Journal of Agricultural Engineering

Journal of Agricultural Engineering 2026; LVII:1941
doi:10.4081/jae.2026.1941

©Copyright: the Author(s), 2026
Licensee PAGEPress, Italy

Detecting bacterial pustules on soybean plants by hyperspectral imaging

Eun Ri Kim,[1] Chan Seok Ryu,[1] Ye Seong Kang[2]

[1]Department of Biosystems Engineering; [2]Department of Smart Agro-Industry, Institute of Agriculture and Life Science, Gyeongsang National University, Jinju, Korea

Abstract

Bacterial pustules are a major threat to soybean cultivation but are difficult to detect early because they manifest on the bottoms of leaves. In this study, hyperspectral imaging was applied to detect bacterial pustules on soybean plants. Images were preprocessed, and representative central wavelengths (*i.e.*, bands) were identified through a two-sample t-test to calculate vegetation indices (VIs) for non-inoculated (*i.e.*, control) and inoculated (*i.e.*, treated) groups. Three machine-learning models were applied to classify infected soybean plants based on the VIs: partial least-squares discriminant analysis, support vector machine, and random forest (RF). The best classification performance was achieved by the RF model using five VIs with an overall accuracy (OA) of 0.89 and kappa coefficient (KC) of 0.77. The RF model also achieved an OA of 0.77 and KC of 0.55 when tested on a dataset before the expression of symptoms. The results of this study can potentially be applied to developing a multispectral image sensor that can be mounted on various platforms for the early detection of bacterial pustules on soybean crops.

Key words: bacterial pustule; hyperspectral imaging; random forest; soybean.

Correspondence: Ye Seong Kang, Department of Smart Agro-Industry, Institute of Agriculture and Life Science, Gyeongsang National University, Jinju, 52725, Korea. E-mail: slow321@gnu.ac.kr

Introduction

Soybeans (*Glycine max* (L.) Merrill) are a globally important commodity and play a crucial role in human nutrition, animal feed, and industry. However, soybean plants are increasingly being affected by bacterial diseases, which have been attributed to rising temperatures and humid weather due to climate change (Schaad, 2008). Such diseases have resulted in a decline in yield and quality with global yield losses of 15-60% (Shea *et al*., 2020). In the United States alone, soybean production suffered a major setback in 2007 with an estimated drop of approximately 8 million tons caused by various diseases including bacterial pustules, wildfire, and bacterial leaf spots (Wrather and Koenning, 2009). Bacterial pustules are caused by *Xanthomonas citri* pv. *Glycines* (syn. *Xanthomonas axonopodis* pv. *glycines*) (Constantin *et al*., 2016), and they are a major bacterial disease (Bertoglio *et al*., 2023) observed in the majority of soybean cultivation regions around the world (Hartman *et al*., 2015). Symptoms manifest as small yellowish-green spots with raised straw-yellow centers that rapidly undergo necrosis (Henning *et al*., 2014) and primarily occur on the underside of the leaves. They greatly reduce photosynthetic efficiency, which directly affects the seed size and yield, and their manifestation on the bottoms of leaves makes early observation a challenge (Jones, 1987). Especially in Korea, bacterial pustules have been widely prevalent since 1990 and have caused substantial damage to farmers. In 1999, bacterial pustules occurred in approximately 90% (94 out of 106) of soybean cultivation fields nationwide (Lee, 1999). In 2005–2006, bacterial pustules were observed in 89.7% (70 out of 78) of surveyed fields (Kang IJ *et al*., 2021). Severe bacterial pustule outbreaks decreased yields by 19.8% in 2006 and 16.8% in 2007 (Hong *et al*., 2010). Therefore, the early field detection of bacterial pustules is crucial to preventing yield losses. Traditional methods for plant disease detection such as artificial visible investigation and biological molecular techniques are limited by their needs for specialized knowledge, time-consuming processes, and manual labor. These methods often identify diseases in later cultivation stages; this leads to a time lag and hinders early detection. These shortcomings have become more pronounced with the rising demand in modern agriculture for real-time and large-scale detection (Zhang *et al*., 2020).

Recent studies have applied hyperspectral imaging to diagnose plant diseases rapidly and nondestructively by capturing subtle spectral variations related to physiological stress. For example, hyperspectral imaging was used to classify potato late blight (Ray *et al*., 2011), fire blight in apple trees (Skoneczny *et al*., 2020), and sugar beet diseases (Mahlein *et al*., 2012). These approaches demonstrate the potential of hyperspectral imaging for early disease detection, but most rely on large spectral datasets that are difficult to adapt to practical multispectral sensors. Because not all spectral bands are informative, selecting representative wavelengths containing critical disease information (Sarhrouni *et al*., 2012; Li *et al*., 2016) is essential for improving efficiency and cost-effectiveness in agricultural monitoring.

In soybean research, hyperspectral analysis has been explored for detecting soybean cyst nematode and sudden death syndrome (Bajwa *et al*., 2017), stress caused by stinkbugs (Marston *et al*., 2022), and stem rot in peanuts (Wei *et al*., 2021). However, to date, no study has specifically addressed the detection of bacterial pus-

tules using hyperspectral imaging, particularly at pre-symptomatic stages when visual observation is not possible. Furthermore, many existing studies have focused primarily on classification accuracy using machine-learning algorithms such as support vector machine (SVM) or random forest (RF) (Widodo and Yang, 2007; Breiman, 2001), rather than on identifying the spectral features or vegetation indices (VIs) that could facilitate sensor optimization. A more targeted selection of sensitive bands and indices is therefore required to enable early detection and practical sensor development. Although hyperspectral imaging has been used in soybean disease studies, most previous work examined other diseases or focused on detection after visible symptoms appeared. In contrast, bacterial pustule has been less studied, and its pre-symptomatic detection remains unclear. Moreover, the potential of wavelengths above 900 nm to capture early physiological changes has not been fully evaluated for this disease.

This study aimed to i) identify representative spectral bands critical for detecting soybean bacterial pustules, ii) derive vegetation indices sensitive to early infection stages, and iii) evaluate their classification performance using machine-learning models. The ultimate goal is to provide foundational spectral information for designing a practical multispectral imaging sensor capable of early detection of soybean bacterial pustules.

Materials and Methods

Experimental design

Experiments were conducted in a glass greenhouse at the National Institute of Crop Science, Gyeongsangnam-do, Republic of Korea (35°29′30.5″ N, 128°44′36.2″ E). The soybean cultivar 'Daechan,' which has high nutritional value and antioxidant activity, was used. Seeds were sown in 1/2000a Wagner pots on April 18, 2022.

Eight plants were used in total: four non-inoculated (control) and four inoculated (treated). Bacterial pustules strain 8ra was cultured on tryptic soy agar (30 g trypticase soy broth, 15 g agar, and 1 L distilled water, pH 7.3) at 28°C for 48 h. Bacterial colonies were suspended in sterile water at a concentration of 1.8×10^7 cfu mL^{-1}, and the suspension was sprayed uniformly over the leaf surfaces of the inoculated plants on May 16, 2022. Control plants were sprayed with sterile water. All plants were covered with plastic wrap overnight to maintain humidity and ensure infection.

Visual inspection was conducted daily from May 16 (inoculation day) to June 3 (diagnosis day). No imaging was conducted on May 21-22 (pre-symptom period) and May 28-29 (post-symptom period). Disease symptoms were first observed on May 25, marking the transition from the pre-symptom to post-symptom phase.

This design resulted in 120 samples in total (8 plants × 15 observation days), comprising 48 samples before (May 16-24, excluding May 21-22) and 72 samples after (May 25-June 3, excluding May 28-29) symptom expression. Each sample corresponded to one hyperspectral observation of a single plant on a given day (plant × day) (Figure 1).

Hyperspectral images

Hyperspectral images were acquired daily at 11:00 am by using an automated imaging acquisition system (FineCube, Hortizen Ltd., Republic of Korea) from May 16 (*i.e.*, the inocula-

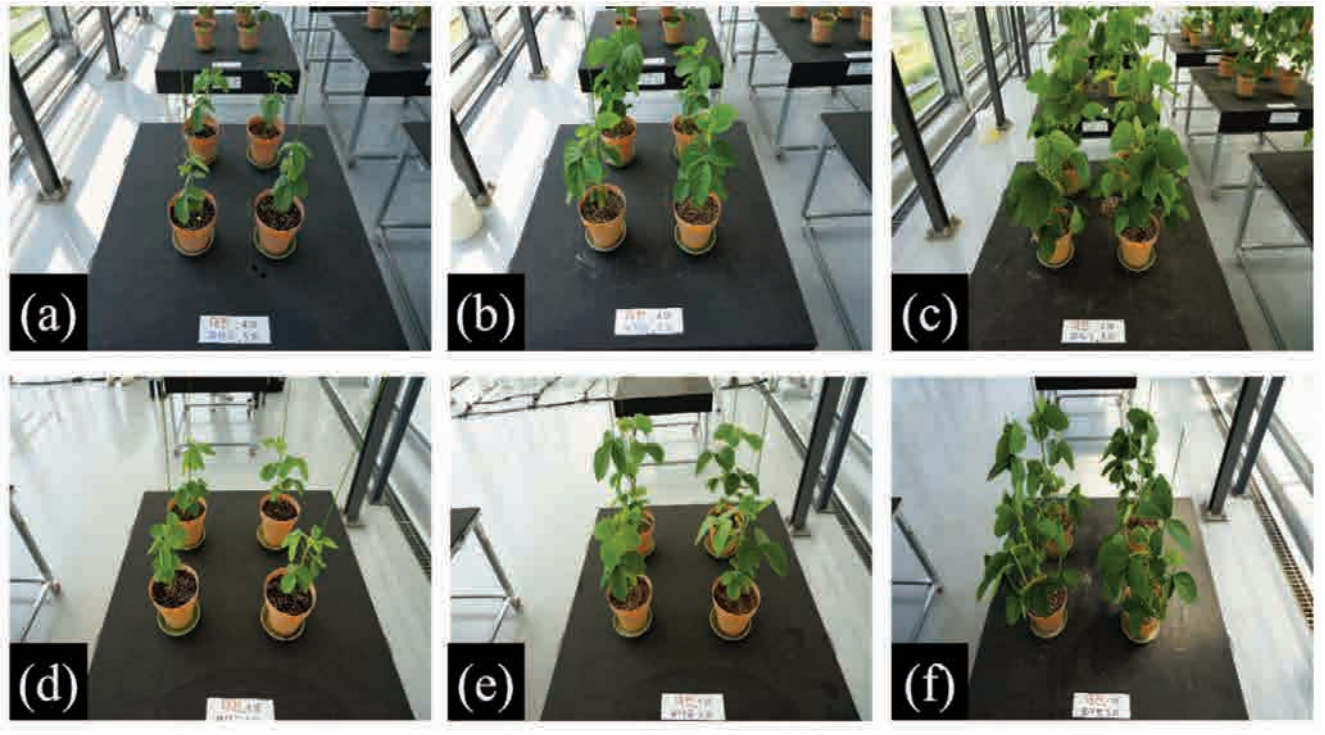

Figure 1. Experimental layout: control group (*i.e.*, not inoculated) (**a**) before (May 17), (**b**) on (May 25), and (**c**) after (June 3) the diagnosis date; treated group *(i.e.*, inoculated) (**d**) before (May 17), (**e**) on (May 25), and (**f**) after (June 3) the diagnosis date.

OPEN ACCESS

tion date) to June 3. A hyperspectral image sensor (FX10e, Specim Spectral Imaging Ltd., Finland) with a spectral resolution of 5.5 nm and 224 bands within a wavelength range of 400-1000 nm was mounted on the automated imaging acquisition system. Shading and insulating curtains were used to minimize the effect of light nonuniformity due to the greenhouse frame. Four halogen lamps (Haloline 118 nm, 500 W, Osram Inc., Germany) were used to compensate for the lack of light due to the curtains. An 18% white reference board (EzyBalance, Lastolite Ltd., England) was placed within the hyperspectral images to normalize different light conditions across time series.

Image processing

The image-processing software ENVI 5.6 (Exelis Visual Information Solutions Inc., USA) was used to apply radiometric correction to the acquired hyperspectral images. The hyperspectral images were then converted into enhanced vegetation index (EVI) images to highlight the vegetation area, as shown in Figure 2. The EVI was calculated as follows:

$$EVI = 2.5\left(\frac{NIR-RED}{NIR+6.0RED-7.5BLUE+1}\right) \quad \text{(Eq. 1)}$$

where the near-infrared (NIR) value was obtained at 780 nm, the red value was obtained at 650 nm, and the blue value was obtained at 450 nm. The reflectance of individual soybean plants was extracted from selected regions of interest (ROI) in the EVI images by the density slice method, during which background soil pixels were excluded so that only the plant canopy was included. Each plant image represented one ROI corresponding to the entire canopy.

A Gaussian filter is a 2D convolution that is commonly used to smooth and enhance images. It is a linear filter similar to a mean filter, but it applies a Gaussian function-based kernel to the image. The Gaussian function is characterized by a bell-shaped curve where values decrease smoothly from the center. The Gaussian filter attenuates high-frequency components such as noise while preserving low-frequency components, and it works by calculating the weighted average of each value with its neighboring value within a specified kernel size:

$$G(x,y) = 1/2\pi\sigma^2\left(e^{-\left(\frac{x^2+y^2}{2e^2}\right)}\right) \quad \text{(Eq. 2)}$$

where G(x, y) represents the value of the Gaussian filter at coordinates (x, y), σ (sigma) is the standard deviation of the Gaussian distribution, determining the curve's width, x and y is spatial coordinates representing the position of a pixel and e is a mathematical constant approximately equal to 2.71828. The Gaussian filter weights values close to the center pixel more. The weighted average then replaces the original value. In this study, a kernel size of 5 was applied, which was determined to effectively remove noise

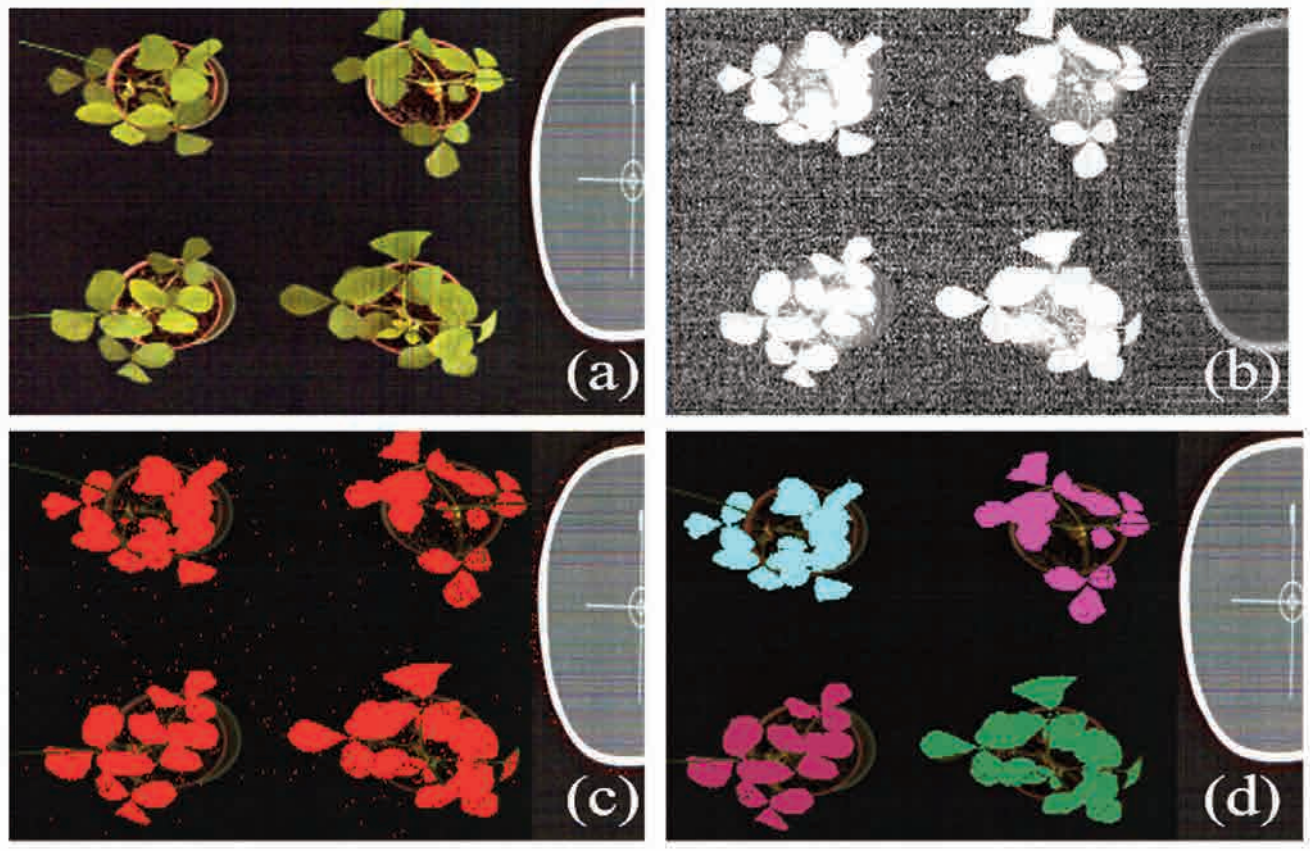

Figure 2. Image processing procedure: (**a**) original RGB image, (**b**) conversion to the enhanced vegetation index, (**c**) extraction of soybean canopies, and (**d**) extraction of regions of interest from individual soybean canopies.

while preserving the spectral characteristics of vegetation. The Gaussian filter was applied in Python 3.7.9 (Python Software Foundation, USA).

Because a hyperspectral image has a narrow bandwidth with many bands, the bands must be merged so that a commercially available bandpass filter can be used, which is necessary for the development of a multispectral image sensor. Merging involves equalizing the reflectance of adjacent bands at specific intervals. In this study, reflectance values with a 5 nm full width at half maximum (FWHM) were merged to 10 nm FWHM at 10-nm interval. Based on the tolerance of the wavelength range of the bandpass filter, the reflectance values were equalized in the range of ±7 nm instead of ±5 nm.

Two-sample *t*-test

In this study, the *t*-value was the magnitude of the difference in mean reflectance values of the two groups:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad \text{(Eq. 3)}$$

where and were the mean reflectance values of the control and treated plants, respectively. is the pooled variance to considers the variability within each group. and are the sample sizes of the control and treated plants, respectively. The p-value was determined according to the degrees of freedom, which were determined by the sample size. Thus, larger sample sizes resulted in higher degrees of freedom. The *p*-value was calculated by using the *t*-value and degrees of freedom to determine the probability in the tail area of the *t*-distribution. The *p*-value ranged between 0 and 1, and a smaller value indicated that the difference in mean values between the two groups was less likely due to chance. Thus, a low *p*-value indicated that the difference between the means was statistically significant. Bands with the lowest p-value in each spectral region (*i.e.*, red, green, blue, NIR1, and NIR2) were considered the most relevant for differentiating between the control and treated groups.

Vegetation indices

VIs were calculated by using selected bands in the blue, green, red, red-edge, NIR1, and NIR2 regions. NIR1 (780-900 nm) and NIR2 (920-980 nm) were distinguished to expand the range of wavelengths beyond 900 nm, which is frequently used to improve the classification accuracy of plant diseases and stress such as soybean mosaic virus (SMV) (Jinendra *et al*., 2010), water stress in gerbera plants (Peñuelas *et al*., 1993), and rice diseases (Wang *et al*., 2017). In total, 121 VIs were calculated, of which Table 1 presents the selected VIs for the modeling process. The simple ratio represents the ratio between reflectance values in given regions. The chlorophyll vegetation index (CVI) is a broadband VI that is specifically sensitive to leaf chlorophyll concentration at the canopy scale under the original experimental conditions (Vincini *et al*., 2008). The Ashburn vegetation index (AVI) is used to monitor green vegetation in the growth stage (Bannari *et al*., 1995). The differenced vegetation index (DVI) is used to monitor vegetation health and distinguishes green vegetation from bare soil or water (Richardson and Wiegand, 1977). The yellow vegetation index (YVI) is calculated from reflectance values in the yellow wavelength range and is used to assess plant health and productivity, with higher values indicating healthier and more productive vegetation (Kauth and Thomas, 1976). The blue-normalized difference vegetation index (BNDVI) is used to assess plant health and productivity based on the difference in reflectance values in green and blue bands (Yang *et al*., 2004). Vegetation indices calculated using the NIR2 band are indicated with the prefix "2-" to distinguish them from those using the NIR1 band.

Classification analysis

Figure 3 shows the flowchart of the classification analysis, which was performed on the image based-agricultural data analysis platform FinePro (Hortizen Co. Ltd., Republic of Korea). Three machine-learning models were tested for their classification per-

Table 1. List of vegetation indices.

Vegetation indices	Equation	Reference
Band ratios	$\frac{\rho_1}{\rho_2}$	Pearson and Miller, 1972
CVI	$\frac{\rho_{NIR} \times \rho_{Red}}{\rho_{Green}^2}$	Vincini *et al*., 2008
AVI	$2 \times \rho_{NIR} - \rho_{Red}$	Bannari *et al*., 1995
DVI	$2.4 \times \rho_{NIR} - \rho_{Red}$	Richardson and Wiegand., 1977
YVI	$-0.899 \times \rho_{Green} - 0.428 \times \rho_{Red} + 0.070 \times \rho_{Red\,edge} + 0.041 \times \rho_{NIR}$	Kauth and Thomas, 1976
BNDVI	$\frac{\rho_{NIR} - \rho_{Blue}}{\rho_{NIR} + \rho_{Blue}}$	Yang *et al*., 2004
BWRDVI	$\frac{0.1 \times \rho_{NIR} - \rho_{Blue}}{0.1 \times \rho_{NIR} + \rho_{Blue}}$	Hancock and Dougherty, 2007

CVI, chlorophyll vegetation index; AVI, Ashburn vegetation index; DVI, differenced vegetation index; YVI, yellow vegetation index; BNDVI, blue-normalized difference vegetation index; BWRDVI, blue-wide dynamic range vegetation index, ρ_1, NIR1, NIR2, Green; ρ_2, blue, red; NIR, NIR1, NIR2; when the NIR2 band was used, the corresponding vegetation index was prefixed with "2–" to indicate substitution of the NIR band.

OPEN ACCESS

formance: partial least-squares discriminant analysis (PLS-DA), support vector machine (SVM), and random forest (RF). Table 2 presents the number of samples used to develop the classification models. To prevent information leakage, data splitting was conducted at the plant level, meaning that all observations from a given plant (across all days) were assigned exclusively to either the training or test dataset. Six plants (three control and three treated; 70% of the total) were used for training, and two plants (one control and one treated; 30%) were reserved for testing. Within the training dataset, five-fold cross-validation was applied to optimize model parameters (e.g., latent variable number in PLS-DA, kernel and gamma in SVM, tree number and depth in RF).

PLS-DA is a supervised multivariate classification method that extracts latent components representing covariance between predictor variables and class labels. This allows dimensionality reduction while preserving discrimination between groups, which is suitable for high-dimensional hyperspectral data.

SVM is a supervised learning algorithm that identifies a decision boundary that maximizes the margin between classes. When linear separation is not possible, kernel functions enable nonlinear discrimination in a transformed feature space, allowing SVM to handle complex class boundaries in high-dimensional data (Widodo and Yang, 2007; Kang YS *et al*., 2021).

RF is an ensemble learning method that constructs multiple decision trees using bootstrap sampling and aggregates their predictions through majority voting. This approach improves robustness and generalization and is well suited for classification tasks involving high-dimensional datasets (Breiman, 2001). Class assignment was determined using the standard majority-vote rule, which corresponds to a probability threshold of 0.5. This provides a clear and reproducible decision criterion for distinguishing between control and inoculated plants.

Shapley additive explanations (SHAP) was used to quantify the contribution of each vegetation index to the classification model. SHAP assigns an importance value to each feature based on how the model's prediction changes when that feature is included or excluded. Because SHAP provides local explainability, it is able to approximate the contribution of each feature even in black-box

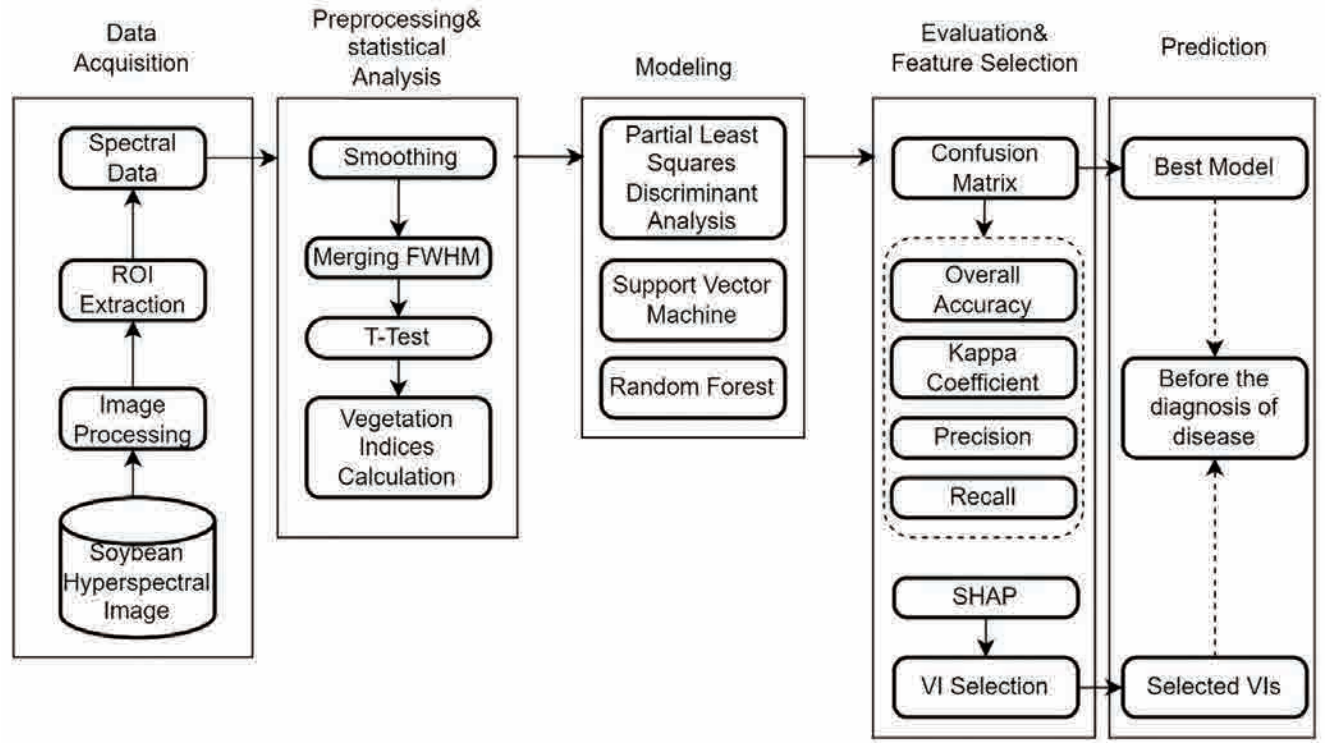

Figure 3. Flowchart for classification of the control and treated groups.

Table 2. Number of samples before and after symptom expression in training and test datasets used for machine learning analysis.

		Number of samples
Training dataset	Before expression	34
	After expression	50
Test dataset	Before expression	14
	After expression	22

machine-learning models by building interpretable surrogate representations. Feature importance rankings were obtained by averaging SHAP values across all observations (Marcílio and Eler, 2020).

Classification performance evaluation

The classification accuracy is commonly evaluated by using a confusion matrix derived from metrics such as the overall accuracy (OA), kappa coefficient (KC), precision score, recall score, and F1 score. Figure 4 shows a confusion matrix representing the number of true positive, true negative, false positive, and false negative predictions made by a classification model. The above metrics can be calculated as follows:

$$\text{Overall Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \quad \text{(Eq. 4)}$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad \text{(Eq. 5)}$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad \text{(Eq. 6)}$$

$$\text{F1 Score} = \frac{2\times Precision\times Recall}{Precision+Recall} \quad \text{(Eq. 7)}$$

$$\text{Kappa} = \frac{Accuracy-P(e)}{1-P(e)}, P(e) = \frac{(TP+FP)\times(TP+FN)\times(FN+TN)\times(FP+TN)}{(TP+FP+FN+TN)^2} \quad \text{(Eq. 8)}$$

OA measures how similar the actual and predicted data are. Precision represents the proportion of true positive cases among the predicted positive cases. Recall (*i.e.*, sensitivity) represents the proportion of actual positive values correctly predicted as positive by the model. The F1 score is the harmonic mean of the precision and recall and is helpful in situations where the data classes are imbalanced. The KC measures the agreement between observed and chance-corrected agreement beyond what is expected by chance alone. It is defined as the ratio of the difference between the observed agreement rate and the chance agreement rate to the maximum possible difference with a range of zero (no agreement) to unity (perfect agreement).

In addition to OA, we used the KC because it accounts for chance agreement and is widely recommended for accuracy assessment in remote sensing and image-based classification (Congalton, 1991). Precision, recall, and F1 score were included to provide class-specific performance evaluation, especially under potential imbalance between the control and treated groups.

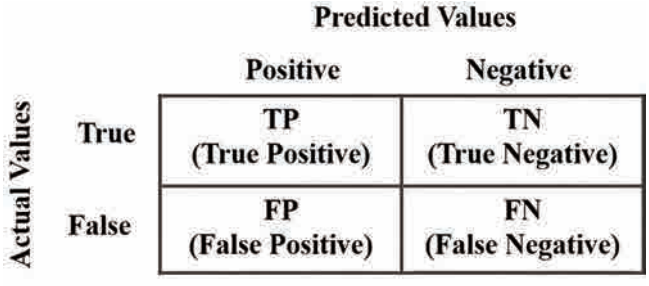

Figure 4. Confusion matrix.

Table 3. Classification performances of the machine-learning models.

	PLS-DA	SVM	RF
Overall accuracy	0.41	0.81	0.89
Kappa coefficient	0.13	0.61	0.77
Precision score	0.43	0.81	0.91
Recall score	0.41	0.81	0.89
F1 score	0.36	0.80	0.89

PLS-DA, partial least-squares discriminant analysis; SVM, support vector machine; RF, random forest.

Table 4. Confusion matrix for the classification performances of the machine-learning models.

			True	
			Control	Treated
Predicted	PLS-DA	Control	13	3
		Treated	4	16
	SVM	Control	12	2
		Treated	5	17
	RF	Control	13	0
		Treated	4	19

PLS-DA, partial least-squares discriminant analysis; SVM, support vector machine; RF, random forest.

OPEN ACCESS

Results

Reflectance curves

Figure 5 presents the smoothed average reflectance curves for the control and treated samples. The control samples showed higher reflectance values than the treated samples except in the NIR region after diagnosis. The decreased difference in the NIR region after diagnosis can be attributed to physiological changes including increased chlorophyll content as the plants progressed in their growth stages (Hansen and Schjoerring, 2003). A distinct difference was observed between the reflectance values of the control and treated samples in the NIR region before diagnosis, which indicates that this region can potentially be used for pre-symptom detection. In the green region (510-560 nm), differences between the control, and treated samples were observed both before and after diagnosis, which can be attributed to the appearance of spots on the leaves as the disease progressed in the treated samples.

Selection of representative bands

After merging, bands were divided into different ranges: 420-480 nm for the blue region, 500-580 nm for the green region, 600-660 nm for the red region, 680-760 nm for the red-edge region, 780-900 nm for the NIR1 region, and 920-980 nm for the NIR2 region. In the t-test, the lowest p-values were at 420 nm in the blue region (1.29e−09), 540 nm in the green region (1.66e−08), 600 nm in the red region (5.95e−06), 700 nm in the red-edge region (3.54e−05), 780 nm in the NIR1 region (0.106), and 940 nm in the NIR2 region (0.126). Consequently, these values were selected as representative bands of these spectral regions, as shown in Figure 6. The representative bands were used to calculate the VIs given in Table 1.

Classification performances of machine learning models

Table 3 presents the classification performance of the machine-learning models. PLS-DA achieved an OA of 0.41, KC of 0.13, precision of 0.43, recall of 0.41, and F1 score of 0.36. SVM achieved an OA of 0.81, KC of 0.61, precision of 0.81, recall of 0.81, and F1 score of 0.80. RF showed the best performance with an OA of 0.89, KC of 0.77, precision of 0.91, recall of 0.89, and F1 score of 0.89. For disease detection models, the most important metric is misclassifying diseased (*i.e.*, treated) soybeans as normal (*i.e.*, control) because plants may not receive the necessary treatment (Kang Y.S. *et al*., 2021). Table 4 presents a confusion matrix showing that RF misclassified zero treated plants as control plants. In contrast, PLS-DA, and SVM misclassified two and three treated plants, respectively.

Table 5 lists the VIs and parameters selected for the machine-learning models. For PLS-DA, the selected VIs were DVI calculated with NIR2, DVI, AVI, AVI calculated with NIR2, and NIR/Blue with two latent variables as parameters. For SVM, the selected VIs were CVI, NIR/Red, CVI calculated with NIR2, NIR/Blue calculated with NIR2, and NIR/Red calculated with NIR2 with the parameters C = 1, gamma = 0.1, and the polynomial kernel. For RF, the selected VIs were YVI calculated with NIR2, Green/Red, NIR/Blue, NIR/Blue calculated with NIR2, and BNDVI calculated with NIR2 with the parameters of max depth = 6 and n estimators = 100. Refer to Table 1 for how these VIs were calculated.

Detection before symptoms expression

Table 6 presents the classification performance of RF with the key VIs and parameters before symptom expression. RF achieved an OA of 0.77, KC of 0.55, precision of 0.80, recall of 0.77, and F1 score of 0.77. This performance indicates that RF can potentially be applied to detecting bacterial pustules before experts can visually identify the disease. The decrease in KC can be attributed to the reduction in sample size.

Effect of the number of vegetation indices on the classification performance

The fewest number of VIs that can be used without sacrificing classification performance is the most computationally efficient approach. In this study, SHAP was used to select the most important VIs. Then, the classification performance of RF was evaluated according to the number of VIs used. As presented in Table 7, five, or nine VIs both resulted in the best performance with an OA of 0.89 and KC of 0.77. Table 8 also indicates that the fewest misclassifications were obtained with five or nine VIs. Therefore, the most efficient choice is to use five VIs as it provides the best classification performance and fewest misclassifications while requiring less computation than nine VIs.

Figure 7 compares the SHAP values of the VIs to determine their importance. Increasing the number of VIs from four (Figure 7c) to five (Figure 7d) increased the classification performance from an OA of 0.67 and KC of 0.32 to an OA of 0.89 and KC of 0.77. The most significant change was the addition of Green/Red, which had the second-highest SHAP value. Decreasing the number of VIs from 10 (Figure 7a) to nine (Figure 7b) increased the classification performance from an OA of 0.75 and KC of 0.49 to an OA of 0.89 and KC of 0.77. The highest SHAP value changed from

Table 5. Vegetation indices and parameters selected by each machine-learning model to detect disease after inoculation.

	PLS-DA	SVM	RF
1	2DVI	CVI	2YVI
2	DVI	NIR/Red	Green/Red
3	AVI	2CVI	NIR/Blue
4	2AVI	2NIR/Blue	2NIR/Blue
5	NIR/Blue	2NIR/Red	2BNDVI
Parameters	Latent variables: 2	C:1 Gamma: 0.1 Kernel: Polynomial	 Max_depth: 6 N_estimators: 100

PLS-DA, partial least-squares discriminant analysis; SVM, support vector machine; RF, random forest; CVI, chlorophyll vegetation index; AVI, Ashburn vegetation index; DVI, differenced vegetation index; YVI, yellow vegetation index; BNDVI, blue-normalized difference vegetation index.

YVI using NIR1 to 2YVI using NIR2. This indicates the significance of the NIR2 to early disease detection, which agrees with the reflectance curves shown in Figure 5. In addition, the VIs NIR/Blue, NIR2/Blue, and 2BNDVI, which involve NIR, and blue bands, were consistently selected regardless of the number of VIs. This confirms that the NIR and blue bands play a significant role in the detection of bacterial pustules. These bands are at opposite ends of the wavelength range, which indicates the necessity of considering all bands for early detection of bacterial disease.

Discussion

As shown in Figure 6, the control, and treated groups showed a significant difference in reflectance values in the NIR region before symptom expression. Previous studies have demonstrated that wavelengths above 900 nm are particularly effective for detecting early physiological changes associated with plant disease and stress. For example, Jinendra *et al*. 2010 detected SMV in latent stages using NIR bands beyond 900 nm, and Peñuelas *et al*. 1993 identified water stress in gerbera plants using reflectance in a similar range. Wang *et al*. 2017 also used NIR and adjacent infrared wavelengths to discriminate rice diseases. These findings support the importance of the NIR2 region (920-980 nm) for early diagnosis and monitoring, consistent with the differences observed between control and treated plants in this study.

The results also indicated that RF trained on data from all stages could use the selected VIs for the early detection of soybean pustules before symptoms appear. In particular, this experiment was conducted during the vegetative growth stage and before the reproductive growth stage, so the model can detect disease early before flowering and pod setting (Purcell *et al*., 2014). However, because canopy structure (e.g., leaf angle, overlapping leaves, and plant shape) may vary more widely under field conditions and could influence reflectance characteristics, evaluating the approach under real cultivation environments will be necessary to

Table 6. Classification performance of the random forest model before symptom expression using selected vegetation indices and parameters.

	Performances
Overall accuracy	0.77
Kappa coefficient	0.55
Precision score	0.80
Recall score	0.77
F_1 score	0.77

Table 7. Classification performance by the random forest model according to the number of vegetation indices.

Number of vegetation indices	Performances	
10	Overall accuracy	0.75
	Kappa coefficient	0.49
9	Overall accuracy	0.89
	Kappa coefficient	0.77
8	Overall accuracy	0.81
	Kappa coefficient	0.60
7	Overall accuracy	0.81
	Kappa coefficient	0.60
6	Overall accuracy	0.83
	Kappa coefficient	0.66
5	Overall accuracy	0.89
	Kappa coefficient	0.77
4	Overall accuracy	0.67
	Kappa coefficient	0.32
3	Overall accuracy	0.78
	Kappa coefficient	0.55
2	Overall accuracy	0.78
	Kappa coefficient	0.55

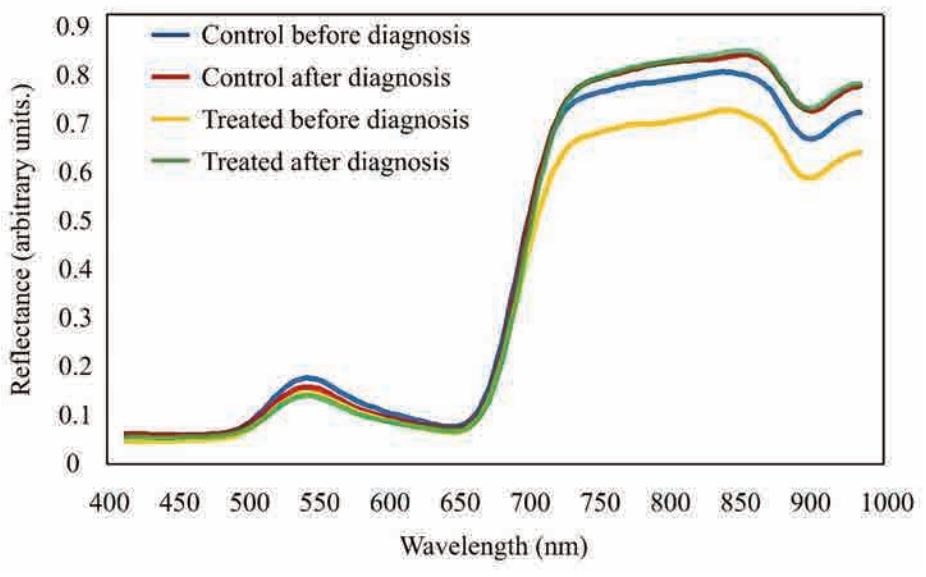

Figure 5. Reflectance curves before (blue: control, yellow: treated) and after (red: control, green: treated) diagnosis.

OPEN ACCESS

confirm its applicability. In practical terms, early detection at the vegetative stage could support routine monitoring and timely disease control decisions before yield loss occurs. In addition, the limited number of plant replicates may affect the robustness and generalizability of the results; therefore, the findings should be interpreted as demonstrating methodological feasibility, with further validation required in future work.

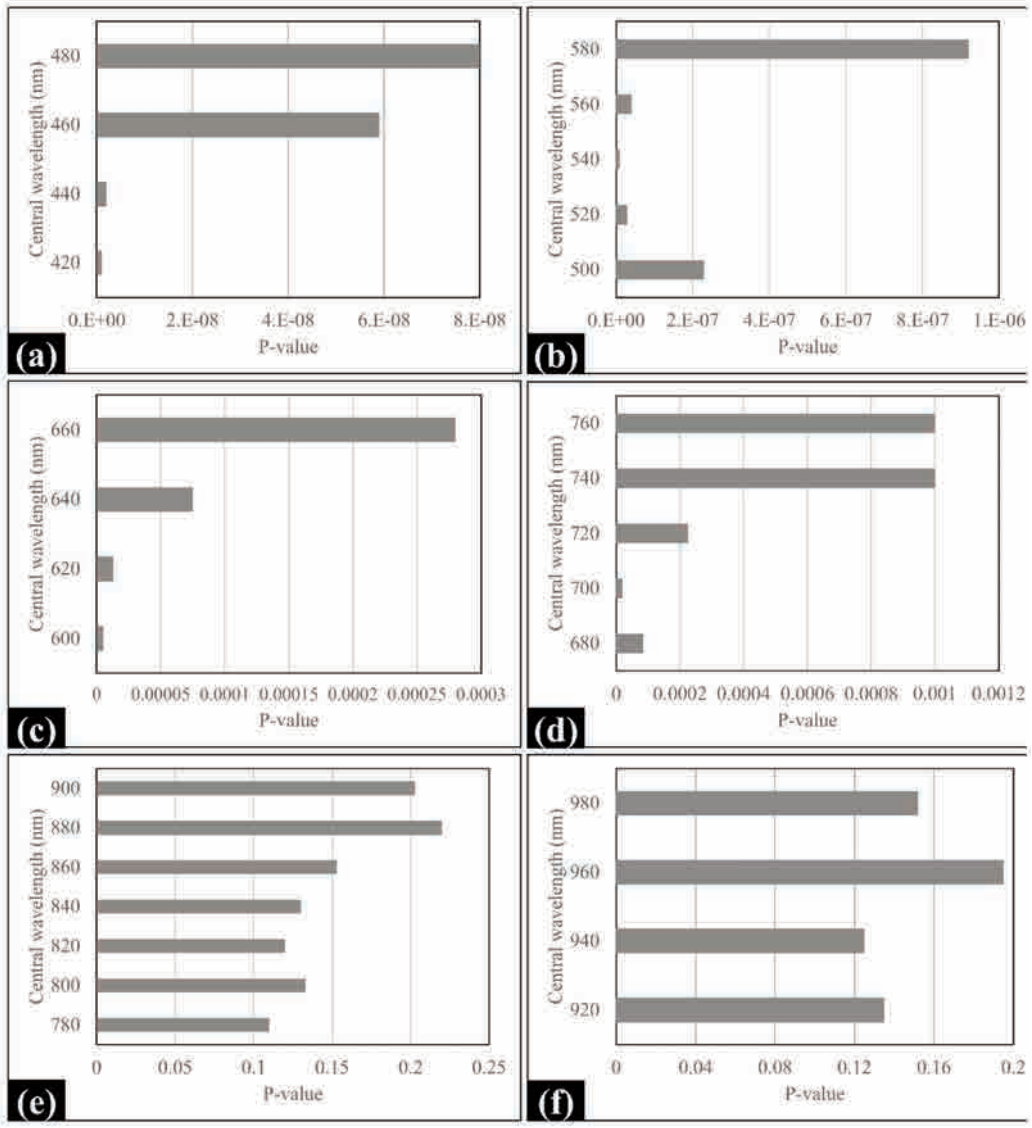

Figure 6. Representative bands selected from the two-sample t-test: (**a**) blue, (**b**) green, (**c**) red, (**d**) red edge, (**e**) NIR1, (f) NIR2.

Table 8. Confusion matrix for the classification performance of the random forest model according to the number of vegetation indices.

			True	
			Control	Treated
Predicted	4	Control	9	4
		Treated	8	15
	5	Control	13	0
		Treated	4	19
	9	Control	13	0
		Treated	4	19
	10	Control	10	2
		Treated	7	17

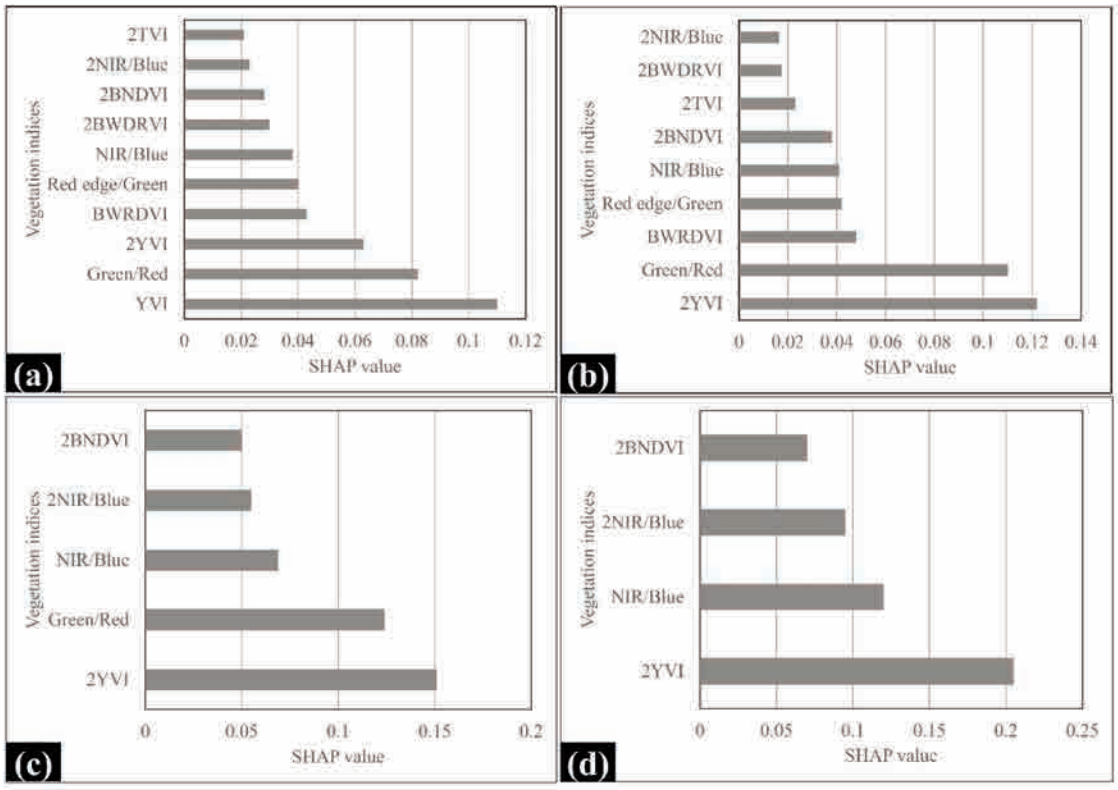

Figure 7. SHAP values according to the number of VIs used by the RF model: (**a**) 10, (**b**) 9, (**c**) 5, and (**d**) 4.

Conclusions

This study proposed an efficient and prospective multispectral sensor for monitoring soybean bacterial pustule disease using hyperspectral images and machine learning analysis. The findings of this study showed that NIR reflectance values could be used to diagnose diseases early, aligning with previous studies. VIs were calculated from representative bands that are identified using a two-sample *t*-test. The RF model, which has five vegetation indices (BNDVI and YVI with NIR2, green/red, NIR1/blue, NIR2/blue), was selected by applying SHAP feature selection, and it demonstrated the best performance with OA 0.89, KC 0.77, precision 0.91, recall 0.89, and F1 score 0.89. When applying this model to data collected before finding the expression of bacterial pustule symptoms from May 16 to May 24, the classification prediction performance was OA 0.77, KC 0.55, precision 0.80, recall 0.77, and F1 score 0.77. The SHAP feature selection identified key VIs with five bands to provide the most efficient model performance. The addition of Green/Red and NIR2 considerably affects the performance of model. NIR and blue bands consistently played an important role in disease detection. This study demonstrates the feasibility of detecting bacterial pustules during the vegetative growth stage before visible symptoms appear, enabling earlier disease management interventions. Such early detection has the potential to reduce yield losses by supporting timely treatment decisions during the period when disease control is most effective. The results also provide foundational spectral information for the design of practical multispectral sensors aimed at real-time disease monitoring. Further evaluation under field conditions will be necessary to confirm robustness across varying canopy structures and environmental factors.

Acknowledgments

The authors appreciate all staff who helped with this study at the Department of Upland Crop Sciences, National Institute of Crop and Food Science, Republic of Korea.

References

Bajwa SG, Rupe JC, Mason J, 2017. Soybean disease monitoring with leaf reflectance. Remote Sens 9:127.

Bannari A, Morin D, Bonn F, Huete AR, 1995. A review of vegetation indices. Remote Sens Rev 13:95-120.

Bertoglio C, Moura Duin I, Netzel de Matos J, Rodrigues Ribeiro N, Pereira Leite R, Balbi-Peña MI, 2023. Comparative study of inoculation methods to determine aggressiveness of Xanthomonas citri pv. glycines isolates. Agronomy 13:1515.

Breiman L, 2001. Random forests. Mach Learn 45:5-32.

Congalton RG, 1991. A review of assessing the accuracy of classifications. Remote Sens Environ 37:35-46.

Constantin EC, Cleenwerck I, Maes M, Baeyen S, Van Malderghem C, De Vos P, Cottyn B, 2016. Genetic characterization of Xanthomonas strains. Plant Pathol 65:792-806.

Hancock DW, Dougherty CT, 2007. Relationships between blue- and red-based vegetation indices and leaf area and yield of alfalfa. Crop Sci 47:2497-2502.

Hansen PM, Schjoerring JK, 2003. Reflectance measurement of canopy biomass and nitrogen status in wheat. Remote Sens Environ 86:542-553.

Hartman GL, Rupe JC, Sikora EJ, Domier LL, Davis JA, Steffey

OPEN ACCESS

KL, 2015. Compendium of soybean diseases and pests. St. Paul, American Phytopathological Society; pp. 56-59.
Henning AA, Almeida ÁMR, Godoy CV, Seixas CDS, Yorinori JT, Costamilan LM, Dias WP, 2014. [Manual de identificação de doenças de soja].[Book in Portuguese]. Brasilia, EMBRAPA Editora.
Hong SJ, Kim YK, Jee HJ, Lee BC, Yoon YN, Park ST, 2010. Selection of bactericides for controlling bacterial pustule. Res Plant Dis 16:266-273.
Jinendra B, Tamaki K, Kuroki S, Vassileva M, Yoshida S, Tsenkova R, 2010. Near infrared spectroscopy for rapid diagnosis of virus-infected soybean. Biochem Biophys Res Commun 397:685-690.
Jones SB, 1987. Bacterial pustule disease microscopy. Phytopathology 77:266-274.
Kang IJ, Kim KS, Beattie GA, Chung H, Heu S, Hwang I, 2021. Population genetics of X. citri pv. glycines in Korea. Plant Pathol J 37:652-661.
Kang YS, Park JW, Jang SH, Song HY, Kang KS, Ryu CS, Kim GH, 2021. Spectral band selection for detecting fire blight. Korean J Agric Forest Meteorol 23:15-33.
Kauth RJ, Thomas GS, 1976. The tasselled cap transformation. Proc LARS Symposium; p. 159.
Lee SD, 1999. Occurrence and characterization of major plant bacterial diseases in Korea. PhD Thesis, Seoul National University.
Li S, Zheng Z, Wang Y, Chang C, Yu Y, 2016. Hyperspectral band selection using multiple classifiers. Pattern Recognit Lett 83:152-159.
Marcílio WE, Eler DM, 2020. Assessing SHAP values for feature selection. Proc 33rd SIBGRAPI Conf IEEE; pp. 340-347.
Mahlein AK, Steiner U, Hillnhütter C, Dehne HW, Oerke EC, 2012. Hyperspectral imaging for sugar beet disease symptoms. Plant Methods 8:3.
Marston ZPD, Cira TM, Knight JF, Mulla D, Alves TM, Hodgson EW, et al., 2022. SVM classification of plant stress. J Econ Entomol 115:1557-1563.
Pearson RL, Miller LD, 1972. Remote spectral measurements as a method for determining plant cover. In: Proc 8th Int Symp Remote Sens Environ. Ann Arbor; pp. 1357-1379.
Peñuelas J, Filella I, Biel C, Serrano L, Savé R, 1993. Reflectance at 950-970 nm as water status indicator. Int J Remote Sens 14:1887-1905.
Purcell LC, Salmeron M, Ashlock L, 2014. Soybean growth and development. In: Arkansas soybean production handbook. Fayetteville, University of Arkansas Press.
Ray SS, Jain N, Arora RK, Chavan S, Panigrahy S, 2011. Hyperspectral potato disease detection. J Indian Soc Remote Sens 39:161-169.
Richardson AJ, Wiegand CL, 1977. Distinguishing vegetation from soil background. Photogramm Eng Remote Sens 43:1541-1552.
Sarhrouni E, Hammouch A, Aboutajdine D, 2012. Mutual-information-based hyperspectral feature selection. arXiv: 1210.0052v1.
Schaad NW, 2008. Emerging plant pathogenic bacteria. In: Fatmi M. (Ed.), Pseudomonas syringae pathovars and related pathogens. Berlin, Springer.
Shea Z, Singer M, Zhang B, 2020. Soybean production and improvement. London, IntechOpen Publ.
Skoneczny H, Kubiak K, Spiralski M, Kotlarz J, Mikiciński A, Puławska J, 2020. Fire blight detection in apple. Remote Sens 12:2101.
Vincini M, Frazzi E, D'Alessio P, 2008. Canopy-scale chlorophyll vegetation index. Precis Agric 9:303-319.
Wang X, Zhang X, Zhou G, 2017. Rice disease detection using NIR spectra. J Indian Soc Remote Sens 45:785-794.
Wei X, Johnson MA, Langston DB, Mehl HL, Li S, 2021. Identifying optimal wavelengths for disease signatures. Remote Sens 13:2833.
Widodo A, Yang BS, 2007. SVM for machine fault diagnosis. Mech Syst Signal Process 21:2560-2574.
Wrather A, Koenning S, 2009. Effects of diseases on soybean yields. Plant Health Prog 10:24.
Yang C, Everitt JH, Bradford JM, Murden D, 2004. Airborne hyperspectral imagery for cotton yield variability. Precis Agric 5:445-461.
Zhang N, Yang G, Pan Y, Yang X, Chen L, Zha, C, 2020. Hyperspectral-based plant disease detection review. Remote Sens 12:3188.

Received: 12 August 2025; Accepted: 11 January 2026.

Contributions: all authors made a substantive intellectual contribution, read and approved the final version of the manuscript and agreed to be accountable for all aspects of the work.

Conflict of interest: the authors declare no competing interests, and all authors confirm accuracy.

Data Availability Statement: the datasets used and analyzed during the current study are available upon reasonable request from the corresponding author.

Funding: this work (GNU-2024-240045) was supported by the research grant of the new professor of the Gyeongsang National University in 2024.

Acknowledgments: the authors appreciate all staff who helped with this study at the Department of Upland Crop Sciences, National Institute of Crop and Food Science, Republic of Korea.

Publisher's note: all claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article or claim that may be made by its manufacturer is not guaranteed or endorsed by the publisher.

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).