AgriEngineering

MDPI

Article

Remote Monitoring of Coffee Leaf Miner Infestation Using Machine Learning

Emerson Ferreira Vilela [1,*], Gabriel Dumbá Monteiro de Castro [2], Diego Bedin Marin [1], Charles Cardoso Santana [3], Daniel Henrique Leite [2], Christiano de Sousa Machado Matos [1], Cileimar Aparecida da Silva [1], Iza Paula de Carvalho Lopes [1], Daniel Marçal de Queiroz [2], Rogério Antonio Silva [1], Giuseppe Rossi [4], Gianluca Bambi [4], Leonardo Conti [4] and Madelaine Venzon [1,*]

1 Minas Gerais Agricultural Research Agency (EPAMIG-Sudeste), Viçosa 36570-000, MG, Brazil
2 Department of Agricultural Engineering, Federal University of Viçosa, Viçosa 36571-900, MG, Brazil
3 Minas Gerais Agricultural Research Agency (EPAMIG), Pitangui Institute of Agricultural Technology (ITAP), Pitangui 35650-000, MG, Brazil
4 Department of Agriculture, Food, Environment and Forestry, University of Florence, 50121 Florence, Italy
* Correspondence: efvilela@yahoo.com.br (E.F.V.); madelaine@epamig.br (M.V.)

Abstract: The coffee leaf miner (*Leucoptera coffeella*) is a key pest in coffee-producing regions in Brazil. The objective of this work was to evaluate the potential of machine learning algorithms to identify coffee leaf miner infestation by considering the assessment period and Sentinel-2 satellite images generated on the Google Earth Engine platform. Coffee leaf miner infestation in the field was measured monthly from 2019 to 2023. Images were selected from the Sentinel-2 satellite to determine 13 vegetative indices. The selection of images and calculations of the vegetation indices were carried out using the Google Earth Engine platform. A database was generated with information on coffee leaf miner infestation, vegetation indices, and assessment times. The database was separated into training data and testing data. Nine machine learning algorithms were used, including Linear Discriminant Analysis, Random Forest, Support Vector Machine, k-nearest neighbors, and Logistic Regression, and a principal component analysis was conducted for each algorithm. After optimizing the hyperparameters, the testing data were used to validate the model. The best model to estimate miner infestation was RF, which had an accuracy of 0.86, a kappa index of 0.64, and a precision of 0.87. The developed models were capable of monitoring coffee leaf miner infestation.

Keywords: Google Earth Engine; *Leucoptera coffeella*; artificial intelligence; multispectral image analysis

Citation: Vilela, E.F.; Castro, G.D.M.d.; Marin, D.B.; Santana, C.C.; Leite, D.H.; Matos, C.d.S.M.; Silva, C.A.d.; Lopes, I.P.d.C.; Queiroz, D.M.d.; Silva, R.A.; et al. Remote Monitoring of Coffee Leaf Miner Infestation Using Machine Learning. *AgriEngineering* **2024**, *6*, 1697–1711. https://doi.org/10.3390/agriengineering6020098

Academic Editor: Murali Krishna Gumma

Received: 17 April 2024
Revised: 28 May 2024
Accepted: 4 June 2024
Published: 13 June 2024

1. Introduction

Brazil is the largest world producer of coffee, and Minas Gerais is the state with the highest production in the country. Therefore, the crop holds significant social and economic importance for Brazil, as well as for Minas Gerais. However, there are several factors threatening the sustainability of coffee production, among which are pest attacks [1].

The coffee leaf miner (*Leucoptera coffeella*) (Guérin-Mèneville and Perrottet) (Lepidoptera: Lyonetiidae) (CLM) is a coffee-exclusive pest, and it is considered a key pest due to the extensive damage it causes in coffee plantations [1–3]. The CLM is present throughout Latin America, causing millions of dollars in losses annually [4]. The damage caused by the CLM results from injuries inflicted by its larvae, which feed on the palisade parenchyma of coffee leaves, leading to the formation of mines that later necrotize. This can cause defoliation and, in more extreme cases, lead to the death of the plant [1,5].

Due to the damage caused by this pest, efforts are being made to monitor the coffee leaf miner (CLM). Considering the specialized labor and time required for visual inspections in pest detection, research has been focusing on developing detection methods through remote sensing. In this context, recent studies have sought to understand the relationships of alteration in the electromagnetic reflectance of coffee plants when infested by the CLM.

Vegetation indices are mostly used for analysis. In this pretext, the authors of [6] used vegetation indices obtained from a multispectral camera mounted on a remotely piloted aircraft to differentiate and detect areas of coffee infested by the CLM. The results of this study show that healthy leaves exhibit higher vegetation index values compared to infected leaves. The Green–Red Normalized Difference Vegetation Index (GRNDVI) was the most effective factor in differentiating between infected and healthy leaves, with average values of 0.32 for healthy leaves and 0.06 for infected leaves. In another study [7], the authors developed a specific vegetation index to estimate coffee leaf miner infestation using Sentinel-2 multispectral images called the Coffee-Leaf-Miner Index (CLMI). The authors employed machine learning (ML) techniques to recognize infestation patterns in coffee plantations. Models based on Random Forest (RF) and Support Vector Machine (SVM) were trained using CLMI and provided an effective approach to detect the coffee leaf miner in coffee plantations [7]. The use of aerial images/photos and artificial intelligence techniques has been shown to be a global trend in modern agriculture, and it is also used to monitor different pests and diseases in coffee plantations [8–10].

However, despite the preceding studies, little has been deeply explored on the topic. Currently, there is a lack of a single monitoring model that identifies and relates the most relevant vegetation indices in CLM detection, and defines, among different machine learning algorithms, the most efficient one for this purpose. Furthermore, for farmers to properly manage CLM, continuous monitoring methods should be chosen to indicate the need for control measures. Taking this into consideration, studies with orbital images obtained from satellites become the most viable alternative for large-scale CLM monitoring.

However, obtaining a historical series of orbital images can be challenging, as the continuous acquisition of orbital data over time is often limited by operational, logistical, cost, and availability issues. These limitations, in addition to potentially affecting long-term analysis, can hinder/mask the understanding of changes in different coffee ecosystems. In this context, Google Earth Engine (GEE) has proven to be a promising multidisciplinary tool with multiple applications as it provides easy and quick access to a vast catalog of aerial images and geospatial datasets for any location on the planet, surpassing traditional barriers to data acquisition [11,12]. Thus, GEE can become an important tool for researchers and farmers, enabling more comprehensive and efficient analyses in monitoring and evaluating agricultural variables, such as CLM infestation in coffee plantations.

In agriculture, GEE has been used to monitor and manage land use [13], map coffee plantations [14], and classify land cover [15]. Therefore, it is possible to observe that the integration of large geospatial databases enabled by GEE brings with it the advantage of enabling the development of robust and applicable methodologies, driving the advancement of scientific and technological research in agriculture. Therefore, it is important to study the application of platforms that enable quick and easy access to a historical series of aerial images for monitoring the CLM on a regional scale.

The need for more comprehensive studies for the application of vegetation indices in conjunction with ML algorithms to monitor CLM infestation is evident since the complex dynamics of this pest require a continuous approach capable of providing information over time and in different regional contexts. The temporal limitation of previously conducted studies highlights the need for more extensive investigations, which enable, in addition to accurate detection, a more comprehensive understanding of seasonal variations and the evolution of CLM infestation in the field. The hypothesis is that machine learning-based models, using spectral data (vegetation indices), can accurately identify CLM infestation in coffee plantations. However, the appropriate selection and validation of algorithms require further in-depth investigations. Thus, this study aimed to evaluate the potential of different ML algorithms to identify CLM infestation by considering data from a four-year historical series of Sentinel-2 satellite images acquired through the Google Earth Engine platform. This research sought to determine a more effective and comprehensive approach to monitor and manage this crucial pest affecting coffee production in Brazil.

2. Materials and Methods

To assess the potential of different machine learning algorithms in identifying CLM infestation using data from a four-year historical series of Sentinel-2 satellite images acquired through the Google Earth Engine platform, this study was conducted in four stages, as demonstrated in the flowchart in Figure 1. In the first stage, CLM infestations were collected/recorded monthly over a period of 4 years. In the second stage, multispectral images were collected, and vegetation indices were obtained. The third stage involved the development and selection of variables and the optimization of hyperparameters of the machine learning models. Finally, the fourth stage comprised the evaluation of the algorithm's performance.

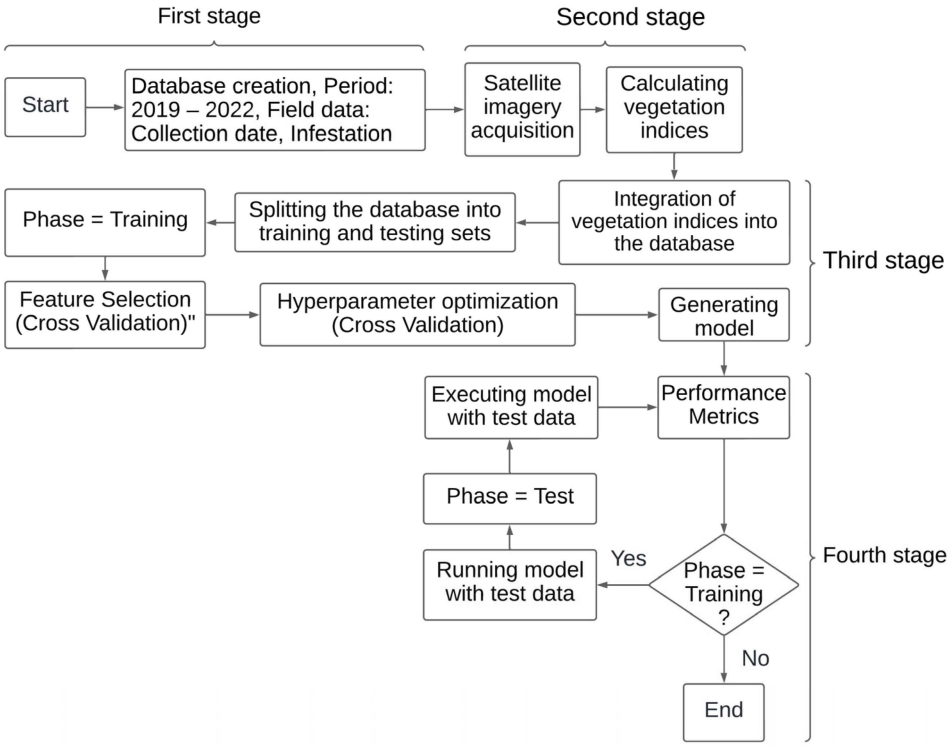


Figure 1. A flowchart of the study stages.

Four experimental fields from the Minas Gerais Agricultural Research Agency (Epamig) located in the municipalities of Três Pontas (CETP), São Sebastião do Paraíso (CESP), Machado (CEMA), and Patrocínio (CETP) in the state of Minas Gerais were selected (Table 1, Figure 2).

Table 1. The coffee cultivars and edaphoclimatic characteristics of the experimental research stations.

Characteristics	CETP	CESP	CEMA	CEPC
Cultivar	Mundo Novo	Catuaí 99	Catuaí 99	Rubi
Spacing (m)	2.30 × 0.70	3.20 × 0.70	2.30 × 0.70	3.50 × 0.70
Planting (year)	1999	2000	1999	1996
Elevation (m)	916	880	970	997
Latitude (S)	21°20′37.014″	20°54′42.023″	21°40′50.848″	18°59′28.284″
Longitude (W)	45°28′59.452″	47°7′20.341″	45°56′38.069″	46°59′21.700″

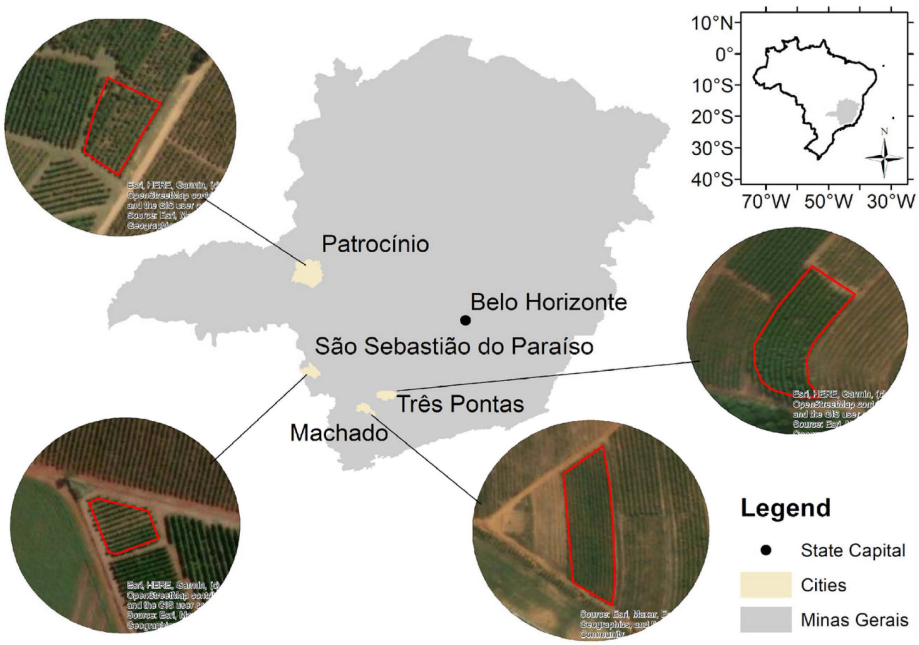


Figure 2. Epamig experimental fields: Três Pontas (CETP), São Sebastião do Paraíso (CESP), Machado (CEMA), and Patrocínio (CEPC). Source: Esri, Maxar, Earthstar Geographics, and GIS User Community.

2.1. Pest Monitoring

CLM infestation was monitored monthly from January 2019 to December 2022, with the data collection date recorded. To monitor the CLM, leaves from the 3rd or 4th pair of two branches on opposite sides, located in the middle third of 50 plants per plot, were evaluated randomly. Counting was only performed on leaves with intact mines.

In order to simplify the model development, the CLM infestation data were categorized into two classes: "infested," which denotes when the infestation record in the experimental field was greater than or equal to 10%, and "healthy", which denotes when the infestation record in the experimental field was less than 10%. The threshold value of 10% was determined based on the work in [2], which indicated that CLM infestation percentages exceeding 10% already indicated productivity losses due to pest infestation. This simplification aimed to make subsequent analyses more objective by transforming the CLM monitoring proposal in the model into a binary classification through the identification of infested areas ($\geq$10% infestation) or healthy areas (<10% infestation).

2.2. Spectral Data Acquisition

The selection of multispectral images from the Sentinel-2 Satellite was carried out monthly from January 2019 to December 2022, corresponding to the dates of the CLM measurements. A maximum interval of 12 days between the evaluation date and the satellite image date was adopted to ensure cloud-free conditions in the area. The acquisition of satellite images for each experimental field, on the date closest to the field infestation record, was carried out by observing the dates of available cloud-free images. Months in which it was not possible to select a satellite image within a 12-day interval after field data collection were excluded from the study. A total of 143 images from the study areas were selected for analysis. From the selected images, 13 vegetation indices were generated (Table 2). The Google Earth Engine platform was used for image selection, and vegetation index calculations were performed through the JavaScript programming language. The vegetation indices were calculated from the mean reflectance in each experimental plot. The delimited areas of the experimental fields were sufficient to meet the proposed analyses in the study. The vegetation indices were calculated from the mean reflectance in each experimental plot, as well as the mean infestation of the plot.

Table 2. Vegetation indices applicable in agricultural areas.

Vegetation Index	Equation	Remarks	Reference
Normalized difference vegetation index (NDVI)	$\frac{NIR-RED}{NIR+RED}$	It measures the greenness and the density of the vegetation	[16]
Enhanced vegetation index (EVI)	$2.5 * \frac{NIR-RED}{NIR+6*RED-7.5*BLUE+1}$	It measures the greenness and accounts for atmospheric influences and the vegetation background signal, and it is sensitive in areas with dense vegetation	[17]
Simple ratio (SR)	$\frac{NIR}{RED}$	It is sensitive to the vegetation presence	[18]
Green normalized difference vegetation index (GNDVI)	$\frac{NIR-GREEN}{NIR+GREEN}$	It is sensitive to the detected chlorophyll concentration	[19]
Infrared percentage vegetation index (IPVI)	$\frac{NIR}{NIR+RED}$	It can be used as an indicator of the photosynthetic activity of the canopy cover and the total chlorophyll content in the leaves	[20]
Modified chlorophyll absorption ratio index (MCARI)	$(RE1 - RED) - (0.2 * (RE1 - GREEN)) * \frac{RE1}{RED}$	It is used to measure chlorophyll concentrations, including variations in the leaf area index	[21]
MERIS terrestrial chlorophyll index (MTCI)	$\frac{NIR-RE1}{RE1-RED}$	It is sensitive to high values of chlorophyll content	[22]
Red edge inflection point (REIP)	$700 + 35 * \frac{\frac{NIR+RED}{2}-RE1}{RE2-RE1}$	It can be used as an indicator of vegetation stress	[23]
Plant senescence reflectance index (PSRI1)	$\frac{RED-GREEN}{RE1}$	It has been proposed to determine the stage of leaf senescence and fruit ripening	[24]
Soil-adjusted vegetation index (SAVI)	$\frac{(1+L)*(NIR-RED)}{(NIR+RED+L)}$	It is a vegetation index that minimizes the effect of the soil's brightness	[25]
Coffee-leaf-miner index (CLMI)	$\frac{(842-490)*(RED-BLUE)-(665-490)*(NIR-BLUE)}{2}$	It was developed to detect coffee leaf miner infestation	[7]
Normalized difference moisture index (NDMI)	$\frac{NIR-SWIR}{NIR+SWIR}$	It is an indicator of water stress in crops	[26]
Normalized difference water index (NDWI)	$\frac{GREEN-NIR}{GREEN+NIR}$	It is an indicator that measures the moisture content	[27]

Near Infrared (NIR); Red-edge 1 (Re1); Red-edge 2 (Re2); Shortwave Infrared (SWIR).

2.3. Model Development

The model development stage of machine learning involved developing and selecting variables and optimizing hyperparameters.

2.3.1. Machine Learning (ML)

After assembling the database comprising infestation data, 13 vegetation indices, the month in which the infestation was measured in the field, and the number of days between the date the data were measured and the date of satellite image generation, various machine learning algorithms were employed. The following algorithms were included: Random Forest (RF), Support Vector Machine (SVM), k-nearest neighbors (KNN), Logistic Regression (LR), and RF with Linear Discriminant Analysis (LDA). These algorithms were also evaluated using a principal component analysis (PCA) to reduce the variable set to just two dimensions. In the end, nine machine learning algorithms were used (RF, RF-PCA, SVM, SVM-PCA, KNN, KNN-PCA, LR, LR-PCA, and RF-LDA). The implementation was carried out in the Python language (version 3.9) utilizing libraries such as Numpy [28], Pandas [29], and SciKit learn [30].

2.3.2. Variable Selection and Hyperparameter Optimization

Variable selection and hyperparameter optimization were conducted for each ML model after the database was split into training data (data from January 2019 to December 2021) and testing data (data from January 2022 to December 2022), as illustrated in Figure 1. The training data accounted for 76%, while the testing data accounted for 24% of the total. During the training stage, variable selection for use in the model was performed, and the most parsimonious algorithms were used. Accuracy was used as the selection criterion by employing cross-validation with K folds equal to 5 (Figure 1). Hyperparameter optimization was carried out using cross-validation with K folds equal to 5 (Table 3). The definitions of hyperparameters and values were based on the SciKit-learn library [30].

Table 3. Optimized hyperparameters in machine learning algorithms.

Algorithm	Hyperparameter	Valor	Algorithm	Hyperparameter	Valor
RF	number of trees	91	RF-PCA	Number of trees	171
	Criterion	Gini		Criterion	gini
	Maximum depth	9		Maximum depth	18
SVM	C	1	SVM-PCA	C	1
	Kernel	Poly		Kernel	Rbf
	Degree	3		Degree	3
	Gamma	Auto		Gamma	scale
LR	Tol	6	LR-PCA	Tol	3
	C	4		C	3
	Max iter	172		Max iter	163
	Intercept scaling	8		Intercept scaling	5
KNN	N neighbors	29	KNN-PCA	N neighbors	8
	Leaf size	10		Leaf size	37
RF-LDA	Number of trees	79			
	Criterion	Gini			
	Maximum depth	14			

2.4. Model Performance Evaluation

The performance of the CLM monitoring models was evaluated based on the main performance metrics used in machine learning algorithms, including confusion matrix, overall accuracy, precision, specificity, recall, and F1, according to Equations (1)–(5):

$$\text{Accuracy} = (TN + TP)/(TN + FP + FN + TP) \quad (1)$$

$$\text{Specificity} = TN/(TN + FP) \quad (2)$$

$$F1 = 2\ (\text{precision} \times \text{recall})/(\text{precision} + \text{recall}) \quad (3)$$

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

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

where True Negative (TN) represents data that were predicted and observed as healthy; True Positive (TP) represents data that were predicted and observed as infested; False Negative (FN) represents data that were predicted as healthy but observed in the field as infested; and False Positive (FP) represents data that were predicted as infested but observed in the field as healthy.

3. Results

3.1. Monitoring of Coffee Leaf Miner Infestation and Image Selection

A descriptive analysis of CLM infestation monitoring in the experimental plots in Três Pontas, São Sebastião do Paraíso, Machado, and Patrocínio from 2019 to 2022 is presented in Table 4.

Table 4. An annual descriptive analysis of coffee leaf miner infestation in the experimental plots in Três Pontas, São Sebastião do Paraíso, Machado, and Patrocínio.

Year	Minimum	Q1	Median	Mean	Q3	Maximum	Se
			%				
2019	0.000	0.500	1.500	4.417	4.750	30.00	1.33
2020	0.000	1.25	8.00	13.25	20.00	76.00	2.59
2021	0.000	0.500	3.00	7.829	9.00	46.00	1.74
2022	0.000	1.00	4.50	10.75	17.50	41.50	2.23

Q1: Quartile 1; Q3: Quartile 3; Se: mean standard error.

The year with the highest CLM infestation was 2020, reaching a maximum value of 76%, while 2019 had the lowest infestation with a maximum of 30% of CLM. The years 2021 and 2022 reached maximum infestation levels of 46% and 41.5%, respectively. The annual average between 2019 and 2022 ranged from 4.41% to 10.75%. The pattern of CLM infestation data was similar in 2021 and 2022, showing similar quantities of points that were considered infested. However, in 2019, there were significantly fewer points with infestation compared to the other years (Table 4).

A Descriptive analysis using measures such as minimum, maximum, quartiles, and the mean standard error provided a comprehensive overview of the distribution of CLM infestation data in the areas over the evaluated period. The minimum and maximum values provided information about the variability of CLM occurrence in the areas. A minimum of 0% was observed in all analyzed years, indicating areas without infestation. However, the maximum values ranged from 30% to 76% during the analyzed period, with the highest being in 2020 at 76%, reflecting significant infestation in at least one of the areas.

The quartiles helped us to understand the distribution of infestation, where Q1 and Q3 evidenced the dispersion of infestation in the areas. In 2020, Q1 was 1.25%, suggesting that 25% of the dataset had low infestation, while Q3 was 20%, indicating that 75% of the dataset had infestation below 20%. In the other years, the maximum values were lower, indicating relative stability. For the mean and median, which are measures of central tendency, it was observed that the mean varied from 4.41% in 2019 to 10.75% in 2022, suggesting a trend of increasing infestation over the years. The median, which is less sensitive to extreme values, remained relatively stable. The mean error reflected the dispersion of data around the mean, showing that moderate values indicate consistency in results over the years.

Following the infestation monitoring, satellite images were selected for vegetation index calculations and subsequently used for identifying infested and non-infested areas (Table 5).

Table 5. A summary of the classification performed on the satellite image database for Três Pontas, São Sebastião, Machado, and Patrocínio.

Infestation	2019	2020	2021	2022	Total
Yes	3	17	10	11	41
No	27	21	31	23	102
Total	30	38	41	34	143

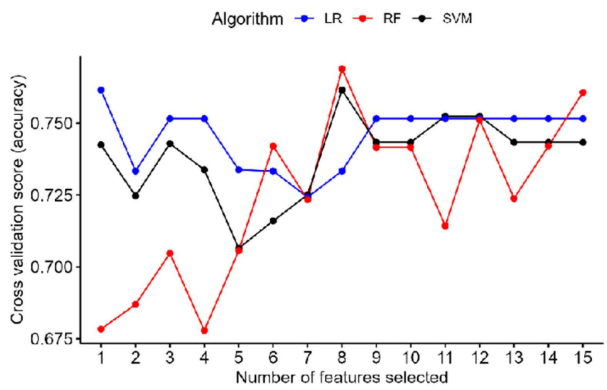
Based on Table 5, it is possible to observe that the year 2021 had the highest number of available images, with 41 records, followed by 2020 and 2022, presenting about 1/3 of the total data framed in the infested class, also known as "Positive". At the end of the period, a database was generated with data on vegetation indices, CLM monitoring, and sampling dates totaling 2304 entries.

3.2. Machine Learning

For the selection of the number of variables in the machine learning algorithms, greater precision and a smaller number of variables were taken into account. For RF and SVM, eight variables were selected for each algorithm. For LR, only one variable was selected (Figure 3A). The most important variables, selected in at least two algorithms, were CLMI, EVI, SAVI, SR, NDVI, and PSRI1 (Figure 3B).

The most important variable for the Logistic Regression model was the CLMI index (100%). For the Random Forest model, three indices stood out, namely the NDVI (18.1%), EVI (16.4%), and NDMI (15.2%), representing approximately 50% of the model's importance. For the Support Vector Machine model, two indices had values exceeding 20%: PSRI1 (33.3%) and EVI (24.2%). For this model, the attributes with the least importance in classification were month (1.52%), SR (4.55%), IPVI (7.58%), NDVI (7.58%), and SAVI (7.58%).

A)



B)

	Feature importance (%)		
Features	LR	RF	SVM
CLMI	100	13.0	-
EVI	-	16.4	24.2
SAVI	-	12.5	7.58
SR	-	13.0	4.55
NDVI	-	18.1	7.58
PSRI1	-	9.03	33.3
NDMI	-	15.2	-
MCARI	-	2.85	-
Days	-	-	16.7
Month	-	-	1.52
IPVI	-	-	7.58
NDWI	-	-	-
MTCI	-	-	-
REIP	-	-	-
GNDVI	-	-	-

Figure 3. Selection of most important variables for estimating coffee leaf miner infestation in coffee plantations. (**A**) Determination of number of variables. Lines represent average of 5-fold cross-validation. (**B**) Variable importance in percentage, with recursive feature elimination. Dashes indicate variables not selected in each model.

Variable selection is important because it reduces the number of input variables, returning the most individually influential variables in a ranking. This simplifies the model, resulting in a shorter computational processing time.

Algorithm Performance

The SVM, LR, and RF algorithms were efficient in discriminating healthy plants with up to 95.83% accuracy. However, these algorithms underestimated CLM infestation in coffee plantations, achieving a maximum accuracy of 63.64% (Figure 4). When a PCA was used with the algorithms, there was improved performance in discriminating coffee plantations with CLM infestation, with accuracy rates ranging from 70% (SVM-PCA) to 90% (KNN-PCA) (Figure 4).

The models based on PCA exhibited a higher number of correct identifications of areas with CLM infestation (True Positives), although they also displayed a higher number of False Positive (FP) errors, where areas were predicted to be infested but observed in the field as healthy, reaching 62.5% for the KNN-PCA model. On the other hand, models that did not utilize PCA were more proficient at identifying True Negative (TN) areas, where the class was predicted and observed as healthy, reaching a 95.83% accuracy for the SVM, LR, and RF models. However, these models also showed a higher False Negative (FN) error rate, where areas were predicted to be healthy but observed in the field as infested.

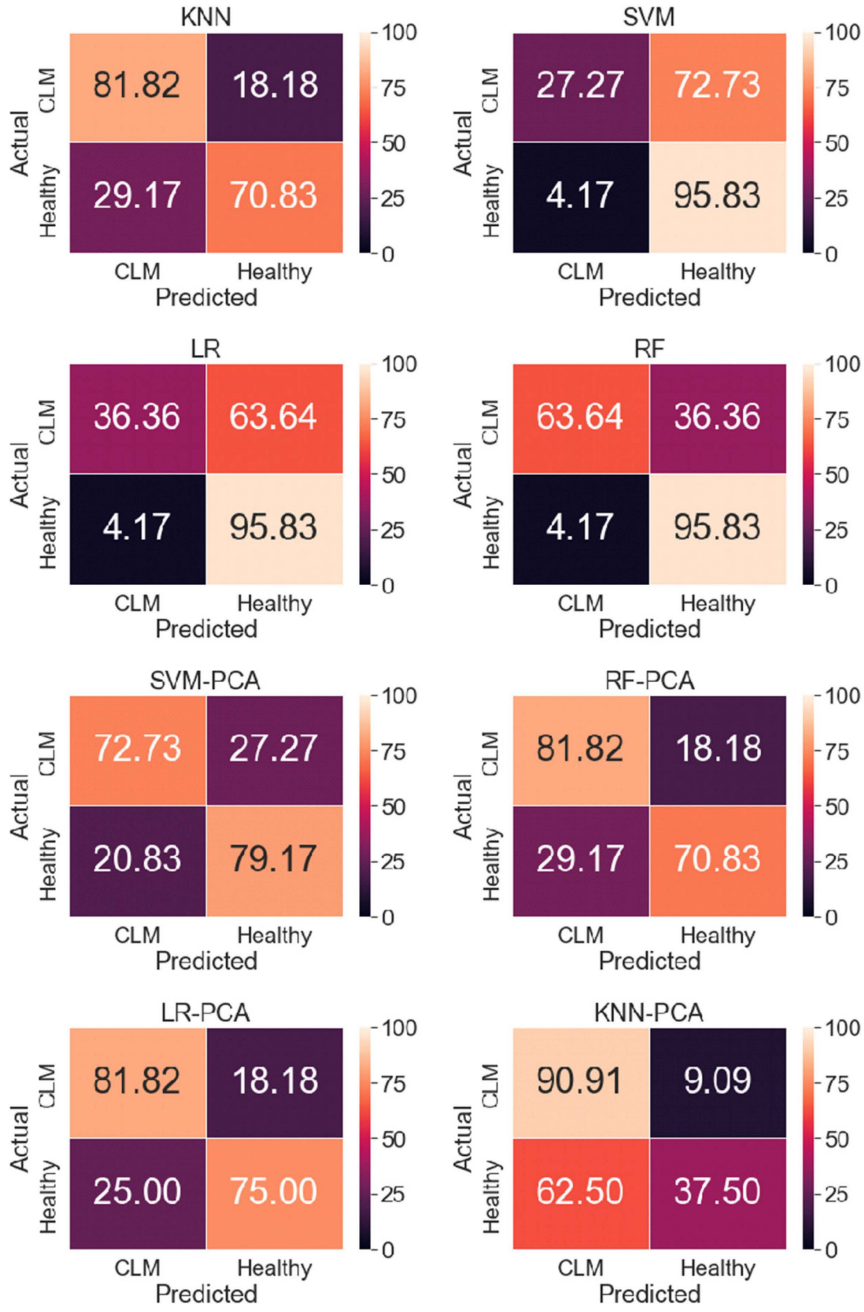


Figure 4. The confusion matrices of machine learning algorithms to predict coffee leaf miner infestation (CLM—coffee plants with leaf miner infestation; healthy—plants without leaf miner infestation). The applied algorithms were Random Forest (RF), k-nearest neighbors (KNN), Support Vector Machine (SVM), Logistic Regression (LR), and PCA. The confusion matrices indicate the numbers of correct and incorrect predictions for each class as percentages.

The accuracy ranged from 0.54 to 0.86. The model that performed best was Random Forest with an accuracy of 0.86, a kappa index of 0.64, and a precision of 0.87 (Figure 5). The model with the lowest performance was the KNN with PCA, with an accuracy of 0.54, a kappa index of 0.21, and a precision of 0.4.

The models that utilized PCA tended to have lower values in performance metrics when compared to their respective models without PCA. Among the models that used PCA, the Random Forest and Logistic Regression algorithms were the ones that exhibited the best performances.

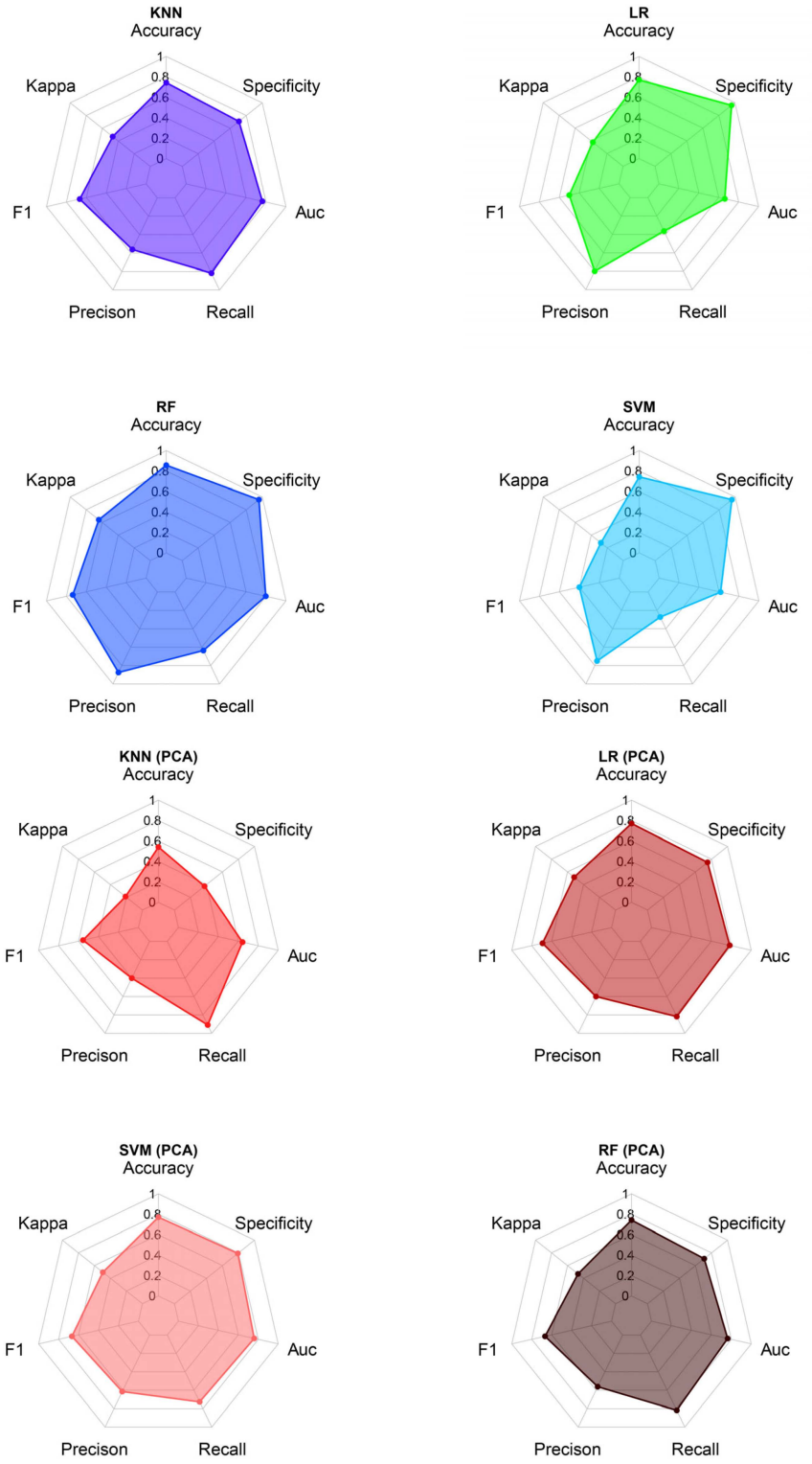


Figure 5. Validation of performance metrics (precision, kappa, AUC, recall, and F1 score) of machine learning algorithms: Random Forest (RF), Logistic Regression (LR), k-nearest neighbors (KNN), and Support Vector Machine (SVM). These algorithms were tested to predict coffee leaf miner infestation based on vegetation indices generated by satellite images.

With the data containing all variables, a PCA was performed, and the first two principal components explained 80.2% of the total variance in the data. The distinct separation between infested and non-infested coffee plantations became evident in the PCA scores

(Figure 6). However, the algorithms did not define the decision boundaries well between the classes, especially the KNN algorithm (Figure 6).

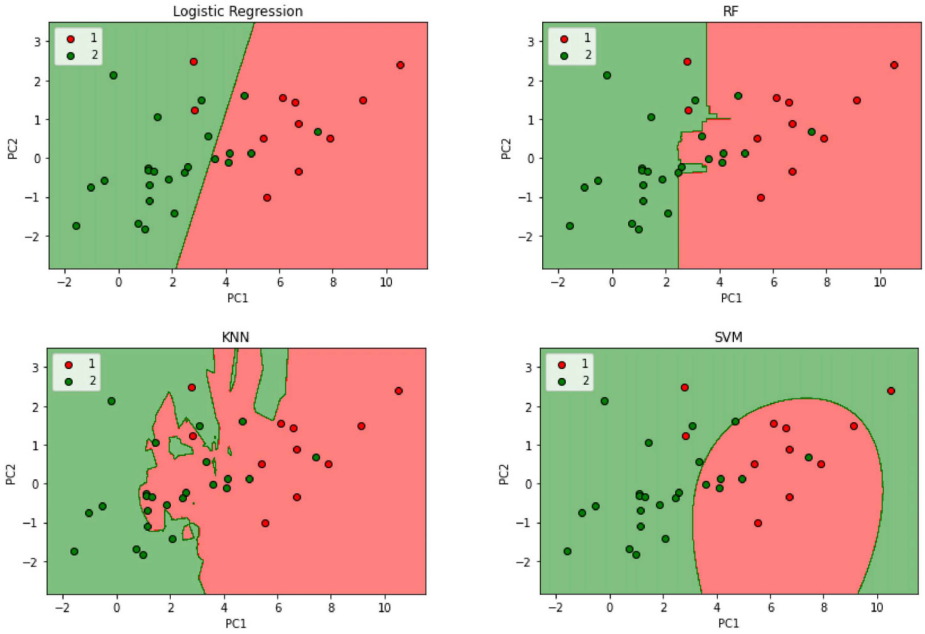


Figure 6. Decision boundaries of machine learning algorithms using principal component analysis (PCA). PCA was applied to vegetation index data, month, and days between assessment and satellite image, with principal components (PC1 and PC2) explaining 69.5% and 10.7% of total data variation, respectively. Applied algorithms were Random Forest (RF), k-nearest neighbors (KNN), Support Vector Machine (SVM), and Logistic Regression. Numbers 1 and 2 indicate coffee with coffee leaf miner infestation (red) and coffee without coffee leaf miner infestation (green), respectively.

The algorithms that presented a better decision limit between classes were the SVM and LR algorithms. It can be seen that it is possible to separate infested areas from non-infested areas, although the models need to be adjusted.

4. Discussion

4.1. Monitoring the Coffee Leaf Miner Infestation and Image Selection

The infestation range varied from a minimum of 0% to maximum values ranging from 30% to 76% during the analyzed period, with the highest occurring in 2020 at 76%. The periods with the highest pest populations are during dry seasons [1]. In large coffee plantations in regions with hot and dry climates, it is advisable to initiate control measures when 20% or even 10% of leaves are infested [2].

For the mean and median, which are measures of central tendency, it was observed that the mean varied from 4.41% in 2019 to 10.75% in 2022, suggesting a trend of increasing infestation over the years. The median, which is less sensitive to extreme values [31], remained relatively stable.

The temporal variation in the CLM infestation levels highlights the complexity of factors influencing the prevalence of this pest [1]. The year 2020 stood out as critical as it had the highest infestation, indicating possible environmental and/or agronomic conditions favoring this increase. On the other hand, the year 2019 revealed the lowest infestation, demonstrating considerable annual variability in infestation conditions during the study period. Understanding these variations is crucial for implementing more targeted management practices and, consequently, more effective ones. The year 2020 can be considered a warning point for the need for specific interventions, while subsequent years indicate the importance of continued monitoring of and adaptations to agricultural practices.

Regarding the selection of satellite images for calculating the vegetation indices for the study period, the year 2021 stood out with the highest number of images (41). This number was higher than those in 2019 and 2022, demonstrating a more extensive temporal coverage for this period. The years 2020 and 2022 had relatively smaller numbers of images, indicating that during this period, there was less coverage compared to 2021. The availability of variables from images over the years directly influences the analysis of CLM infestation conditions, as a broader range of images can affect the accuracy and representativeness of models, considering that more extensive coverage over time tends to provide a more complete view of infestation fluctuations.

Table 3 shows that even with the variation in the number of images, there was a considerable number of periods when the areas were classified as being infested in all years. In this context, the developed models not only took into account the quantity of images but also the temporal representativeness to ensure that the results were the most robust, generalizable, and reliable in identifying infested and non-infested areas.

4.2. Machine Learning

The careful selection of the number of variables in ML algorithms plays a crucial role in the effectiveness and interpretability of the models. The adopted approach aimed to maximize the model's accuracy with the fewest possible variables. When analyzing Figure 3, interesting results are observed regarding the number and importance of selected variables. Regarding the choice of the number of variables, the RF and SVM algorithms were more parsimonious with the selection of eight variables each, demonstrating to be more efficient and less susceptible to overfitting. However, LR stood out by selecting only one variable, indicating an even more simplified approach, possibly based on the choice of the most informative variable.

When using a large number of variables, the algorithm may lose its learning generalization (overfitting) [32]. ML models can be simplified by selecting a limited number of variables and not only contribute to interpretability, but also to the reduction in the computational processing time. The processing time is a relevant factor in the agricultural context, considering that operational efficiency is crucial in intervention management.

Among the 15 considered variables, it can be highlighted that 6 variables (CLMI, EVI, SAVI, SR, NDVI, and PSRI1) were the most relevant for classifications. Each of these variables plays a specific role in characterizing coffee leaf miner infestation conditions. The CLMI was created to be a more sensitive index to identify coffee leaf miner infestation [7] and was the only index selected by the LR model. The CLMI possibly relates to the mean chlorophyll in the maximum entropy index, thus indicating that variations in the chlorophyll concentration are affected by the presence of CLM. Other indices, such as the NDVI, which is widely used to monitor agricultural crops as a standard indicator of vegetation health; the SAVI, which seeks to reduce soil influence; as well as EVI, which aims to optimize the vegetation signal [33], were other indices selected in at least two algorithms. All of these indices present NIR and RED bands in their compositions.

The ranking of variable importance (Figure 3B) highlights that certain variables, such as CLMI, EVI, and NDVI, are the most considered in the algorithms, and these vegetation indices play key roles in the classification process of areas that are infested or not infested by CLM.

Algorithm Performance

ML algorithms using a principal component analysis (PCA) achieved 90.91% True Positives, indicating that they were more sensitive to identifying plants infested with CLM; however, they showed a lower precision, meaning they had the highest number of errors in False Positive cases, where the model indicated the area as infested, but in reality, the area was not infested. These cases are less concerning, as classifying an area without infestation as infested can serve as an alert to the producer.

Recall was higher when ML and a PCA were used, indicating that among all observed infestation cases, the average prediction of these models was correct in 82%.

On the other hand, ML algorithms without PCA achieved the majority, namely 95.83% True Negatives, indicating that they were more sensitive to identifying plants without CLM infestation and showed higher precision. However, they presented lower recall, indicating that they made more errors in identifying plants infested with CLM; in other words, the model predicted the area as a non-infested area, but in reality, infestation was occurring in the area. This error would be more problematic, as it may deter the rural producer from going to the field to assess the area when, in reality, infestation with CLM is already occurring. These results indicate that ML-based approaches can detect patterns to identify CLM-infested plants that were not found in other analyses.

The Kappa value ranged from 0.21 to 0.64. Values of Kappa between 0.4 and 0.6 are considered good [34]. Cohen suggested interpreting the Kappa result as follows: 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement [35]. Accuracy (0.86) was higher in the RF algorithm, indicating that among all predictions the model made, considering areas with and without CLM infestation, the model was correct 86% of the time. The precision (0.88), Kappa index (0.64), and AUC (0.8) were also the highest for the RF model. Among the ML algorithms used to identify CLM infestation, RF was the algorithm that presented the best results among most of the metrics used.

5. Conclusions

The development of models, such as the one in this work, contributes substantially to the advancement of knowledge on the control and monitoring of pests and diseases of large crops around the world. The developed models were able to monitor CLM infestation in coffee areas, serving as a warning of CLM infestation on a regional scale. The model that showed the best results was RF. However, there are limitations to the use of these models, such as different crop management practices, soil variations, and the effects of other diseases and pest species, among other factors. This makes the detection of CLM through remote sensing complex. Therefore, further studies on the remote monitoring of CLM infestation in coffee plants that consider the inclusion of more variables and that take into account environmental factors should be carried out to improve the monitoring of CLM infestation.

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

Funding: This research was funded by "Fundação de Amparo à Pesquisa de Minas Gerais" (FAPEMIG), "Conselho Nacional de Desenvolvimento Científico e Tecnológico" (CNPq), and "Consórcio Brasileiro de Pesquisa e Desenvolvimento do Café" (CBP&D-Café).

Data Availability Statement: The original contributions presented in the study are included in the article; further inquiries can be directed to the corresponding authors.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Venzon, M. Agro-ecological Management of Coffee Pests in Brazil. *Front. Sustain. Food Syst.* **2021**, *5*, 721117. [CrossRef]
2. Reis, P.R.; Souza, J.C.; Silva, R.A.; Santa-Cecília, L.V.C. Principais pragas do cafeeiro no Cerrado Mineiro: Reconhecimento e manejo. In *Cafeicultura do Cerrado*; Carvalho, G.R., Ferreira, A.D., Andrade, V.T., Botelho, C.E., Carvalho, J.P.F., Eds.; EPAMIG: Belo Horizonte, Brazil, 2021; pp. 321–346.
3. Picanço Filho, M.C.; Lima, E.; Carmo, D.d.G.d.; Pallini, A.; Walerius, A.H.; da Silva, R.S.; Sant'Ana, L.C.d.S.; Lopes, P.H.Q.; Picanço, M.C. Economic Injury Levels and Economic Thresholds for *Leucoptera coffeella* as a Function of Insecticide Application Technology in Organic and Conventional Coffee (*Coffea arabica*), Farms. *Plants* **2024**, *13*, 585. [CrossRef] [PubMed]
4. Pantoja-Gomez, L.M.; Corrêa, A.S.; Oliveira, L.O.; Guedes, R.N.C. Common origin of Brazilian and Colombian popula-tions of the Neotropical coffee leaf miner, *Leucoptera coffeella* (Lepidoptera: Lyonetiidae). *J. Econ. Entomol.* **2019**, *112*, 924–931. [CrossRef] [PubMed]
5. Dantas, J.; Motta, I.O.; Vidal, L.A.; Nascimento, E.F.M.B.; Bilio, J.; Pupe, J.M.; Veiga, A.; Carvalho, C.; Lopes, R.B.; Rocha, T.L.; et al. A Comprehensive Review of the Coffee Leaf Miner *Leucoptera coffeella* (Lepidoptera: Lyonetiidae)—A Major Pest for the Coffee Crop in Brazil and Others Neotropical Countries. *Insects* **2021**, *12*, 1130. [CrossRef] [PubMed]
6. Santos, L.M.d.; Ferraz, G.A.e.S.; Marin, D.B.; Carvalho, M.A.d.F.; Dias, J.E.L.; Alecrim, A.d.O.; Silva, M.d.L.O.e. Vegetation Indices Applied to Suborbital Multispectral Images of Healthy Coffee and Coffee Infested with Coffee Leaf Miner. *AgriEngineering* **2022**, *4*, 311–319. [CrossRef]
7. Vilela, E.F.; Ferreira, W.P.M.; Castro, G.D.M.d.; Faria, A.L.R.d.; Leite, D.H.; Lima, I.A.; Matos, C.d.S.M.; Silva, R.A.; Venzon, M. New Spectral Index and Machine Learning Models for Detecting Coffee Leaf Miner Infestation Using Sentinel-2 Multispectral Imagery. *Agriculture* **2023**, *13*, 388. [CrossRef]
8. Pereira, F.V.; Martins, G.D.; Vieira, B.S.; de Assis, G.A.; Orlando, V.S.W. Multispectral Images for Monitoring the Physiological Parameters of Coffee Plants Under Different Treatments Against Nematodes. *Precis. Agric.* **2022**, *23*, 2312–2344. [CrossRef]
9. de Castro, G.D.M.; Vilela, E.F.; de Faria, A.L.R.; Silva, R.A.; Ferreira, W.P.M. New vegetation index for monitoring coffee rust using sentinel-2 multispectral imagery. *Coffee Sci.* **2023**, *18*, e182170. [CrossRef]
10. Velásquez, D.; Sánchez, A.; Sarmiento, S.; Toro, M.; Maiza, M.; Sierra, B. A Method for Detecting Coffee Leaf Rust through Wireless Sensor Networks, Remote Sensing, and Deep Learning: Case Study of the Caturra Variety in Colombia. *Appl. Sci.* **2020**, *10*, 697. [CrossRef]
11. Velastegui-Montoya, A.; Montalván-Burbano, N.; Carrión-Mero, P.; Rivera-Torres, H.; Sadeck, L.; Adami, M. Google Earth Engine: A Global Analysis and Future Trends. *Remote Sens.* **2023**, *15*, 3675. [CrossRef]
12. Amani, M.; Ghorbanian, A.; Ahmadi, S.A.; Kakooei, M.; Moghimi, A.; Mirmazloumi, S.M.; Moghaddam, S.H.A.; Mahdavi, S.; Ghahremanloo, M.; Parsian, S.; et al. Google Earth Engine Cloud Computing Platform for Remote Sensing Big Data Applications: A Comprehensive Review. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 5326–5350. [CrossRef]
13. Liu, L.; Xiao, X.; Qin, Y.; Wang, J.; Xu, X.; Hu, Y.; Qiao, Z. Mapping Cropping Intensity in China Using Time Series Landsat and Sentinel-2 Images and Google Earth Engine. *Remote Sens. Environ.* **2020**, *239*, 111624. [CrossRef]
14. Kelley, L.C.; Pitcher, L.; Bacon, C. Using Google Earth Engine to Map Complex Shade-Grown Coffee Landscapes in Northern Nicaragua. *Remote Sens.* **2018**, *10*, 952. [CrossRef]
15. Phan, T.N.; Kuch, V.; Lehnert, L.W. Land Cover Classification using Google Earth Engine and Random Forest Classifier—The Role of Image Composition. *Remote Sens.* **2020**, *12*, 2411. [CrossRef]
16. Rouse, J.W.; Haas, R.H.; Schell, J.A.; Deering, D.W. Monitoring Vegetation Systems in the Great Plains with ERTS. *NASA Spec. Publ.* **1974**, *351*, 309.
17. Justice, C.O.; Vermote, E.; Townshend, J.R.; Defries, R.; Roy, D.P.; Hall, D.K.; Salomonson, V.V.; Privette, J.L.; Riggs, G.; Barnsley, M.J.; et al. The Moderate Resolution Imaging Spectroradiometer (MODIS): Land Remote Sensing for Global Change Research. *IEEE Trans. Geosci. Remote Sens.* **1998**, *36*, 1228–1249. [CrossRef]
18. Jordan, C.F. Derivation of Leaf Area Index from Quality of Light on the Forest Floor. *Ecology* **1969**, *50*, 663–666. [CrossRef]
19. Gitelson, A.A.; Kaufman, Y.J.; Merzlyak, M.N. Use of a green channel in remote sensing of global vegetation from EOS-MODIS. *Remote Sens. Environ.* **1996**, *58*, 289–298. [CrossRef]
20. Crippen, R. Calculating the Vegetation Index Faster. *Remote Sens. Environ.* **1990**, *34*, 71–73. [CrossRef]
21. Daughtry, C.S.T.; Walthall, C.L.; Kim, M.S.; de Colstoun, E.B.; McMurtrey, J.E. Estimating Corn Leaf Chlorophyll Concentration from Leaf and Canopy Reflectance. *Remote Sens. Environ.* **2000**, *74*, 229–239. [CrossRef]
22. Dash, J.; Curran, P.J. The Meris Terrestrial Chlorophyll Index. *Int. J. Remote Sens.* **2004**, *25*, 5403–5413. [CrossRef]
23. Guyot, G.; Baret, F.; Major, D.J. High Spectral Resolution: Determination of Spectral Shifts between the Red and Infrared. *Int. Arch. Photogramm. Remote Sens.* **1988**, *11*, 750–760.
24. Fernández-Manso, A.; Fernández-Manso, O.; Quintano, C. Sentinel-2A Red-Edge Spectral Indices Suitability for Discriminating Burn Severity. *Int. J. Appl. Earth Obs. Geoinf.* **2016**, *50*, 170–175. [CrossRef]
25. Roujean, J.L.; Breon, F.M. Estimating PAR Absorbed by Vegetation from Bidirectional Reflectance Measurements. *Remote Sens. Environ.* **1995**, *51*, 375–384. [CrossRef]
26. McFeeters, S.K. The use of the Normalized Difference Water Index (NDWI) in the delineation of Open Water Features. *Int. J. Remote Sens.* **1996**, *17*, 1425–1432. [CrossRef]

27. Gao, B. NDWI—A Normalized Difference Water Index for Remote Sensing of Vegetation Liquid Water from Space. *Remote Sens. Environ.* **1996**, *58*, 257–266. [CrossRef]
28. Harris, C.R.; Millman, K.J.; Van der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; et al. Array Programming with Numpy. *Nature* **2020**, *585*, 357–362. [CrossRef] [PubMed]
29. McKinney, W. Data structures for statistical computing in Python. In Proceedings of the 9th Python in Science Conference (SciPy 2010), Austin, TX, USA, 28–30 June 2010; pp. 56–61. [CrossRef]
30. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
31. Frost, J. *Introduction to Statistics: An Intuitive Guide for Analyzing Data and Unlocking Discoveries*; Jim Publishing: Costa Mesa, CA, USA, 2019; Volume 2020, 256p.
32. Han, H.; Jiang, X. Overcome Support Vector Machine Diagnosis Overfitting. *Cancer Inf.* **2014**, *13*, 145–158. [CrossRef]
33. Ponzoni, F.J.; Shimabukuro, Y.E.; Kuplich, T.M. *Sensoriamento Remoto Aplicado ao Estudo da Vegetação*, 2nd ed.; Parêntese: São José dos Campos, Brazil, 2012; 160p.
34. Landis, J.R.; Koch, G.G. The Measurement of Observer Agrément for Categorical Data. *Biometrics* **1977**, *33*, 159–174. [CrossRef]
35. McHugh, M.L. Interrater Reliability: The Kappa Statistic. *Biochem. Med.* **2012**, *22*, 276–282. [CrossRef]