Heliyon 8 (2022) e12256

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

# Classifying cadmium contaminated leafy vegetables using hyperspectral imaging and machine learning

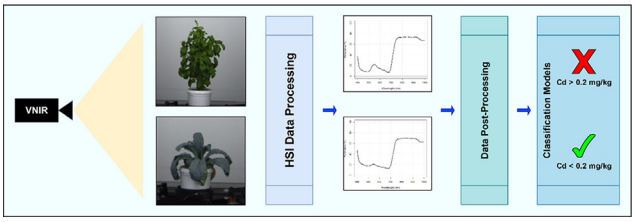
Augusto Souza [a], Maria Zea Rojas [b], Yang Yang [a], Linda Lee [c], Lori Hoagland [b,*]

[a] Institute for Plant Sciences, Purdue University, West Lafayette, IN, USA
[b] Horticulture and Landscape Architecture Department, Purdue University, West Lafayette, IN, USA
[c] Agronomy Department, Purdue University, West Lafayette, IN, USA

## HIGHLIGHTS

- Classification modeling can be used to estimate Cd concentrations in leafy greens.
- Neural network model was the best algorithm in this study for classifying plants.
- Cadmium induced changes in wavelengths linked with chlorophyll and leaf structure.

## GRAPHICAL ABSTRACT




## ARTICLE INFO

*Keywords:*
ReliefF
Ensemble learning
Support vector machine
Neural networks basil
Kale

## ABSTRACT

Cadmium (Cd) is a toxic element that can accumulate in edible plant tissues and negatively impact human health. Traditional Cd quantification methods are time-consuming, expensive, and generate a lot of toxic waste, slowing development of methods to reduce uptake. The objective of this study was to determine whether hyperspectral imaging (HSI) and machine learning (ML) can be used to predict Cd concentrations in plants using kale (*Brassica oleracea*) and basil (*Ocimum basilicum*) as model crops. The experiments were conducted in an automated phenotyping facility where all environmental conditions except soil Cd concentration were kept constant. Cd concentrations were determined at harvest using traditional methods and used to train the ML models with data collected from the imaging sensor. Visible/near infrared (VNIR) images were also collected at harvest and processed to calculate reflectance at 473 bands between 400 to 998 nm. All reflectance spectra were subject to the feature selection algorithm ReliefF and Principal Component Analysis (PCA) to generate data and provide input to evaluate three ML classification models: artificial neural network (ANN), ensemble learning (EL), and support vector machine (SVM). Plants were categorized according to Cd concentrations higher or lower than the safety threshold of 0.2 mg $kg^{-1}$ Cd. Wavelengths with the highest ranks for Cd detection were between 519 and 574, and 692 and 732 nm, indicating that Cd content likely altered the plants' chlorophyll content and altered leaf internal structure. All models were able to sort the plants into groups, though the model with the best F1 score was the ANN for the validation subset that utilized reflectance from all wavelengths. This study demonstrates that HSI and ML are promising technologies for the fast and precise diagnosis of Cd in leafy green plants, though additional studies are needed to adapt this approach for more complex field environments.

* Corresponding author.
*E-mail address:* lhoaglan@purdue.edu (L. Hoagland).

https://doi.org/10.1016/j.heliyon.2022.e12256
Received 6 September 2022; Received in revised form 18 October 2022; Accepted 2 December 2022

## 1. Introduction

Dietary exposure to the heavy metal cadmium (Cd) can harm human health (Ismael et al., 2019). The Food and Agriculture Organization (FAO) determined that Cd concentrations greater than just 0.2 mg $kg^{-1}$ in leafy vegetables pose a serious human health risk (FAO and WHO, 2015). For example, vegetables such as kale (*Brassica oleracea*) and basil (*Ocimum basilicum*) can accumulate high concentrations of Cd in their edible tissues when grown in contaminated soil (Zea et al., 2022). Identifying fields that are contaminated and developing management practices that can prevent leafy greens from accumulating Cd at levels that exceed this safety threshold are critical to protecting human health. For example, some soil amendments can bind Cd in soil, reducing bioavailability and uptake into plants (Zea et al., 2022). There are also mechanisms within plants that can be exploited to prevent uptake of heavy metals like Cd into edible plant tissues. For example, Tang et al. (2017) recently used CRISPR/Cas9 to knock out a metal transport gene in rice, allowing for the creation of new lines that minimize Cd risk in grains. Others are using more traditional breeding strategies to develop improved varieties that restrict heavy metal uptake (Zea et al., 2022). However, the challenge in identifying contaminated fields and/or developing remediation solution is that while Cd is toxic to most plant species, plants generally display few visible symptoms of stress (Ismael et al., 2019; Sánchez-Pardo et al., 2013; Zea et al., 2022). This makes it difficult to determine if plants have accumulated dangerous levels of this element.

Currently, the most common way to quantify Cd in plant tissues is using destructive, post-harvest wet chemical methods that rely on analytical tools such as inductively coupled plasma mass spectrometry (ICP-MS). While highly effective, these methods are time consuming and expensive, and they also generate a lot of toxic waste since plant tissues must be digested in concentrated acid prior to quantification of Cd using ICP-MS. This is particularly problematic when trying to screen thousands of plants in a breeding program to advance those with low heavy metal uptake. Alternatively, new technologies such as hyperspectral imaging (HSI) could be explored to quantify Cd-induced stress responses and predict uptake. For example, we previously demonstrated that it is possible to quantify Cd-induced stress responses in basil and kale using HSI that were impossible to detect with the human eye (Zea et al., 2022). HSI combines digital imaging techniques with spectroscopic analysis algorithms that allow for faster and more accurate non-destructive plant physiological process measurements. This technique analyzes a broad spectrum of light instead of just assigning primary colors (red, green, blue) to each pixel. It works in the visible (VIS) and Near Infrared (NIR) bands, which cover 400 nm–1400 nm. Changes in reflectance in these wavelengths have been explored to capture differences in leaf pigmentation (400–700 nm) and mesophyll cell structure (700–1300 nm) in plants (Knipling, 1970), which can be altered by toxic heavy metals like Cd (He et al., 2015; Ruffing et al., 2021).

With the right training, HSI has potential to go beyond simply quantifying plant stress responses to estimating concentrations of potentially toxic elements (PTEs) like Cd in plant tissues and even soils. For example, Liu et al. (2011) used hyperspectral reflectance data to monitor copper (Cu) (ranging from 20.4 to 68.2 mg $kg^{-1}$) and Cd (ranging from 0.093 to 0.465 mg $kg^{-1}$) in rice crops using Vegetation Indices (VI), which were developed using reflectance at two or more wavelengths (Liu et al., 2011). Tan et al. (2020) used aerial hyperspectral images to estimate the spatial distribution of PTEs in agricultural soils (Tan et al., 2020). In both cases, authors took advantage of machine learning algorithms for predicting PTE occurrence. Their data showed that the Random Forest method had the best results in predicting concentrations of chromium (Cr), Cu, or lead (Pb). More recently, Liu et al. (2019) used a combination of Particle Swarm Optimization (PSO) with a backpropagation neural network (BPNN) to create an integrated method called PSO-BPNN to estimate Cd, mercury (Hg), and arsenic (As) in soil (Liu et al., 2019). The model achieved $R^2$ values between 0.742 to 0.811, which was a significant improvement compared to other standard regression models. These examples demonstrate that combining HSI with modern machine learning algorithms has potential to estimate heavy metal concentration accurately. However, to our knowledge, these methods have not yet been tested for their potential to estimate heavy metal contamination in edible leafy green vegetables, which have high particularly high food safety risks (Baldantoni et al., 2016). In addition, most existing studies to date have been conducted in the field using unmanned arial vehicles (UAV's) equipped with cameras, or hand-held spectroscopic sensors that are in direct contact with plant samples. While the results have clearly been valuable in developing models to predict the uptake of PTE's, we predict that by conducting studies in highly controlled environments and state-of-the-art, high-resolution cameras, it will be possible to develop models with much greater accuracy.

The primary goal of this study was to determine whether HSI can be used as a non-destructive method to classify plants according to the Cd concentration in two distinct types of leafy greens crops: kale and basil. Secondly, different machine learning algorithms were compared with the aim of identifying an optimal classification model for detecting kale and basil plants with Cd concentrations higher than the FAO safety threshold value of 0.2 mg $kg^{-1}$ of fresh plant weight (FAO and WHO, 2015).

## 2. Materials and methods

### 2.1. Experimental design

A total of 64 pots were prepared using an artificial "soil" mix with equals parts in volume (1:1:1) of field soil, sand, and BM8 potting mix (Berger, Saint-Modeste, Quebec, Canada). Initial concentrations of the following metals were determined at the Midwest Soil Testing Laboratory in Omaha, NE, using the EPA 6010b protocol and ICP-MS: arsenic (As), Cd, Cr, cobalt (Co), Cu, mercury (Hg), molybdenum (Mo), nickel (Ni), Pb, selenium (Se) and zinc (Zn). The lab report indicated that these metals were either not detectable or present at low concentrations verifying that artificial soil was not contaminated.

The experiment was conducted using pots without drainage holes to prevent contamination and related hazards. For irrigation, each pot received a certain amount of fertigation solution (see below) to maintain the target weight of 5 kg and keep the pots near field capacity. As the plants grew over time, the target weight was slightly increased to compensate for the increase in plant biomass. For more details, see (Zea et al., 2022).

For the Cd application, pots were amended with an aqueous solution of $CdCl_2$ (99.995% purity, Sigma Aldrich) to reach total soil Cd concentrations of 0, 5, 10, and 15 mg $kg^{-1}$, respectively. These rates were selected because they represent realistic levels of Cd that can be found in low to moderately contaminated agricultural soils. All of the pots were then irrigated with 1200 ml of water and set aside to equilibrate for two weeks, giving enough time for Cd to adsorb onto soil particles. After the incubation period, 32 pots were planted with basil (cv. *Genovese basil*), and the other half with kale (cv. *Lacinato kale*) that had been sown four weeks earlier in potting media (Berger, Ca). For each Cd concentration level, eight replicates were created for basil and kale, respectively.

The plants were cultivated at Purdue University's Ag Alumni Seeds Controlled Environment Phenotyping Facility (AAPF) in West Lafayette Indiana, U.S. during spring 2019. The facility has an automated growth chamber, where the plants were randomly set up in conveyor belts, where they were watered – as described above – and received 14 h of light daily, at 25 °C and 60% of relative air humidity (RH). The fertigation regiment consisted of a mixture of water and 80 mg kg-1 Peters 15-5-15 Ca–Mg fertilizer. Basil plants matured earlier and were harvested 62 days after transplanting (when they started to flower), and kale plants were harvested 84 days after transplanting for aboveground biomass determination. All plant materials were dried in an oven at 60 °C for approximately 3 days until no weight change was observed to obtain

their dry biomass. Afterwards, the materials were ground to 1 mm in size using a UDY cyclone sample mill (UDY Corp., Boulder, CO, USA) for elemental analyses.

Total Cd concentrations in plant tissues were determined using ICP optimal emissions spectroscopy (ICP-OES) (Shimadzu ICPE-9820, Tokyo, Japan) following digestion using a Mars 6 (CEM, Charlotte NC, USA) with Xpress vessels. Briefly, 0.5 g samples were placed in 10 ml $HNO_3$ and subject to a temperature of 200 °C, a pressure of 800 psi, and a power of 900-1050 W. The samples were run alongside periodic table mix 1 for ICP (TraceCERT grade, Sigma Aldrich, St. Louis, MO, USA) as the reference standard, and blanks and additional standard checks were run periodically as a quality assurance measure. Additional information regarding quality control measures to quantify Cd concentrations in plant tissues can be found in Zea et al. (2022).

### 2.2. Hyperspectral imaging set-up

The HSI imaging system at Purdue's AAPF has two Visible/Near Infrared (VNIR) imagers in two different orientations: one scans from the side of a plant and the other scans from the top (Figure S1). Each imager assembly comprises an MSV-500 VNIR scanner (Middleton Spectral Vision, Middleton, WI) with a 2000-pixel linear sensor which scans in 473 bands between 400 to 998 nm. These cameras work as a linear scanner, where they sweep the plant from the bottom to the top (for the side-view camera), or from the back to the front (for the top-view camera).

White and dark calibration reference tests must be done to calculate the relative light reflectance spectra for any scanned object. The white reference was done using a proper board with known spectral characteristics (High Reflectance White PVC VNIR Reference, Middleton WI, USA) and halogen lights on, while the dark reference test was done without any lights and the camera shutter closed. The relative light reflectance for a single wavelength was calculated using Eq. (1).

$$R_{cor}(\%) = \frac{R - R_w}{R_w - R_d} \times 100 \quad (1)$$

where, $R_{cor}$ is the relative light reflectance for the scanned object, R is the raw digital number acquired from the object, $R_w$ is the raw digital number acquired from the white reference test, $R_d$ is the raw digital number acquired from the dark calibration reference test.

Scans were made on the same days that the plants were harvested for biomass and Cd determination. Both side and top-view scans of each plant were made. The acquired scans were later processed using a script developed in MATLAB. Following similar methods introduced by Zhang et al. (2019), the script first segments the plant out of the background and then calculates the mean reflectance spectrum of the plant.

### 2.3. Spectral feature selection and spectral data dimension reduction: ReliefF and Principal Component Analysis

For some wavelengths in a plant's reflectance spectrum, especially in the beginning and end of the VNIR light spectrum, the signal to noise ratios were low. Therefore, these wavelengths were not included in the following analysis, leaving a spectrum of 441 wavelengths from 418.9 to 978.9 nm.

ReliefF is a feature selection method that enables key feature identification from high-dimensional data sets (Kira and Rendell, 1992). ReliefF takes into consideration the interaction between variables. Therefore, if or when there is a difference in a feature value when observing two close predictors, the weight (or the scores) of that feature increases. Principal Component Analysis (PCA) is widely established as a method for reducing the dimension of a data set while still preserving innate information, thereby facilitating easier data visualization and analysis (Müller et al., 2006).

In our study, the predictor variables were reflectance (dimensionless) at specific wavelengths, while the response variable was the Cd concentration in mg per kg of plant fresh weight. For model training in the next step, we created three sets of spectral data sets: the original reflectance spectra, the ReliefF spectra, and the ReliefF + PCA combination. The ReliefF spectra consisted of the key wavelengths identified using the ReliefF algorithm. Using this calculation, the key wavelengths were identified as the ones with rankings $>8 \times 10^{-3}$. This score was pre-defined from a preliminary analysis of implementing this algorithm and finding this score led to a satisfactory selection of wavelengths. In the ReliefF + PCA scores combination, the ReliefF features were first identified in the same way as in the ReliefF data set, then the PCA was conducted on the ReliefF spectra. Scores from the first three principle components (PCs) were used as predictor variables for classification modeling.

### 2.4. Training the classification models

Using the three reflectance datasets described above, we evaluated the performance of three machine learning (ML) algorithms as classifiers in categorizing plant tissue samples with high (greater than 0.20 mg $kg^{-1}$) or low Cd concentrations (lower than 0.20 mg $kg^{-1}$) using data generated using ICP-OES (Figure 1). The three ML algorithms evaluated were Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Ensemble Learning (EL). The samples were divided into 75% for training and 25% for testing for all three models.

ANN is a supervised-learning computational model inspired by the functionality of the human brain. The ANN built in this study had five hidden layers with ten nodes each. The validation test was used during the training to minimizing overfitting, while the testing data was only used to calculate metrics to measure the model success.

SVM is another supervised-learning model for prediction or classification. This technique formulates a hyperplane that can separate data points of two or more classes. This hyperplane can take several forms; thus, this method is very useful for non-linear datasets. The hyperplane chosen was a fourth-order polynomial and the data set was run as a classification model.

EL is another machine learning technique that takes advantage of several classification models, defined as weak learners, to obtain a single well-correlated predictive model called strong learner. For this model, one hundred random-forest decision trees were generated and LogitBoost was used as a boost algorithm (Friedman et al., 2000).

Each of these models were generated and analyzed ten times with a different and random division of the samples as the training or validation data. For each trial and the three different data types (all wavelengths, selected wavelengths, and PCA scores), Recall (R), Precision (P), and the F1-Scores were calculated respectively (Eqs. (2), (3), and (4)), to assess the performance of these models in Cd sample classification using the spectral data sets as inputs. Comparing the result of these metrics between the training and validations subsets allowed us to determine if the models were overfitting.

$$R = \frac{True\ Positives}{True\ Postives + False\ Negatives} \times 100 \quad (2)$$

$$P = \frac{True\ Positives}{True\ Positives + False\ Positives} \times 100 \quad (3)$$

$$F1 = 2 \times \frac{P \times R}{P + R} \times 100 \quad (4)$$

## 3. Results and discussion

### 3.1. Cadmium concentrations in leaves and potential plant responses

As predicted, Cd concentrations in plant tissues generally increased with soil Cd treatment levels (Figure 2). Using traditional wet-chemical

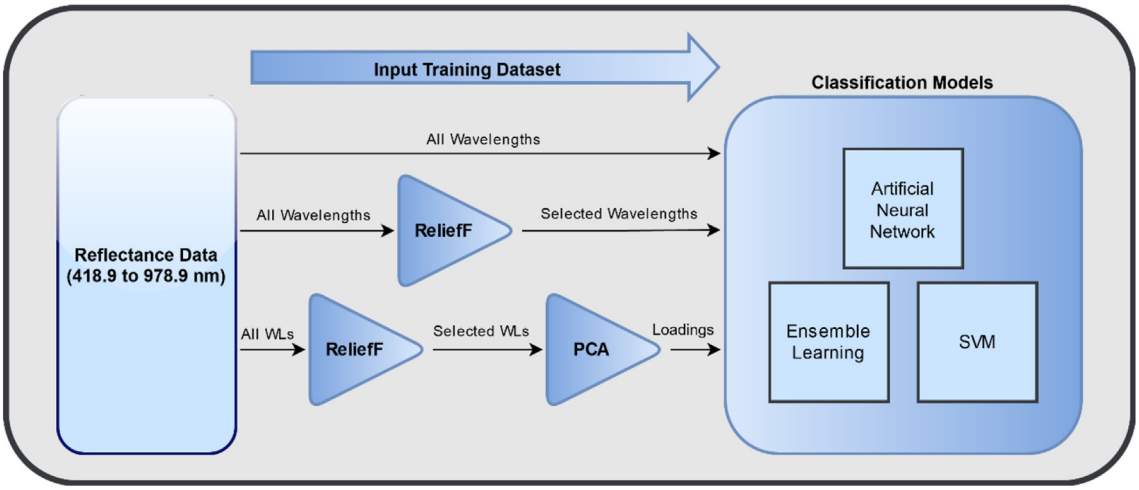



**Figure 1.** Flow diagram of the dataset processing used as input for training the classification models to quantify cadmium concentrations in kale and basil fresh biomass.

methods, we determined that most of the plants in the 0 mg $kg^{-1}$ Cd soil treatment group had Cd concentrations <0.20 mg $kg^{-1}$ of fresh weight, ranging from 0.01 to a little less than 0.20 mg $kg^{-1}$. A few samples had Cd concentrations >0.20 mg $kg^{-1}$, which could have been due to cross-contamination during the Cd treatment or post-harvest quantification process. All plants in the other three soil Cd treatments had Cd tissue concentrations above 0.20 mg $kg^{-1}$ (the concentration FAO considers to pose a human health risk (FAO and WHO, 2015)). For the plants in the 15 mg $kg^{-1}$ soil Cd treatment, the average Cd concentration was slightly higher than 2.0 mg $kg^{-1}$; however, some plants had concentrations as high as 7.0 mg $kg^{-1}$.

Differences in the mean VNIR reflectance spectra of samples with Cd tissue concentrations below and above the 0.2 mg $kg^{-1}$ Cd safety threshold set by the FAO can be observed in both the visible (400–700 nm) and near-infrared (700–1000 nm) bands (Figure 3). Pigments, especially chlorophyll, absorb strongly in the visible band, while leaf internal structure and canopy structure are the major factors driving differences in green vegetation reflectance in the NIR band (Knipling, 1970). These differences in the plants' reflectance indicate that the soil Cd treatment may have altered the chlorophyll content and canopy structure of the kale and basil plants in this study, which is consistent with other studies evaluating plant stress responses to Cd (He et al., 2015; Ruffing et al., 2021; Song et al., 2019). Results of SPAD readings reported in our previous study (see Zea et al., 2022) and reported here in Figures S2 and S3, support this assertion.

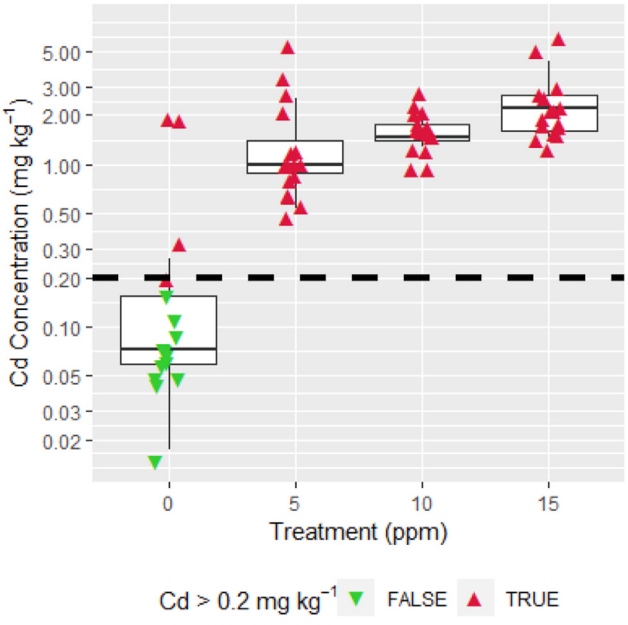



**Figure 2.** Cadmium (Cd) concentration (mg $kg^{-1}$ of plant fresh weight) in kale and basil aboveground biomass in response to soil Cd treatments. The green upside down triangles are plants with < 0.20 mg $kg^{-1}$ Cd and the red triangles are plants with> 0.20 mg $kg^{-1}$. The black dashed line shows the FAO safety threshold for leafy greens. The y-axis is a log 10 scale

### 3.2. Spectral processing

To identify key spectral features (i.e. wavelengths) with strong correlations to Cd induced stress responses, the ReliefF weight spectrum was computed using the spectra and Cd concentration measurements (Figure 4). Two major wavelength ranges in the ReliefF weight spectrum are above the assigned selection threshold. The first range is located in the green band between 519 to 574 nm, while the second range is between 692 and 732 nm, which overlaps with the red-edge band (Peñuelas and Filella, 1998). Within the visible band of light (380–700 nm), chlorophyll absorption of light in the green band is relatively weaker than in the blue and red bands (Greg Mitchell and Kiefer, 1988), indicating that reflectance in the green wavelengths is more sensitive to minor changes in chlorophyll content by Cd in these two plant species. Reflectance in the red-edge has also been shown to be correlated with changes in chlorophyll concentration caused by biotic or abiotic plant stress (Mutanga and

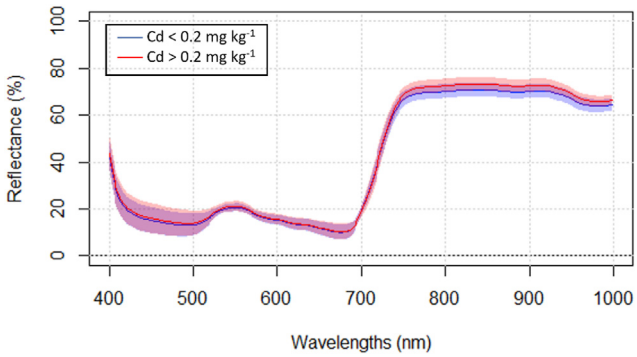



**Figure 3.** Average reflectance spectrum for two cadmium concentration groups (below and above the 0.2 mg $kg^{-1}$ Cd safety threshold set by the FAO). The shaded area represents the standard deviation for each wavelength.

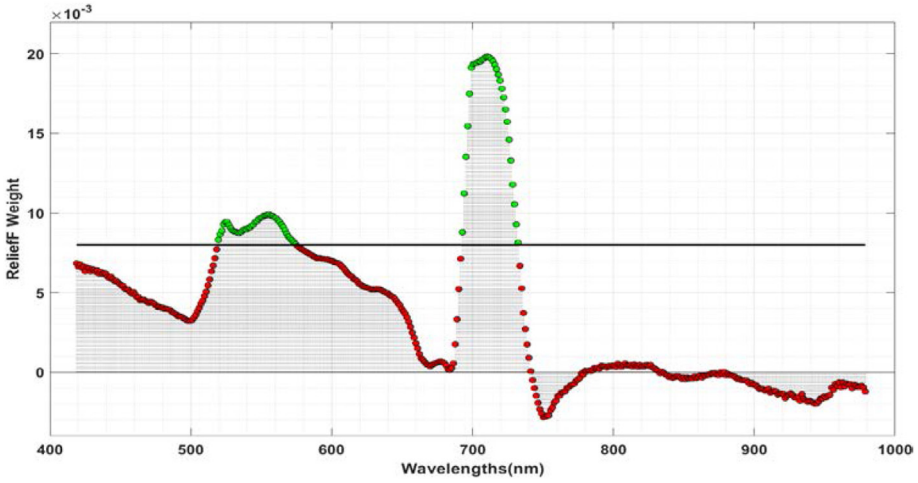



**Figure 4.** ReliefF weights for the light reflectance responses at different wavelengths to cadmium concentration in kale and basil plants. The black line at 8 ($\times 10^{-3}$) ReliefF weight is the selection threshold. The green and red circles represent the wavelengths that were selected and not selected, respectively, for use in the analysis.

Skidmore, 2007). Thus, spectral features identified by the ReliefF algorithm in this study again seem to indicate that the Cd treatment likely caused changes in the chlorophyll content of the plants.

The PCA conducted using wavelengths selected by the ReliefF indicated that the first three PCs were candidates for deeper investigation. The variation in data explained by each of the first three PCs was 63.2%, 20.1%, and 16.1%, respectively, totaling 99.4%. The PCA loadings depict how much each predictor variable contributes to a principal component (PC), and whether they have a positive or negative correlation. From the results of the PC loadings shown in Figure 5, the wavelengths had a mix of positive and negative effects in response to Cd contamination. For PC1, some positive peaks are seen at 555 and 720 nm, which coincides with peaks from the ReliefF results. PC2 had its highest loadings in the first range of wavelengths, while PC3 loadings had maximum positive values at 519 and 732 nm, and 706 nm for the maximum negative loading.

Using the PCA scores, it was possible to differentiate between samples collected from the two Cd treatment groups (below and above the 0.2 mg $kg^{-1}$ Cd safety threshold set by the FAO) (Figure 6). For example, within 6 (a) and (b), clustering between scores for the plants in the two Cd level groups can be observed. Furthermore, when analyzing the PC1 and PC3 (Figure 6c) or PC2 and PC3 (Figure 6d) scores, scores for plants with low Cd concentration are positioned primarily in the quadrant where the PC3 scores are < zero, while the PC1 and PC2 scores are >0.2. While results were promising, there was nonlinearity in separation using PCA. Therefore, we chose a machine learning algorithm that can handle the nonlinearity in the data and provide the clustering capability needed to improve differentiation of spectral features between treatments.

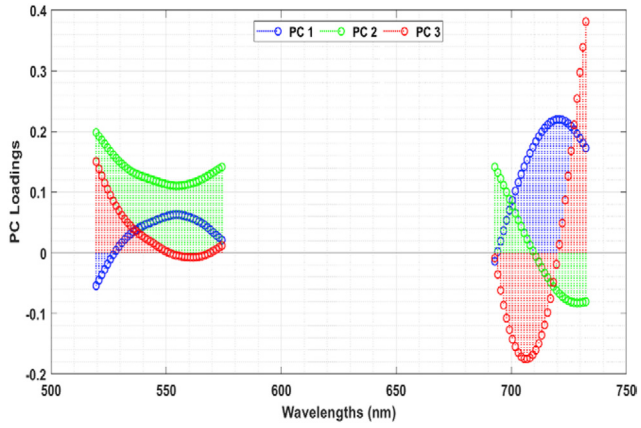



**Figure 5.** Principal Components' Loadings for the wavelengths selected by the ReliefF weight spectrum correlated with changes based on exposure of kale and basil plants to cadmium.

### 3.3. Classification model training and avoiding false positives

Results of three metrics (Precision, Recall, and F1- core) for evaluating the performance of the three ML classification models using the three different data types (all wavelengths, PCA scores, selected wavelengths) are summarized in Figures 7, 8, and 9, respectively. These are important calculations since they can help minimize issues related to large, high-dimensional data sets. For example, too many predictor variables in classification models can lead to overfitting the data. In our case, identifying spectral features with strong correlations with the variable of interest (Cd concentration in kale and basil tissues) will allow us to establish a better understanding of the relationship between a plant's reflectance spectra, as well as its responses to changes in environmental conditions, and possibly even the physiological mechanism driving the plant stress responses.

During the training process, all three ML algorithms using the three types of training data sets resulted in precisions with medians equaling 100% except for the ANN using all wavelengths, and the SVM using the selected wavelengths (Figure 7a). The validation results indicated that the classification precision decreased for almost all cases compared the training runs (Figure 7b). Yet, all the ANN runs, trained using the three data sets, still demonstrated a median precision at 100%, even though the variance in the predictions were. Furthermore, the SVM model trained using the PCA dataset also seemed to be able to classify samples with very high precision. The most considerable differences were for the EL training with the range in precision including as low as 60% when using all wavelengths. This indicates that the EL model was overfitting during the training data set and cannot classify the validation data as precisely. SVM precision also decreased when using all wavelengths with the median equaling 76.0%, the lowest median for all models and data set types. This indicated the use of too many potentially autocorrelated features reduced the predictive power of this model.

Recall is another critical metric for classification models since it is related to false negatives. For this study, a false negative indicates that a highly Cd contaminated plant will be classified as healthy, which is a dangerous situation that should be avoided. Similar to the precision results, the EL runs had the best recall results in the training process (Figure 8a). The SVM trained with all wavelengths also seemed to perform well. The performance of the ANN models was worse than the models trained using the other two. The only exception was the SVM using the PCA scores. However, when analyzing the validation results (Figure 8b), the best model regarding recall was the ANN using the PCA

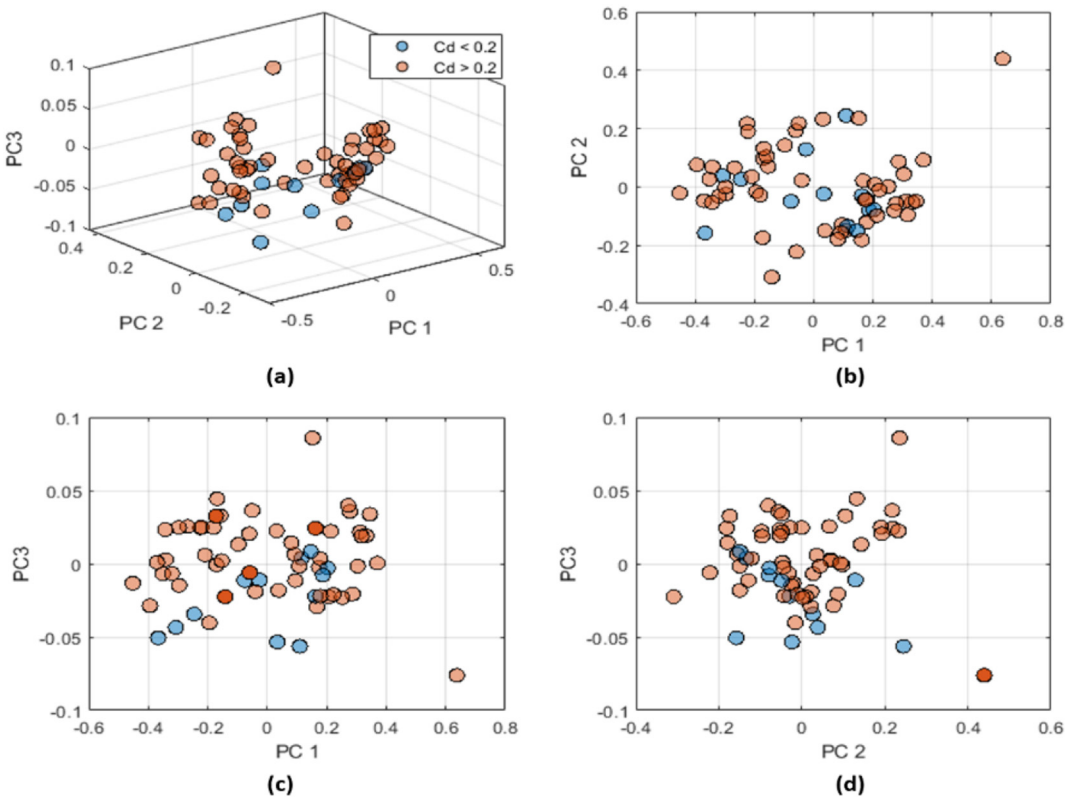



**Figure 6.** PCA score for the first three principal components (a), for PC1 and PC2 (b), PC1 and PC3 (c), and PC2 and PC3 (d). The blue circles represent the scores of the plants with Cd $< 0.20$ mg kg$^{-1}$, while the red markers are for those plants with concentration $> 0.20$ mg kg$^{-1}$.

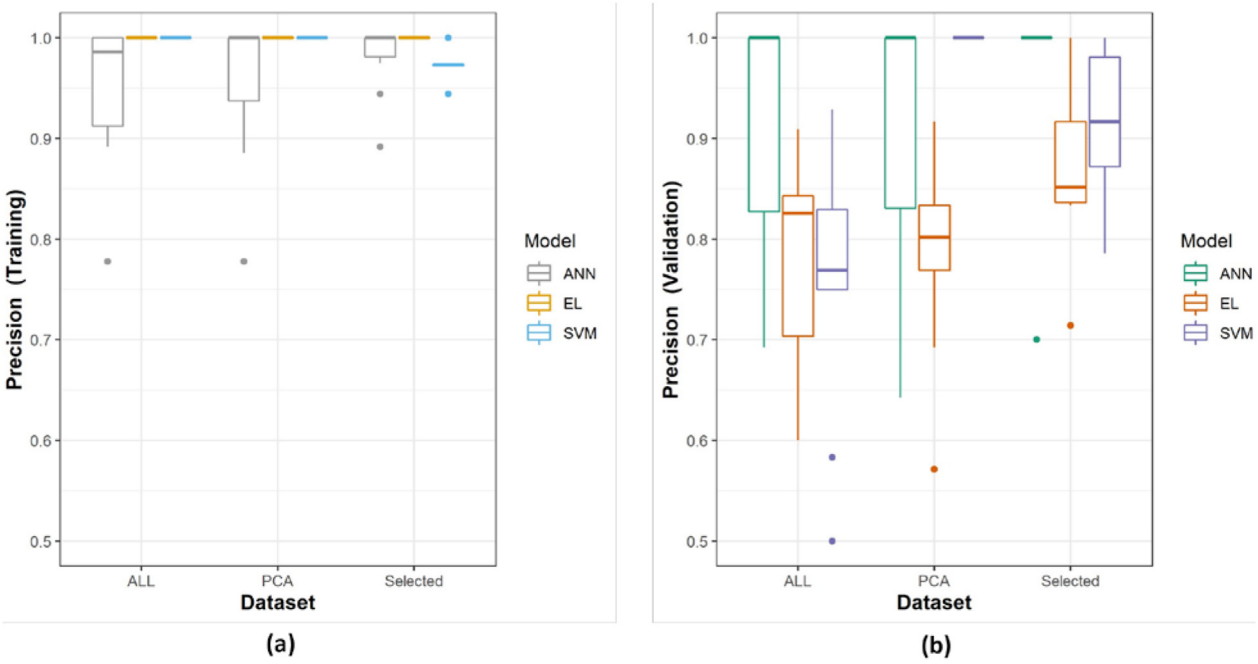



**Figure 7.** Precision results for all three data sets and classification models using different training (left graph) and validation data subsets (right graph) to differentiate between cadmium concentrations in kale and basil biomass.

scores, while the worst was the SVM using all wavelengths. Overall, the ANN was more consistent when comparing the training and validation data sets since the recall median was similar between these two fractions of the data. This result again indicated that there might be overfitting issues in the models trained using the EL or the SVM.

Since avoiding false negatives is crucial in this study, the recall metric is one of the most important to consider when deciding the best model to use, but precision is also essential because it is also important to avoid discarding healthy plants. The F1-score is a metric related to both Recall and Precision that can help determine the best classification model. From

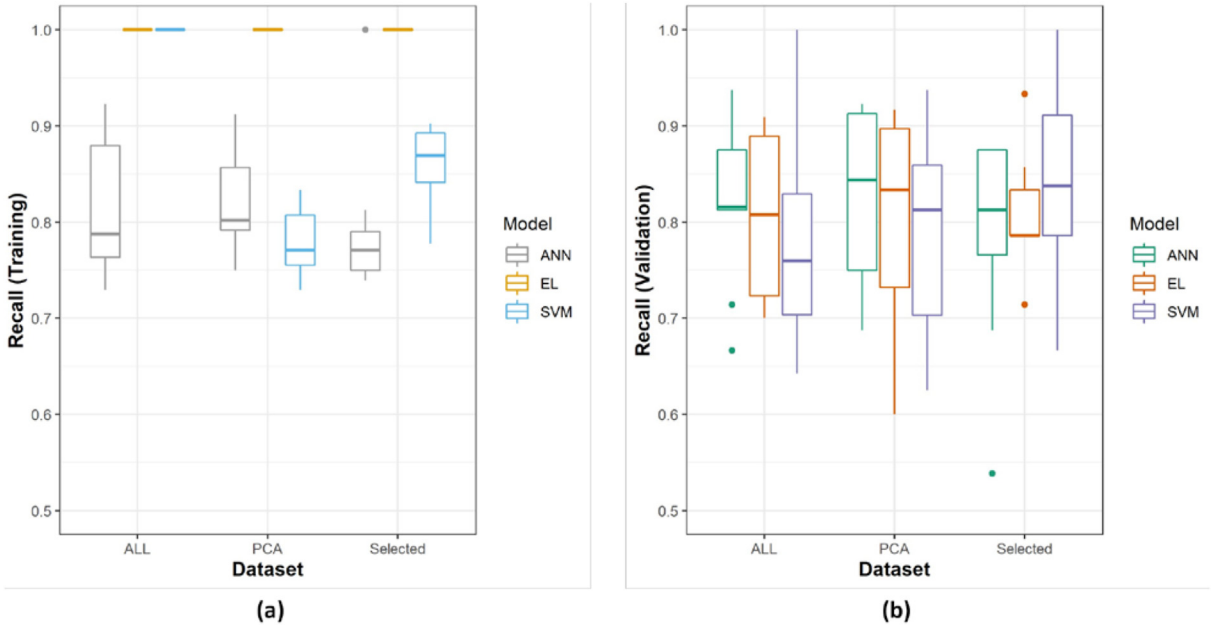



**Figure 8.** Recall results for all three data sets and three classification models using different training (left graph) and validation (right graph) data subsets.

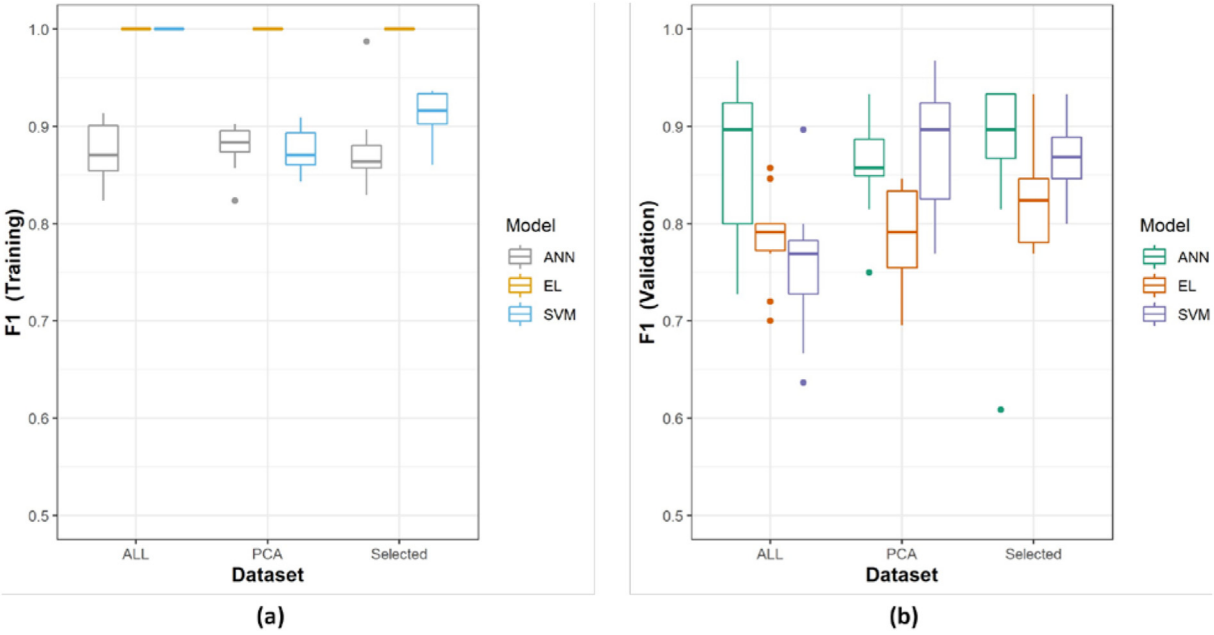



**Figure 9.** F1-score results for all three data sets and three classification models using different training (left graph) and validation (right graph) data subsets.

the training subset, EL was the best when using all data types together and SVM when using all wavelengths (Figure 9a). However, performance of these models was worse for the validation dataset than ANN and SVM using the other two datasets (Figure 9b). The models with the highest F1 medians for the validation set were ANN using all selected wavelengths, and SVM using the PCA scores.

### *3.4. Identifying the best classification model*

The best model to use for a particular study depends on the dataset used as input, as well as which metric is selected for the evaluation. For this study, since false negatives are more important to avoid than false positives, the recall metric turned out to be the most viable. Also, since the validation set was not used during the models' training, the results of this metric reflect more on how these models would behave for other basil and kale plants. Consequently, the best model overall in this study was the ANN, especially the one trained using the PCA scores, since it had one of the highest medians and lowest variations in F1-score compared to SVM and EL. When considering the F1-score validation results, ANN also showed some of the best results, even though SVM using the PCA scores had lower variation compared to ANN using all wavelengths, and a higher median than ANN using the PCA scores. The ANN runs also had better recall results when using the same dataset, except when using selected wavelengths data. This contradicts our prediction that SVM would be the best classifier since the separation between the data points is evident on some plants (Figure 6). When comparing the performance of ANN and SVM for binary classification in a previous study, ANN had a better result since it had a better learning rate compared to SVM (Kim et al., 2010). The poor performance for the EL training in this study could be explained by the overfitting of the data, since the validation results

were much lower than the training subsets. Thus, using EL might not be the best option for this purpose.

It is interesting to observe that the models that were trained using specific wavelengths (with a lower data dimension), generally speaking, performed as well as ones trained with all wavelengths when analyzing the three metrics and two validation datasets. In fact, the selected wavelength models had better predictive results for the validation data set compared to the other model. This was also reported when using Feature Selection algorithms to classify healthy and unhealthy plants due to *Cercospora* and rust presence on sugar beet leaves (Alsuwaidi et al., 2016). The F1 for the validation also indicates that the model chosen had the best values, which indicates that these less inclusive models did not overfit compared to the models that used all wavelengths. Thus, the models did not lose predictive power when using only the wavelengths identified in the feature selection (ReliefF) step.

In summary, the ANN model had the best results regardless of which wavelengths were used, and therefore appears to be the best model for predicting whether kale and basil plants are above critical safety thresholds. The other models performed better when using only the selected wavelengths, indicating how feature selection can improve prediction of these models by using key predictor variables and taking steps to minimize overfitting. It is important to emphasize that the frequency of false negatives is a key factor in evaluating the accuracy of these models. Training the model several times using different subsets of data and comparing results with the validation data set helped reduced the potential for false negatives in this study. However, future studies should also consider validating the training models by measuring Cd concentrations in new samples to further reduce the potential for false negatives. In addition, other feature selection methods such as competitive adaptive reweighted sampling (CARS), random frog (RF), and successive projections (SPA) could be evaluated to see if they could better fit the data.

## 4. Implications and future directions

All classification models and datasets evaluated in this study successfully predicted which plants were above or below the Cd safety threshold for safe consumption of leafy greens, demonstrating that HSI + ML are promising technologies for the fast and precise diagnosis of food safety risks. Since we were able to identify spectral peaks correlated with Cd-induced stress responses that have been linked with critical physiological processes, we expect that this could also be a valuable tool in helping researchers better understand how metals impact plants, thereby leading to the development of more effective remediation strategies. We acknowledge that many other stress factors such as heat, water deficits, nutrient deficiencies, salinity and other PTE's can also influence leaf pigmentation and mesophyll cell structure, and hence, hyperspectral reflectance values, and these responses could vary among crop varieties. Thus, future studies that combine Cd contamination with other types of stress responses and include multiple crop varieties will be needed to adapt the models and apply them under real-world conditions. In addition, while we expect that HSI and ML could someday provide a faster and cheaper method to quantify the presence of PTE in plants, while also generating less toxic waste relative to other detection methods, we acknowledge that this is still a relatively expensive and intricate technology. Consequently, future efforts to design cheaper, custom-imaging sensors that are more affordable and easier to implement, such as multispectral sensors widely used in remote sensing, will also be needed to adapt this technology for PTE quantification.

## Declarations

### *Author contribution statement*

Augusto Souza: Performed the experiments, Analyzed and interpreted the data, Contributed reagents, materials, analysis tools or data, Wrote the paper.

Maria Zea Rojas: Conceived and designed the experiments, Performed the experiments, Analyzed and interpreted the data, Contributed reagents, materials, analysis tools or data, Wrote the paper.

Yang Yang: Analyzed and interpreted the data, Contributed reagents, materials, analysis tools or data, Wrote the paper.

Linda Lee: Analyzed and interpreted the data, Contributed reagents, materials, analysis tools or data, Wrote the paper.

Lori Hoagland: Conceived and designed the experiments, Performed the experiments, Analyzed and interpreted the data, Contributed reagents, materials, analysis tools or data, Wrote the paper.

### *Funding statement*

This work was supported by NIFA-HATCH [Hatch project 1015999].

### *Data availability statement*

Data will be made available on request.

### *Declaration of interest's statement*

The authors declare no conflict of interest.

### *Additional information*

Supplementary content related to this article has been published online at https://doi.org/10.1016/j.heliyon.2022.e12256.

## Acknowledgement

We would like to acknowledge the Purdue Libraries Open Access Fund that covered the expenses related to publication.

## References

Alsuwaidi, A., Veys, C., Hussey, M., Grieve, B., Yin, H., 2016. Hyperspectral selection based algorithm for plant classification. In: IST 2016 – 2016 IEEE International Conference on Imaging Systems and Techniques, Proceedings, pp. 395–400.

FAO, WHO, 2015. General Standard for Contaminants and Toxins in Food and Feed (Codex Stan 193-1995). International Food Standards.

Friedman, J., Tibshirani, R., Hastie, T., 2000. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). Ann. Stat. 28, 337–407.

Greg Mitchell, B., Kiefer, D.A., 1988. Chlorophyll α specific absorption and fluorescence excitation spectra for light-limited phytoplankton. Deep-Sea Res., Part A 35, 639–663.

He, S., He, Z., Yang, X., Stoffella, P.J., Baligar, V.C., 2015. Soil biogeochemistry, plant physiology, and phytoremediation of cadmium-contaminated soils. In: Advances in Agronomy, pp. 135–225.

Ismael, M.A., Elyamine, A.M., Moussa, M.G., Cai, M., Zhao, X., Hu, C., 2019. Cadmium in plants: uptake, toxicity, and its interactions with selenium fertilizers. Metallomics 11, 255–277.

Kim, K., Ko, K., Kim, W., Yu, S., Han, C., 2010. Performance comparison between neural network and SVM for terrain classification of legged robot. In: Proceedings of the SICE Annual Conference, pp. 1343–1348.

Kira, K., Rendell, L.A., 1992. The feature selection problem: traditional methods and a new algorithm. In: AAAI-92 Proceedings. San Jose, California, pp. 129–134.

Knipling, E.B., 1970. Physical and physiological basis for the reflectance of visible and near-infrared radiation from vegetation. Remote Sens. Environ. 1, 155–159.

Liu, M., Liu, X., Ding, W., Wu, L., 2011. Monitoring stress levels on rice with heavy metal pollution from hyperspectral reflectance data using wavelet-fractal analysis. Int. J. Appl. Earth Obs. Geoinf. 13, 246–255.

Liu, P., Liu, Z., Hu, Y., Shi, Z., Pan, Y., Wang, L., Wang, G., 2019. Integrating a hybrid back propagation neural network and particle swarm optimization for estimating soil heavy metal contents using hyperspectral data. Sustainability 11.

Müller, W., Nocke, T., Schumann, H., 2006. Enhancing the Visualization Process with Principal Component Analysis to Support the Exploration of Trends.

Mutanga, O., Skidmore, A.K., 2007. Red edge shift and biochemical content in grass canopies. ISPRS J. Photogrammetry Remote Sens. 62, 34–42.

Peñuelas, J., Filella, L., 1998. Technical focus: visible and near-infrared reflectance techniques for diagnosing plant physiological status. Trends Plant Sci. 3, 151–156.

Ruffing, A.M., Anthony, S.M., Strickland, L.M., Lubkin, I., Dietz, C.R., 2021. Identification of metal stresses in Arabidopsis thaliana using hyperspectral reflectance imaging. Front. Plant Sci. 12.

Sánchez-Pardo, B., Carpena, R.O., Zornoza, P., 2013. Cadmium in white lupin nodules: impact on nitrogen and carbon metabolism. J. Plant Physiol. 170, 265–271.
Song, X., Yue, X., Chen, W., Jiang, H., Han, Y., Li, X., 2019. Detection of cadmium risk to the photosynthetic performance of hybrid pennisetum. Front. Plant Sci. 10.
Tan, K., Wang, H., Chen, L., Du, Q., Du, P., Pan, C., 2020. Estimation of the spatial distribution of heavy metal in agricultural soils using airborne hyperspectral imaging and random forest. J. Hazard Mater. 382, 120987.
Zea, M., Souza, A., Yang, Y., Lee, L., Nemali, K., Hoagland, L., 2022. Leveraging high-throughput hyperspectral imaging technology to detect cadmium stress in two leafy green crops and accelerate soil remediation efforts. Environ. Pollut. 292.
Zhang, L., Maki, H., Ma, D., Sánchez-Gallego, J.A., Mickelbart, M.v., Wang, L., Rehman, T.U., Jin, J., 2019. Optimized angles of the swing hyperspectral imaging system for single corn plant. Comput. Electron. Agric. 156, 349–359.