Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2021 doi:10.20944/preprints202102.0498.v1

Machine Learning-Based Approach to Predict Insect-Herbivory-Damage and Insect-Type Attack in Maize Plants Using Hyperspectral Data

Danielle Elis Garcia Furuya[1], Mayara Maezano Faita Pinheiro[1], Felipe David Georges Gomes[1], Wesley Nunes Gonçalvez[2], José Marcato Junior[2], Diego de Castro Rodrigues3[3], Maria Carolina Blassioli-Moraes[4], Mirian Fernandes Furtado Michereff[4], Miguel Borges[4], Raúl Alberto Alaumann[4], Ednaldo José Ferreira[5], Ana Paula Marques Ramos[1], Lucas Prado Osco[6,✉], and Lúcio André de Castro Jorge[5]

[1]Post-Graduate Program of Environment and Regional Development, University of Western São Paulo (UNOESTE), Rodovia Raposo Tavares, km 572-Limoeiro, 19067-175, Presidente Prudente, São Paulo, Brazil
[2]Faculty of Engineering, Architecture, and Urbanism and Geography, Federal University of Mato Grosso do Sul (UFMS), Avenida Costa e Silva, 79070-900, Campo Grande, Mato Grosso do Sul, Brazil
[3]Department of Computing, Federal Institute of Education of Tocantins (IFTO), Dianópolis, Tocantins, Brazil
[4]National Research Center of Development of Genetic Research and Biotechnology, Brazilian Agricultural Research Agency (EMBRAPA), Parque Estação Biológica W5 Asa Norte, 70770-917, Brasília, DF, Brazil
[5]National Research Center of Development of Agricultural Instrumentation, Brazilian Agricultural Research Agency (EMBRAPA), Rua XV de Novembro, 1452, 13560-970, São Carlos, São Paulo, Brazil
[6]Faculty of Engineering and Architecture and Urbanism, University of Western São Paulo (UNOESTE), Rodovia Raposo Tavares, km 572-Limoeiro, 19067-175, Presidente Prudente, São Paulo, Brazil

A strategy to reduce qualitative and quantitative losses in crop-yields refers to early and accurate detection of insect-damage caused in plants. Remote sensing systems like hyperspectral proximal sensors are a promising strategy for managing crops. In this aspect, machine learning predictions associated with clustering techniques may be an interesting approach mainly because of its robustness to evaluate high dimensional data. In this paper, we model the spectral response of insect-herbivory-damage in maize plants and propose an approach based on machine learning and a clustering method to predict whether the plant is herbivore-attacked or not using leaf reflectance measurements. We differentiate insect-type damage based on the spectral response and indicate the most contributive wavelengths to perform it. For this, we used a maize experiment in semi-field conditions. The maize plants were submitted to three different treatments: control (health plants); plants submitted to Spodoptera frugiperda herbivory-damage, and; plants submitted to Dichelops melacanthus herbivory-damage. The leaf spectral response of all plants (controlled and submitted to herbivory) was measured with a FieldSpec 3.0 Spectroradiometer from 350 to 2500 nm for eight consecutive days. We evaluated the performance of different learners like random forest (RF), support vector machine (SVM), extreme gradient boost (XGB), neural networks (MLP), and measured the impact of a day-by-day analysis into the prediction. We proposed a novel framework with a ranking strategy, based on the accuracy returned by predictions, and a clusterization method based on a self-organizing map (SOM) to identify important regions in the reflectance measurement. Our results indicated that the RF-based framework algorithm is the overall best learner to deal with this type of data. After the 5th day of analysis, the accuracy of the algorithm improved substantially. It separated the three treatments into different groups with an F-measure equal to 0.967, 0.917, and 0.881, respectively. We also verified that the most contributive spectral regions are situated in the near-infrared domain. We conclude that the proposed approach with machine learning methods is adequate to monitor herbivory-damage of S. frugiperda and stink bugs like Dichelops melacanthus in maize, differentiating the types of insect-attack early on. We also demonstrate that the framework proposed for the analysis of the most contributive wavelengths is suitable to highlight spectral regions of interest.

proximal hyperspectral sensing | precision agriculture | random forest.
Correspondence: *lucasosco@unoeste.br*

Introduction

One of the major factors that impact a country's economic development is its agronomic sector since it is responsible for, among others, raw material providing, employment generation, and both human and animal food production. Several issues can impact a crop yield rate like chemical fertilizer overutilization, presence of chemicals in water supply, uneven distribution of rainfall, soil fertility difference, and the attack of pests or diseases in plants (1). Plant diseases are described as some form of modification that hampers the normal processes in their healthy development (1). As such, not only the disease but also insect-damage occurrences significantly endangers agriculture around the world (2) being usually associated with huge economic losses. To illustrate this scenario, for 12 major maize-growing countries, insect-damage costs a total of 1–4 billion dollars in lost crops per year (3).

Brazil is the third-largest producer of maize cultivar in the world. The last crop season (2019/2020) represents a record production with 105 million tons approximately, resulting in an increase of 2.6 percent in relation to the previous one (4). In China, the second-largest producer of maize in the world, caterpillars that ravage crops are advancing across fields and threatening this nation's vast supply of maize (3). Africa, where the pest arrived in 2016, and southern Asia have also reported a recent outbreak of bugs, causing maize yield losses surpassing 50 percent (3). A strategy to minimize both qualitative and quantitative losses in crop yield refers to early and accurate detection of insect-damage caused in plants (5). However, the traditional approach for monitoring plants in the field is labor-intensive, being prone to be subjective, and generally shows low efficiency (2, 5).

Remote sensing systems are a promising strategy for managing crops because they can provide directly noncontact and spatially continuous monitoring of diseases and pests efficiently (2, 6). The principle of remote sensing science is that all targets (e.g. soil, vegetation, water, etc.) on the terrestrial surface reflect or emit electromagnetic energy in specific wavelengths owing to a difference in their chemical composition, inner physical structure, and surface properties (roughness) (7). In a hyperspectral context, this means measuring hundreds of narrow bands within the electromagnetic spectrum. In this regard, the spectroscopy area refers to the method of obtaining the hyperspectral characteristics of a target regarding radiation flux intensity emitted or reflected by its constituents at different wavelengths to provide a precise fingerprint of a target (e.g. a plant) (7). For the last decades, many studies have proved the potential of remote sensing systems in the precision agriculture area, mainly for plant disease detection (2, 5, 6, 8–17).

Methods for modeling insect-damage caused in plants can be divided into traditional statistical analysis to even innovative machine learning approaches (2). Machine learning may be an interesting approach mainly because of its robustness to evaluate high dimensional data such as hyperspectral collected from proximal sensing equipment. One study (10) investigated the potential of near-infrared hyperspectral (1000 to 1600 nm) images processed by linear discriminant analysis and quadratic discriminant analysis for the detection of insect-damaged wheat kernels and pointed out that methods correctly classified 85–100% healthy and insect-damaged wheat kernels. Another research (11) adopted hyperspectral (400 to 720nm) images processed with the stepwise discriminant analysis for the detection of external insect damage in jujube fruits, and the overall classification accuracy was about 97.0%. An investigation (14), applying the partial least squares discriminant analysis in hyperspectral (1100 to 1700 nm) images of the short-wave infrared region, was able to demonstrate, with an accuracy upper to 96%, aflatoxin contamination on corn kernels.

One laboratory research (15) measured the hyperspectral reflectance (350 to 2,500 nm) of symptomatic and asymptomatic rice leaves infected by four different diseases. Based on probabilistic neural network classifiers, it was concluded, with the mean overall accuracy upper to 91%, that symptomatic and asymptomatic rice leaves can be discriminated using hyperspectral reflectance measurements only. Another study (18) applied two machine learning algorithms, radial basis function (RBF) and K-nearest neighbor (KNN), in hyperspectral (400 to 1,000 nm) images for the detection of citrus canker in several disease development stages (i.e., asymptomatic, early, and late symptoms) on Sugar Belle leaves and immature (green) fruit, and the overall classification accuracy of both methods was higher than 94%. One recent study (19) developed decision-tree machine learning algorithms to predict the level of P. truncatus infestation and associated damage of maize grain in smallholder farmer stores. To P. truncatus population size prediction, the model performance was weak (r = 0.43) because of the complicated sampling and detection of the pest and eight-week long period between sampling events. To grain damage prediction, the model had a stronger correlation coefficient (r = 0.93) being considered a good estimator of damages in grain caused by insects. Another recent study (20) investigated several machine learning algorithms to predict the cotton leafworm (Spodoptera littoralis) plant infestation in the greenhouses and found that the XGBoost algorithm was the most effective algorithm achieving a prediction accuracy of 84%.

As mentioned, remote sensing systems often provide high-dimensional datasets. This is due to spectral, temporal, and spatial features of remote systems, which are merged into a data vector, and occasionally require the application of techniques for datasets reduction or clustering. To accomplish this demand, a clustering method like the Self-Organizing Map (SOM) is a promissory alternative. SOM can dimensionally organize complex data into clusters, according to their relationships, being a highly appropriate method to solve difficult high-dimensional and nonlinear problems such as feature extraction and image classification such as those acquired by remote sensing systems (21). A main feature of the SOM is to compose a nonlinear mapping of a high-dimensional input space to a typically 2-D grid of artificial neural units (22, 23). For that, SOM is based on an artificial neural network trained based on unsupervised learning, consisting of a two-layer, an input layer and an output layer known as the Kohonen layer (22, 23). The literature review presents many studies using the SOM architecture to attend different applications, including remote sensing and agriculture-related problems (21, 24–29).

Although the SOM method is in widespread use across several disciplines (23), there is still a lack of investigation to date in the hyperspectral remote sensing area, especially for indicating the most relevant spectral regions to identify insect-damage in crops, such as maize. Predicting insect-herbivory-damage in plants with spectral data is an actual and important practice since it can assist agricultural management in a rapid and in-situ manner. However, methods that appropriately model high dimensional data from hyperspectral proximal sensors need further evaluation, mainly when the ranking and SOM approach is applied. The ranking approach has been adopted, for example, to identify the individual contribution of each spectral information, collected by remote sensing system, included in a machine learning model to solve precision agriculture problems (30–32). The ranking method calculates the increased or decreased difference in the performance of the algorithm against the performance of a baseline method concerning a given variable, and this returns a metric score for the individual input variables,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2021 doi:10.20944/preprints202102.0498.v1

thus indicating the contribution of each index into the model. In this regard, we propose an approach based on machine learning to predict whether the plant is attacked or not by insects using leaf reflectance measurements. The results obtained here showed that the reflectance measures differentiate the herbivore-type of damage, i.e, differentiate the herbivory provoked by larvae of Spodoptera frugiperda, a chewing insect, from the herbivory provoked by the stink bug Dichelops melacanthus, a sucking feeding insect. In short, here we present:

1. The performance of different learners;
2. The impact of a day-by-day analysis into the prediction, and;
3. A framework to identify important spectral regions for this prediction using the ranking and SOM approach.

Materials and Method

The method (Figure 1) was divided into the following main phases: 1) proximal sensing data acquisition; collected from different maize plants during different days in-field conditions; 2) data process and organization; separated into multiple datasets to be evaluated by the models; 3) machine learning evaluation; used to indicate the more appropriate learner to predict the insect-damage in this type of data; 4) daily analysis comparison; implemented to determine the impact on an individual analysis of the overall best learner, predefined in the previous step; 5) ranking and clustering with SOM of the contribution of wavelengths to the models' performance; proposed to help to define the appropriate spectral regions to separate insect-damage from healthy plants and to differentiate the insect-type damage in maize plants.

Insects. Spodoptera frugiperda were maintained in separate environmental rooms at 27±1 °C, with 65±10% relative humidity and a 14 h photoperiod. S. frugiperda larvae were obtained from a laboratory colony maintained at Embrapa Genetic Resources and Biotechnology in Brasília, DF, Brazil. The larvae were reared in plastic containers on an artificial diet based on beans (Phaseolus vulgaris). Second instar larvae (33) were used in experiments and starved for 24 h before the experiment. Dichelops melacanthus individuals were obtained from a laboratory colony started from adults collected in soybean fields near Embrapa Genetic Resources and Biotechnology, Brasília, Brazil (15°47′0″S, 47°55′0″W). Stink bugs were reared in 8 L plastic containers on a diet of soybean seeds (cv Conquista), sunflower seeds (Helianthus annuus), raw peanuts (Arachis hypogaea), fresh green beans (Phaseolus vulgaris), and water. The food supply was renewed twice a week. To provide an oviposition substrate and shelter for the bugs, a 15 cm^2 piece of nylon mesh screen was placed inside the cages. They were kept in a controlled-environment room at L14: D10 photoperiod, 26 ± 0.3 °C and 70 ± 10% r.h.

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

Experimental Area and Data Acquisition. The semi-field experiments were conducted in an external area of our laboratory in Brasilia with natural light. The plants of maize received one of the following treatments: 1) undamaged plants (UDP) (did not receive the treatment), 2) two (2) 2nd instar larvae of S. frugiperda herbivory damaged plants (Sf-HDP) (N = X for each treatment) and 3) two (2) adult females of Dichelops melcanthus herbivory damaged plants (Dm-HDP). Reflectance data from plants under these three treatments were collected from 09 to 15 h. The data was acquired over 8 days, except for day 3 and 7, which were collected outside the 09 to 15 h interval, thus, not used in this study.
The spectral reflectance from the plants was collected with a compact, field-portable, and precision instrument with a spectral range of 3,50-2,500 nm, FieldSpec 3.0 spectroradiometer, at the daylight conditions, in a rapid data collection time of 0.1 seconds per spectrum. The sampling interval is 1.4 nm for the spectral region 3,50-1,000 nm and 2 nm for the spectral region 1,000-2,500 nm. It was used a small size of the pistol grip and 8 grades fore optics around 50 cm far from the samples and material with approximately 100% reflectance across the entire spectrum as a white reference panel or white reference standard. At each time of acquisition was calibrated the reflectance with the white standard.
The processed data, in reflectance value, was organized into separated subsets according to the required analysis, both detailed in the following subsections. Before the analysis, we reordered the spectral wavelengths into columns, to be used separately by the models. This assured that the wavelengths, from 350 to 2500 nm, were used as an input parameter for the algorithms. We then identified low signal-to-noise spectral regions, mostly related to atmospheric conditions and equipment interference, from 1,350 to 1,410 nm, 1,820 to 1,940 nm, and 2,460 to 2,500 nm. This resulted in 1,934 attributes to be incorporated into the analysis as input variables.

Insect Herbivory-Damage Classification. To determine the overall best learner to model the spectral configuration-sets, we choose 8 algorithms based upon their theoretical characteristics and state-of-the-art usage. The algorithms were: ExtraTree (ExT); k-Nearest Neighbour (kNN); Logistic Regression (LoR); Multi-Layer Perceptron (MLP); Naive Bayes (NB); Random Forest (RF); Support Vector Machine (SVM) and; Extreme Gradient Boost (XGB). How they were applied is described in the following subsection. However, in an experimental initial phase, we evaluated the individual performance of the algorithms to determine whether a fine-tuning of its parameters was necessary. Upon multiple comparisons of fine-tuning meta methods like Random Search,

 doi:10.20944/preprints202102.0498.v1

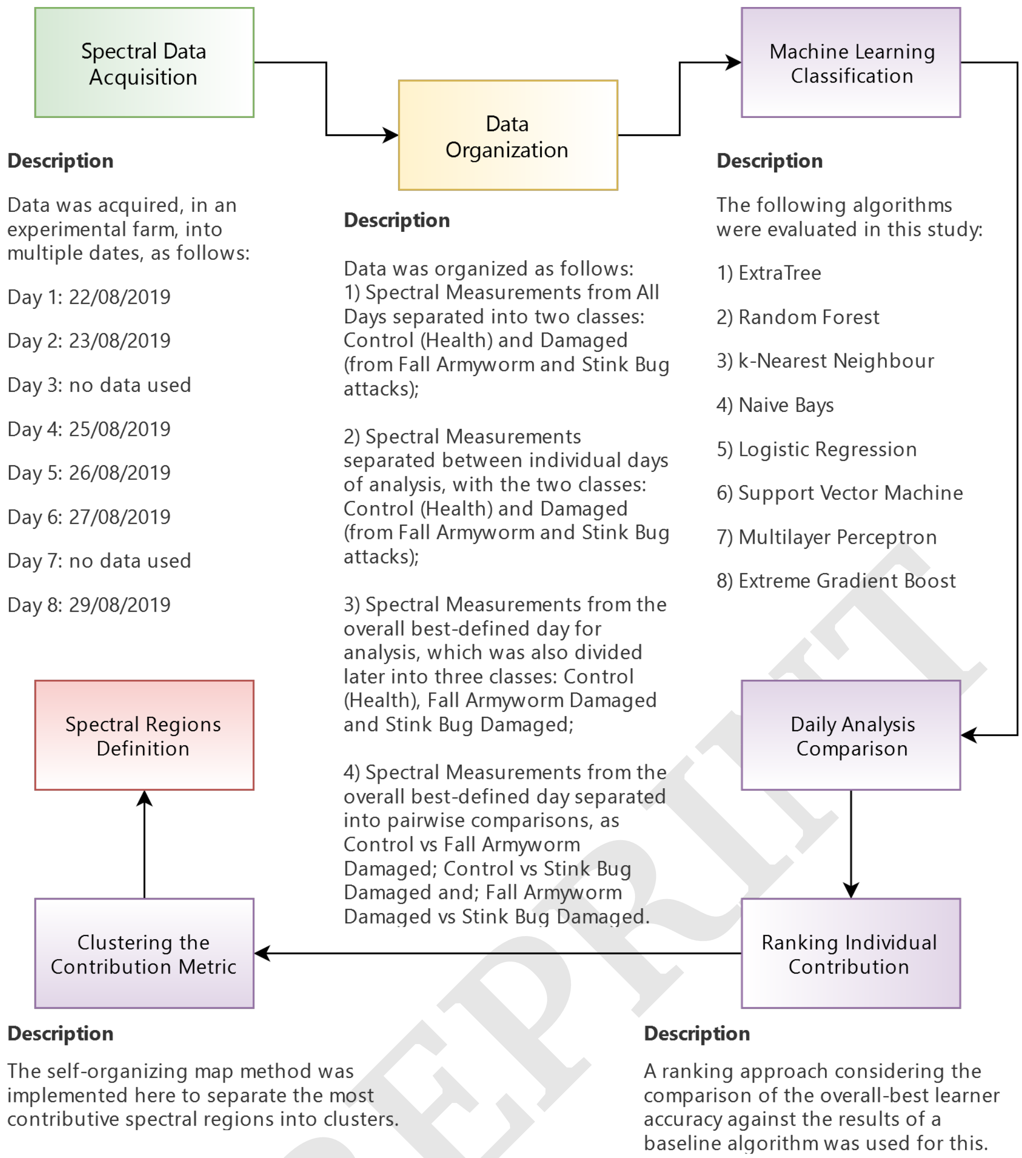

Fig. 1. The summarized steps of the framework developed in this study.

Grid Search, and others against the results of the baseline of each algorithm, we verified that no gain in accuracy was obtained in relation to the processing time needed to perform the classifications. Because of that, the default values of the implemented libraries were adopted in our experiment.

The ExT and RF learners are algorithms based on decision trees, where the ExT (34) is an ensemble method that builds randomized trees with independent structures, while RF (35) combines tree predictors in a manner that each tree depends on values of a random independent vector. In our environment, the number of attributes decided at each node of the ExT was defined as random, and the RF used 100% of the training set as bagging size. The LoR and the NB learner are both based on a probabilistic concept, where LoR (36) is based on a sigmoid function, and NB (37) uses a naïve approach based on the Bayes Theorem, disregarding the correlation between input variables. Here we adopted a Ridge value equal to 0.00000001 in the log-likelihood for the LoR learner, and did not use any Kernel estimator nor supervised discretization for the NB method.

The MLP (38) uses hidden layers to perform a classification task, and executes it in a feed-forward manner, being dependent on its activation functions and solver adapted for optimizing its weights. The MLP used here adopted a learning rate of 0.05, a momentum of 0.1, Adam solver, and sigmoid functions. The SVM (39) separates an attribute space using a hyperplane and calculates a linear function while maximizing the margins between instances. We implemented a C-SVC type, with an eps and gamma equal to 0.001 and exp(-gamma*|u-v|²), respectively, using a radial basis function as Kernel. The kNN (40) verifies the proximities of the data by adopting a set of weights and distance metrics. Here we set the number of neighbors to be equal to 5 and used a euclidean distance approach to measure it. Lastly, the XGB, which is one of the most recognized algorithms by the machine learning community (41), implements a forward stagewise ensemble method and computes second-order gradients of a loss function. In our experiment, the default parameters of its library were adopted.

Protocol. The machine learning models were implemented through the open-source software Weka 3.9.4, using integrated libraries from both Weka, R, and Scikit-Learn packages. A workstation equipped with an Intel(R) i7-8550U 1.80 GHz CPU and 12 Gb 2,333 MHz RAM was used for. The computational analysis was conducted in two different phases: In the first phase we determine the overall best learner to model the spectral data, and investigated the impact of different daily measurements on this learner; in the second phase, we modeled the damage according to its origin type (S. frugiperda or D. melacanthus). The datasets and analyses performed are summarized in Table 1.
The comparison with multiple machine learning algorithms was performed using all of the reflectance measurements acquired during 8 days of analysis. For this, the 8 algorithms were compared after 100 validation results. This was achieved with a 10-fold cross-validation randomized strategy, executed with 10 repetitions. With this approach, the algorithm is validated with "unknown" data to the learner, not used at its training phase. Since this procedure was run 10 times, the models were built from scratch for each repetition. Later, each model was also evaluated with 10% of its data in a separated testing phase. The same data subsets were considered for every classification. The classification metrics evaluated in this study were Precision, Recall, and F-measure. We also used the True-Positive and False-Positive Rates and the Receiver Operating Characteristic (ROC) curve for each class.
After defining the overall best learner, a daily comparison with data collected from the beginning to the end of the experiment (days 1 through 8) was used as separated inputs for the classification task. The strategy, described in the previous step, for processing and evaluating the performance of the algorithm, was also adopted. This helped define the impact of the continuous attack of the insets into the analysis of the spectral behavior of the maize plants. With that, it was possible to indicate the most discrepant days of analysis since the beginning of the infestation. We then performed a Shapiro-Wilk normality test at a 95% confidence interval. As it returned p-values both above and under 0.05, the ANOVA analysis followed by a Tukey's pairwise test was used to compare the mean values of each prediction when data was considered normal (parametric), and the Mann-Whitney pairwise test was adopted when data was non-parametric.
Once the overall best learner and possible best day of analysis were both defined, we evaluated the algorithms' performance according to the three different classes available: control (health plants); caterpillar, and; bug attacked. This analysis was conducted in two steps: firstly considering the full dataset with all classes confronting each-other in a "one-against-all" fashion, with the intent to produce some difficulties to the model and measure its robustness, and later in a pairwise comparison to indicate differences aside of each test. Lastly, to help ascertain our predictions, we also calculated the average and standard deviation of spectral curves and used it to discuss its implications and possible challenges.

Ranking and Clustering of Spectral Data. To calculate the potential of every wavelength used as input for the overall best classifier, we adopted a ranking approach. This ranking approach consists of a direct comparison between the used classifiers' accuracy, obtained with a specific input variable (i.e. the individual wavelength), against the performance obtained at the same conditions with a baseline algorithm. The baseline algorithm used for this comparison was the ZeroR learner, which calculates the average value of the measured variables and uses it as a prediction. This algorithm is considered the baseline for the Weka library of machine learning classifiers. A Metric score, related to this difference in performance between algorithms is obtained from this approach. In this regard, this score can be positive or negative, and even return a number above 1 (since the increase may exceed 100%).
We used the Metric score to indicate the most contributive spectral wavelengths for the prediction. The intention behind it is to provide information related to the importance of these variables in separating healthy plants from the different insect-type damaged plants, evaluated in our dataset. To help ascertain the most contributive spectral regions instead of only the individual contribution of our data, we implemented a clustering algorithm, based on an unsupervised artificial neural network, known as the Self-Organizing Map (SOM). The SOM applies a competitive learning approach using a neighborhood function. This helps to preserve the topological properties of the input variables, and it is useful for evaluating it as it creates a low-dimensional visualization of high-dimensional data. The SOM was executed with 1000 and 2000 epochs in, respectively, the ordering and convergence phases. A height of lattice equal to 2, a learning rate of 1.0, and the normalization of the attributes were also used in this task. With that, we plotted feature maps of the Metric score and identified the highest contributive spectral regions used by the machine learning algorithm to model it.

Results and Discussion

The results of the conducted approach were generated into a sequential form of analysis, where data or information gathered at the previous analysis was incorporated into the subsequent evaluations. In this aspect, this section is divided into two segments: the first being related to an overview of the algorithms performance and day-by-day analysis, and; the second presenting the outcomes of the ranking and clustering approach to indicate the most contributive wavelength intervals to separate each class of insect-damage.

Modeling Insect-Damaged with Machine Learning and Hyperspectral Data. The initial dataset was composed of all the measure variables within the eight days of analysis and separated into two classes: Undamaged plants (UDP) and herbivory-damaged plants (with S. frugiperda larvae and D.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2021 doi:10.20944/preprints202102.0498.v1

Table 1. Configuration-sets used in this study to predict insect-damage and separate insect-damage type.

Dataset	UDP (n)	Herbivory damaged (n)	Total (n)	Sf-HDP (n)	Dm-HDP (n)	Experiment
Total Analysis	464	855	1319	505	350	Multiple Algorithm Comparison
Day 1	104	180	284	100	80	Single Algorithm Analysis
Day 2	150	265	415	185	80	Single Algorithm Analysis
Day 4	50	120	170	70	50	Single Algorithm Analysis
Day 5	60	90	150	50	40	Single Algorithm Analysis
Day 6	50	100	150	50	50	Single Algorithm Analysis
Day 8	50	100	150	50	50	Single Algorithm Analysis
UDP vs Sf-HDP	60	50	110	50	0	Ranking and Self-Organizing Map
UDP vs Dm-HDP	60	40	100	0	40	Ranking and Self-Organizing Map
Sf-HDP vs Dm-HDP	0	90	90	50	40	Ranking and Self-Organizing Map

The "Herbivory Damaged" group amount corresponds with the sum from data of larvae and stink-bugs groups. The Total Analysis group is the sum from Day 1 to 8 groups.

melacanthus). The averaged and the standard deviation values of every wavelength indicated that both groups differentiate each other, in amplitude terms, in most of the spectrum space, with a possible exception in the red-edge region (Figure 2). Another observation is that, in the visible spectrum, the damaged group had a higher deviation from the averaged values than the control group. This indicates that this region may not be interesting to separate both groups, even if the averaged curve was slightly higher for the damaged group than the control.

The prediction using the described dataset was executed with 8 machine learning algorithms, and the results indicated a significant overall better performance with the Random Forest (RF) learner (Figure 3 and Table 2). Here, we compared both the Precision, Recall, and F-Measures among the algorithms, and adopted Tukey's pairwise test to indicate the differences between mean values of each prediction. Since F-Measure is a harmonic mean between Precision and Recall (42), we considered it the most important parameter to compare the models. The kNN and XGB algorithms also returned good accuracies, and the Recall mean value obtained with the kNN was higher than RFs'. However, since the Precision values of RF were higher, the harmonic measure (F) was higher for this classifier.

The ROC curve for the RF was the highest of all learners, indicating that the algorithm returned high true-positives and low false-positives values with more consistency than the others. SVM and NB returned the worst results, and although SVM presented a Precision equal to 1 in all of the validations' set (Figure 3), which is due to an overestimate of one of the classes (damage group) above the other (control group), this scenario resulted in the lowest Recall possible. In the testing phase (Table 2) the SVM method presented a more leveled classification. Regardless, it returned one of the worst possible outcomes. Some classifiers, including SVM, are sensitive to imbalanced training data sets, in which some classes are represented by a much smaller number of samples than other classes (43).

The RF algorithm is considered one of the most powerful learners in use, and its capability of learning from multiple input variables is something that is benefited from a highly-dimensional dataset such as this one (35)(Belgiu and Drăgu, 2016; Breiman, 2001). In other studies related to spectral readings and agronomic-related predictions with machine learning methods, RF was able to infer both macro and micronutrients in hyperspectral readings with satisfactory performances (31). The algorithm was also used in the hyperspectral imagery-domain to predict weed presence in maize-crops (44) and vegetable crop biomass with Unmanned Aerial Vehicle (UAV) type of data (45). Regarding insect-damage detection in crops, the accuracy achieved here was approximate from the values obtained by the other methods (10, 11, 14, 15, 18–20), as it will be demonstrated in the subsequent results.

Table 2. Algorithms' test comparison considering health and insect-damaged maize plants at all days of analysis. Letters positioned after the metric value indicate the differences between each algorithm at training.

Algorithm	Precision	Recall	F-Measure	ROC Area
ExT	0.698 e	0.698 c	0.698 d	0.686 e
kNN	0.748 c	0.746 a	0.747 b	0.741 c
LogR	0.637 f	0.647 d	0.628 e	0.715 d
MLP	0.617 g	0.619 f	0.618 f	0.670 f
NB	0.568 h	0.473 g	0.434 h	0.533 g
RF	0.785 a	0.787 a	0.783 a	0.854 a
SVM	0.771 b	0.630 e	0.515 g	0.539 g
XGB	0.724 d	0.728 b	0.722 c	0.792 b

We used the RF algorithm, with the same preset configurations from the previous analysis, to evaluate its prediction capability in a day-to-day approach. In this regard, the RF learner was capable of achieving higher accuracies (F-Measure) than when considering all of the datasets, which indicated a possible noise hindrance on the data. This may be because, from the initial time of the experiment (day 1), not much of a difference in the spectral behavior of the maize plants may be noticeable. This could be explained by the short amount of time that the plants were under attack by the insects, thus not evidencing it as much as in the subsequent days. Still, the classification was better than a random guess, which is an indication of how robust hyperspectral data and machine learning analysis are. To ensure this comparison and highlight some of these aspects, we evaluated both the multiple validations sets returned by our consecutive runs during the training phase, as well as the F-measure returned at the testing phase of the algorithm (Figure 4).

Although from day 4 and beyond the RF model returned F-measure values close to 1.0 in some of the different training runs, the analysis on day 5 achieved the overall best prediction, with some outliers below 0.80. The testing results were also slightly higher than the averaged values of the training runs, and for the 1st day, it was outside the first interquartile interval. Regardless, it stayed inside the 95% range, which validates it. The Mann-Whitney pairwise test indicated that

Fig. 2. Spectral wavelengths for health plants (control) and insect-damaged maize plant classes after 8 days of the attack. The averaged is represented by lines and the standard deviation values by the colored areas. UDP= undamaged plants, HDP= herbivory damaged plants.

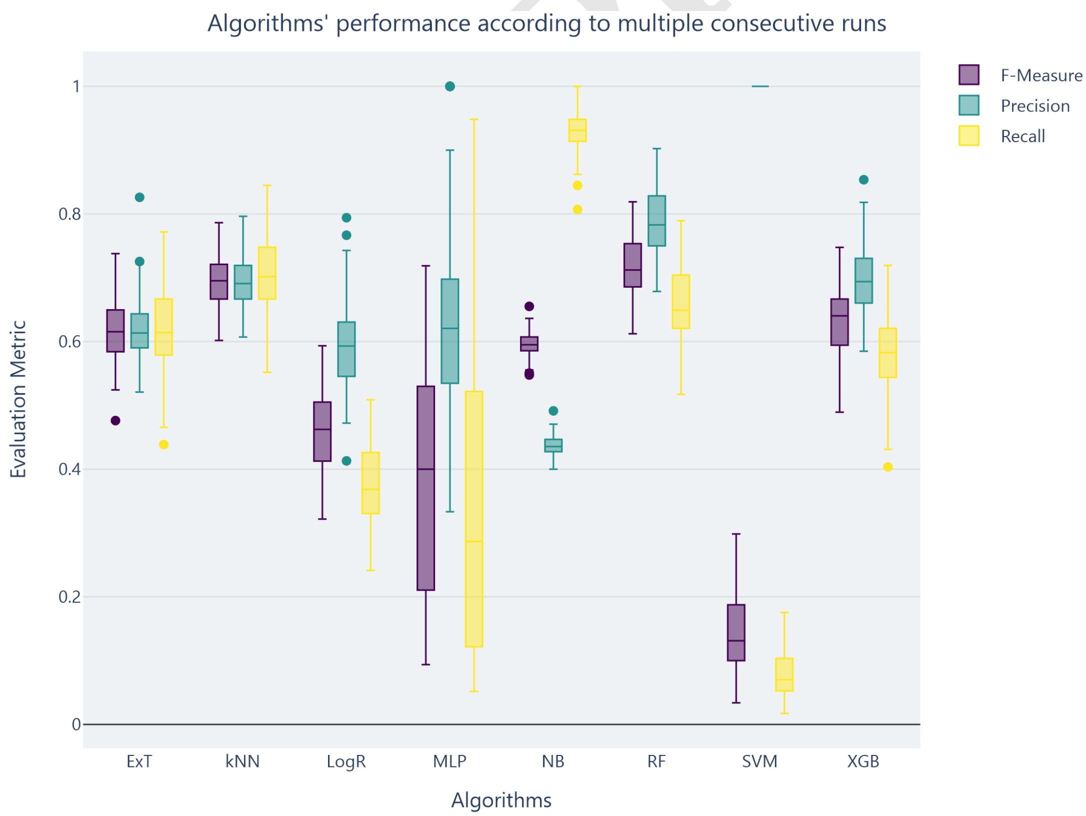

Fig. 3. Machine learning models' performance comparison considering health and insect-damaged maize plants with 8 consecutive days of insect-damage.

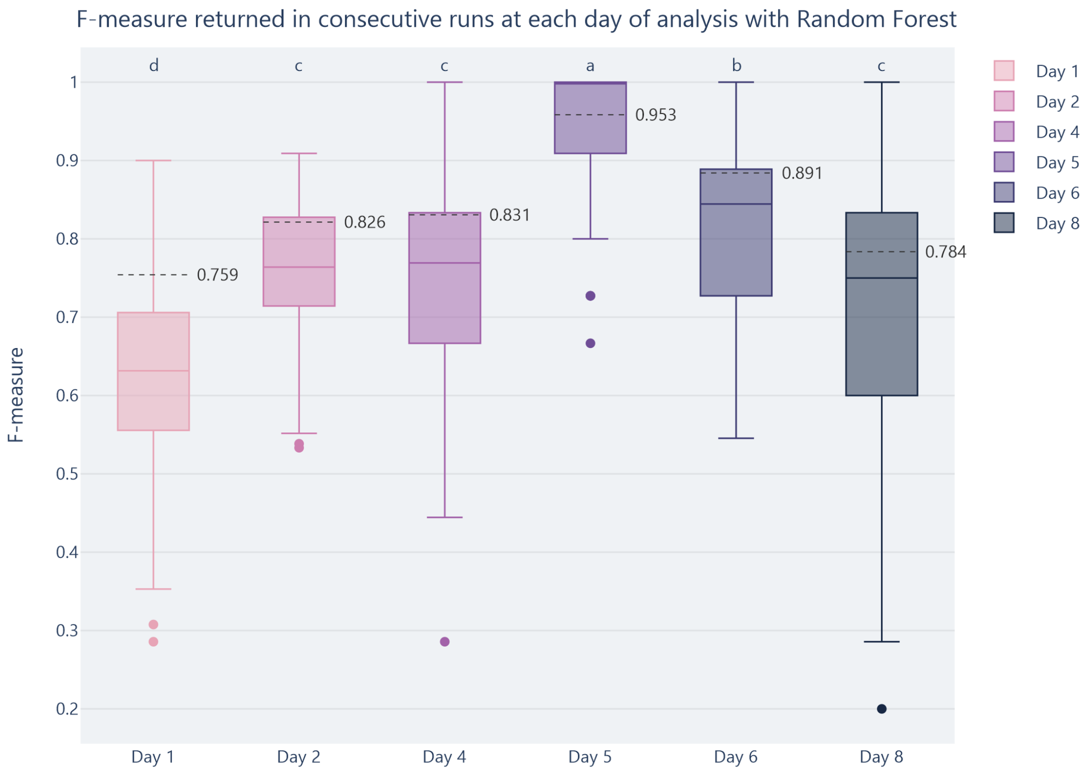

Fig. 4. Random Forest (RF) performance metric comparison between days of analysis considering health and insect-damaged maize plants. Letters positioned above the metric value indicate the differences between each day's prediction. The value highlighted inside the box-plot regions corresponds with the F-measure returned at the testing phase.

days 2, 4, and 8 returned non-statistically significant differences, which is a good indicator that, from the 2nd day of the insect-attack, it is possible to achieve high prediction values with RF to separate health from insect-damaged maize plants. This information is interesting since it is an indication of how the reflectance measurements from proximal sensing, alongside the robustness of the RF algorithm, are sensitive to the effects of the insect-attack in maize plants.

Differentiating Insect-Damage Types. Since RF returned the overall best predictions on the 5th day of analysis, we chose this configuration set to evaluate the capability of the combination between spectral behavior and machine learning approach to separate different types of insect-damage. We should point out that, when evaluating the spectral average and its standard deviation of the individually measured wavelengths (Figure 5), similar characteristics by considering the two-classes problem with every day of analysis (Figure 2) are present. In this aspect, interesting enough, the treatment that presented the highest deviation from the mean was the measures obtained from maize plants with S. frugiperda herbivory. The most visible aspect in this set (Figure 5), however, is that the averaged spectral behavior of both treatments with herbivory damaged, stink bug, and larvae of S. frugiperda is almost equal in the green region, around the 550 nm band. This indicates that it may be difficult for a sensor in this region to differentiate one attack from the other.

In a "one-against-all" type of approach, the RF algorithm was able to separate with high accuracies the three treatments (undamaged plants, Sf-HDP and Dm-HDP). The undamaged plants returned better metrics overall, followed by the Sf-GDP and, later, Dm-HDP (Table 3). This was important to indicate that, even considering similarities between the spectral curves, the model was able to overcome most of it and indicate the correct group. For real plantation conditions, where the producer may or may not know the type of plague in a specific maize plant, this approach may be adequate to assist in defining it.

Table 3. Evaluation metrics returned by the Random Forest algorithm for separating all the classes on day 5.

Group	Precision	Recall	F-Measure	ROC Area
UDP	0.967	0.967	0.967	0.989
Sf-HDP	0.957	0.880	0.917	0.970
Dm-HDP	0.841	0.925	0.881	0.970

To determine the individual performance of the RF learner when confronting the different classes in a pairwise manner, we used different subsets with a two-class approach, as indicated in Table 1. This approach demonstrated that it is easier for the algorithm to separate undamaged maize plants from Dm-HDP (Figure 6). It was also capable of differentiating between maize plants from caterpillar to bug attack, much like when grouping the three classes (Table 3).

The testing metrics also indicated interesting information for the different scenarios considered (Tables 4, 5, and 6). Al-

Fig. 5. Spectral wavelengths for the three classes (undamaged plants - UDP, Sf-HDP, and Dm-HDP) measured on day 5 of the analysis.

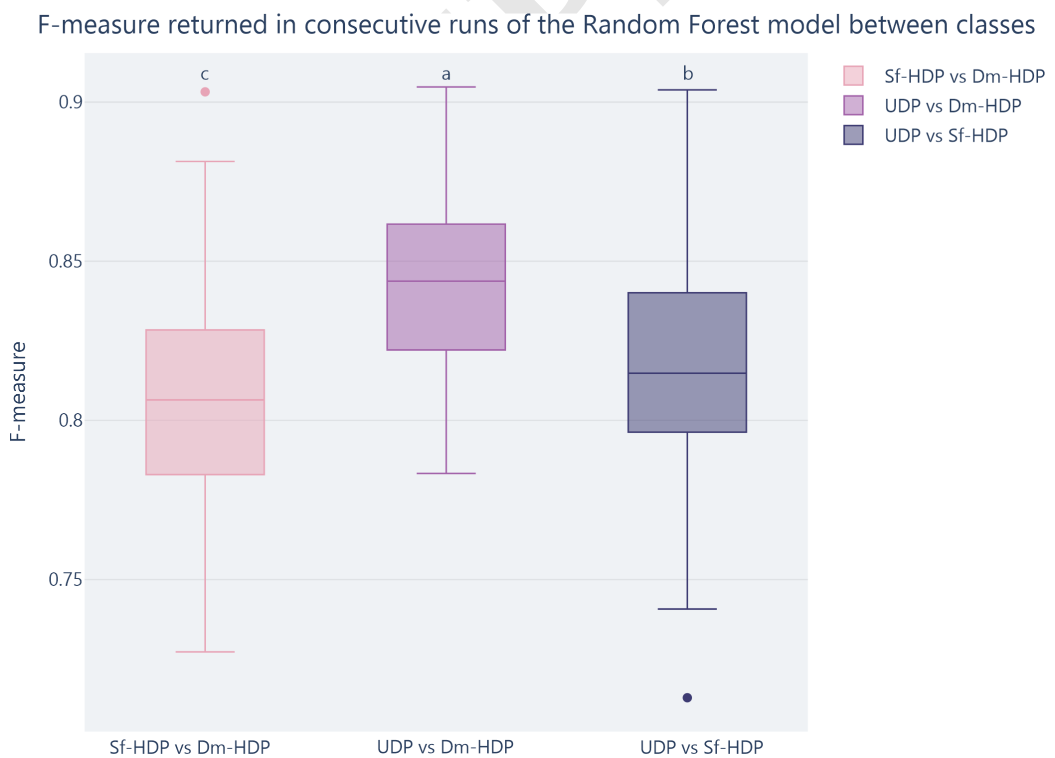

Fig. 6. Evaluation metric returned at training the Random Forest algorithm for separating classes at day 5 in a pairwise comparison.

though these results were higher (F-Measure wise) in comparison with the averaged consecutive runs, they corroborated the pattern observed at the training phase (Figure 6). They were also within the 95% interquartile range calculated. The pattern returned by these analyses indicates that it is easier for the algorithm to separate bug attacks than caterpillar attacks from healthy maize plants. And although it is possible to achieve high accuracies in separating the insect-type of attack by comparing the bug group against the caterpillar group, it is, as expected, more difficult than when in comparison with healthier plants. This is probably related to how both spectral curves (Figure 5) behave for the different types of classes. The bug and control averaged curves are far apart from each other, while the averaged curve from caterpillar attacked maize plants is in between them, with a high standard deviation.
Also, in some regions (Figure 5), the averaged spectral curves of bug and caterpillar groups are almost near each other. In general, lepidopteran larvae can induce higher levels of injury in plant tissues when compared to stink bugs, which are sucking insects. Studies have shown that maize plants can have their direct defense response suppressed by S. frugiperda larvae (46). S. frugiperda can manipulate the plant defense in its favor (47) minimizing the production of toxic compounds. On the other hand, herbivory injury of D. melacanthus herbivory in maize plants induces direct plant defense during the first 24 hours of herbivory. Similar results were observed by (47) when S. frugiperda larvae at the fourth instar feed on maize plants. The higher changes in the chemical profile of direct defense in maize plants injury by herbivory of the stink bug compared to larvae of S. frugiperda support the better separation obtained by the algorithm.

Table 4. Testing metrics of the classification for separating undamaged maize plants from Sf-HDP.

Treatments	TP Rate	FP Rate	Precision	Recall	F-Measure
UDP	0.983	0.060	0.952	0.983	0.967
Sf-HDP	0.940	0.017	0.979	0.940	0.959
Averaged	0.964	0.040	0.964	0.964	0.964

Table 5. Testing metrics of the classification for separating undamaged plants from Dm-HDP.

Treatments	TP Rate	FP Rate	Precision	Recall	F-Measure
UDP	0.967	0.05	0.967	0.967	0.967
Sf-HDP	0.950	0.033	0.950	0.950	0.950
Averaged	0.960	0.043	0.960	0.960	0.960

Table 6. Testing metrics of the classification for separating Sf-HDP from Dm-HDP.

Treatments	TP Rate	FP Rate	Precision	Recall	F-Measure
UDP	0.820	0.125	0.891	0.820	0.854
Sf-HDP	0.875	0.180	0.795	0.875	0.833
Averaged	0.844	0.149	0.849	0.844	0.845

The ranking approach combined with the Self-Organizing Map (SOM) clustering method is, in the presented sense, a newly developed approach that can help with the analysis to indicate the most contributive wavelengths and spectral regions used for the classification performed by the machine learning algorithm. This highlights the importance of the input data (wavelengths) and how well they respond to the algorithms' modeling. Since the RF learner performs multiple combinations of the wavelengths used, it is difficult to evaluate its predictions pattern. In this sense, the ranking approach applied in the described manner may help to ascertain its relationship with the input variables. Here, this framework was implemented with the subsets separated into the pairwise comparison manner (Figure 7).

The ranking approach in the machine learning context is normally used as a pre-processing step to reduce the number of input variables to the models by selecting only the most important data. In agricultural related problems, we implemented this type of approach with the RF and other learners (30–32, 48), and it returned important data to monitor maize-yield, canopy nitrogen content in citrus and leaf nitrogen concentration and plant height in maize plants too. Also in this aspect, when implementing this type of approach for proximal sensing, a different concept with the Relief-F method (31) was considered for mapping both macro and micronutrients in citrus-trees. Yet, by adopting the Metric score calculation after the algorithms' classification, it is possible to measure how well each wavelength relates to the performance of the algorithm.

The addition of the SOM method helped to indicate which regions should be isolated by considering the cluster constructed with the highest Metric values (cluster 3). These regions can be defined by their higher contribution with the RF models prediction, and also with lesser interference from the wavelengths grouped into inferior clusters (clusters 2, 1, and 0). To summarize the metric values related to the defined regions with the help of the SOM method, we calculate a descriptive analysis of the spectral regions (Table 7). Here, the highest average metric values were obtained, interestingly enough, for the comparison between control and caterpillar groups instead of control and bug groups comparison.

The identification of isolated spectral regions is an important feature to be incorporated into studies that aim to evaluate different types of behavior in plants. The main idea behind it is to propose more direct and clear spectral bands to be associated with the respective problem. Our model focused on insect-damage in maize plants, however, the proposed framework should be possible to be implemented in related research. It could also be considered into novel studies that aim to develop simpler and direct methods to estimate these variables, such as spectral vegetation indices, or even sensors and equipment that focus on these particular spectral regions. And although it may be related to the model predictor (being this case, the RF learner), we intend to perform further investigations to compare more traditional methods with the machine learning algorithm to better ascertain the impact related to this reduction in data-dimensionality.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2021 doi:10.20944/preprints202102.0498.v1

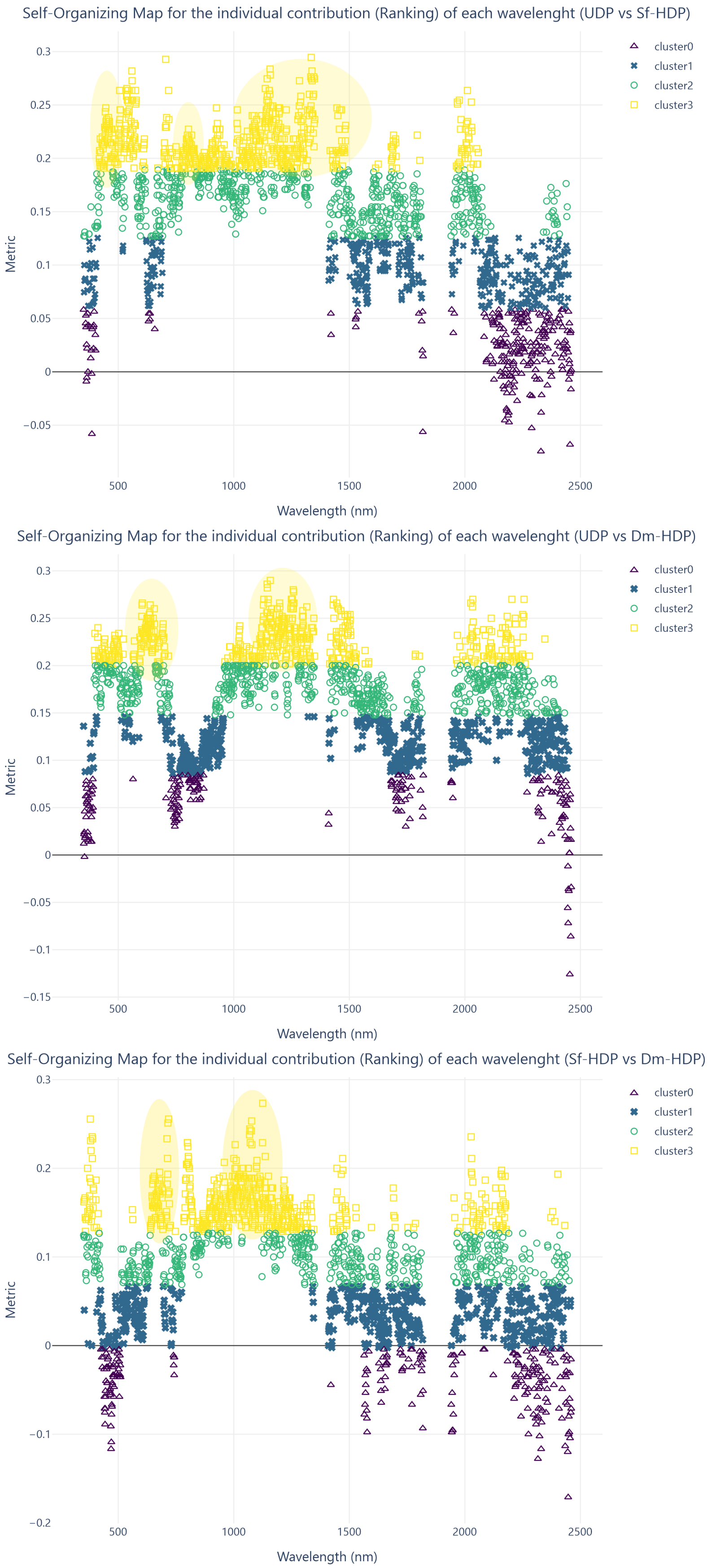

Fig. 7. Ranking metric and SOM clustering method indicating the importance of wavelengths for the three maize plant experiments. The highlighted areas in yellowish-circles indicate the most contributive regions, with less interference from other clusters.

Table 7. Returned metric values for the spectral regions defined with the help of Ranking + SOM methods using the spectral measures from the three treatments - Undamaged plants (UDP), Plants with an injury of two S. frugiperda larvae (SF-HDP), and from plants with an injury of two D. melacanthus.

Comparison	Spectral Regions	Min. Metric	Average Metric	Max. Metric
UDP vs Sf-HDP	420-480, 500-580, and 780-1350 nm	0.191	0.215 +- 0.020	0.294
UDP vs Dm-HDP	600-660 and 1100-1350 nm	0.148	0.177 +- 0.015	0.2
Sf-HDP vs Dm-HDP	640-700 and 900-1250 nm	0.098	0.098 +- 0.017	0.069

Conclusions

The main contribution of this study was to present an approach with machine learning models to detect and separate insect-damaged plants from healthy maize plants using only the reflectance measurements obtained with a proximal hyperspectral sensing approach. We also indicated which learner was more efficient to evaluate this and the impact of a day-by-day analysis into the prediction. Lastly, we proposed a novel framework to identify important spectral regions from visible to short-wave infrared bands (from 350 to 2500 nm) using a combination of ranking and self-organizing map (SOM) approaches. Our results indicated that the RF algorithm is the overall best learner to deal with this type of data. After the 5th day of analysis, the accuracy of the RF algorithm improved substantially. It separated the control, caterpillar, and bug groups with an F-measure equal to 0.967, 0.917, and 0.881, respectively. We also verified that the most contributive spectral regions are situated in the near-infrared domain and, on a small scale, at red, green, and blue, in this respective order.

We conclude that the approach with machine learning methods is adequate to monitor insect-damage in maize plants, differentiating the types of insect-attack early on. We demonstrate that the framework proposed for this analysis, indicating the most contributive wavelengths, is suitable to highlight spectral regions of interest. We hope that novel research adopts the proposal presented herein other types of cultivars and cultures. We suggest that the information presented here, obtained with proximal measurements at wavelength scale, can be implemented in other projects that aim to evaluate the impact of the spectral regions on detecting insect-damage in imagery sensors embedded in UAV platforms. In future research, with larger datasets, we also intend to adopt deep learning-based methods to establish an overview of its performance over insect-damage classification with non-image type of data.

ACKNOWLEDGEMENTS

The authors acknowledge the support of the UFMS (Federal University of Mato Grosso do Sul) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) (Finance Code 001).

FUNDING

This research was partially funded by National Council for Scientific and Technological Development (CNPq), project number: 433783/2018-4, 303559/2019-5, and 304052/2019-1, and received financial support from the Brazilian Corporation of Agricultural Research (EMBRAPA), project number: 11.14.09.001.04.00.

CONFLICTS OF INTEREST

The authors declare no conflict 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.

Bibliography

1. Vijai Singh, Namita Sharma, and Shikha Singh. A review of imaging techniques for plant disease detection. *Artificial Intelligence in Agriculture*, 4:229–242, 2020. doi: 10.1016/j.aiia.2020.10.002.
2. Jingcheng Zhang, Yanbo Huang, Ruiliang Pu, Pablo Gonzalez-Moreno, Lin Yuan, Kaihua Wu, and Wenjiang Huang. Monitoring plant diseases and pests through remote sensing technology: A review. *Computers and Electronics in Agriculture*, 165:104943, oct 2019. doi: 10.1016/j.compag.2019.104943.
3. Andrew Silver. Caterpillar's devastating march across china spurs hunt for native predator. *Nature*, 570(7761):286–287, June 2019. doi: 10.1038/d41586-019-01867-3.
4. CONAB. Monitoring of the brazilian harvest 2019/2020, 2020.
5. Anne-Katrin Mahlein. Plant disease detection by imaging sensors – parallels and specific demands for precision agriculture and plant phenotyping. *Plant Disease*, 100(2):241–251, February 2016. doi: 10.1094/pdis-03-15-0340-fe.
6. Nesreen M. Abd El-Ghany, Shadia E. Abd El-Aziz, and Shahira S. Marei. A review: application of remote sensing as a promising strategy for insect pests and diseases management. *Environmental Science and Pollution Research*, 27(27):33503–33515, jun 2020. doi: 10.1007/s11356-020-09517-2.
7. J R Jensen. *Remote sensing of the environment: an earth resource perspective second edition*, volume 1. Prentice Hall, 2014. ISBN 9780131889507.
8. Dimitrios Moshou, Cédric Bravo, Jonathan West, Stijn Wahlen, Alastair McCartney, and Herman Ramon. Automatic detection of 'yellow rust' in wheat using reflectance measurements and neural networks. *Computers and Electronics in Agriculture*, 44(3):173–188, September 2004. doi: 10.1016/j.compag.2004.04.003.
9. K. Steddom, M. W. Bredehoeft, M. Khan, and C. M. Rush. Comparison of visual and multispectral radiometric disease evaluations of cercospora leaf spot of sugar beet. *Plant Disease*, 89(2):153–158, 2005. doi: 10.1094/PD-89-0153. PMID: 30795217.
10. C.B. Singh, D.S. Jayas, J. Paliwal, and N.D.G. White. Detection of insect-damaged wheat kernels using near-infrared hyperspectral imaging. *Journal of Stored Products Research*, 45(3):151–158, July 2009. doi: 10.1016/j.jspr.2008.12.002.
11. J. Wang, K. Nakano, S. Ohashi, Y. Kubota, K. Takizawa, and Y. Sasaki. Detection of external insect infestations in jujube fruit using hyperspectral reflectance imaging. *Biosystems Engineering*, 108(4):345–351, April 2011. doi: 10.1016/j.biosystemseng.2011.01.006.
12. Anne-Katrin Mahlein, Ulrike Steiner, Christian Hillnhütter, Heinz-Wilhelm Dehne, and Erich-Christian Oerke. Hyperspectral imaging for small-scale analysis of symptoms caused by different sugar beet diseases. *Plant Methods*, 8(1):3, 2012. doi: 10.1186/1746-4811-8-3.
13. A.-K. Mahlein, T. Rumpf, P. Welke, H.-W. Dehne, L. Plümer, U. Steiner, and E.-C. Oerke. Development of spectral indices for detecting and identifying plant diseases. *Remote Sensing of Environment*, 128:21–30, January 2013. doi: 10.1016/j.rse.2012.09.019.
14. Lalit Mohan Kandpal, Sangdae Lee, Moon S. Kim, Hyungjin Bae, and Byoung-Kwan Cho. Short wave infrared (SWIR) hyperspectral imaging technique for examination of aflatoxin b1 (AFB1) on corn kernels. *Food Control*, 51:171–176, May 2015. doi: 10.1016/j.foodcont.2014.11.020.
15. Zhan-Yu Liu, Jia-Guo Qi, Nan-Nan Wang, Zeng-Rong Zhu, Ju Luo, Li-Juan Liu, Jian Tang, and Jia-An Cheng. Hyperspectral discrimination of foliar biotic damages in rice using principal component analysis and probabilistic neural network. *Precision Agriculture*, 19(6):973–991, apr 2018. doi: 10.1007/s11119-018-9567-4.
16. Nikrooz Bagheri. Application of aerial remote sensing technology for detection of fire blight infected pear trees. *Computers and Electronics in Agriculture*, 168:105147, January 2020. doi: 10.1016/j.compag.2019.105147.
17. Dongyan Zhang, Yang Ding, Pengfei Chen, Xiangqian Zhang, Zhenggao Pan, and Dong Liang. Automatic extraction of wheat lodging area based on transfer learning method and deeplabv3/mathplus network. *Computers and Electronics in Agriculture*, 179:105845, dec 2020. doi: 10.1016/j.compag.2020.105845.
18. Jaafar Abdulridha, Ozgur Batuman, and Yiannis Ampatzidis. UAV-based remote sensing technique to detect citrus canker disease utilizing hyperspectral imaging and machine learning. *Remote Sensing*, 11(11):1373, jun 2019. doi: 10.3390/rs11111373.
19. Tinashe Nyabako, Brighton M. Mvumi, Tanya Stathers, Shaw Mlambo, and Macdonald Mubayiwa. Predicting prostephanus truncatus (horn) (coleoptera: Bostrichidae) populations and associated grain damage in smallholder farmers' maize stores: A machine learning approach. *Journal of Stored Products Research*, 87:101592, may 2020. doi: 10.1016/j.jspr.2020.101592.
20. Ahmed Tageldin, Dalia Adly, Hassan Mostafa, and Haitham S Mohammed. Applying machine learning technology in the prediction of crop infestation with cotton leafworm in greenhouse. *bioRxiv*, 2020. doi: 10.1101/2020.09.17.301168.
21. Matthieu Molinier, Jorma Laaksonen, and Tuomas Hame. Detecting man-made structures and changes in satellite imagery with a content-based information retrieval system built on self-organizing maps. *IEEE Transactions on Geoscience and Remote Sensing*, 45(4):861–874, April 2007. doi: 10.1109/tgrs.2006.890580.
22. Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69, 1982. doi: 10.1007/bf00337288.
23. Teuvo Kohonen. *Self-Organizing Maps*. Springer Berlin Heidelberg, 2001. doi: 10.1007/978-3-642-56927-2.
24. BC Hewitson and RG Crane. Self-organizing maps: applications to synoptic climatology. *Climate Research*, 22:13–26, 2002. doi: 10.3354/cr022013.

25. Y. Hong, Y.-M. Chiang, Y. Liu, K.-L. Hsu, and S. Sorooshian. Satellite-based precipitation estimation using watershed segmentation and growing hierarchical self-organizing map. *International Journal of Remote Sensing*, 27(23):5165–5184, December 2006. doi: 10.1080/01431160600763428.
26. Amir Houshang Ehsani and Friedrich Quiel. Application of self organizing map and SRTM data to characterize yardangs in the lut desert, iran. *Remote Sensing of Environment*, 112 (7):3284–3294, July 2008. doi: 10.1016/j.rse.2008.04.007.
27. Yingjie Li, Jing Chen, Qingmiao Ma, Hankui K. Zhang, and Jane Liu. Evaluation of sentinel-2a surface reflectance derived using sen2cor in north america. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(6):1997–2021, jun 2018. doi: 10.1109/jstars.2018.2835823.
28. Yufeng Li, Alan Wright, Hongyu Liu, Juan Wang, Gang Wang, Yuqing Wu, and Lingjun Dai. Land use pattern, irrigation, and fertilization effects of rice-wheat rotation on water quality of ponds by using self-organizing map in agricultural watersheds. *Agriculture, Ecosystems & Environment*, 272:155–164, February 2019. doi: 10.1016/j.agee.2018.11.021.
29. David Rivas-Tabares, Ángel de Miguel, Bárbara Willaarts, and Ana M. Tarquis. Self-organizing map of soil properties in the context of hydrological modeling. *Applied Mathematical Modelling*, 88:175–189, December 2020. doi: 10.1016/j.apm.2020.06.044.
30. Lucas Prado Osco, Ana Paula Marques Ramos, Érika Akemi Saito Moriya, Maurício de Souza, José Marcato Junior, Edson Takashi Matsubara, Nilton Nobuhiro Imai, and José Eduardo Creste. Improvement of leaf nitrogen content inference in valencia-orange trees applying spectral analysis algorithms in UAV mounted-sensor images. *International Journal of Applied Earth Observation and Geoinformation*, 83:101907, nov 2019. doi: 10.1016/j.jag.2019.101907.
31. Lucas Prado Osco, Ana Paula Marques Ramos, Mayara Maezano Faita Pinheiro, Érika Akemi Saito Moriya, Nilton Nobuhiro Imai, Nayara Estrabis, Felipe Ianczyk, Fábio Fernando de Araújo, Veraldo Liesenberg, Lúcio André de Castro Jorge, Jonathan Li, Lingfei Ma, Wesley Nunes Gonçalves, José Marcato Junior, and José Eduardo Creste. A machine learning framework to predict nutrient content in valencia-orange leaf hyperspectral measurements. *Remote Sensing*, 12(6):906, mar 2020. doi: 10.3390/rs12060906.
32. Ana Paula Marques Ramos, Lucas Prado Osco, Danielle Elis Garcia Furuya, Wesley Nunes Gonçalves, Dthenifer Cordeiro Santana, Larissa Pereira Ribeiro Teodoro, Carlos Antonio da Silva Junior, Guilherme Fernando Capristo-Silva, Jonathan Li, Fábio Henrique Rojo Baio, José Marcato Junior, Paulo Eduardo Teodoro, and Hemerson Pistori. A random forest ranking approach to predict yield in maize with uav-based vegetation spectral indices. *Computers and Electronics in Agriculture*, 178:105791, nov 2020. doi: 10.1016/j.compag.2020.105791.
33. LILIAN SCHMIDT, ULRICH SCHURR, and URSULA S. R. RÖSE. Local and systemic effects of two herbivores with different feeding mechanisms on primary metabolism of cotton leaves. *Plant, Cell & Environment*, 32(7):893–903, jul 2009. doi: 10.1111/j.1365-3040.2009.01969.x.
34. Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, March 2006. doi: 10.1007/s10994-006-6226-1.
35. Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/a:1010933404324.
36. S. Le Cessie and J. C. Van Houwelingen. Ridge estimators in logistic regression. *Applied Statistics*, 41(1):191, 1992. doi: 10.2307/2347628.
37. George H John. Estimating Continuous Distributions in Bayesian Classi ers. *Robotics*, 1995.
38. Simon Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice-Hall, 1993.
39. Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory - COLT '92*. ACM Press, 1992. doi: 10.1145/130385.130401.
40. N. S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, aug 1992. doi: 10.1080/00031305.1992.10475879.
41. Sergio González, Salvador García, Javier Del Ser, Lior Rokach, and Francisco Herrera. A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. *Information Fusion*, 64:205–237, dec 2020. doi: 10.1016/j.inffus.2020.07.007.
42. Jiawei Han and Micheline Kamber. *Data Mining Concept and Tehniques.* San Fransisco: Morgan Kauffman, 2006.
43. Aaron E. Maxwell, Timothy A. Warner, and Fang Fang. Implementation of machine-learning classification in remote sensing: an applied review. *International Journal of Remote Sensing*, 39(9):2784–2817, February 2018. doi: 10.1080/01431161.2018.1433343.
44. Junfeng Gao, David Nuyttens, Peter Lootens, Yong He, and Jan G. Pieters. Recognising weeds in a maize crop using a random forest machine-learning algorithm and near-infrared snapshot mosaic hyperspectral imagery. *Biosystems Engineering*, 170:39–50, June 2018. doi: 10.1016/j.biosystemseng.2018.03.006.
45. Thomas Astor, Supriya Dayananda, Sunil Nautiyal, and Michael Wachendorf. Vegetable crop biomass estimation using hyperspectral and RGB 3d UAV data. *Agronomy*, 10(10):1600, oct 2020. doi: 10.3390/agronomy10101600.
46. Felipe C. Wouters, Blair Blanchette, Jonathan Gershenzon, and Daniel G. Vassão. Plant defense and herbivore counter-defense: benzoxazinoids and insect herbivores. *Phytochemistry Reviews*, 15(6):1127–1151, November 2016. doi: 10.1007/s11101-016-9481-1.
47. Gaétan Glauser, Guillaume Marti, Neil Villard, Gwladys A. Doyen, Jean-Luc Wolfender, Ted C.J. Turlings, and Matthias Erb. Induction and detoxification of maize 1, 4-benzoxazin-3-ones by insect herbivores. *The Plant Journal*, 68(5):901–911, October 2011. doi: 10.1111/j.1365-313x.2011.04740.x.
48. Lucas Prado Osco, Ana Paula Marques Ramos, Danilo Roberto Pereira, Érika Akemi Saito Moriya, Nilton Nobuhiro Imai, Edson Takashi Matsubara, Nayara Estrabis, Maurício de Souza, José Marcato Junior, Wesley Nunes Gonçalves, Jonathan Li, Veraldo Liesenberg, and José Eduardo Creste. Predicting canopy nitrogen content in citrus-trees using random forest algorithm associated to spectral vegetation indices from UAV-imagery. *Remote Sensing*, 11(24):2925, dec 2019. doi: 10.3390/rs11242925.