validation. Naïve Bayes was evaluated using the default GaussianNB configuration and was not hyperparameter-optimized.

A detailed descriptive and statistical characterization of the original HLB-positive and HLB-negative reflectance spectra at both leaf and plant levels has been previously reported using the same spectral dataset [32]. That study included wavelength-wise comparisons of the original reflectance profiles and statistical evaluation of spectral separability between HLB classes. Therefore, these analyses were not repeated here, as the present study specifically focuses on machine learning classification, validation across FIELD partitions, and the stability of model-associated spectral contributions.

We also conducted experiments in which dimensionality reduction was applied using Principal Component Analysis. This procedure was employed to address the high dimensionality and strong correlation among hyperspectral variables, generating a reduced set of orthogonal components used as input features for the classification models [35]. The reduced feature space was subsequently used to train classification algorithms, including linear and kernel-based methods such as Logistic Regression [40] and Support Vector Machines (SVM) [41], as well as tree-based approaches such as Random Forest [42], Decision Tree [43], Gradient Boosting [44], XGBoost [45] and LightGBM [46], distance-based methods such as KNN [47], and probabilistic methods such as Naïve Bayes [48]. Model performance was evaluated during internal cross-validation and subsequently on the corresponding withheld FIELD using the metrics described in Section 2.2.

Furthermore, internal model outputs were analyzed to investigate the relative importance of spectral wavelengths in the classification process. This analysis enabled the identification of spectral regions with greater influence on model decision-making, contributing to the understanding of wavelength-specific discriminatory behavior associated with HLB infection.

Supervised machine learning algorithms were trained to learn the statistical relationships between spectral reflectance and infection status. The modeling strategy assumed that HLB induces measurable alterations in foliar biochemical composition and internal leaf structure—such as pigment degradation, imbalances in water and carbohydrate content, and disruption of mesophyll organization—which, in turn, manifest as systematic variations in reflectance across specific wavelength regions [7].

The experimental design was organized to separately evaluate the influence of biological spectral representation and algorithmic refinement on HLB classification performance. Five complementary experiments were therefore conducted. The first established the baseline classification performance using individual leaf spectra. The second assessed the effect of plant-level spectral aggregation by averaging the reflectance of all sampled leaves from each tree. The remaining experiments investigated whether classifier performance could be further improved through class balancing, hyperparameter optimization with Optuna, and dimensionality reduction using PCA. This design enabled the observed performance patterns associated with data representation and algorithmic refinements to be descriptively examined under the same leave-one-FIELD-out evaluation framework.

2.2. Model Evaluation Metrics

Model performance was evaluated using complementary confusion-matrix-based metrics [49,50]. In addition to the conventional global metrics of accuracy, weighted precision, weighted recall, and weighted F1-score, class-specific and imbalance-sensitive measures were used to explicitly characterize performance for the minority HLB-positive class. These included HLB-positive precision (positive predictive value, PPV), HLB-positive sensitivity (recall), HLB-positive F1-score, HLB-negative specificity, balanced

accuracy, and Matthews correlation coefficient (MCC). The inclusion of these metrics was particularly important because support-weighted measures may be dominated by the majority class and therefore may not fully represent classification performance for HLB-positive samples.

Accuracy measures the proportion of correctly classified samples:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}, \quad (3)$$

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

Precision quantifies the proportion of predicted positive samples that are correctly classified:

$$Precision = \frac{\text{TP}}{\text{TP+FP}}, \quad (4)$$

Recall (also known as sensitivity) measures the proportion of actual positive samples that are correctly identified:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}, \quad (5)$$

The F1-score represents the harmonic mean between precision and recall:

$$\text{F1} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}, \quad (6)$$

All class-specific metrics reported in the Results section were calculated directly from the corresponding confusion-matrix counts, considering HLB-positive samples as the positive class and HLB-negative samples as the negative class. [48–51]:

$$Sensitivity\ (HLB+) = \frac{TP}{TP + FN}, \quad (7)$$

HLB-negative Specificity quantifies the proportion of HLB-negative observations correctly classified as negative and complements Sensitivity by describing performance for the negative class [48,49]:

$$Specificity\ (\text{HLB}-) = \frac{\text{TN}}{\text{TN + FP}}, \quad (8)$$

Balanced Accuracy (*B. Accuracy*) gives equal importance to the positive and negative classes by averaging Sensitivity and Specificity. It is therefore more informative than conventional Accuracy when class frequencies are unequal [49,51]:

$$B.Accuracy = \frac{Sensitivity + Specificity}{2}, \quad (9)$$

The Matthews Correlation Coefficient (MCC) summarizes the relationship between observed and predicted classes using all four elements of the confusion matrix. MCC ranges from −1 to +1, where +1 indicates perfect classification, 0 indicates performance with no overall correlation between prediction and observation, and −1 indicates complete disagreement. Because it incorporates TP, TN, FP, and FN simultaneously, MCC provides a complementary measure of classification performance under class imbalance [49]:

$$MCC = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{[(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})]}}, \quad (10)$$

To explicitly characterize predictive performance for the HLB-positive class, HLB-positive Precision (positive predictive value, PPV) was calculated as the proportion of samples predicted as HLB-positive that were correctly classified [51]:

$$\text{Precision}_{\text{HLB+}} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (11)$$

where TP represents the number of HLB-positive samples correctly classified as positive and FP represents the number of HLB-negative samples incorrectly classified as HLB-positive. This metric therefore quantifies the reliability of positive predictions, with higher values indicating a lower proportion of false-positive HLB detections [51].

The HLB-positive F1-score was additionally calculated as the harmonic mean of HLB-positive Precision and HLB-positive Sensitivity (Recall), providing a single class-specific measure that jointly accounts for false-positive and false-negative predictions [51]:

$$\mathrm{F1}_{\mathrm{HLB+}} = 2\frac{\mathrm{Precision}_{HLB+} \times \mathrm{Sensitivity}_{HLB+}}{\mathrm{Precision}_{HLB+} + \mathrm{Sensitivity}_{HLB+}}, \quad (12)$$

which is equivalently expressed directly from the confusion matrix as:

$$\mathrm{F1}_{\mathrm{HLB+}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \quad (13)$$

These class-specific metrics complement HLB-positive Sensitivity and HLB-negative Specificity by explicitly describing both the reliability of HLB-positive predictions and the balance between Precision and Sensitivity for the minority HLB-positive class.

All class-specific metrics reported in the Results section directly from the corresponding confusion-matrix counts, considering HLB-positive samples as the positive class and HLB-negative samples as the negative class. Weighted metrics were calculated separately using class support and should therefore not be interpreted as equivalent to the corresponding positive-class metrics.

Weighted Precision, weighted Recall, and weighted F1-score were retained as global summaries of performance. In the weighted formulation, each class contributes according to its number of observations (support) [51]. The class-specific and imbalance-sensitive metrics above were reported alongside these global measures to make performance for HLB-positive and HLB-negative samples explicit.

To characterize the variability of model performance across spatially withheld datasets, the mean and sample standard deviation (SD) of each performance metric were additionally calculated across the four leave-one-FIELD-out evaluations. These statistics were used descriptively to quantify cross-FIELD performance stability and were not interpreted as formal tests of statistical significance between classifiers, because the four withheld FIELD differed in sample size and class composition.

2.3. Logistic Regression Coefficient Stability Analysis

To evaluate model-associated spectral contributions while accounting for the strong collinearity among adjacent hyperspectral wavelengths, Logistic Regression coefficients were analyzed across the four leave-one-FIELD-out configurations. The analysis used standardized spectral predictors and considered HLB-positive samples as the positive class (HLB+ = 1) and HLB-negative samples as the negative class (HLB− = 0). Accordingly, positive coefficients indicate an increase in the model log-odds of the HLB-positive class, whereas negative coefficients indicate an association toward the HLB-negative class. Because correlated predictors can redistribute coefficient magnitude among neighboring wavelengths, individual coefficients were not interpreted as direct measures of independent wavelength importance; instead, their magnitude, variability, and directional consistency across FIELD holdouts were evaluated [39].

For each wavelength λ, the mean standardized coefficient across the four FIELD holdouts was calculated as:

$$\beta_\lambda = (1/K)\ \Sigma K_{k=1}\ \beta_{\lambda,k}, \quad (14)$$

where $\beta_{\lambda,k}$ is the standardized Logistic Regression coefficient associated with wavelength λ in holdout configuration k, and K = 4 represents the four leave-one-FIELD-out configurations. The mean coefficient summarizes the average direction and magnitude of the model association for each wavelength across the spatially blocked evaluations.

Coefficient variability across FIELD holdouts was quantified using the sample standard deviation:

$$SD_{\beta\lambda} = \sqrt{[\Sigma K_{k=1}\ (\beta_{\lambda,k} - \bar{\beta_\lambda})^2/(K-1)]}, \quad (15)$$

Lower $SD_{\beta\lambda}$ values indicate greater stability of coefficient magnitude across FIELD holdouts, whereas larger values indicate greater variation among spatial partitions.

Because coefficient direction may be stable even when its magnitude varies, the mean absolute coefficient was also calculated to characterize the overall magnitude of the model-associated contribution independently of sign:

$$mean\ |\beta|_{\lambda} = (1/K)\ \Sigma K_{k=1}\ |\beta_{\lambda,k}|, \quad (16)$$

Finally, directional stability was quantified using sign consistency (SC), defined as the proportion of FIELD holdouts in which a wavelength retained its dominant coefficient sign:

$$SC_{\lambda} = max(n_{\beta}>0,\ n_{\beta}<0)/K, \quad (17)$$

where $n_{\beta}>0$ and $n_{\beta}<0$ represent the number of FIELD holdouts in which the coefficient was positive or negative, respectively.

A wavelength was considered directionally stable when the coefficient retained the same sign in all four FIELD holdouts ($SC_{\lambda}=1.0$). To avoid relying on an arbitrary absolute-weight threshold, stable model-associated spectral contributions were defined using two simultaneous criteria: (i) identical coefficient sign across all four FIELD holdouts and (ii) mean absolute coefficient within the upper 10% of the wavelength-specific coefficient distribution. The overall data organization, preprocessing, and model-evaluation workflow adopted in this study is summarized in Figure 2.

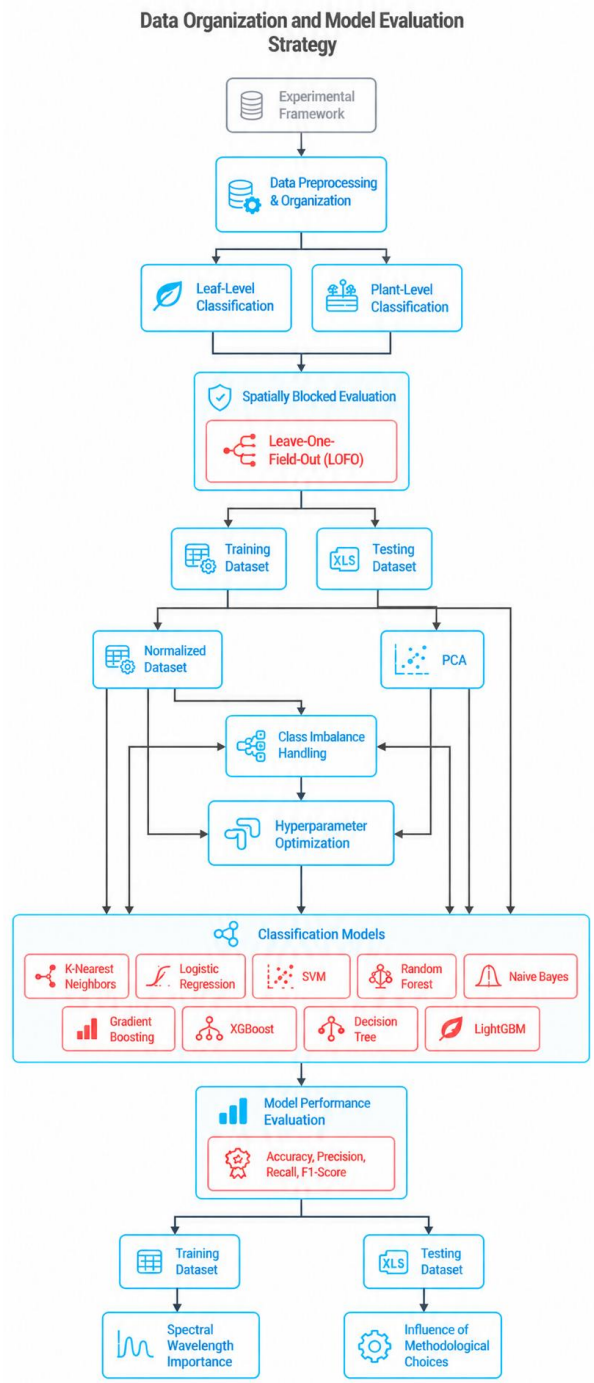


Figure 2. Workflow of data organization and model evaluation adopted in this study. Leaf- and plant-level representations were evaluated using a spatially blocked leave-one-FIELD-out framework, in which three FIELDs were used for model development and the remaining FIELD was withheld for final evaluation. Data-dependent preprocessing and model optimization were restricted to the model-development data.

3. Results

3.1. Baseline Model Performance

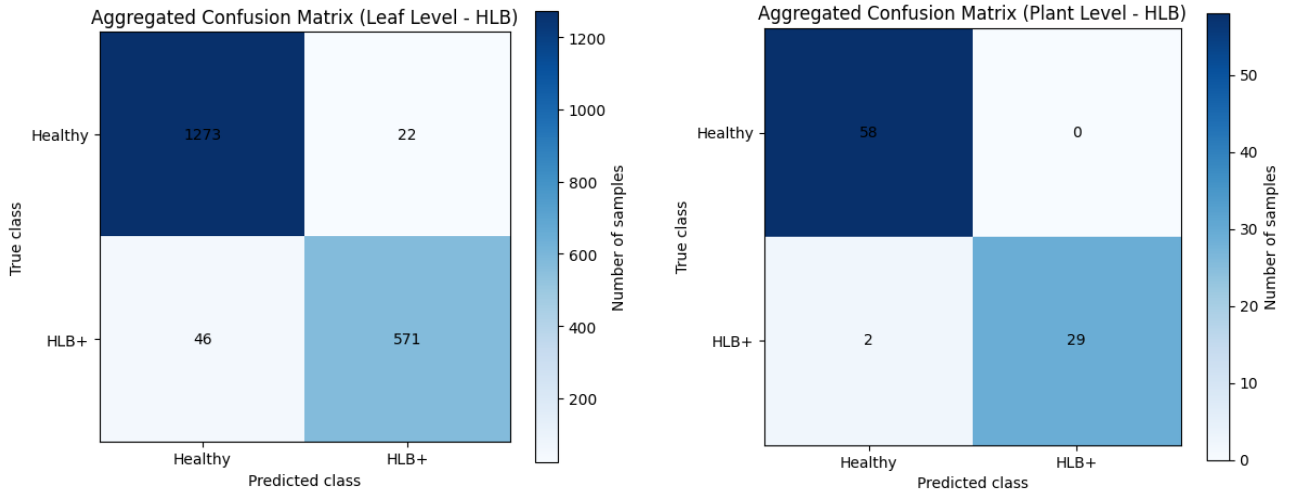
The overall classification performance at the leaf level is summarized in Table 3, where consistently high metric values are observed across most algorithms. Among the evaluated classifiers, Logistic Regression showed the highest observed test performance in most FIELD configurations, with accuracy and weighted F1-score exceeding 0.94, indicating strong discrimination between HLB-negative and HLB-positive samples within

the evaluated dataset. Support Vector Machine and ensemble-based methods, including Random Forest, Gradient Boosting, LightGBM, and XGBoost, also exhibited robust and comparable performance, albeit with slightly greater variability across spatial partitions. In contrast, Naïve Bayes showed substantially lower performance, suggesting that its underlying assumptions were less suitable for the highly correlated hyperspectral feature space evaluated in this study.

Table 3. Classification performance of machine-learning models for HLB detection at the leaf level, evaluated on the test set under a class-balanced training scheme and leave-one-FIELD-out validation. Metrics include accuracy, weighted precision, weighted recall, and weighted F1-score.

Classifier	Accuracy	Precision (Weighted)	Recall (Weighted)	f1 Score (Weighted)
Logistic Regression	0.96801 ± 0.01945	0.96826 ± 0.01921	0.96801 ± 0.01945	0.96778 ± 0.01984
Support Vector Machine	0.94569 ± 0.02281	0.94879 ± 0.02193	0.94569 ± 0.02281	0.94500 ± 0.02298
K-Nearest Neighbors	0.91052 ± 0.02675	0.91438 ± 0.02657	0.91052 ± 0.02675	0.90869 ± 0.02806
Naive Bayes	0.64041 ± 0.19783	0.78216 ± 0.12432	0.64041 ± 0.19783	0.62049 ± 0.22635
Random Forest	0.92869 ± 0.03586	0.93401 ± 0.03290	0.92869 ± 0.03586	0.92865 ± 0.03531
Gradient Boosting	0.92730 ± 0.02825	0.93195 ± 0.02643	0.92730 ± 0.02825	0.92702 ± 0.02781
Decision Tree	0.87667 ± 0.04333	0.88341 ± 0.04079	0.87667 ± 0.04333	0.87780 ± 0.04282
LightGBM	0.92683 ± 0.02434	0.93218 ± 0.02253	0.92683 ± 0.02434	0.92678 ± 0.02374
XGBoost	0.93235 ± 0.02356	0.93668 ± 0.02228	0.93235 ± 0.02356	0.93222 ± 0.02337

A more detailed representation of classification behavior is provided by the aggregated confusion matrices shown in Figure 3a-b, which synthesize model performance across FIELD combinations. Although the number of correctly classified samples is high, the matrices reveal the presence of both false positives and false negatives at the leaf level. These errors are not uniformly distributed, indicating that model performance is influenced by the specific spatial configuration of training and test sets. This behavior suggests that, despite the strong discriminatory power of hyperspectral data, local spectral variability and spatial heterogeneity introduce subtle shifts in classification boundaries when evaluated under strict leave-one-FIELD-out validation conditions.



(**a**) (**b**)

Figure 3. (**a**) Aggregated confusion matrices for HLB detection at the leaf level across selected FIELD combinations, evaluated on the test set and displayed using a blue-scale colormap; (**b**) Aggregated confusion matrices for HLB detection at the plant level across selected FIELD combinations, evaluated on the test set.

3.2. Effect of Analysis Scale (Leaf vs. Plant)

The comparison between leaf-level and plant-level modeling reveals a clear scale-dependent effect in classification performance. At the leaf level, hyperspectral signatures contained sufficient information to discriminate between samples from HLB-negative and HLB-positive plants, although performance varied across FIELD partitions. This variability may be associated with local physiological heterogeneity and leaf-to-leaf spectral variation within individual plants.

This contrast is evident when comparing leaf-level results (Table 3) with plant-level performance (Table 4), where plant-level models exhibited more consistent classification performance across the evaluated FIELD partitions. While leaf-level models exhibit variability across spatial partitions, plant-level models consistently achieve high and stable performance, with near-perfect recall in several configurations.

Table 4. Classification performance of machine learning models for HLB detection at the plant level, evaluated on the test set using normalized and class-balanced training data. Metrics include accuracy, weighted precision, weighted recall, and weighted F1-score.

Classifier	Accuracy	Precision (Weighted)	Recall (Weighted)	f1 Score (Weighted)
Logistic Regression	0.94138 ± 0.09564	0.96056 ± 0.07170	0.94138 ± 0.09564	0.94262 ± 0.09293
Support Vector Machine	0.95803 ± 0.03263	0.96186 ± 0.02936	0.95803 ± 0.03263	0.95704 ± 0.03378
K-Nearest Neighbors	0.90104 ± 0.08331	0.92531 ± 0.05499	0.90104 ± 0.08331	0.89877 ± 0.08335
Naive Bayes	0.74899 ± 0.22070	0.85660 ± 0.11171	0.74899 ± 0.22070	0.72442 ± 0.27023
Random Forest	0.94138 ± 0.09564	0.96056 ± 0.07170	0.94138 ± 0.09564	0.94262 ± 0.09293
Gradient Boosting	0.92471 ± 0.12862	0.95447 ± 0.07900	0.92471 ± 0.12862	0.92579 ± 0.12621
Decision Tree	0.92471 ± 0.12862	0.95447 ± 0.07900	0.92471 ± 0.12862	0.92579 ± 0.12621
LightGBM	0.92471 ± 0.12862	0.95481 ± 0.07843	0.92471 ± 0.12862	0.92599 ± 0.12629
XGBoost	0.90805 ± 0.16174	0.94869 ± 0.08655	0.90805 ± 0.16174	0.91019 ± 0.15718

This improvement is further illustrated in the aggregated confusion matrices presented in Figure 3b. In contrast to the leaf-level results, where classification errors are visibly distributed across classes, plant-level matrices exhibit strong diagonal dominance, with negligible false negatives and minimal false positives. This pattern indicates more consistent classification outcomes at the plant level across the evaluated FIELD partitions.

3.3. Effect of Algorithmic Refinements

3.3.1. Effect of Class Balancing

The impact of class balancing was evaluated through a direct comparison between balanced and unbalanced training configurations across all FIELD-based experiments at both leaf and plant levels (Tables 5 and 6). Prior to model training, spectral variables were standardized to place measurements acquired at different wavelengths on a common numerical scale, preventing variables with larger numerical ranges from dominating the learning process.

Table 5. Observed performance under each leave-one-FIELD-out evaluation for the classifier reported in the corresponding experiment at the leaf level, evaluated on the withheld test FIELD under balanced and unbalanced training conditions without hyperparameter optimization. Metrics include weighted precision, weighted recall, weighted F1-score, HLB-positive precision (PPV), HLB-positive sensitivity (recall), HLB-positive F1-score, HLB-negative specificity, balanced accuracy, and Matthews correlation coefficient (MCC).

Experiment (Test)	Precision (Weighted)	Recall (Weighted)	F1 (Weighted)	Sensitivity (HLB+)	Specificity (HLB−)	Balanced Accuracy	MCC	Precision (HLB+)	F1 (HLB+)	Configuration
FIELD_7_LEAFS	0.990	0.990	0.990	0.990	0.990	0.990	0.978	0.980	0.985	Unbalanced

FIELD_9_LEAFS	0.967	0.967	0.967	0.953	0.973	0.963	0.924	0.944	0.948	Unbalanced
FIELD_2_LEAFS	0.944	0.943	0.942	0.859	0.984	0.921	0.869	0.962	0.907	Unbalanced
FIELD_1_LEAFS	0.972	0.972	0.972	0.947	0.984	0.966	0.936	0.966	0.956	Unbalanced
FIELD_7_LEAFS	0.990	0.990	0.990	0.990	0.990	0.990	0.978	0.980	0.985	Balanced
FIELD_9_LEAFS	0.967	0.967	0.967	0.953	0.973	0.963	0.924	0.944	0.948	Balanced
FIELD_2_LEAFS	0.944	0.943	0.942	0.859	0.984	0.921	0.869	0.962	0.907	Balanced
FIELD_1_LEAFS	0.972	0.972	0.972	0.947	0.984	0.966	0.936	0.966	0.956	Balanced

Table 6. Observed performance under each leave-one-FIELD-out evaluation for the classifier reported in the corresponding experiment at the plant level, evaluated on the withheld test FIELD under balanced and unbalanced training conditions without hyperparameter optimization. Metrics include weighted precision, weighted recall, weighted F1-score, HLB-positive precision (PPV), HLB-positive sensitivity (recall), HLB-positive F1-score, HLB-negative specificity, balanced accuracy, and Matthews correlation coefficient (MCC).

Experiment (Test)	Precision (Weighted)	Recall (Weighted)	F1 (Weighted)	Sensitivity (HLB+)	Specificity (HLB−)	Balanced Accuracy	MCC	Precision (HLB+)	F1 (HLB+)	Configuration
FIELD_7_LEAFS	1	1	1	1.000	1.000	1.000	1.000	1.000	1.000	Unbalanced
FIELD_9_LEAFS	0.939	0.933	0.931	0.800	1.000	0.900	0.853	1.000	0.889	Unbalanced
FIELD_2_LEAFS	0.967	0.966	0.965	0.900	1.000	0.950	0.925	1.000	0.947	Unbalanced
FIELD_1_LEAFS	1	1	1	1.000	1.000	1.000	1.000	1.000	1.000	Unbalanced
FIELD_7_LEAFS	1	1	1	1.000	1.000	1.000	1.000	1.000	1.000	Balanced
FIELD_9_LEAFS	0.939	0.933	0.931	0.800	1.000	0.900	0.853	1.000	0.889	Balanced
FIELD_2_LEAFS	0.967	0.966	0.965	0.900	1.000	0.950	0.925	1.000	0.947	Balanced
FIELD_1_LEAFS	1	1	1	1.000	1.000	1.000	1.000	1.000	1.000	Balanced

As shown in Table 4, variations in performance are better explained by the specific combination of FIELD used in each experiment than by the application of class balancing.

The class-specific metrics provide additional insight into the behavior of the classifiers beyond the support-weighted performance measures. At the leaf level, HLB-positive precision ranged from 0.944 to 0.980, while HLB-positive F1-score ranged from 0.907 to 0.985 across the four withheld FIELDs. The lowest HLB-positive F1-score occurred when FIELD 2 was withheld (0.907), primarily reflecting the lower HLB-positive sensitivity observed in this partition (0.859), despite a high positive predictive value (0.962). Importantly, both HLB-positive precision and HLB-positive F1-score remained unchanged between balanced and unbalanced configurations, reinforcing the observation that class balancing did not materially affect the classification outcomes in these experiments.

At the plant level, HLB-positive precision was 1.000 in all evaluated FIELDs, indicating that no HLB-negative plants were incorrectly predicted as HLB-positive. HLB-positive F1-score ranged from 0.889 to 1.000, with the lowest value observed when FIELD 9 was withheld, where sensitivity decreased to 0.800. Thus, the remaining classification errors at the plant level were attributable to false-negative rather than false-positive predictions. As at the leaf level, the identical class-specific metrics obtained under balanced and unbalanced configurations indicate that class balancing did not modify the final plant-level predictions.

The class-specific metrics further indicate that the apparent stability of the global weighted metrics does not imply identical performance for both diagnostic classes. At the plant level, HLB-negative specificity remained 1.000 across the evaluated FIELD-based configurations, whereas HLB-positive sensitivity ranged from 0.800 to 1.000. Consequently, balanced accuracy ranged from 0.900 to 1.000 and MCC from 0.853 to 1.000. These results emphasize the importance of reporting class-specific and imbalance-sensitive metrics in addition to support-weighted performance measures.

3.3.2. Hyperparameter Optimization

Hyperparameter optimization produced limited and configuration-dependent effects. At the leaf level, Optuna did not alter the final classification outcomes across the evaluated FIELD-based partitions. At the plant level, three of the four FIELD holdouts showed unchanged confusion matrices, whereas the configuration in which FIELD 9 was withheld showed an improvement in the optimized unbalanced model, eliminating the single false-negative prediction observed in the corresponding baseline configuration (Tables 7 and 8).

Table 7. Observed performance under each leave-one-FIELD-out evaluation for the classifier reported in the corresponding experiment at the leaf level after hyperparameter optimization with Optuna, evaluated on the withheld test FIELD under balanced and unbalanced training configurations. Metrics include weighted precision, weighted recall, weighted F1-score, HLB-positive precision (PPV), HLB-positive sensitivity (recall), HLB-positive F1-score, HLB-negative specificity, balanced accuracy, and Matthews correlation coefficient (MCC).

Experiment (Test)	Precision (Weighted)	Recall (Weighted)	F1 (Weighted)	Sensitivity (HLB+)	Specificity (HLB−)	Balanced Accuracy	MCC	Precision (HLB+)	F1 (HLB+)	Configuration
FIELD_7_LEAF	0.990	0.990	0.990	0.990	0.990	0.990	0.978	0.980	0.985	Unbalanced
FIELD_9_LEAF	0.967	0.967	0.967	0.953	0.973	0.963	0.924	0.944	0.948	Unbalanced
FIELD_2_LEAF	0.944	0.943	0.942	0.859	0.984	0.921	0.869	0.962	0.907	Unbalanced
FIELD_1_LEAF	0.972	0.972	0.972	0.947	0.984	0.966	0.936	0.966	0.956	Unbalanced
FIELD_7_LEAF	0.990	0.990	0.990	0.990	0.990	0.990	0.978	0.980	0.985	Balanced
FIELD_9_LEAF	0.967	0.967	0.967	0.953	0.973	0.963	0.924	0.944	0.948	Balanced
FIELD_2_LEAF	0.944	0.943	0.942	0.859	0.984	0.921	0.869	0.962	0.907	Balanced
FIELD_1_LEAF	0.972	0.972	0.972	0.947	0.984	0.966	0.936	0.966	0.956	Balanced

Table 8. Observed performance under each leave-one-FIELD-out evaluation for the classifier reported in the corresponding experiment at the plant level after hyperparameter optimization with Optuna, evaluated on the withheld test FIELD under balanced and unbalanced training configurations. Metrics include weighted precision, weighted recall, weighted F1-score, HLB-positive precision (PPV), HLB-positive sensitivity (recall), HLB-positive F1-score, HLB-negative specificity, balanced accuracy, and Matthews correlation coefficient (MCC).

Experiment (Test)	Precision (Weighted)	Recall (Weighted)	F1 (Weighted)	Sensitivity (HLB+)	Specificity (HLB−)	Balanced Accuracy	MCC	Precision (HLB+)	F1 (HLB+)	Configuration
FIELD_7_PLANT	1	1	1	1.000	1.000	1.000	1.000	1.000	1.000	Unbalanced
FIELD_9_PLANT	1	1	1	1.000	1.000	1.000	1.000	1.000	1.000	Unbalanced
FIELD_2_PLANT	0.967	0.966	0.965	0.900	1.000	0.950	0.925	1.000	0.947	Unbalanced
FIELD_1_PLANT	1	1	1	1.000	1.000	1.000	1.000	1.000	1.000	Unbalanced
FIELD_7_PLANT	1	1	1	1.000	1.000	1.000	1.000	1.000	1.000	Balanced
FIELD_9_PLANT	0.939	0.933	0.931	0.800	1.000	0.900	0.853	1.000	0.889	Balanced
FIELD_2_PLANT	0.967	0.966	0.965	0.900	1.000	0.950	0.925	1.000	0.947	Balanced
FIELD_1_PLANT	1	1	1	1.000	1.000	1.000	1.000	1.000	1.000	Balanced

The class-specific metrics confirm that the effect of hyperparameter optimization was limited to a specific plant-level configuration. At the leaf level, HLB-positive precision and HLB-positive F1-score remained unchanged after optimization across all withheld FIELDs, consistent with the identical final classification outcomes. At the plant level, the optimized unbalanced configuration for the withheld FIELD 9 eliminated the false-negative prediction observed in the corresponding baseline model, increasing HLB-positive sensitivity from 0.800 to 1.000 and HLB-positive F1-score from 0.889 to 1.000, while HLB-positive precision remained 1.000. The remaining FIELD configurations showed no corresponding change in class-specific performance, indicating that the benefit of hyperparameter optimization was localized rather than systematic across spatial partitions.

3.4. Effect of Dimensionality Reduction (PCA)

Across all FIELD combinations, the application of Principal Component Analysis (PCA) resulted in consistently high predictive performance, although slightly lower than that observed using the full hyperspectral feature space. Accuracy ranged from approximately 0.90 to 0.95, whereas F1-scores varied between 0.83 and 0.92 (Table 9), indicating that the reduced feature space retained most of the discriminative information required for HLB detection.

All data-dependent preprocessing steps were estimated exclusively from the corresponding training data. Within internal plant-grouped cross-validation, standardization parameters and, when applicable, PCA transformations were fitted using the internal training partition and subsequently applied to the corresponding validation partition without refitting.

Table 9. Performance of PCA-based models for HLB detection at the leaf level under each leave-one-FIELD-out evaluation, evaluated on the withheld test FIELD and comparing configurations without and with Optuna hyperparameter optimization. Metrics include weighted precision, weighted recall, weighted F1-score, HLB-positive precision (PPV), HLB-positive sensitivity (recall), HLB-positive F1-score, HLB-negative specificity, balanced accuracy, and Matthews correlation coefficient (MCC).

Experiment (Test)	Precision (Weighted)	Recall (Weighted)	F1 (Weighted)	Sensitivity (HLB+)	Specificity (HLB−)	Balanced Accuracy	MCC	Precision (HLB+)	F1 (HLB+)	Configuration
FIELD_7_LEAF	0.967	0.967	0.966	0.898	1.000	0.949	0.923	1.000	0.946	Without Optuna
FIELD_9_LEAF	0.922	0.922	0.920	0.821	0.969	0.895	0.817	0.926	0.870	Without Optuna
FIELD_2_LEAF	0.913	0.913	0.910	0.780	0.977	0.879	0.799	0.941	0.853	Without Optuna
FIELD_1_LEAF	0.921	0.919	0.917	0.793	0.979	0.866	0.814	0.948	0.864	Without Optuna
FIELD_7_LEAF	0.967	0.967	0.966	0.898	1.000	0.949	0.923	1.000	0.946	With Optuna
FIELD_9_LEAF	0.922	0.922	0.920	0.821	0.969	0.895	0.817	0.926	0.870	With Optuna
FIELD_2_LEAF	0.913	0.913	0.910	0.780	0.977	0.879	0.799	0.941	0.853	With Optuna
FIELD_1_LEAFS	0.921	0.919	0.917	0.793	0.979	0.866	0.814	0.948	0.864	With Optuna

The number of principal components required to retain at least 95% of the cumulative spectral variance showed limited variation among the leave-one-FIELD-out iterations. At the leaf level, five principal components were retained in all four iterations, explaining 95.69–96.53% of the cumulative variance, whereas at the plant level, four to five components were retained, explaining 95.21–97.46%. Despite this substantial dimensionality reduction, PCA-based models maintained high classification performance, although positive-class performance showed greater FIELD-dependent variation at the leaf level. HLB-positive precision remained high across all leaf-level holdouts (0.926–1.000), while sensitivity ranged from 0.780 to 0.898 and HLB-positive F1-score from 0.853 to 0.946. The lowest sensitivity occurred when FIELD 2 was withheld (0.780), followed by FIELD 1 (0.793), whereas FIELD 7 showed the highest sensitivity (0.898) and perfect HLB-positive precision (1.000). At the plant level, PCA-based models were more stable, with HLB-positive precision of 1.000 across all FIELDs and HLB-positive F1-scores ranging from 0.947 to 1.000 (Table 10). Overall, the PCA representation preserved high predictive performance while revealing greater FIELD-dependent variability in HLB-positive classification at the leaf level than at the plant level.

Table 10. Performance of PCA-based models for HLB detection at the plant level under each leave-one-FIELD-out evaluation, evaluated on the withheld test FIELD and comparing configurations without and with Optuna hyperparameter optimization. Metrics include weighted precision, weighted recall, weighted F1-score, HLB-positive precision (PPV), HLB-positive sensitivity (recall), HLB-positive F1-score, HLB-negative specificity, balanced accuracy, and Matthews correlation coefficient (MCC).

Experiment (Test)	Precision (Weighted)	Recall (Weighted)	F1 (Weighted)	Sensitivity (HLB+)	Specificity (HLB−)	Balanced Accuracy	MCC	Precision (HLB+)	F1 (HLB+)	Configuration
FIELD_7_PLANT	1	1	1	1.000	1.000	1.000	1.000	1.000	1.000	Without Optuna
FIELD_9_PLANT	1	1	1	1.000	1.000	1.000	1.000	1.000	1.000	Without Optuna
FIELD_2_PLANT	0.967	0.966	0.965	0.900	1.000	0.950	0.925	1.000	0.947	Without Optuna
FIELD_1_PLANT	1	1	1	1.000	1.000	1.000	1.000	1.000	1.000	Without Optuna
FIELD_7_PLANT	1	1	1	1.000	1.000	1.000	1.000	1.000	1.000	With Optuna
FIELD_9_PLANT	1	1	1	1.000	1.000	1.000	1.000	1.000	1.000	With Optuna
FIELD_2_PLANT	0.967	0.966	0.965	0.900	1.000	0.950	0.925	1.000	0.947	With Optuna
FIELD_1_PLANT	1	1	1	1.000	1.000	1.000	1.000	1.000	1.000	With Optuna

Across both scales, PCA-based models maintain performance levels comparable to full-spectrum models, indicating that the most relevant spectral information is retained in a reduced set of components.

3.5. Spectral Contribution and Logistic Regression Coefficient Stability

Among the evaluated classifiers, Logistic Regression exhibited consistently high performance across the FIELD-based experiments and was therefore examined as an interpretable linear model for investigating model-associated spectral contributions. This post hoc interpretive analysis was not used for model selection or hyperparameter tuning.

Because adjacent hyperspectral wavelengths are strongly correlated, individual Logistic Regression coefficients were not interpreted as independent measures of wavelength importance. Instead, spectral contribution was evaluated according to coefficient stability across the four leave-one-FIELD-out configurations. For each wavelength, the mean standardized coefficient, coefficient standard deviation, mean absolute coefficient, and sign consistency were calculated.

A total of 195 wavelengths satisfied the adopted stability criterion, retaining the same coefficient sign across all four FIELD holdouts while also ranking within the upper 10% of the mean absolute coefficient distribution (Figure 4). The most extensive stable spectral intervals occurred at 645–655, 1378–1391, 1414–1439, 1564–1628, 1891–1916, 2017–2025, 2029–2046, 2049–2053, and 2119–2123 nm. The 1564–1628 nm interval represented the longest continuous region of stable coefficients, comprising 65 wavelengths.

Several of the stable intervals occurred within the SWIR domain, including wavelength regions associated with water-sensitive absorption features. Stable contributions around 1378–1439 nm and 1891–1916 nm were particularly noteworthy because these spectral regions had been deliberately retained during preprocessing to preserve potentially relevant information related to leaf water status and biochemical composition.

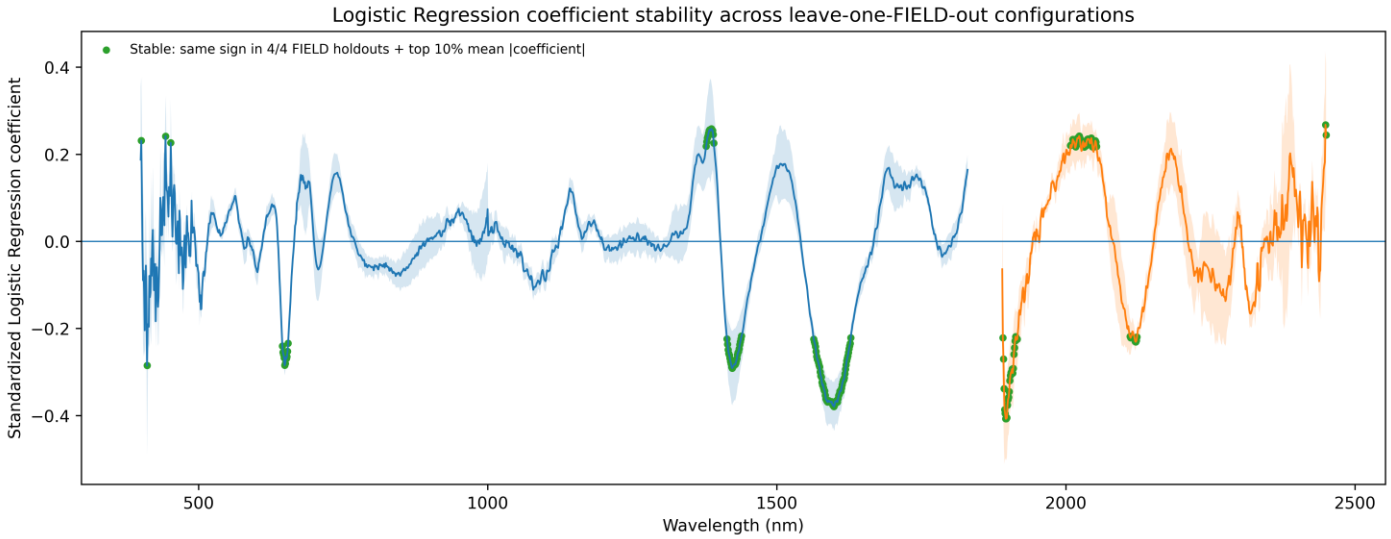


Figure 4. Stability of standardized Logistic Regression coefficients across the four leave-one-FIELD-out configurations. The solid line represents the mean coefficient at each wavelength, whereas the shaded region represents ±1 standard deviation across FIELD holdouts. Highlighted wavelengths retained the same coefficient sign across all four FIELD holdouts and exhibited a mean absolute coefficient within the upper 10% of the coefficient distribution. Line colors distinguish the retained spectral segments for visualization and do not represent different models or classes. The gap between 1831 and 1889 nm corresponds to the spectral interval excluded during preprocessing.

Figure 5 illustrates the distribution of Logistic Regression coefficients as a function of wavelength at the leaf and plant levels. The leaf-level model exhibits pronounced oscillations in spectral weights across the entire spectral range, with several regions showing strong positive and negative deviations from the zero-weight reference line. This behavior indicates a higher sensitivity of the leaf-based model to localized spectral variations, reflecting the intrinsic heterogeneity at the leaf scale. In contrast, the plant-level model presents a smoother weight profile, with lower-amplitude fluctuations concentrated around zero, suggesting that averaging spectral responses at the plant level effectively attenuates noise and reduces the influence of leaf-to-leaf variability. Despite differences in coefficient magnitude, both models consistently identify similar spectral regions with non-zero weights, particularly within the red-edge and shortwave infrared domains. Positive coefficients indicate wavelengths that contribute to the identification of HLB-positive samples, whereas negative coefficients are associated with the discrimination of non-HLB-infected samples, while values close to zero denote low discriminative relevance. These regions are therefore identified as having greater discriminatory power between HLB-positive and HLB-negative samples, indicating that the underlying physiological alterations induced by Huanglongbing are preserved across scales. Despite differences in coefficient magnitude, both representations showed contributions within similar spectral domains, particularly in the red-edge and shortwave infrared regions. This cross-scale consistency suggests that model-associated spectral information relevant to HLB discrimination was preserved after plant-level aggregation. Given the strong collinearity among adjacent hyperspectral wavelengths, these patterns should be interpreted at the spectral-domain level rather than as evidence of independent wavelength importance.

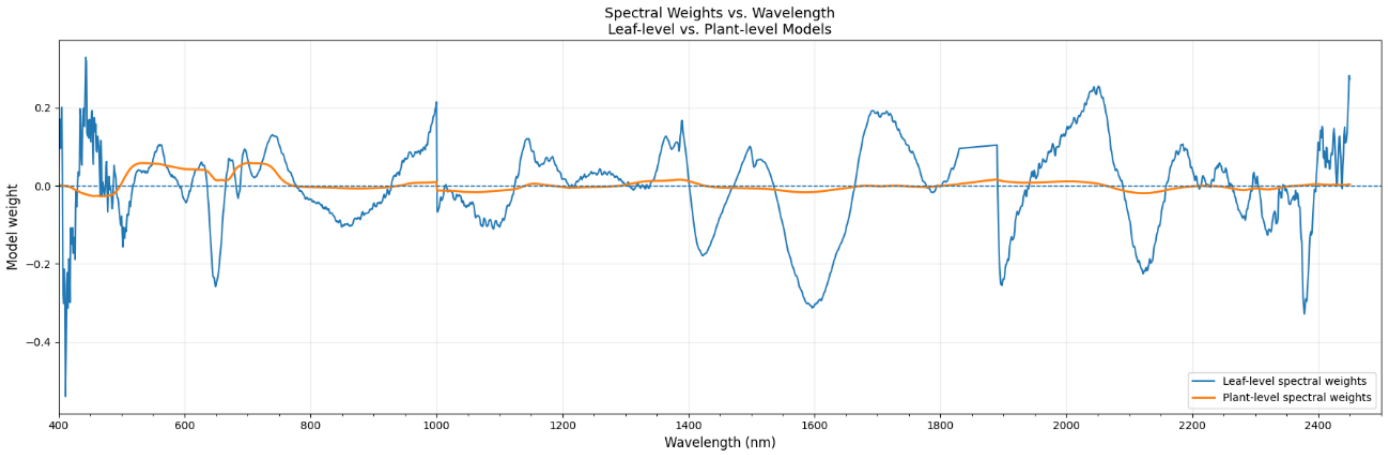


Figure 5. Comparison of Logistic Regression coefficient patterns at the leaf and plant levels as a function of wavelength. The curves represent the standardized spectral coefficients obtained for the two organizational scales, allowing comparison of model-associated spectral contributions across the evaluated wavelength range.

4. Discussion

The results demonstrate that hyperspectral reflectance contains sufficiently structured information to discriminate PCR-confirmed HLB-positive and HLB-negative citrus samples under a leave-one-field-out validation framework. Logistic Regression achieved an accuracy (weighted) of approximately 0.968 at the leaf level, whereas Support Vector Machine reached approximately 0.958 at the plant level. These values are consistent with previous studies demonstrating that HLB-related physiological changes can be detected from visible, near-infrared, and hyperspectral observations [4–6,21]. Li et al. [4] reported distinguishable spectral characteristics between non-HLB-infected and HLB-infected citrus leaves, while Sankaran et al. [5] demonstrated the feasibility of VIS–NIR spectroscopy for HLB detection under orchard conditions. At the canopy scale, Deng et al. [6] obtained high HLB detection performance using UAV-based hyperspectral imagery and a multi-input neural-network framework. Together, these findings indicate that HLB produces detectable optical alterations across different sensing platforms and organizational scales. The present study extends this evidence by showing that high discrimination can also be achieved using parsimonious and interpretable classifiers when training and testing samples originate from different orchard fields.

Furthermore, HLB-negative status indicates a negative molecular diagnosis for HLB and should not be interpreted as confirmation of the absence of other biotic or abiotic stresses. Accordingly, the reported discrimination should be understood as separation between HLB-positive and HLB-negative samples within the conditions represented by the present dataset.

The strong performance obtained with Logistic Regression suggests that increasing algorithmic complexity is not necessarily required when the spectral feature space already provides substantial class separation. This result agrees with studies showing that linear or moderately complex classifiers can perform competitively in plant-disease spectroscopy when physiologically relevant wavelengths are adequately represented [10,12,13,25].

Residual FIELD-specific variability likely reflects a combination of biological and spatial differences among orchard areas and residual acquisition-related variability. Although spectral measurements were obtained using a fixed portable setup with artificial tungsten-halogen illumination and a consistent sensor–sample geometry, field-based proximal spectroscopy may still be affected by subtle variations in sample positioning,

leaf properties, instrument response, and other acquisition-related factors. In contrast, UAV-based studies frequently rely on convolutional or multi-input neural networks because image observations include additional variability caused by canopy geometry, soil background, shadows, illumination conditions, viewing geometry, and mixed pixels [6,17,18,22]. Therefore, the difference between the parsimonious model favored in the present study and the more complex architectures used in UAV studies is likely associated with the nature of the input data. Direct foliar measurements reduce background and geometric effects, increasing the relative contribution of biochemical and structural information to classification. Although the favored model differs from the complex neural-network architectures adopted in some UAV-based studies, this difference does not necessarily indicate disagreement in the underlying biological response. Instead, it likely reflects differences in sensing scale, background interference, spatial complexity, and input dimensionality.

The present results were obtained from leaf-level hyperspectral measurements and should not be directly extrapolated to canopy-, UAV-, or satellite-based observations. At broader observational scales, spectral responses are additionally influenced by canopy architecture, shadowing, background contributions, illumination geometry, atmospheric effects, and the spectral and spatial characteristics of the imaging sensor. Therefore, the spectral domains identified here should be regarded as candidate regions for subsequent evaluation rather than as directly transferable HLB detection bands. Future studies should determine whether these spectral patterns remain discriminative when moving from leaf spectroscopy to canopy-level and imaging-based observations.

A major contribution of this study is the evaluation of HLB classification at two complementary organizational scales. Individual-leaf models provided high classification performance but showed greater variation across the evaluated FIELD partitions, whereas plant-level models based on mean spectral signatures exhibited consistently high performance. These differences should be interpreted in the context of the distinct observational units represented by the two analyses rather than as directly comparable error rates. Leaves collected from the same plant are not statistically independent because they share physiological, structural, and environmental characteristics. Moreover, PCR diagnosis was performed using pooled material at the plant level, meaning that all leaves inherited the same diagnostic label even though pathogen distribution, symptom expression, and associated spectral alterations may vary within the canopy. Consequently, an HLB-positive plant may contain leaves with strong spectral alterations as well as leaves with weak or undetectable responses, introducing label uncertainty into leaf-level classification. Plant-level averaging integrates multiple leaf observations into a single representation of each biological individual, thereby reducing the contribution of leaf-specific spectral variability. However, because the leaf- and plant-level analyses differ both in spectral representation and in the biological unit used for evaluation, the observed performance differences cannot be attributed exclusively to spectral averaging.

Importantly, plant-level aggregation preserved the physiologically meaningful spectral domains identified at the leaf level, including the visible, red-edge, near-infrared, and shortwave infrared regions. The smoother coefficient profile observed at the plant level should therefore not be interpreted as a loss of disease-related information, but as evidence that aggregation attenuated high-frequency leaf-to-leaf variability while retaining the most consistent HLB-related signals. This result motivates subsequent evaluation at broader observational scales because UAV and satellite observations generally integrate canopy-level responses rather than isolated leaves; however, transferability of the identified spectral patterns was not evaluated here.

The spectral domains identified by the classifiers are consistent with known physiological responses of vegetation to disease. The importance of visible wavelengths

can be associated primarily with changes in photosynthetic pigments, including chlorophyll and carotenoids, whereas the recurrent contribution of the red-edge region reflects the transition between strong chlorophyll absorption in the red domain and structural scattering in the near-infrared [7,26,28]. HLB disrupts phloem transport, promotes carbohydrate accumulation, affects chloroplast function, and induces progressive chlorosis and asymmetric mottling [2,3]. These processes can modify absorption in the blue, green, and red regions and alter the position and slope of the red edge. The persistence of red-edge contributions at both scales indicates that pigment-related HLB alterations were sufficiently systematic to remain detectable after plant-level aggregation.

The contribution of the near-infrared domain is plausibly associated with changes in mesophyll organization, cellular integrity, and multiple scattering within the leaf. Although these responses are not necessarily specific to HLB, disease-induced structural disruption can modify the internal interfaces that control NIR reflectance. The shortwave infrared coefficients are also physiologically meaningful because this region includes absorption features related to water, carbohydrates, proteins, cellulose, lignin, and other biochemical constituents [27–30]. HLB affects carbon allocation and promotes starch accumulation in infected leaves [2,3]; therefore, the relevance assigned to SWIR wavelengths may reflect combined changes in water status and biochemical composition. The agreement between these domains and those identified in previous statistical analyses supports their physiological relevance, although direct biochemical measurements would be required to confirm the specific mechanisms responsible for each spectral response.

No improvement in classification performance was observed after class balancing under the evaluated configurations. This result should be interpreted in the context of the moderate class imbalance represented in the present dataset, with HLB-positive samples accounting for 34.8% of plants and 32.3% of leaf spectra. One possible explanation is that the HLB-positive samples in this dataset exhibited sufficient spectral contrast for the principal classifiers to identify them reliably, thereby limiting the observed effect of class balancing on the resulting decision boundaries. This finding should not be generalized as evidence that balancing is unnecessary for HLB detection; rather, it indicates that its effectiveness depends on spectral overlap, sample distribution, classifier characteristics, and validation design. In datasets containing less developed symptoms, greater environmental variability, or more severe class imbalance, weighting or resampling may still provide benefits.

Hyperparameter optimization with Optuna produced limited and configuration-dependent effects. Most FIELD-based evaluations were unchanged relative to their baseline configurations; however, a performance improvement was observed for one plant-level holdout. Therefore, no consistent advantage of hyperparameter optimization was observed across the evaluated configurations. A plausible explanation is that the default or initially established models were already operating within a stable region of the parameter space and that performance was constrained primarily by the information contained in the spectra. When class separation is strong, several hyperparameter combinations may converge to similar decision boundaries. Overall, the comparatively small and configuration-dependent effects of hyperparameter optimization, together with the larger observed differences associated with spectral representation, are consistent with the study hypothesis within the present dataset. This conclusion, however, should not be extrapolated to more heterogeneous datasets or image-based observations, for which optimization may be substantially more relevant.

The present machine learning analysis complements the wavelength-wise statistical characterization previously reported for the same spectral dataset [32]. Whereas the previous study focused on the statistical separability of HLB-positive and HLB-negative

spectral responses at leaf and plant levels, the present study evaluates their predictive behavior under spatially blocked FIELD-wise validation and examines the stability of spectral contributions within multivariate classification models. It should be noted that the wavelength-wise statistical tests reported in our previous study [32] used an unadjusted significance threshold of $p < 0.05$, without correction for multiple comparisons. Therefore, the overlap observed here should be interpreted as exploratory methodological convergence rather than as independent statistical validation of the identified spectral regions.

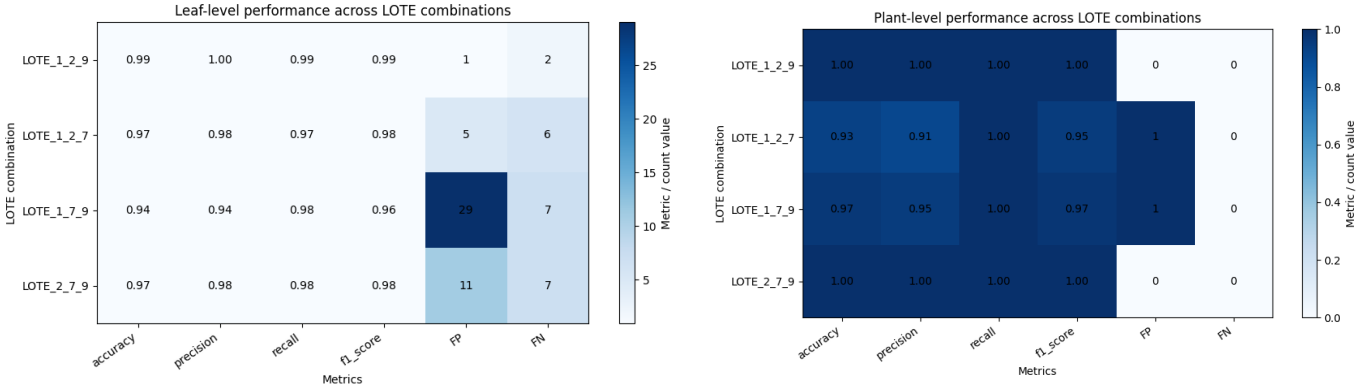
The partial agreement between the approaches is expected because univariate tests evaluate each wavelength independently, whereas multivariate classifiers exploit combinations of correlated spectral variables. A wavelength with a modest individual difference may contribute meaningfully when combined with information from other spectral regions. Conversely, a statistically significant wavelength may provide little additional predictive information when it is highly redundant with neighboring bands. Therefore, statistical significance and predictive importance should be regarded as complementary rather than interchangeable criteria. The overlap observed in Figure 6 indicates methodological convergence between the multivariate classification analysis and the previously reported univariate statistical results, while spectral regions identified only by the machine learning analysis illustrate the ability of multivariate models to integrate distributed and correlated spectral responses.

Figure 6. Comparison between stable model-associated spectral contributions and wavelength regions previously identified by univariate statistical analysis. (**A**) Spectral weights obtained from logistic regression models at leaf and plant levels. (**B**) Wavelength regions identified as statistically significant ($p < 0.05$) based on univariate analysis. (**C**) Comparison between machine learning-derived spectral importance and statistically significant wavelength intervals, highlighting regions of agreement and complementary information.

The comparison between full-spectrum and PCA-based models provides additional insight into the role of dimensionality reduction in hyperspectral classification. PCA substantially reduced the original hyperspectral feature space while retaining at least 95% of the cumulative spectral variance within each leave-one-FIELD-out iteration. Five principal components were retained across all leaf-level iterations, whereas four to five components were required at the plant level, demonstrating that a compact representation of the spectral variability was consistently obtained despite changes in the model-development FIELD combination. PCA-based models maintained high predictive performance, although the leaf-level results showed greater FIELD-dependent variation, particularly in HLB-positive sensitivity. This behavior is consistent with the unsupervised nature of PCA, which maximizes total spectral variance rather than class separability; consequently, disease-related information associated with comparatively low-variance spectral patterns may be attenuated during dimensionality reduction. At the plant level, this effect was less pronounced, suggesting that spectral aggregation may concentrate recurring spectral patterns and reduce within-plant variability in the retained low-dimensional representation.

Despite the high classification performance achieved, particularly at the plant level, the near-perfect metrics warrant careful interpretation. Accuracy and recall values approaching unity may reflect strong class separability within the experimental dataset, as illustrated in Figure 7, but may also be associated with the limited range of environmental conditions, disease expressions, host responses, and acquisition conditions represented in the study. In addition, the use of PCR-based labels at the plant level

introduces an inherent aggregation effect because infection status was assigned from pooled leaf material. This procedure may reduce label variability at the plant scale while masking intra-plant heterogeneity, potentially contributing to the observed performance difference between leaf- and plant-level models.



(**a**) (**b**)

Figure 7. (**a**) Heatmap of classification performance metrics across FIELD-based leave-one-FIELD-out validation scenarios at the leaf level, evaluated on the test set. Rows represent distinct training–testing partitions, and columns correspond to accuracy, precision, recall, F1-score, and error counts (FP and FN); (**b**) Heatmap of classification performance metrics across FIELD-based leave-one-FIELD-out validation scenarios at the plant level, evaluated on the test set. Rows represent distinct training–testing partitions, and columns correspond to accuracy, precision, recall, F1-score, and error counts (FP and FN).

From an operational perspective, plant-level aggregation provided a more stable basis for HLB discrimination, particularly under heterogeneous symptom distribution. The persistence of visible, red-edge, NIR, and SWIR information after aggregation suggests that the main physiological signals observed at the leaf level may remain detectable when measurements are integrated at the plant or canopy scale. The combination of leaf- and plant-level analyses may therefore support hierarchical sensing frameworks in which proximal spectroscopy is used to interpret physiological responses and canopy-oriented platforms are used for spatial monitoring.

The coefficient-stability analysis provides an additional perspective on the spectral domains associated with HLB discrimination. Stable model-associated contributions were observed not only in the visible region but predominantly across several SWIR intervals, including 1378–1391, 1414–1439, 1564–1628, and 1891–1916 nm. These regions are influenced by combinations of leaf water content, internal structure, and biochemical constituents, suggesting that HLB-related spectral discrimination may extend beyond pigment-associated responses in the visible and red-edge domains. The occurrence of stable coefficients near water-sensitive absorption features is particularly relevant because these wavelengths were intentionally retained during preprocessing whenever reliable measurements were available. However, because contiguous hyperspectral bands are highly collinear, these results should be interpreted as stable model-associated spectral regions rather than as evidence that individual wavelengths independently drive disease discrimination.

The Logistic Regression coefficient stability analysis across FIELD holdouts (Figure 4) and the comparison of leaf- and plant-level spectral coefficient patterns (Figure 5) provide complementary perspectives on model-associated spectral behavior. The stability analysis identifies wavelength regions whose contributions remain consistent across spatial partitions, whereas the cross-scale comparison shows that plant-level aggregation

reduces coefficient variability while preserving contributions within several of the same spectral domains. Together, these findings suggest that aggregation attenuates fine-scale spectral variability without eliminating the broader spectral regions consistently associated with HLB discrimination.

These stable regions also overlap partially with spectral intervals previously identified as statistically discriminative using the same experimental dataset [32]. This agreement should be interpreted as methodological convergence between statistical and machine learning analyses rather than as independent physiological validation.

Model performance was evaluated using a leave-one-field-out framework in which observations from each withheld field were used exclusively for testing. Compared with random sample partitioning, this strategy reduces the likelihood that local similarities among neighboring observations artificially inflate classification performance. Nevertheless, all four fields were located within the same commercial orchard and were sampled during a single acquisition period. Therefore, the results demonstrate generalization across internal spatially separated FIELD partitions but not transferability to independent orchards. The variation in performance among withheld FIELD should also not be attributed exclusively to physiological differences among plants. Hyperspectral reflectance is sensitive to illumination geometry, scattering, sensor–target configuration, and other acquisition-related factors, and residual differences among FIELD may therefore represent an acquisition-related domain shift in addition to biological heterogeneity. The variation in performance among withheld FIELDs should also not be attributed exclusively to physiological differences among plants. Hyperspectral reflectance may be influenced by illumination geometry, scattering, sensor–target configuration, and other acquisition-related factors. In the present study, these sources of variability were reduced by using a fixed portable measurement setup, artificial tungsten-halogen illumination, a constant sensor-to-sample distance, and a consistent viewing geometry; however, residual acquisition-related variability cannot be completely excluded. Additional limitations include the relatively small number of independent plant units (89 plants), the assignment of plant-level PCR labels to individual leaves, the absence of temporal evaluation, and the use of proximal foliar spectra, which exclude canopy architecture, shadows, background effects, viewing geometry, and atmospheric influences. Future studies should evaluate the proposed strategy across independent orchards, seasons, citrus cultivars, disease expressions, and canopy-level sensing platforms.

Overall, the comparison with previous research indicates broad agreement regarding the detectability of HLB-related optical changes [4–6,19–21]. Within the present dataset, plant-level spectral representation was associated with greater classification stability while preserving physiologically relevant spectral domains. In addition, the observed performance differences associated with spectral representation were larger and more consistent than those associated with class balancing and hyperparameter optimization. These conclusions are supported within the experimental dataset but require validation across independent locations and canopy-level platforms before operational application.

5. Conclusions

This study demonstrated that hyperspectral foliar reflectance combined with supervised machine learning enabled accurate discrimination between samples obtained from HLB-positive and HLB-negative citrus plants, as determined by plant-level PCR diagnosis. Logistic Regression achieved an accuracy of 0.968 and a weighted F1-score of 0.968 at the leaf level, whereas Support Vector Machine achieved an accuracy of 0.958 and a weighted F1-score of 0.958 at the plant level. Classification performance was consistently high at both organizational scales, although leaf-level models exhibited greater variation

across FIELD-based evaluations, whereas plant-level models based on aggregated spectral signatures showed more consistent performance. Because these analyses involve different observational units, the difference in performance should not be interpreted as a direct quantitative estimate of the improvement attributable exclusively to spectral averaging.

The visible (400–700 nm), red-edge (700–750 nm), near-infrared (750–1300 nm), and shortwave infrared (1500–2500 nm) domains contributed to HLB discrimination at both scales. PCA preserved high predictive performance after substantial dimensionality reduction, retaining at least 95% of the cumulative spectral variance within each leave-one-FIELD-out iteration. Leaf-level PCA models showed greater FIELD-dependent variability in HLB-positive classification than plant-level models, particularly in sensitivity, whereas plant-level performance remained highly stable. Within the present dataset, the observed differences associated with spectral representation were more pronounced than those associated with class balancing or hyperparameter optimization.

Class balancing did not alter the observed classification outcomes, whereas hyperparameter optimization produced limited and configuration-dependent effects, with improvement observed in only one plant-level FIELD holdout.

These findings are supported within a dataset comprising 1912 leaves from 89 plants collected in a single commercial orchard. Therefore, future studies should evaluate transferability across independent orchards, citrus cultivars, seasons, disease expressions, and canopy-level remote-sensing platforms.

The spectral domains identified in this study provide candidate regions for future evaluation using canopy-level, UAV, and satellite observations. However, their transferability across sensing scales and platforms remains to be experimentally validated.

Author Contributions: Conceptualization, M.M.B.J., L.P.O. and A.P.M.R.; methodology, M.M.B.J., L.P.O., L.A.d.C.J. and A.P.M.R.; software, M.M.B.J., L.P.O. and A.P.M.R.; validation, M.M.B.J., L.P.O. and A.P.M.R.; formal analysis, M.M.B.J.,L.P.O. and A.P.M.R.; investigation, M.M.B.J., L.P.O, L.A.d.C.J. and A.P.M.R.; resources, L.A.d.C.J., L.P.O. and A.P.M.R.; data curation, M.M.B.J.; writing—original draft preparation, M.M.B.J.; writing—review and editing, M.M.B.J., L.P.O., L.A.d.C.J. and A.P.M.R.; visualization, M.M.B.J.; supervision, L.P.O. and A.P.M.R.; project administration, A.P.M.R.; funding acquisition, A.P.M.R., L.A.d.C.J. and L.P.O. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Coordination for the Improvement of Higher Education Personnel (CAPES), Brazil — Finance Code 001 (Grant No. 88887.254022/2026-00) and by Embrapa Instrumentação, Brazil (data avalyability),

Data Availability Statement: The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bové, J.M. Huanglongbing: A destructive, newly emerging, century-old disease of citrus. *J. Plant Pathol.* **2006**, *88*, 7–37. https://doi.org/10.4454/jpp.v88i1.828.
2. Wang, N.; Pierson, E.A.; Setubal, J.C.; Xu, J.; Levy, J.G.; Zhang, Y.; Li, J.; Rangel, L.T.; Martins, J. The Candidatus Liberibacter–Host Interface: Insights into Pathogenesis Mechanisms and Disease Control. Annu. Rev. Phytopathol. 2017, 55, 451–482. https://doi.org/10.1146/annurev-phyto-080516-035513.
3. Li, W.; Duan, Y.; Li, D.; Lee, W.S.; Yang, C. Spectral characteristics and detection of citrus huanglongbing infected leaves. *Trans. ASABE* **2012**, *55*, 711–720. https://doi.org/10.13031/2013.41518.

4. Sankaran, S.; Mishra, A.; Maja, J.M.; Ehsani, R. Visible–near infrared spectroscopy for detection of Huanglongbing in citrus orchards. Comput. Electron. Agric. 2011, 77, 127–134. https://doi.org/10.1016/j.compag.2011.03.004.
5. Deng, X.; Zhu, Z.; Yang, J.; Zheng, Z.; Huang, Z.; Yin, X.; Wei, S.; Lan, Y. Detection of citrus Huanglongbing based on multi-input neural network model of UAV hyperspectral remote sensing. *Remote Sens.* **2020**, *12*, 2678. https://doi.org/10.3390/rs12172678.
6. Mahlein, A.K. Plant disease detection by imaging sensors—Parallels and specific demands for precision agriculture and plant phenotyping. *Plant Dis.* **2016**, *100*, 241–251. https://doi.org/10.1094/PDIS-03-15-0340-FE.
7. Zhang, M.; Qin, Z.; Liu, X.; Ustin, S.L. Detection of stress in tomatoes induced by late blight disease in California, USA, using hyperspectral remote sensing. Int. J. Appl. Earth Obs. Geoinf. 2003, 4, 295–310. https://doi.org/10.1016/S0303-2434(03)00008-4.
8. Sankaran, S.; Mishra, A.; Ehsani, R.; Davis, C. A review of advanced techniques for detecting plant diseases. *Comput. Electron. Agric.* **2010**, *72*, 1–13. https://doi.org/10.1016/j.compag.2010.02.007.
9. Zhang, N.; Yang, G.; Pan, Y.; Yang, X.; Chen, L.; Zhao, C. A review of advanced technologies and development for hyperspectral-based plant disease detection in the past three decades. *Remote Sens.* **2020**, *12*, 3188. https://doi.org/10.3390/rs12193188.
10. Abdulridha, J.; Batuman, O.; Ampatzidis, Y. UAV-based remote sensing technique to detect citrus canker disease utilizing hyperspectral imaging and machine learning. *Remote Sens.* **2019**, *11*, 1373. https://doi.org/10.3390/rs11111373.
11. Appeltans, S.; Pieters, J.G.; Mouazen, A.M. Detection of leek rust disease under field conditions using hyperspectral proximal sensing and machine learning. *Remote Sens.* **2021**, *13*, 1341. https://doi.org/10.3390/rs13071341.
12. Pandey, P.; Payn, K.G.; Lu, Y.; Heine, A.J.; Walker, T.D.; Acosta, J.J.; Young, S. Hyperspectral imaging combined with machine learning for the detection of fusiform rust disease incidence in loblolly pine seedlings. *Remote Sens.* **2021**, *13*, 3595. https://doi.org/10.3390/rs13183595.
13. Li, W.; Hartung, J.S.; Levy, L. Quantitative real-time PCR for detection and identification of Candidatus Liberibacter species associated with citrus huanglongbing. *J. Microbiol. Methods* **2006**, *66*, 104–115. https://doi.org/10.1016/j.mimet.2005.10.018.
14. Terentev, A.; Dolzhenko, V.; Fedotov, A.; Eremenko, D. Current state of hyperspectral remote sensing for early plant disease detection: A review. *Sensors* **2022**, *22*, 757. https://doi.org/10.3390/s22030757.
15. Wan, L.; Li, H.; Li, C.; Wang, A.; Yang, Y.; Wang, P. Hyperspectral sensing of plant diseases: Principle and methods. *Agronomy* **2022**, *12*, 1451. https://doi.org/10.3390/agronomy12061451.
16. Neupane, K.; Baysal-Gurel, F. Automatic identification and monitoring of plant diseases using unmanned aerial vehicles: A review. *Remote Sens.* **2021**, *13*, 3841. https://doi.org/10.3390/rs13193841.
17. Kouadio, L.; El Jarroudi, M.; Belabess, Z.; Laasli, S.-E.; Roni, M.Z.K.; Amine, I.D.I.; Mokhtari, N.; Mokrini, F.; Junk, J.; Lahlali, R. A review on UAV-based applications for plant disease detection and monitoring. *Remote Sens.* **2023**, *15*, 4273. https://doi.org/10.3390/rs15174273.
18. Chang, A.; Yeom, J.; Jung, J.; Landivar, J. Comparison of canopy shape and vegetation indices of citrus trees derived from UAV multispectral images for characterization of citrus greening disease. *Remote Sens.* **2020**, *12*, 4122. https://doi.org/10.3390/rs12244122.
19. Garza, B.N.; Ancona, V.; Enciso, J.; Perotto-Baldivieso, H.L.; Kunta, M.; Simpson, C. Quantifying citrus tree health using true color UAV images. *Remote Sens.* **2020**, *12*, 170. https://doi.org/10.3390/rs12010170.
20. Dong, R.; Shiraiwa, A.; Ichinose, K.; Pawasut, A.; Sreechun, K.; Mensin, S.; Hayashi, T. Hyperspectral Imaging and Machine Learning for Huanglongbing Detection on Leaf-Symptoms. Plants 2025, 14, 451. https://doi.org/10.3390/plants14030451.
21. Shahi, T.B.; Xu, C.-Y.; Neupane, A.; Guo, W. Recent advances in crop disease detection using UAV and deep learning techniques. *Remote Sens.* **2023**, *15*, 2450. https://doi.org/10.3390/rs15092450.
22. Kuswidiyanto, L.W.; Noh, H.-H.; Han, X. Plant disease diagnosis using deep learning based on aerial hyperspectral images: A review. *Remote Sens.* **2022**, *14*, 6031. https://doi.org/10.3390/rs14236031.
23. Meng, R.; Lv, Z.; Yan, J.; Chen, G.; Zhao, F.; Zeng, L.; Xu, B. Development of spectral disease indices for southern corn rust detection and severity classification. *Remote Sens.* **2020**, *12*, 3233. https://doi.org/10.3390/rs12193233.
24. Wei, X.; Johnson, M.A.; Langston, D.B., Jr.; Mehl, H.L.; Li, S. Identifying optimal wavelengths as disease signatures using hyperspectral sensor and machine learning. *Remote Sens.* **2021**, *13*, 2833. https://doi.org/10.3390/rs13142833.
25. Mahlein, A.-K.; Rumpf, T.; Welke, P.; Dehne, H.-W.; Plümer, L.; Steiner, U.; Oerke, E.-C. Development of spectral indices for detecting and identifying plant diseases. *Remote Sens. Environ.* **2013**, *128*, 21–30. https://doi.org/10.1016/j.rse.2012.09.019.
26. Asner, G.P. Biophysical and biochemical sources of variability in canopy reflectance. *Remote Sens. Environ.* **1998**, *64*, 234–253. https://doi.org/10.1016/S0034-4257(98)00014-5.

27. Curran, P.J. Remote sensing of foliar chemistry. *Remote Sens. Environ.* **1989**, *30*, 271–278. https://doi.org/10.1016/0034-4257(89)90069-2.
28. Delalieux, S.; van Aardt, J.; Keulemans, W.; Schrevens, E.; Coppin, P. Detection of biotic stress (*Venturia inaequalis*) in apple plants using hyperspectral sensors. *Remote Sens. Environ.* **2009**, *113*, 2722–2734. https://doi.org/10.1016/j.rse.2009.09.002.
29. Dou, S.; Li, M.; Qi, X.; Hou, Y.; Yuan, S.; Song, Y.; Mei, Z.; Qi, G. Investigating the Association Between Citrus Huanglongbing and Chlorophyll Content Using Hyperspectral Detection. Sensors 2025, 25, 7292. https://doi.org/10.3390/s25237292. .
30. Calderón, R.; Navas-Cortés, J.A.; Lucena, C.; Zarco-Tejada, P.J. High-resolution airborne hyperspectral and thermal imagery for early detection of *Verticillium* wilt of olive using fluorescence, temperature and narrow-band spectral indices. *Remote Sens. Environ.* **2013**, *139*, 231–245. https://doi.org/10.1016/j.rse.2013.07.031.
31. Jolliffe, I.T.; Cadima, J. Principal component analysis: A review and recent developments. *Philos. Trans. R. Soc. A* **2016**, *374*, 20150202. https://doi.org/10.1098/rsta.2015.0202.
32. Morata Bortoloto Junior, M.; Osco, L.P.; Jorge, L.A.C.; Ramos, A.P.M. Discriminative Spectral Regions for Detecting Huanglongbing in Citrus Plants through Statistical Analysis. ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci. 2026, X-3/W4-2025, 255–262. https://doi.org/10.5194/isprs-annals-X-3-W4-2025-255-2026.
33. Wulff, N.A.; Daniel, B.; Sassi, R.S.; Moreira, A.S.; Bassanezi, R.B.; Sala, I.; Coletti, D.A.B.; Rodrigues, J.C. Incidence of *Diaphorina citri* carrying *Candidatus Liberibacter asiaticus* in Brazil's citrus belt. *Insects* **2020**, *11*, 672. https://doi.org/10.3390/insects11100672.
34. ASD Inc. *FieldSpec Spectroradiometer Technical Specifications*; ASD Inc.: Boulder, CO, USA, 2024.
35. Jolliffe, I.T. *Principal Component Analysis*, 2nd ed.; Springer: New York, NY, USA, 2002.
36. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: New York, NY, USA, 2009.
37. Roberts, D.R.; Bahn, V.; Ciuti, S.; Boyce, M.S.; Elith, J.; Guillera-Arroita, G.; Hauenstein, S.; Lahoz-Monfort, J.J.; Schröder, B.; Thuiller, W.; et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **2017**, *40*, 913–929. https://doi.org/10.1111/ecog.02881.
38. He, H.; Garcia, E.A. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284. https://doi.org/10.1109/TKDE.2008.239.
39. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; ACM: New York, NY, USA, 2019; pp. 2623–2631. https://doi.org/10.1145/3292500.3330701.
40. Cox, D.R. The regression analysis of binary sequences. *J. R. Stat. Soc. Ser. B* **1958**, *20*, 215–242.
41. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. https://doi.org/10.1007/BF00994018.
42. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. https://doi.org/10.1023/A:1010933404324.
43. Breiman, L.; Friedman, J.; Stone, C.; Olshen, R. *Classification and Regression Trees*; Wadsworth International Group: Belmont, CA, USA, 1984.
44. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232.
45. 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, San Francisco, CA, USA, 13–17 August 2016*; ACM: New York, NY, USA, 2016; pp. 785–794.
46. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
47. Fix, E.; Hodges, J.L. *Discriminatory Analysis: Nonparametric Discrimination*; USAF School of Aviation Medicine: Randolph Field, TX, USA, 1951.
48. Hand, D.J.; Yu, K. Idiot's Bayes—Not so stupid after all? *Int. Stat. Rev.* **2001**, *69*, 385–398.
49. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. https://doi.org/10.1016/j.patrec.2005.10.010.

50. Powers, D.M.W. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *J. Mach. Learn. Technol.* **2011**, *2*, 37–63.
51. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. https://doi.org/10.1016/j.ipm.2009.03.002.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.