Spectral Analysis Review, 2026, 9(1), 1-18
https://www.scirp.org/journal/sar
ISSN Online: 2331-2106
ISSN Print: 2331-2092

Early Detection of Cocoa Swollen Shoot Using Hyperspectral Reflectance Spectroscopy

Nimongon Seydou Silué[1], Penétjiligué Adama Soro[1], Amara Kamate[1], Emma Georgina Hueva Zoro[1], Kouabenan Anicet Kouakou[2], Adjo Viviane Adohi-Krou[1]

[1]Unité de Formation et de Recherche (UFR) des Sciences des Structures de la Matière et de Technologie (SSMT), Laboratoire des Sciences de la Matière, de l'Environnement et de l'Énergie Solaire (LASMES), Université Félix Houphouët Boigny, Abidjan, Côte D'Ivoire
[2]Unité de Formation et de Recherche (UFR) des Sciences Fondamentales et Appliquées (SFA), Laboratoire de Physique Fondamentale et Appliquée (LPFA), Université Nangui Abrogoua, Abidjan, Côte D'Ivoire
Email: adams.soro@gmail.com

How to cite this paper: Silué, N.S., Soro, P.A., Kamate, A., Zoro, E.G.H., Kouakou, K.A. and Adohi-Krou, A.V. (2026) Early Detection of Cocoa Swollen Shoot Using Hyperspectral Reflectance Spectroscopy. *Spectral Analysis Review*, **9**, 1-18.
https://doi.org/10.4236/sar.2026.91001

Received: November 11, 2025
Accepted: January 27, 2026
Published: January 30, 2026

Abstract

Cocoa swollen shoot is a viral disease prevalent mainly in West Africa. It is responsible for significant yield losses. Effective monitoring and accurate detection of this disease are essential to ensure stable and reliable cocoa production, as well as the income of thousands of farmers. Current standard methods often rely on visual inspection for disease symptoms. This method is costly, time-consuming, and prone to errors due to the subjectivity of inspectors. Recent advances in precision agriculture, using spectral data at different scales (leaf, canopy, and space), offer the potential to solve these problems at low cost and with high efficiency. Spectral vegetation indices are widely used for the indirect detection of plant diseases. Therefore, this work aims to evaluate the potential of hyperspectral reflectance spectroscopy, using vegetation indices, for the early detection of cocoa swollen shoot. Thus, reflectance spectra of healthy, asymptomatic and symptomatic cocoa tree leaves were collected using an Ocean Optics USB 4000 spectrometer, with a wavelength range from 350 to 1100 nm. To reduce scattering and noise effects in the collected data, we applied three essential spectroscopic preprocessing methods: the Multiplicative Scattering Correction (MSC), the Standard Normal Variable (SNV), and the first order Savitzky-Golay derivative (SG-D1). Principal Component Analysis (PCA) identified MSC as the most effective preprocessing method. Subsequently, eight vegetation indices (NDVI, GNDVI, Datt5, Ctr2, PSRI, SIPI, PRI, Lic2) selected from the literature were calculated. Using classification algorithms such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT) and Random Forest (RF) applied to all eight vegetation indices, the performance of each index was evaluated for the infection detection and the clas-

sification of the three types of cocoa trees. The results obtained indicate that the most efficient model is RF combined with MSC pretreatment with an overall accuracy of 97.59%, thus showing the strong potential of reflectance spectroscopy for the early detection of cocoa tree infection by the swollen shoot virus.

Keywords

Swollen Shoot Virus, Early Detection, Spectroscopy, Cocoa Tree, Hyperspectral Reflectance, Machine Learning, Vegetation Indices

1. Introduction

The cocoa tree (*Theobroma cacao L.*) is a plant native to the tropical regions of Central and South America [1]. It is cultivated in several tropical and subtropical areas of the world. The main cocoa producers are Côte d'Ivoire, Ghana, Indonesia, Ecuador, Nigeria, Cameroon, and Brazil [2]. Côte d'Ivoire is the world's leading producer of cocoa beans, followed by Ghana [2]. The production of cocoa and its derivatives is in high demand in the food market, but cocoa cultivation is subject to numerous biotic and abiotic pressures, intensified by climate change [3] [4]. However, the environment in which cocoa trees grow is prone to various diseases that can damage the plant and the fruit, leading to significant production losses [5] [6]. Swollen shoot is the most endemic disease of the cocoa tree in West Africa. It is transmitted by several species of cochineal insects and remains a major limiting factor for cocoa production in Côte d'Ivoire, Ghana, and other West African producing countries [7]. This disease can reduce yields by 30 to 50 percent and even lead to the death of cocoa trees within two to three years of infection [5]. Symptoms of the disease are variable. It manifests mainly through abnormal swelling of shoots, branches and roots; discoloration, deformation, and premature leaf fall; stunted pods and a significant yield reduction, which can lead to the death of the tree [8].

Current methods of swollen shoot infection detection rely on visual inspection of symptoms, enzyme-linked immunosorbent assays (ELISA) and polymerase chain reaction (PCR) [9]-[14] based on biochemical techniques. However, sample processing is lengthy and complex. On the other hand, visual inspection is laborious, time-consuming, prone to numerous errors and is unsuitable for large scale plantations. In recent years, several studies have used imaging data to detect diseases in cocoa crops. These studies generally were based on images of symptomatic and healthy cocoa pods [15]-[17]. However, very few studies focused on the detection of swollen shoots using hyperspectral reflectance data from cocoa leaves [18]. Diffuse reflectance spectroscopy experienced considerable growth in improving and simplifying the complex tasks related to crop assessment. Using machine learning algorithms, it becomes a tool enabling farmers to make better decisions and is also used to automatically identify and classify plant diseases [19]-

[21]. This technique is based on the principle that stress affects the physical structure and photosynthesis of plants, as well as the absorption and the reflectance of light [22] [23]. The use of diffuse reflectance spectral data for disease diagnosis allows detection before disease symptoms are visible and offers the possibility of non-destructive sampling. Furthermore, diffuse reflectance spectroscopy is flexible to implement due to the compact size and low weight of the sensors. Data can be collected from a single leaf spot or across large orchards by mounting the sensors on drones, enabling accurate and real-time results in the field [24]. Spectral measurements of diffuse leaf reflectance can reveal differences between healthy and infected plants [25] [26]. Furthermore, reflectance variations can serve as a basis for the future design of optical sensors for detecting swollen shoot disease in cocoa plants. Therefore, identifying spectral bands and vegetation indices sensitive to this disease is essential for the effective application of reflectance data analysis techniques.

This study aims to detect early infection of cocoa trees by the swollen shoot virus by discriminating healthy plants from asymptomatic and symptomatic infected plants using hyperspectral reflectance spectroscopy. The potential of the machine learning algorithms used was evaluated from spectral vegetation indices susceptible to the disease. Early detection of swollen shoot infection will help to prevent potential epidemics and avoid yield losses through the implementation of an appropriate and timely management strategy.

2. Materials and Methods

2.1. Biological Samples

In this study, we used healthy, asymptomatic, and symptomatic infected cocoa leaves as biological samples. The *Theobroma cacao* variety we studied is named Amelonado. These leaves were collected with the support of the Central Biotechnology Laboratory of the Centre National de Recherche Agronomique (CNRA) in two fields located in Côte d'Ivoire. The healthy samples were collected on a CNRA experimental field located at Adiopodoumé Km 17, while the collection of leaves infected with the swollen shoot virus was carried out in a plantation located at Garango in Bouaflé department where swollen shoot disease is widespread [27]. These experimental cocoa trees plantations were established at the same period and are subject to the same cultivation practices. However, rainfall in Adiopodoume is slightly higher than in Bouaflé. Each type of leaf was collected from 20 trees with 3 leaves per tree, for a total of 60 leaves per type. So the biological samples consisted of 60 healthy leaves, 60 asymptomatic infected leaves, and 60 symptomatic infected leaves. **Figure 1** shows the three leaf types we used.

2.2. Data Collection and Preprocessing

The hyperspectral reflectance signatures of cocoa leaves were acquired using an Ocean Optics USB 4000 spectrometer, which has a wavelength range from 350 to 1100 nm and a spectral resolution of 0.22 nm. The setup includes a bifurcated

optical fiber (QR400-7-VIS-BX), a 20 W halogen source (HL-2000-HP-FHSA), a RPH probe tip, and a laptop computer for data storage and processing. Specialized data acquisition software (SpectraSuite) was used, and the spectrometer was calibrated with a diffuse reflectance reference surface (WS-1). Further details of the experimental setup can be found in previous work [21] [28]. **Figure 2** illustrates the experimental setup used.

Figure 1. Types of leaves submitted for study. (a) healthy leaves; (b) asymptomatic leaves; (c) symptomatic leaves.

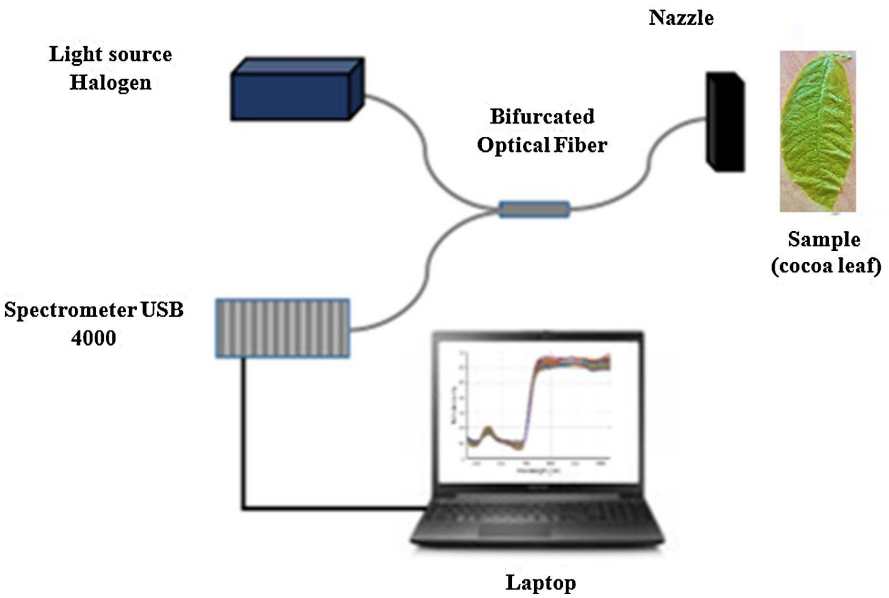

Figure 2. Experimental device for acquiring hyperspectral data.

Measurements were performed on 180 cocoa leaves. For each leaf sample, three mean reflectance spectra were measured on the upper, middle, and lower parts of the leaf's adaxial surface. A total of 540 spectra were then collected, with 180 spectra per leaf type. To reduce the influence of noise inherent to the hyperspectral reflectance device at the limits of its spectral range, only spectral reflectance data between 400 and 1000 nm were used for processing and analysis. These data were then preprocessed using three methods: the Multiplicative Scattering Correction (MSC), the Standard Normal Variable (SNV), and the first order Savitzky-Golay

derivative (SG-D1).

The MSC method effectively eliminates scattering effects while preserving useful spectral information, so that the spectra are as close as possible to a reference spectrum. Generally, the average spectrum of the entire dataset is considered as the reference spectrum [29] [30]. SNV performs standard normalization on each spectral data point to reduce light scattering effects [31]. SG-D1 is highly sensitive to noise. It is an effective technique for suppressing background effects and enhancing subtle, weak spectral features useful for evaluating target parameters [32] [33].

2.3. Principal Component Analysis

Principal component analysis (PCA) is a widely used technique for reducing the dimensionality of high-dimensional data. It transforms a set of potentially correlated variables into a new set of uncorrelated variables, called principal components, through orthogonal transformation [34]. PCA of raw and preprocessed spectra can provide crucial information on data separation [35]. We generated the principal components of the spectral data and visualized the distribution of samples of healthy, asymptomatic, and symptomatic leaves in three dimensions. This visualization will identify the different clusters, thus providing a better understanding of the underlying structure of the raw and preprocessed spectral data.

2.4. Calculated Vegetation Indices

Vegetation indices (VIs), obtained by mathematical transformation from spectral bands, generally highlight the spectral characteristics of leaves [36]. Numerous studies have demonstrated the effectiveness of VIs for monitoring crop diseases [21] [37]-[39]. Taking into account the spectral range of our spectrometer, we selected eight vegetation indices from the literature related to leaf pigments or structure and crop diseases. These indices were used as variables in the development of classification models for the three types of cocoa leaves. Using spectral data, a program developed in MATLAB allowed us to calculate these eight vegetation indices. **Table 1** presents the definitions, calculation formulas, and bibliographic references for each VI.

The statistical analysis of significant differences applied to the independent variables consisted of verifying the differences between healthy and infected (asymptomatic and symptomatic) cocoa leaves based on all the eight VIs. Normality was tested using the Shapiro-Wilk test [48] and equality of variances with the Leneve test [49]. Since non-normality and non-homogeneity of variances were observed in the data for asymptomatic and symptomatic leaves, a statistical difference analysis was performed using the non-parametric Kruskal-Wallis test [50]. This test does not require assumptions about normality and homoscedasticity, making it a robust option for comparing multiple independent samples. Finally, Dunn's multiple comparisons test [51] was applied with a significance level of 5% using MATLAB software.

Table 1. Vegetation indices used.

Vegetation indices	Formula	References
Normalized Difference Vegetation Index (NDVI)	$(R_{800}-R_{670})/(R_{800}+R_{670})$	[40]
Green Normalized Difference Vegetation Index (GNDVI)	$(R_{800}-R_{550})/(R_{800}+R_{550})$	[41]
Datt5	R_{672}/R_{550}	[42]
Carter Indices 2 (Ctr2)	R_{695}/R_{760}	[43]
Plant Senescence Reflectance Index (PSRI)	$(R_{680}-R_{500})/R_{750}$	[44]
Structure Intensive Pigment Index (SIPI)	$(R_{800}-R_{445})/(R_{800}-R_{680})$	[45]
Photochemical Reflectance Index (PRI)	$(R_{531}-R_{570})/(R_{531}+R_{570})$	[46]
Lichtenthaler Index 2 (Lic2)	R_{440}/R_{690}	[47]

2.5. Classification Models

To compare the performance of the classifiers based on the all eight vegetation indices evaluated in this study, four machine learning algorithms frequently used for classification were employed, namely: K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF). These methods were chosen because of their high classification accuracy in similar studies [52]-[55].

The KNN algorithm is a supervised machine learning algorithm widely used for classification and predictive regression problems. It is capable to describe nonlinear relationships between collected samples and hyperspectral data. Specifically, for given test samples, a specific distance evaluation method is used to determine the k closest samples [56]. Then, a predictive classification is performed on these k samples. In general, KNN is easy to use and works well with little prior knowledge of the data distribution. Euclidean distance is the most common method for calculating the variations between samples, represented by input vectors [57].

SVM is a supervised algorithm used in machine learning. It performs classification by finding the hyperplane that maximizes the margin between data. The vectors that define the hyperplane are called support vectors. SVM is based on statistical approaches, enabling data classification and assigning each class a specific score that serves as the basis for evaluation. SVM can also be used for regression tasks even if it is more commonly used for classification purposes [58]. Intuitively, good separation is achieved by the hyperplane that has the greatest distance from the nearest training data points of any class. Indeed, in general, the larger the margin is, the smaller the classifier's generalization error is [59].

A decision tree (DT) is a widely used supervised machine learning technique for classification and regression tasks. The classifier breaks down a dataset into progressively smaller subsets based on their attribute values [60]. The decision tree algorithm has a tree structure where internal nodes represent features, while

terminal nodes contain class labels or predicted values. A prediction is made by traversing the path from the root to one of the terminal nodes based on the input feature values. This process divides the data until no further division is possible or when all values of the target variable are identical. Decision trees are easy to understand and interpret. They can handle numerical and categorical variables and are robust to outliers and missing values. Most decision trees consist of a random forest-type classifier that determines the category based on the classes in a particular tree.

Random Forest (RF) is a classification model composed of multiple decision trees, combined to obtain a more accurate and stable prediction. Each tree depends on the values of a randomly sampled vector that follows the same distribution for all trees in the forest. The final classification result is obtained by averaging the classification results of all individual binary decision trees [61]. The RF method has many advantages over other machine learning methods; the most notable one is its high accuracy when a large dataset is used for training [62]. Furthermore, it is simple and quick to implement.

2.6. Performance Metrics

In order to better assess the performance of the classification models used and their stability for the swollen shoot infection detection, 540 observations of the eight vegetation indices were divided into two datasets: a training set (70%) and a test set (30%). The split of samples was performed at the leaf level, so all spectra acquired on the same leaf belong to the same dataset. A confusion matrix was provided for each model, comparing the predicted classes of the test set to the real classes in order to evaluate the performance of classification models based on True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN) [63]. Accuracy, precision, recall and F1-Score were computed as parameters for evaluating model performance. The formulas for these performance indicators are respectively given by Equation (1), Equation (2), Equation (3) and Equation (4) as follows:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

This metric reveals the model's predictive performance, reflecting its overall ability to correctly classify healthy, asymptomatic, and symptomatic leaves.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

Precision represents the proportion of real positive cases among all samples predicted to be positive. The higher the Precision is, the better the model performs at distinguishing between healthy, asymptomatic, and symptomatic leaves. In other words, high accuracy means fewer incorrect positive predictions.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

This indicator shows the proportion of samples correctly predicted to be posi-

tive and which actually are. Thus, Recall can be used to assess the completeness of the prediction of healthy, asymptomatic, and symptomatic leaves.

$$\text{F1-Score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{4}$$

This is the harmonic mean between Precision and Recall, taking into account the overall balance between the two parameters. The F1-Score represents a balanced average between positive predictive value and sensitivity. An optimal value for this metric indicates better algorithm efficiency.

3. Results and Discussion

3.1. Spectral Reflectance Signature of Cocoa Leaves

Spectral preprocessing is an essential step in spectral analysis. This approach yields reliable and more accurate spectral data, facilitating subsequent analyses and applications. Preprocessing modifies spectra in various ways. Figure 3 provides an overview of the raw and preprocessed reflectance spectra of healthy, asymptomatic, and symptomatic cocoa leaf samples.

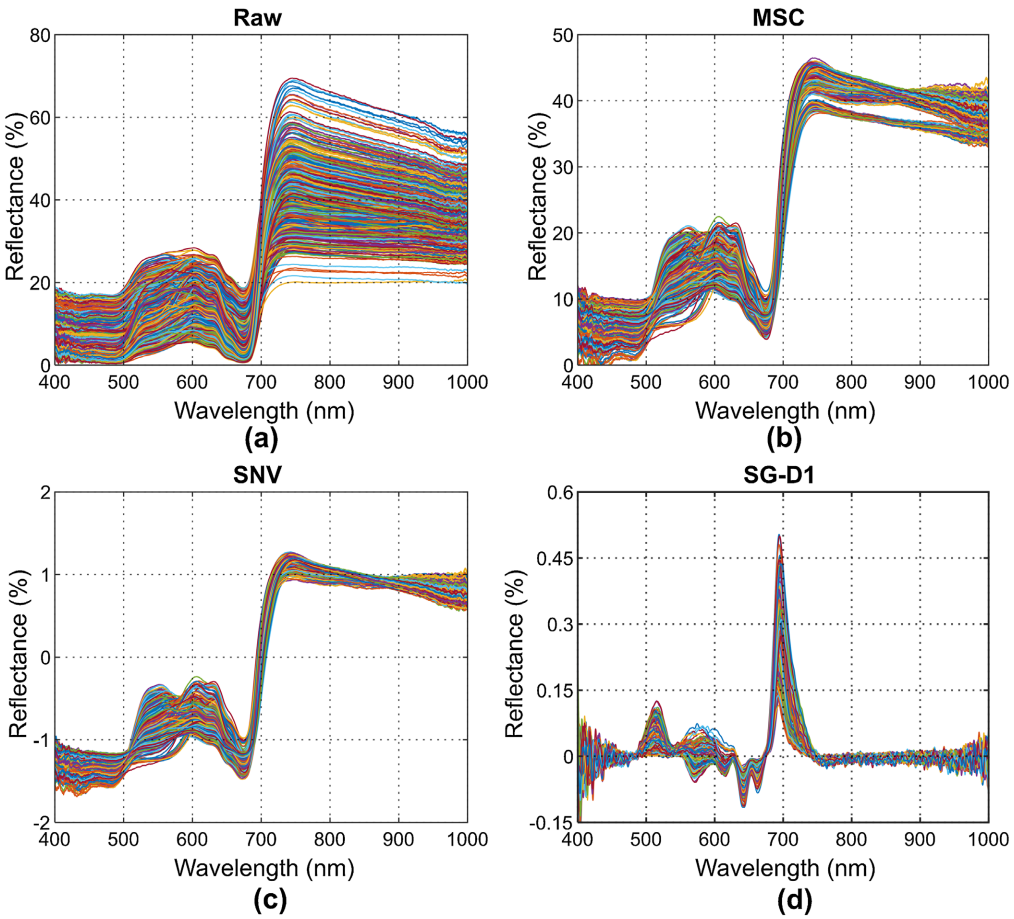

Figure 3. Raw reflectance spectra (a) and preprocessed: MSC (b), SNV (c) and SG-D1 (d).

Observation of the raw spectra in Figure 3(a) reveals that the spectral signatures

of our samples are similar, with low reflectance in the visible region, a sharp increase in the far-infrared and higher reflectance in the near-infrared. However, these raw reflectance spectra of healthy, asymptomatic, and symptomatic cocoa leaves exhibit strong dispersion in the 400 - 1000 nm wavelength range. The MSC method corrects the scattering effect in spectral data to bring these data closer to the real spectral information. The spectral signatures in Figure 3(b) are more consistent and the spectral differences due to scattering are corrected. SNV homogenizes the spectra by standardizing each spectrum, eliminating multiplicative noise and reflectance variations. The spectra in Figure 3(c) are normalized to the same scale to facilitate comparison and subsequent analysis. The SG-D1 highlights changes in the position of emission or absorption peaks in the spectral reflectance signature. Figure 3(d) provides more information on the positions of these peaks but this may weaken the relative variation in their intensity.

3.2. Principal Component Analysis of Raw and Preprocessed Spectra

The spectral curves obtained by pretreatment, while each exhibiting their own characteristics, cannot be intuitively compared in terms of their effects on the reflectance data of our samples. Therefore, in this study, principal component analysis (PCA) was used to analyze the results of the different pretreatment methods in order to identify any clustering trends in the healthy and infected cocoa leaf samples. The PCA results obtained from the raw and pretreated spectra showed that the first three principal components provided useful clustering for healthy, asymptomatic, and symptomatic cocoa leaves. Among the three pretreatment methods, MSC proved to be the best method compared to the others, as shown in Figure 4. Indeed, it yielded clearer clustering trends in the space of the first three principal components.

The best grouping by leaf type (healthy, asymptomatic, and symptomatic) was obtained by combining MSC pretreatment and PCA. This could be attributed to the ability of the MSC technique to correct the scattering effects and eliminate unwanted scattering from the spectral data matrix. Thus, the MSC pretreatment technique is best suited to our raw reflectance data for better classification of the three types of cocoa plants.

3.3. Analysis of Significant Differences between Cocoa Leaves Types Using Vegetation Indices

The Kruskal-Wallis nonparametric test and Dunn's multiple comparisons test were performed to evaluate the differences between healthy and infected cocoa leaves using various vegetation indices. The results are summarized in Table 2. They indicate that all vegetation indices of infected plants are significantly influenced by viral infection. Indeed, the calculated vegetation indices show significant differences between healthy and asymptomatic leaves, and also between healthy and symptomatic leaves. Regarding the comparison between the vegetation indices

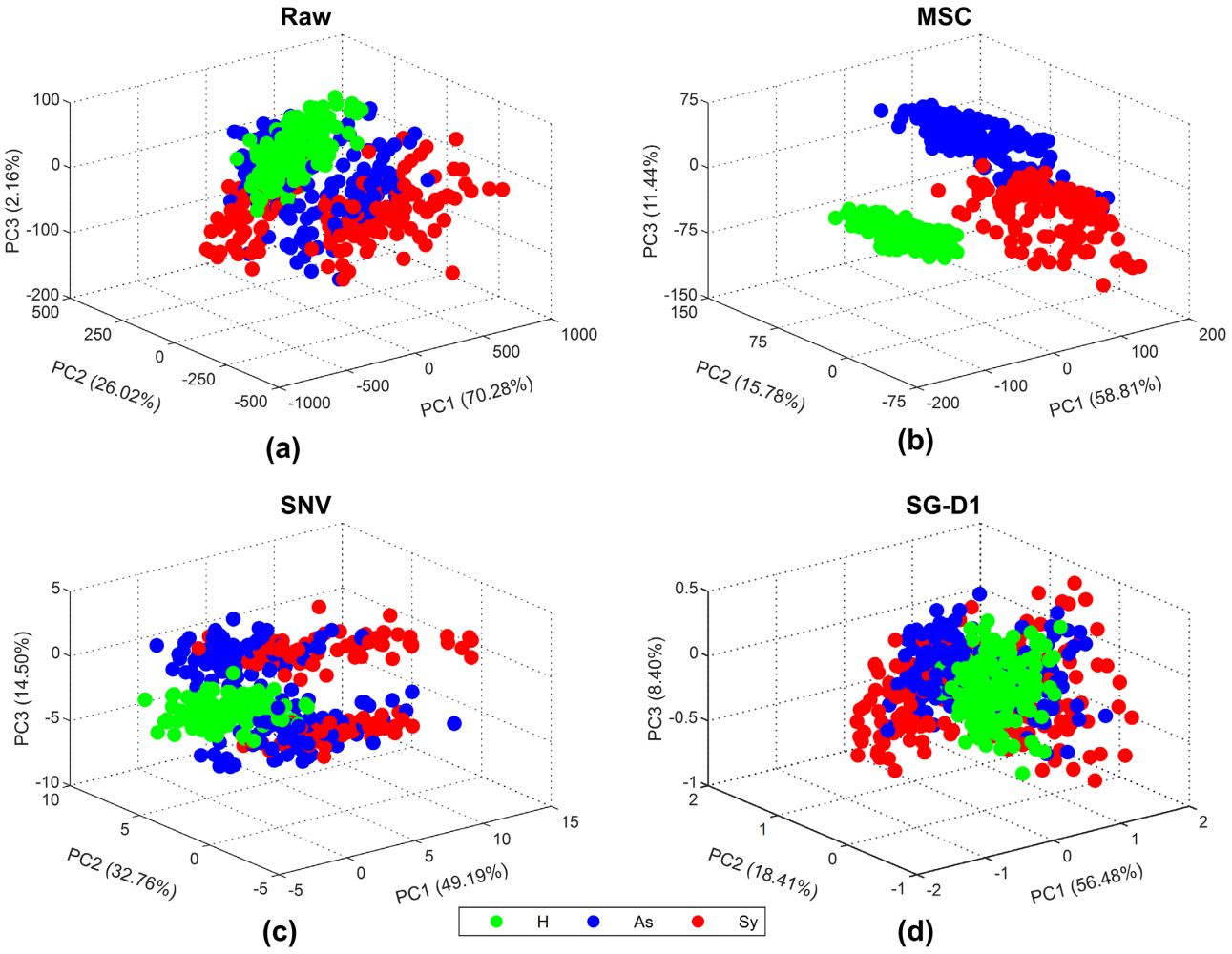

Figure 4. Representation of the first three principal components of raw reflectance spectra (a), pretreated by: MSC (b), SNV (c), SG-D1 (d) of healthy (H), asymptomatic (As) and symptomatic (Sy) cocoa leaves.

Table 2. Significant differences in vegetation indices of healthy (H), asymptomatic (As) and symptomatic (Sy) cocoa leaves, obtained by the Kruskal-Wallis test for a statistical significance threshold $p < 0.05$.

	NDVI	GNDVI	Datt5	Ctr2	PSRI	SIPI	PRI	Lic2
H-As	*	*	*	*	*	*	*	*
H-Sy	*	*	*	*	*	*	*	*
As-Sy	ns	*	ns	*	*	ns	*	*

(*) significant difference, (ns) no significant difference

of asymptomatic and symptomatic leaves, only the GNDVI, Ctr2, PSRI, PRI, and Lic2 indices show significant differences. These five vegetation indices are significantly different for the three cocoa leaves types, so they well discriminated them. This is explained by the fact that these vegetation indices are closely linked to changes in leaf pigments. This interdependence influences their values [64]-[66]. These changes are manifested by the symptoms of chlorosis observed on the

symptomatic leaves of cocoa plants infected by the swollen shoot virus.

The distribution of each vegetation index for the three cocoa leaves states is illustrated by the box plots in Figure 5.

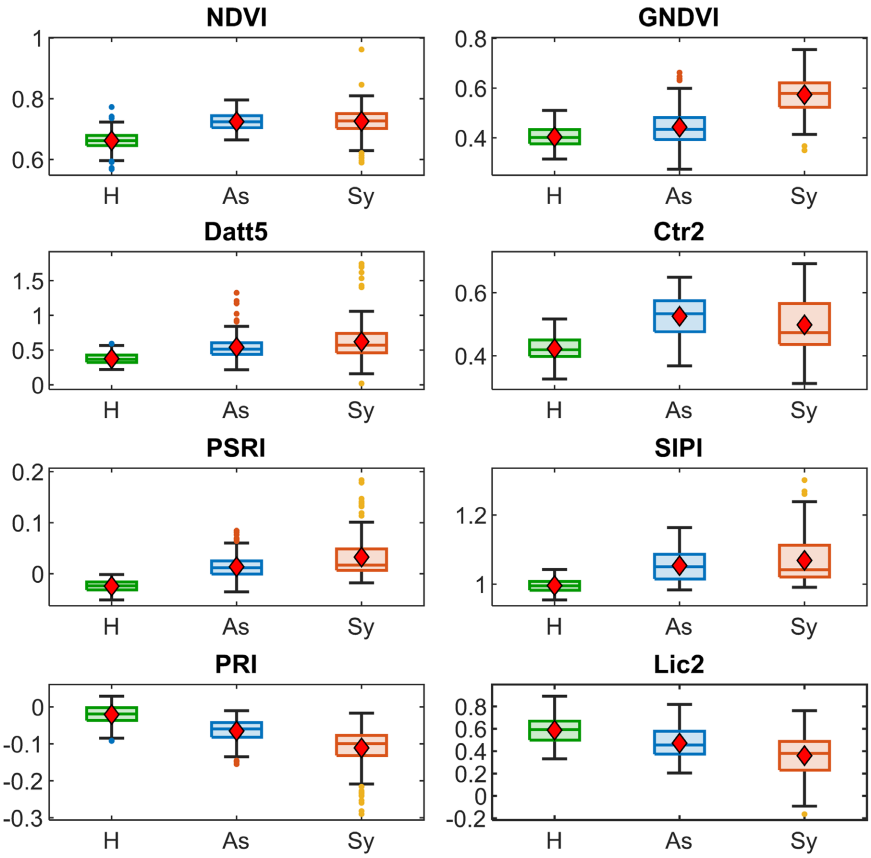

Figure 5. Box plot of vegetation indices of healthy (H), asymptomatic (As) and symptomatic (Sy) cocoa leaves.

Compared to healthy leaves, the median values of the vegetation indices NDVI, GNDVI, Datt5, PSRI, and SIPI increase with infection, while those of PRI and Lic2 decrease. There is virtually no overlap between the box plots of healthy and symptomatic leaves vegetation indices. In contrast, there is generally some overlap between the boxes for asymptomatic and symptomatic cocoa leaves. Similarly, for GNDVI, Dattt5 and Lic2, there is slight overlap between the boxes for healthy and asymptomatic cocoa leaves. This demonstrates the difficulty to detect early infection by cocoa swollen shoot virus using threshold values based on individual vegetation indices.

3.4. Classification Models for Healthy and Infected Cocoa Leaves

The overall performance of the four evaluated algorithms (KNN, SVM, DT, RF) is excellent, as shown in Table 3. RF stands out as the most efficient algorithm, with an accuracy of 97.59%, a Precision of 97.62%, a Recall of 97.59%, and an F1-Score of 97.59%. These results demonstrate extremely reliable classification and

excellent consistency between accuracy and Recall, thus highlighting the robustness of the model. The KNN algorithm also performed well, with an accuracy of 95.93%, a Precision of 96.25%, a Recall of 95.93%, and an F1-Score of 95.92%, making it a competitive alternative. The performance of the SVM algorithm is also remarkable, with an accuracy of 93.89%, a Precision of 94.1%, a Recall of 93.89%, and an F1-Score of 93.91%, but significantly less efficient than the RF and KNN models. In contrast, DT achieved more modest results, with 90.74%, 90.85%, 90.74%, and 90.74% respectively for accuracy, Precision, Recall, and the F1-Score. This indicates a strong classification capacity, but significantly lower stability and reliability than the other three models. These results confirm the relevance of MSC preprocessing for improving the quality of our spectral data, the effectiveness of the computed vegetation indices and reveal the robustness of the RF. Indeed, in this study, RF proved to be the most accurate and robust model for classifying healthy, asymptomatic and symptomatic cocoa leaves. Therefore, we can conclude that cocoa swollen shoot disease can effectively be diagnosed at an early stage using a combination of sensitive vegetation indices as input variables in a classification model.

Table 3. Algorithms performance evaluation.

Models	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
KNN	95.93	96.25	95.93	95.92
SVM	93.89	94.10	93.89	93.91
DT	90.74	90.86	90.74	90.75
RF	97.59	97.67	97.59	97.59

To further illustrate the predictive performance of the different classification models, for healthy, asymptomatic, and symptomatic cocoa leaves, we present in **Figure 6** the confusion matrices of the KNN, SVM, DT and RF classifications.

As illustrated in **Figure 6**, for a total of 180 samples, the RF model correctly predicted 176 samples, while the KNN, SVM, and DT models predicted 172, 170, and 163 samples, respectively. Despite some differences in predictive accuracy among these models when classifying the three cocoa leaves categories, all four models performed satisfactorily.

The performance obtained with the RF model is similar to that of Kouassi *et al.* (2024) [17], who used convolutional neural networks combined with different classifiers. The combinations with the XGBOOST classifier to differentiate between healthy and infected cocoa leaves yielded the best performance, with overall accuracies exceeding 95%.

In contrast, Batoo *et al.* (2025) [18] used a three-layer convolutional neural network to classify infected and healthy cocoa tree samples. This classification reached an accuracy of 88%, lower than that of our models.

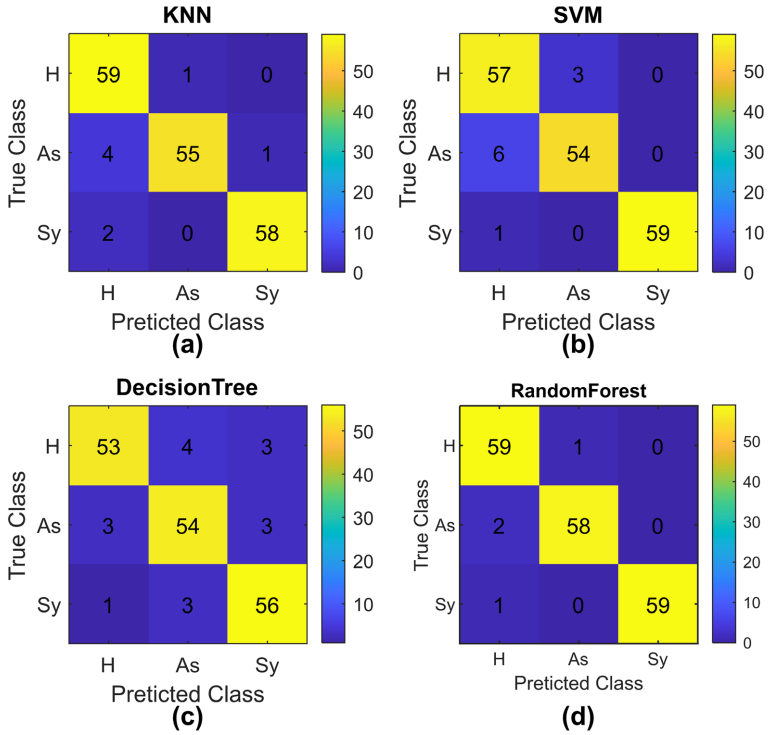

Figure 6. Confusion matrices: (a) KNN; (b) SVM; (c) DT and (d) RF.

4. Conclusion

In this paper, spectral reflectance data were used to discriminate between healthy and swollen shoot virus-infected cocoa leaves. This study successfully constructed four classification models for healthy, asymptomatic, and symptomatic cocoa leaves based on eight vegetation indices and explored the impact of spectral data preprocessing. Indeed, principal component analysis of the raw and preprocessed spectral data using the MSC, SNV, and SG-D1 methods showed that the MSC preprocessing method best discriminated between the three cocoa plant types. All eight vegetation indices, calculated from the spectral data preprocessed by the MSC method, were used as input data for the four classification models. The results indicate that the Random Forest model performs best, with an accuracy of 97.59%, surpassing the KNN, SVM, and DT models. However, KNN remains a robust alternative model with an accuracy of 95.93%. These results confirm that machine learning, applied to carefully selected vegetation indices, is a promising approach for automating the detection of cocoa swollen shoot virus infection in leaves, thus offering a new approach for controlling this viral disease. Future work will focus on optimizing pretreatment methods and exploring other vegetation indices to increase the accuracy and efficiency of the models. Furthermore, we plan to collect our biological samples at the same experimental site.

Acknowledgements

We would like to thank the Centre National de Recherche Agronomique (CNRA) for its technical collaboration.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Cuatrecasas, J. (1964) Cacao and Its Allies: A Taxonomic Revision of the Genus Theobroma. In: *Contributions from the United States National Herbarium*, Smithsonian Institution Press, 379-614.

[2] Food and Agriculture Organization of the United Nations (2025) FAOSTAT: Crops and Livestock Products. https://www.fao.org/faostat/en/#data/QCL/visualize

[3] Sultan, B., Defrance, D. and Iizumi, T. (2019) Evidence of Crop Production Losses in West Africa Due to Historical Global Warming in Two Crop Models. *Scientific Reports*, **9**, Article No. 12834. https://doi.org/10.1038/s41598-019-49167-0

[4] Morillo, F., Sánchez, P., Herrera, B., Liendo-Barandiaran, C., Muñoz, W. and Vicente Hernández, J. (2009) Pupal Development, Longevity and Behavior of *Carmenta theobromae* (Lepidoptera: Sesiidae). *Florida Entomologist*, **92**, 355-361. https://doi.org/10.1653/024.092.0222

[5] Muller, E. (2016) Cacao Swollen Shoot Virus (CSSV): History, Biology, and Genome. In: *Cacao Diseases*: *A History of Old Enemies and New Encounters*, Springer, 337-358.

[6] Sena, K., Alemanno, L. and Gramacho, K.P. (2014) The Infection Process of *Moniliophthora perniciosa* in Cacao. *Plant Pathology*, **63**, 1272-1281. https://doi.org/10.1111/ppa.12224

[7] Posnette, A.F. and Strickland, A.H. (1948) Virus Diseases of Cacao in West Africa. *Annals of Applied Biology*, **35**, 53-63. https://doi.org/10.1111/j.1744-7348.1948.tb07350.x

[8] Ofori, A., Padi, F.K., Ameyaw, G.A., Dadzie, A.M., Opoku-Agyeman, M., Domfeh, O., *et al.* (2022) Field Evaluation of the Impact of Cocoa Swollen Shoot Virus Disease Infection on Yield Traits of Different Cocoa (*Theobroma cacao* L.) Clones in Ghana. *PLOS ONE*, **17**, e0262461. https://doi.org/10.1371/journal.pone.0262461

[9] Dzahini-Obiatey, H., Ameyaw, G.A. and Ollennu, L.A. (2006) Control of Cocoa Swollen Shoot Disease by Eradicating Infected Trees in Ghana: A Survey of Treated and Replanted Areas. *Crop Protection*, **25**, 647-652. https://doi.org/10.1016/j.cropro.2005.09.004

[10] Kouakou, K., Kébé, B.I., Kouassi, N., Anno, A.P., Aké, S. and Muller, E. (2011) Impact de la Maladie Virale du Swollen Shoot du Cacaoyer sur la Production de Cacao en Milieu Paysan à Bazré (Côte d'Ivoire). *Journal of Applied Biosciences*, **43**, 2947-2957. http://www.m.elewa.org/JABS/2011/43/7.pdf

[11] Ramos-Sobrinho, R., Kouakou, K., Bi, A.B., Keith, C.V., Diby, L., Kouame, C., *et al.* (2021) Molecular Detection of Cacao Swollen Shoot Badnavirus Species by Amplification with Four PCR Primer Pairs, and Evidence That Cacao Swollen Shoot Togo B Virus-Like Isolates Are Highly Prevalent in Côte d'Ivoire. *European Journal of Plant Pathology*, **159**, 941-947. https://doi.org/10.1007/s10658-021-02203-0

[12] Muller, E., Jacquot, E. and Yot, P. (2001) Early Detection of Cacao Swollen Shoot Virus Using the Polymerase Chain Reaction. *Journal of Virological Methods*, **93**, 15-22. https://doi.org/10.1016/s0166-0934(00)00241-x

[13] Sagemann, W., Lesemann, D.E., Paul, H.L., Adomako, D. and Owusu, G.K. (1985) Detection and Comparison of Some Ghanaian Isolates of Cacao Swollen Shoot Virus

(CSSV) by Enzyme-Linked Immunosorbent Assay (ELISA) and Immunoelectron Microscopy (IEM) Using an Antiserum to CSSV Strain 1a. *Journal of Phytopathology*, **114**, 79-89. https://doi.org/10.1111/j.1439-0434.1985.tb04339.x

[14] Ameyaw, G.A., Dzahini-Obiatey, H.K. and Domfeh, O. (2014) Perspectives on Cocoa Swollen Shoot Virus Disease (CSSVD) Management in Ghana. *Crop Protection*, **65**, 64-70. https://doi.org/10.1016/j.cropro.2014.07.001

[15] Coulibaly, M., Kouassi, K.H., Kolo, S. and Asseu, O. (2020) Detection of "Swollen Shoot" Disease in Ivorian Cocoa Trees via Convolutional Neural Networks. *Engineering*, **12**, 166-176. https://doi.org/10.4236/eng.2020.123014

[16] Kumi, S., Kelly, D., Woodstuff, J., Lomotey, R.K., Orji, R. and Deters, R. (2022) Cocoa Companion: Deep Learning-Based Smartphone Application for Cocoa Disease Detection. *Procedia Computer Science*, **203**, 87-94. https://doi.org/10.1016/j.procs.2022.07.013

[17] Kouassi, K.S., Diarra, M., Edi, K.H. and Koua, B.J. (2024) Detection of Cocoa Leaf Diseases Using the CNN-Based Feature Extractor and XGBOOST Classifier. *Open Journal of Applied Sciences*, **14**, 2955-2972. https://doi.org/10.4236/ojapps.2024.1410193

[18] Batool, T., Allainguillaume, J., Zhang, W. and Bell, M.J. (2025) Application of Machine Learning to Screen Hyperspectral Data of Cacao Plants to Identify Cacao Swollen Shoot Virus (CSSV). In: *Precision Agriculture* 25 (*SET TWO VOLUMES*), Brill Wageningen Academic, 105-112. https://doi.org/10.1163/9789004725232_012

[19] Lacotte, V., Peignier, S., Raynal, M., Demeaux, I., Delmotte, F. and da Silva, P. (2022) Spatial-Spectral Analysis of Hyperspectral Images Reveals Early Detection of Downy Mildew on Grapevine Leaves. *International Journal of Molecular Sciences*, **23**, Article 10012. https://doi.org/10.3390/ijms231710012

[20] Khan, I.H., Liu, H., Li, W., Cao, A., Wang, X., Liu, H., *et al.* (2021) Early Detection of Powdery Mildew Disease and Accurate Quantification of Its Severity Using Hyperspectral Images in Wheat. *Remote Sensing*, **13**, Article 3612. https://doi.org/10.3390/rs13183612

[21] Kamate, A., Soro, P.A., Zoro-Diama, E.G., Diomandé, K.S. and Adohi-Krou, A.V. (2023) Detection of Rice Yellow Mottle at the Asymptomatic Stage by Hyperspectral Fluorescence and Reflectance Spectroscopies. *Optics and Photonics Journal*, **13**, 63-78. https://doi.org/10.4236/opj.2023.134005

[22] Falcioni, R., Antunes, W.C., Demattê, J.A.M. and Nanni, M.R. (2023) Reflectance Spectroscopy for the Classification and Prediction of Pigments in Agronomic Crops. *Plants*, **12**, Article 2347. https://doi.org/10.3390/plants12122347

[23] Schaepman, M.E., Ustin, S.L., Plaza, A.J., Painter, T.H., Verrelst, J. and Liang, S. (2009) Earth System Science Related Imaging Spectroscopy—An Assessment. *Remote Sensing of Environment*, **113**, S123-S137. https://doi.org/10.1016/j.rse.2009.03.001

[24] Liu, N., Zhang, W., Liu, F., Zhang, M., Du, C., Sun, C., *et al.* (2022) Development of a Crop Spectral Reflectance Sensor. *Agronomy*, **12**, Article 2139. https://doi.org/10.3390/agronomy12092139

[25] Skoneczny, H., Kubiak, K., Spiralski, M., Kotlarz, J., Mikiciński, A. and Puławska, J. (2020) Fire Blight Disease Detection for Apple Trees: Hyperspectral Analysis of Healthy, Infected and Dry Leaves. *Remote Sensing*, **12**, Article 2101. https://doi.org/10.3390/rs12132101

[26] Song, Z., Liu, Y., Yu, J., Guo, Y., Jiang, D., Zhang, Y., *et al.* (2024) Estimation of Chlorophyll Content in Apple Leaves Infected with Mosaic Disease by Combining

Spectral and Textural Information Using Hyperspectral Images. *Remote Sensing*, **16**, Article 2190. https://doi.org/10.3390/rs16122190

[27] Aka Romain, A., Klotioloma, C., Pierre N'Guessan, W., Kouakou, K., Gnion Mathias, T., F. N'Guessan, K., *et al.* (2020) Cocoa Swollen Shoot Disease in Côte d'Ivoire: History of Expansion from 2008 to 2016. *International Journal of Sciences*, **9**, 52-60. https://doi.org/10.18483/ijsci.2203

[28] Kamate, A., Soro, P.A., Zoro-Diama, E.G., Diomande, K.S. and Adohi-Krou, A.V. (2023) Water Stress Early Detection of Eggplant Plants by Hyperspectral Fluorescence Spectroscopy. *Open Journal of Applied Sciences*, **13**, 343-354. https://doi.org/10.4236/ojapps.2023.133028

[29] Isaksson, T. and Næs, T. (1988) The Effect of Multiplicative Scatter Correction (MSC) and Linearity Improvement in NIR Spectroscopy. *Applied Spectroscopy*, **42**, 1273-1284. https://doi.org/10.1366/0003702884429869

[30] Yang, H., Chen, Q., Qian, J., Li, J., Lin, X., Liu, Z., *et al.* (2024) Determination of Dry-Matter Content of Kiwifruit before Harvest Based on Hyperspectral Imaging. *AgriEngineering*, **6**, 52-63. https://doi.org/10.3390/agriengineering6010004

[31] Barnes, R.J., Dhanoa, M.S. and Lister, S.J. (1989) Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra. *Applied Spectroscopy*, **43**, 772-777. https://doi.org/10.1366/0003702894202201

[32] Savitzky, A. and Golay, M.J.E. (1964) Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, **36**, 1627-1639. https://doi.org/10.1021/ac60214a047

[33] Demetriades-Shah, T.H., Steven, M.D. and Clark, J.A. (1990) High Resolution Derivative Spectra in Remote Sensing. *Remote Sensing of Environment*, **33**, 55-64. https://doi.org/10.1016/0034-4257(90)90055-q

[34] Gewers, F.L., Ferreira, G.R., Arruda, H.F.D., Silva, F.N., Comin, C.H., Amancio, D.R., *et al.* (2021) Principal Component Analysis: A Natural Approach to Data Exploration. *ACM Computing Surveys*, **54**, 1-34. https://doi.org/10.1145/3447755

[35] Vigni, M.L., Durante, C. and Cocchi, M. (2013) Exploratory Data Analysis. In: *Data Handling in Science and Technology*, Elsevier, 55-126. https://doi.org/10.1016/b978-0-444-59528-7.00003-x

[36] Thorp, K.R. (2024) Vegspec: A Compilation of Spectral Vegetation Indices and Transformations in Python. *SoftwareX*, **28**, Article 101928. https://doi.org/10.1016/j.softx.2024.101928

[37] Bi, P., Yu, L., Zhou, Q., Kuang, J., Tang, R., Ren, L., *et al.* (2024) Early Detection of Dendroctonus Valens Infestation with UAV-Based Thermal and Hyperspectral Images. *Remote Sensing*, **16**, Article 3840. https://doi.org/10.3390/rs16203840

[38] Gao, B., Yu, L., Ren, L., Zhan, Z. and Luo, Y. (2023) Early Detection of Dendroctonus Valens Infestation at Tree Level with a Hyperspectral UAV Image. *Remote Sensing*, **15**, Article 407. https://doi.org/10.3390/rs15020407

[39] Vera-Esmeraldas, A., Pizarro-Oteíza, S., Labbé, M., Rojo, F. and Salazar, F. (2025) UAV-Based Spectral and Thermal Indices in Precision Viticulture: A Review of NDVI, NDRE, SAVI, GNDVI, and CWSI. *Agronomy*, **15**, Article 2569. https://doi.org/10.3390/agronomy15112569

[40] Rouse, J.W., Haas, R.H., Deering, D.W., Schell, J.A. and Harlan, J.C. (1974) Monitoring the Vernal Advancement and Retrogradation (Green Wave Effect) of Natural Vegetation. Rapport Final NASA/GSFC Type III, NASA-CR-144661, Greenbelt, Maryland, USA: NASA/GSFC, 390p.

https://skclivinglandscapes.org/remote_sensing/resources/Section6Resources/Rouse_et_al.1974NDVI.pdf

[41] Gitelson, A.A. and Merzlyak, M.N. (1997) Remote Estimation of Chlorophyll Content in Higher Plant Leaves. *International Journal of Remote Sensing*, **18**, 2691-2697. https://doi.org/10.1080/014311697217558

[42] Datt, B. (1998) Remote Sensing of Chlorophyll A, Chlorophyll B, Chlorophyll A+B, and Total Carotenoid Content in Eucalyptus Leaves. *Remote Sensing of Environment*, **66**, 111-121. https://doi.org/10.1016/s0034-4257(98)00046-7

[43] Carter, G.A. (1994) Ratios of Leaf Reflectances in Narrow Wavebands as Indicators of Plant Stress. *International Journal of Remote Sensing*, **15**, 697-703. https://doi.org/10.1080/01431169408954109

[44] Merzlyak, M.N., Gitelson, A.A., Chivkunova, O.B. and Rakitin, V.Y. (1999) Non-Destructive Optical Detection of Pigment Changes during Leaf Senescence and Fruit Ripening. *Physiologia Plantarum*, **106**, 135-141. https://doi.org/10.1034/j.1399-3054.1999.106119.x

[45] Penuelas, J., Baret, F. and Filella, I. (1995) Semi-Empirical Indices to Assess Carotenoids/Chlorophyll A Ratio from Leaf Spectral Reflectance. *Photosynthetica*, **31**, 221-230.

[46] Gamon, J.A., Peñuelas, J. and Field, C.B. (1992) A Narrow-Waveband Spectral Index That Tracks Diurnal Changes in Photosynthetic Efficiency. *Remote Sensing of Environment*, **41**, 35-44. https://doi.org/10.1016/0034-4257(92)90059-s

[47] Lichtenthaler, H.K., Lang, M., Sowinska, M., Heisel, F. and Miehé, J.A. (1996) Detection of Vegetation Stress via a New High Resolution Fluorescence Imaging System. *Journal of Plant Physiology*, **148**, 599-612. https://doi.org/10.1016/s0176-1617(96)80081-2

[48] Shapiro, S.S. and Wilk, M.B. (1965) An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, **52**, 591-611. https://doi.org/10.1093/biomet/52.3-4.591

[49] Schultz, B.B. (1985) Levene's Test for Relative Variation. *Systematic Zoology*, **34**, 449-456. https://doi.org/10.2307/2413207

[50] Kruskal, W.H. and Wallis, W.A. (1952) Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, **47**, 583-621. https://doi.org/10.1080/01621459.1952.10483441

[51] Dunn, O.J. (1964) Multiple Comparisons Using Rank Sums. *Technometrics*, **6**, 241-252. https://doi.org/10.1080/00401706.1964.10490181

[52] Zhang, C., Liu, Y. and Tie, N. (2023) Forest Land Resource Information Acquisition with Sentinel-2 Image Utilizing Support Vector Machine, K-Nearest Neighbor, Random Forest, Decision Trees and Multi-Layer Perceptron. *Forests*, **14**, Article 254. https://doi.org/10.3390/f14020254

[53] Anku, K., Percival, D. and Heung, B. (2025) Field Assessment Strategies: Assessing and Classifying Blight Disease in Wild Blueberry Populations Using Multispectral and Hyperspectral Sensors. *Remote Sensing*, **17**, Article 3074. https://doi.org/10.3390/rs17173074

[54] Macedo, F.L., Nóbrega, H., de Freitas, J.G.R. and Pinheiro de Carvalho, M.A.A. (2025) Assessment of Vegetation Indices Derived from UAV Imagery for Weed Detection in Vineyards. *Remote Sensing*, **17**, Article 1899. https://doi.org/10.3390/rs17111899

[55] Marin, D.B., Ferraz, G.A.E.S., Santana, L.S., Barbosa, B.D.S., Barata, R.A.P., Osco,

L.P., *et al.* (2021) Detecting Coffee Leaf Rust with UAV-Based Vegetation Indices and Decision Tree Machine Learning Models. *Computers and Electronics in Agriculture*, **190**, Article 106476. https://doi.org/10.1016/j.compag.2021.106476

[56] Akbulut, Y., Sengur, A., Guo, Y. and Smarandache, F. (2017) NS-k-NN: Neutrosophic Set-Based K-Nearest Neighbors Classifier. *Symmetry*, **9**, Article 179. https://doi.org/10.3390/sym9090179

[57] Hatem, M.Q. (2022) Skin Lesion Classification System Using a K-Nearest Neighbor Algorithm. *Visual Computing for Industry*, *Biomedicine*, *and Art*, **5**, Article No. 7. https://doi.org/10.1186/s42492-022-00103-6

[58] Ahmad, M., Aftab, S., Salman, M., Hameed, N., Ali, I. and Nawaz, Z. (2018) SVM Optimization for Sentiment Analysis. *International Journal of Advanced Computer Science and Applications*, **9**, 393-398. https://doi.org/10.14569/ijacsa.2018.090455

[59] Müller, K., Mika, S., Tsuda, K. and Schölkopf, K. (2018) An Introduction to Kernel-Based Learning Algorithms. In: *Handbook of Neural Network Signal Processing*, CRC Press, 94-133. https://doi.org/10.1201/9781315220413-4

[60] Pradhan, B. (2013) A Comparative Study on the Predictive Ability of the Decision Tree, Support Vector Machine and Neuro-Fuzzy Models in Landslide Susceptibility Mapping Using GIS. *Computers & Geosciences*, **51**, 350-365. https://doi.org/10.1016/j.cageo.2012.08.023

[61] Zuntz, J., Lanusse, F., Malz, A.I., Wright, A.H., Slosar, A., Abolfathi, B., *et al.* (2021) The LSST-DESC 3x2pt Tomography Optimization Challenge. *The Open Journal of Astrophysics*, **4**, 30. https://doi.org/10.21105/astro.2108.13418

[62] Sheykhmousa, M., Mahdianpari, M., Ghanbari, H., Mohammadimanesh, F., Ghamisi, P. and Homayouni, S. (2020) Support Vector Machine versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **13**, 6308-6325. https://doi.org/10.1109/jstars.2020.3026724

[63] Alabede, T.E., Bunakiye, J.R., Ishima, M.D. and Abdulazeez, A.O. (2025) A Machine Learning Approach to Predicting High Blood Pressure Using Predictive Modeling on Local and Global Datasets to Enhance Patient Safety. *International Journal of Applied Information Systems*, **13**, 72-85. https://doi.org/10.5120/ijais2025452036

[64] Sims, D.A. and Gamon, J.A. (2002) Relationships between Leaf Pigment Content and Spectral Reflectance across a Wide Range of Species, Leaf Structures and Developmental Stages. *Remote Sensing of Environment*, **81**, 337-354. https://doi.org/10.1016/s0034-4257(02)00010-x

[65] Ashourloo, D., Mobasheri, M.R. and Huete, A. (2014) Evaluating the Effect of Different Wheat Rust Disease Symptoms on Vegetation Indices Using Hyperspectral Measurements. *Remote Sensing*, **6**, 5107-5123. https://doi.org/10.3390/rs6065107

[66] Song, G. and Wang, Q. (2022) Developing Hyperspectral Indices for Assessing Seasonal Variations in the Ratio of Chlorophyll to Carotenoid in Deciduous Forests. *Remote Sensing*, **14**, Article 1324. https://doi.org/10.3390/rs14061324