AMERICAN UNIVERSITY OF BEIRUT

CLASSIFICATION AND PREDICTION USING MACHINE LEARNING ALGORITHMS BASED ON HYPERSPECTRAL DATA TOWARDS DETECTION OF CONTAMINATION IN SOIL AND VEGETATION

by
SAMER IMAD ABED EL RAHIM

A thesis
submitted in partial fulfillment of the requirements
for the degree of Master of Engineering
to the Department of Mechanical Engineering
of the Faculty of Maroun Semaan of Engineering and Architecture
at the American University of Beirut

Beirut, Lebanon
September 2022

AMERICAN UNIVERSITY OF BEIRUT

CLASSIFICATION AND PREDICTION USING MACHINE LEARNING ALGORITHMS BASED ON HYPERSPECTRAL DATA TOWARDS DETECTION OF CONTAMINATION IN SOIL AND VEGETATION

by

SAMER IMAD ABED EL RAHIM

Approved by:

Samir Mustapha

Digitally signed by Samir Mustapha
DN: cn=Samir Mustapha, o=American University of Beirut, ou=Mechanical Engineering, email=sm154@aub.edu.lb, c=LB
Date: 2022.09.07 09:50:38 +03'00'

Dr. Samir Mustapha, Associate Professor — Advisor
Department of Mechanical Engineering

Dr. Darine Salam, Associate Professor — Co-Advisor and Member of Committee
Department of Civil and Environmental Engineering

Ali Tehrani

Digitally signed by Ali Tehrani
DN: cn=Ali Tehrani, c=FI, o=Aalto University, ou=Bioproducts and Biosystems, email=ali.tehrani@aalto.fi
Date: 2022.09.07 09:32:29 +03'00'

Dr. Ali Tehrani, Associate Professor — Co-Advisor and Member of Committee
Department of Chemical Engineering

Date of thesis defense: September 2, 2022

AMERICAN UNIVERSITY OF BEIRUT

THESIS RELEASE FORM

Student Name: Abed El Rahim, Samer, Imad

I authorize the American University of Beirut, to: (a) reproduce hard or electronic copies of my thesis; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes:

- ☒ As of the date of submission
- ☐ One year from the date of submission of my thesis.
- ☐ Two years from the date of submission of my thesis.
- ☐ Three years from the date of submission of my thesis.

Samer Abed El Rahim September 8, 2022

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Dr. Samir Mustapha, for his patience, guidance, and support. I have benefited greatly from his wealth of knowledge and meticulous editing. I am extremely grateful that he took me on as a student and continued to have faith in me over the years.

Thank you to my co-advisors and committee members, Dr. Darine Salam and Dr. Ali Tehrani. Your encouraging words and thoughtful, detailed feedback have been very important to me.

Thanks to the American University of Beirut for awarding me the MEPI-TLG LiDs scholarship and providing me with the financial means to complete this project.

ABSTRACT
OF THE THESIS OF

Samer Imad Abed El Rahim for Master of Engineering
Major: Mechanical Engineering

Title: Classification and Prediction using Machine Learning Algorithms Based on Hyperspectral Imaging Towards Detection of Contamination in Soils and Vegetation

Hyperspectral imaging is a powerful technique in remote sensing that found its application in various fields such as agriculture, health monitoring, target detection, etc. Heavy metal contamination in soil and food crops has been considered a major problem for several decades. Owing to their toxicity and persistence, heavy metals in soils are one of the most hazardous pollutants in the environment. Therefore, hyperspectral remote sensing of heavy metal contamination in soils and food crops has been widely examined for both qualitative and quantitative detection. This work will explore the application of artificial intelligence and hyperspectral imaging as a robust tool towards the classification and prediction of heavy metals.

Multiple preprocessing (Savitzky Golay, Standard Normal Variate, and Multiplicative Scatter Correlation), data reduction (Principal Component Analysis and Linear Discriminant Analysis), and estimation models (a Random Forests, Support Vector Machines, and K-Nearest Neighbors) were implemented on different datasets to evaluate their effect on classification and prediction results. Three case studies were investigated on the Sorghum Plant, Salinas A-Scene (vegetation), and Paint Condition Assessment to establish an estimation and classification model based on the correlation between the selected features and the full-spectrum features respectively. Results showed that the RF model achieved the highest accuracy in comparison with the rest of the models, with accuracies of 98 %, 90 %, and 98 % for the three case studies respectively. The high accuracy of the results and predictions showed that, in combination with the appropriate spectra-preprocessing and data reduction techniques, machine learning algorithms with hyperspectral imaging and remote sensing stand out as an advanced prediction and classification system for soil and agricultural products.

Keywords: Heavy metal contamination, Hyperspectral imaging, Machine learning, Classifications

TABLE OF CONTENTS

ACKNOWLEDGEMENTS 1
ABSTRACT 2
ILLUSTRATIONS 5
TABLES 6
ABBREVIATIONS 7
LITERATURE REVIEW 8

A. Contamination in soil and impact on environment and health 8
B. Contamination in Lebanon 10
C. Common techniques used for soil characterization 13
1. Study Area & Soil Sampling 13
2. Chemical Analysis 15
D. Hyper Spectroscopy 17

MACHINE LEARNING APPLICATION FOR SOILS CONTAMINATION DETECTION BASED ON HYPERSPECTRAL DATA 28

A. Introduction 28
B. Preprocessing 30
1. Scatter Corrections 30
2. Spectral Derivatives 32
C. Vegetation Indexing 36
D. Feature Extraction and Dimensionality Reduction 37
1. Principle Component Analysis (PCA) 38
2. Linear Discriminant Analysis (LDA) 39

E. Estimation Models 39

1. Multivariate Linear Regression (MLR) & Partial Least Square Regression (PLSR) 41
2. Random Forest (RF) 43
3. Support Vector Machines (SVM) 44
4. Artificial Neural Networks (ANN) 46

PROBLEM STATEMENT & OBJECTIVES 49

METHODOLOGY & SPECTRAL MEASUREMENTS 50

A. Soil Preparation & Sampling 51

B. Soil Contamination Process 51

C. UV VIS-Measurements 52

D. Preprocessing & Machine Learning Algorithms 54

E. Testing on Various Datasets 55

APPLICATION OF MACHINE LEARNING TO HYPERSPECTRAL IMAGING 57

A. Sorghum Plant Dataset 57

B. Salinas A-Scene Dataset 64

C. Paint Condition Assessment of Sydney Harbor Bridge 68

CONCLUDING REMARKS AND FUTURE WORK 75

REFERENCES 77

ILLUSTRATIONS

Figure

Figure 1 Atomic absorption spectroscopy instrumentation [23] 16
Figure 2 Inductively coupled plasma-atomic emission spectroscopy [24] 17
Figure 3 Hyperspectral Remote Sensing [31] 19
Figure 4 Original spectra of 30 soil samples [36] 22
Figure 5 Spectral reflectance of soils with different heavy metal concentrations [39, 47] .. 23
Figure 6 Spectral reflectance of wheat plants treated with four levels of Cu [5] 24
Figure 7 Correlation coefficient between spectral response and six levels of metal [5] 25
Figure 8 Raw reflectance spectra and smoothed spectra by the Savitzky-Golay filter [45]. 33
Figure 9 VNIR/SWIR spectrum recorded by the ASD pro FR portable spectroradiometer corresponding to soil samples with different clay content [58] 35
Figure 10 Machine learning model infrastructure [52] 40
Figure 11 Artificial neural network architecture [66] 47
Figure 12 Proposed framework for heavy metal estimation 50
Figure 13– Spectral reflectance of soil contaminated with multiple levels of Cr 52
Figure 14 Spectral reflectance of soil contaminated with multiple levels of Zn 53
Figure 15 Spectral reflectance of soil contaminated with multiple levels of Cr and Zn 53
Figure 16 Sorghum plant hyperspectral image [68] 58
Figure 17 Spectral reflectance of sorghum plant dataset 59
Figure 18 Confusion matrix for RF without normalization and using LDA 60
Figure 19 Confusion matrix for SVM without normalization and using LD 61
Figure 20 Pixel prediction using RF without normalization and using LDA 62
Figure 21 Pixel prediction using SVM without normalization and using LDA 62
Figure 22 Salinas A Scene hyperspectral image [69] 64
Figure 23 Confusion matrix for RF with normalization 65
Figure 24 Pixel prediction using RF with normalization 66
Figure 25 Ground truth 67
Figure 26 Prediction result for long-range image using Decision Trees 72
Figure 27 Prediction result for long-range image using Decision Trees 72
Figure 28 Prediction result for short-range image using Decision Trees 73
Figure 29 Prediction result for short-range image using Decision Trees 73

TABLES

Table

Table 1 Average concentrations of heavy metals in soil collected from Deir Kanoun dump and canal [12] 11
Table 2 The effect of different preprocessing methods on modeling accuracy by partial least squares regression model [59] 36
Table 3 Model performance statistics for selected bands [22, 26] 44
Table 4 Classification detection accuracy results of sorghum dataset 60
Table 5 Classification detection accuracy results of Salinas A-Scene dataset 65
Table 6 Model metric results for long-range images 70
Table 7 Model metric results for short-range images 70

ABBREVIATIONS

AAS	Atomic Absorption Spectroscopy
ANN	Artificial Neural Networks
CR	Continuum Removal
EF	Enrichment Factor
FD	First Derivative
ICP - AES	Inductively Coupled Plasma-Atomic Emission
KNN	K - Nearest Neighbors
LDA	Linear Discriminate Analysis
LG	Logistic Regression
MCARI	Modified Chlorophyll Absorption Reflectance Index
MLR	Multivariate Linear Regression
MSC	Multiplicative Scatter Correlation
NDVI	Normalized Difference Vegetation Index
PCA	Principal Component Analysis
PCR	Principal Component Regression
PLSR	Partial Least Square Regression
R^2	Coefficient of Determination
RF	Random Forests
RMSE	Root Mean Squared Error
RVI	Ratio Vegetation Index
SAVI	Soil Adjusted Vegetation Index
SD	Second Derivative
SG	Savitzky Golay
SNV	Standard Normal Variate
SVM	Support Vector Machines

CHAPTER I

LITERATURE REVIEW

A. Contamination in soil and impact on environment and health

Heavy metal contamination refers to the excessive accumulation of toxic heavy metals in the soil as a result of human activities [1]. With the evolution of the global economy over the recent years, the content and type of heavy metals in the soil and food crops have constantly increased, causing harm to the environment [2, 3]. Heavy metals are extremely pollutant to the environment as they can easily get enriched into the food chain. Once the soil suffers from heavy metal contamination, it becomes difficult to be remediated [1].

With the constant development of societies, heavy metal contamination has become a serious threat in the world. The excess metal contaminants found in the soil originate from several sources, including the extreme use of fertilizers and pesticides, sewage irrigation, mining activities, and untreated industrial solid waste [4]. Heavy metals in the soil include some significant metals of biological toxicity, such as Pb, Cd, Hg, As, and Cr. Other heavy metals of certain biological toxicity include Zn, V, Cu, Ni, and Sn. They all share a common feature of being harmful to human beings and the environment [5].

For example, Pb has several carcinogenic effects on human health. Pb toxicity causes intestinal cancer, lung cancer, and central nervous system [6]. Cd on the other hand, in addition to Pb, can reach the brain and cause Alzheimer's disease. Also, high levels of Cd can accumulate in the kidneys and the gastric system leading to gastric, breast, and lung

cancer, in addition to renal cancer. Moreover, As's carcinogenic effect touch the prostate glands and causes prostate cancer, leukemia, and liver cancer [6].

Heavy metals have a huge impact on soil microorganisms and enzymatic activity, plants, and humans. Enzymatic and microbial activity of the soil can sensitively affect the soil's quality and the microbial biomass played an important role as an indicator for determining the level of soil contamination [7, 8]. The process of organic matter decomposition and nutrient cycling is also highly related to the enzymes in the soil and the activities of enzymes decreased significantly with the increase of heavy metal contamination [9]. A low concentration of heavy soil metals doesn't disturb the plants' growth, however, when exceeded its tolerance threshold, the plant will be poisoned which might lead to its death [10]. Research also showed that heavy metals enter the human body through inhalation of dust, skin absorption, etc., damaging human health[1]. Humans' health is also affected indirectly by polluting food, water, and the atmosphere [11].

The severity of heavy metals lies within the fact that once discharged into the environment, they are difficult to remove, unlike organic pollutants which degrade to water and carbon dioxide. These metals persist in the environment by binding to soils and sediments until deployed by alterations in overlying vegetation, hydrology, and/or weather patterns [12]. Moreover, polluted soil act as a sink for organic compounds such as bisphenols, polycyclic aromatic hydrocarbons (PAHs), and phthalates, which have a high affinity to soil organic matter, low mobility, and high durability [13]. Furthermore, water resources and agricultural areas are affected by the presence of heavy metals and organic compounds, where the soil-crop system plays a major role in the exposure of humans to

contaminants [14, 15]. Soil pollutants reach humans through the food chain, where tissue accumulation and serious toxicity and diseases result after the consumption of crops with critical values of contamination [12]. For instance, higher risks of lung, skin, liver, kidney, and bladder cancer, in addition to cardiovascular diseases, are associated with high exposure to high levels of heavy metals [12].

B. Contamination in Lebanon

Lebanon is one of the developing countries that has been a worldwide concern when associated with heavy metals and organic compounds contamination, due to Lebanon's poor solid waste management [12]. Random disposal, open burning, and dumping of wastes across Lebanon resulted in the flow of leachates from hundreds of landfills into the soil and water resources [16]. Deir Kanoun Ras El Ain is one of the several Lebanese villages that have reported solid waste mismanagement for years. Household, medical, and industrial wastes of the dump in Deir Kanoun leachate influx into a running canal irrigating nearby agricultural lands [17]. Table 1 presents the average concentrations of heavy metals in soil collected from the dump and canal. Pb, Cd, As, and Hg were abundant in all of the collected samples from the dump and canal, with concentrations ranging from 504.3–1365 mg/kg, 77–131.1 mg/kg, 51–603.3 mg/kg, and 0.16–6.48 mg/kg, respectively [12].

Table 1 Average concentrations of heavy metals in soil collected from Deir Kanoun dump and canal [12]

Sample Site

Heavy Metal	**D1**	**D2**	**D3**	**D4**	**C1**	**C2**
Pb	504.3 ± 73.02	622.1 ± 82.89	1206.1 ± 80.35	1306.1 ± 12.62	1365 ± 31.5	728.4 ± 83.68
Cd	131.1 ± 32.21	118.7 ± 15.03	117 ± 24.44	77 ± 27.22	78 ± 9.46	104.4 ± 5.5
As	51 ± 15.54	471.97 ± 77.42	521.7 ± 74.68	95.87 ± 15.75	603.3 ± 91.08	281.6 ± 86.81
Hg	3.587 ± 1.27	0.459 ± 0.156	2.39 ± 0.416	0.16 ± 0.044	6.48 ± 0.761	0.38 ± 0.073

Baroudi et al [18] studied the distribution of organochlorine pesticides and heavy metals in Lebanese agricultural soil from six villages, namely Kobbet Choumra, Tal Mehyen, Mqaitaa, Al Mhamra, Qlaiaat, and Qaabarine, located in the plain of Akkar, North Lebanon. Results showed that Cd, Zn, and Pb led had high EF, which is a metric for determining how much the presence of an element in a sampling medium has increased relative to average natural abundance because of human activity greater than 1, with Cd being the highest among them all. This suggests that these heavy metals' sources are human activities, in particular, irrigation with contaminated water and local addition of fertilizers and pesticides containing Cd and Zn. Also, the concentrations of As, Cu, Ni, Pb, and Zn were found in the soils with averages of 4.1, 34.8, 85.51, 36.4, and 76.3 mg/kg, respectively [18].

Another study was performed by Darwish et al. [19] focused on Central Bekaa plain, which is considered the main region with prime agricultural land in Lebanon. In the Central Bekaa, farmers are obliged to irrigate with polluted water to compensate for the water shortage during peak crop demands. A preliminary study of contamination hazards of land resources showed that Cr and Ni were present in the soil at concentrations above tolerance levels [19]. The soil contamination by heavy metals threatened land resources, such as groundwater. The main factors contributing to this contamination were the intensive use of fertilizers, irrigation techniques entrenched by poor agricultural practices, and land use policy [19].

An assessment of heavy metal contamination of fertilizer products and phosphogypsum waste produced by the Lebanese Chemical Company was performed in 2002 [20]. The article cited that the presence of Cd, in addition to gamma-emitting radionuclides, at elevated concentrations in the produced fertilizers is of greatest concern as a result of its toxicity and ability to accumulate in soils and bio-accumulate in humans, animals, and plants. It can be readily available for uptake by a range of crops once it is present in the soil. In addition, Cd has no biochemical or nutritional function, adding to its high toxicity to humans and animals [20]. The study also showed that Cd is found as a fraction of the phosphorus available in fertilizers, and the addition of these fertilizers containing Cd impurities, in particular simple super phosphate, would lead to elevated Cd concentration in soil [20].

C. Common techniques used for soil characterization

In the past, soil contamination was not treated with the same degree of importance compared to water and air pollution [1]. This is because soil contamination often takes place on a wide range and is more difficult to be controlled and directed in comparison to air and water pollution [1]. However, metal contamination has been exceeding the environmental tolerance, thus causing ecological damage [1]. Also, some soils take one or two hundred years to be remediated, since the remediation costs are high and the remediation cycle is relatively long [21]. Therefore, heavy metal contamination became a hot topic of environmental protection worldwide. It is important to quantify and monitor the levels of heavy metals in soils and crops to discern the presence of heavy metal contamination and assess its level in the soil.

Soil contamination by heavy metals can be examined through various methods, after soil sampling, that is based on spectroscopy analysis performed in the laboratory. Below is a brief description of the characterization approach, highlighting the main steps involved and the associated challenges.

1. Study Area & Soil Sampling

Samples for the detection of metal contamination in soils are usually collected from areas located close to industrial, agricultural, or human activities that are suspected to pollute the soil through the heavy metal deposition. The number of soil samples and their relative metal contamination differs from one study area to another. Worldwide, samples are collected from depths of 10-20 cm and then sieved through a mesh of less than 2 mm

[1]. It is then processed and cleared up using a mixed acid such as HF, HNO_3, H_2O_2, etc. [1].

Tan et al. [22] collected 30 samples from a mining zone located 20 km north of Xuzhou, China. 10 samples were extracted from each site, and each sample is divided into 0 cm, 20 cm, 40 cm, or 60 cm interface layers [22]. The four samples are then mixed to form one sample and then divided into two parts after removing sundries, such as roots, leaves, stones, etc. One part was sent to the chemistry laboratory to measure the metal contamination using traditional methods, and the other part was sent to the darkroom for soil spectral signatures measurement [22]. Kemper et al [23] examined an area 40 km west of Seville, where a tailings dam collapsed at the Aznalco'llar Mine. A total of 214 soil samples were taken at four different levels (0-2, 2-20, 20,40, 40-60 cm) from six test sites [23]. In addition to the collected samples, 32 artificially contaminated soils were produced, but not affected by the accident, and sludge was added in increasing weight percentages [23]. Sayyed et al [3] studied Chitgar industrial area, which is an area under intense human interference, industrialization, and growing urbanization. Five sampling points in a grid of 0.5x1km at each sampling stating were selected and a total of 70 topsoil samples (0 – 15 cm) were taken from each sub-cell. Three soil samples were taken, to have the background heavy metal concentration levels, at a depth of 100 cm to make sure that they are at least affected by anthropogenic activities [3]. These samples were brought to the laboratory after being placed in clean polythene bags [3].

2. Chemical Analysis

Determining the heavy metal concentration in soils and food crops is a crucial step. It is required to develop statistical models to relate the spectral signatures to the heavy metal content and then allow for future predictions of contamination levels based on the spectral signature of the samples. Several methods for determining the metal soil's contamination in the laboratory are widely used, including atomic absorption spectroscopy (AAS), inductively coupled plasma-atomic emission spectroscopy (ICP-AES), and inductively coupled plasma-mass spectroscopy (ICP-MS) [1].

After air drying the samples, removing the debris and other objects, and grounding on mortar to remove aggregates and lumps, soil samples were passed through a 2mm sieve to collect granulometric fraction [3]. To extract and separate the heavy metal (analyte) from the sample matrix, wet acid digestion is commonly used. Sayyed et al [3] took 5g of sample into a 300ml polypropylene jar and distilled water was added to make a total of 200ml. The solution was then acidified with 10ml HF, 5ml $HClO_4$, 2.5ml HCL, and 2.5ml HNO_3 [3]. The solution was stored and heavy metal constituents were analyzed by AAS [3]. AMA254, an atomic absorption spectrometer (AAS), is constructed specific for mercury detection [23]. As and Cd were analyzed in solution by AAS with hydride generation and AAS in a graphite furnace respectively, obtained by microwave acid digestion [23].

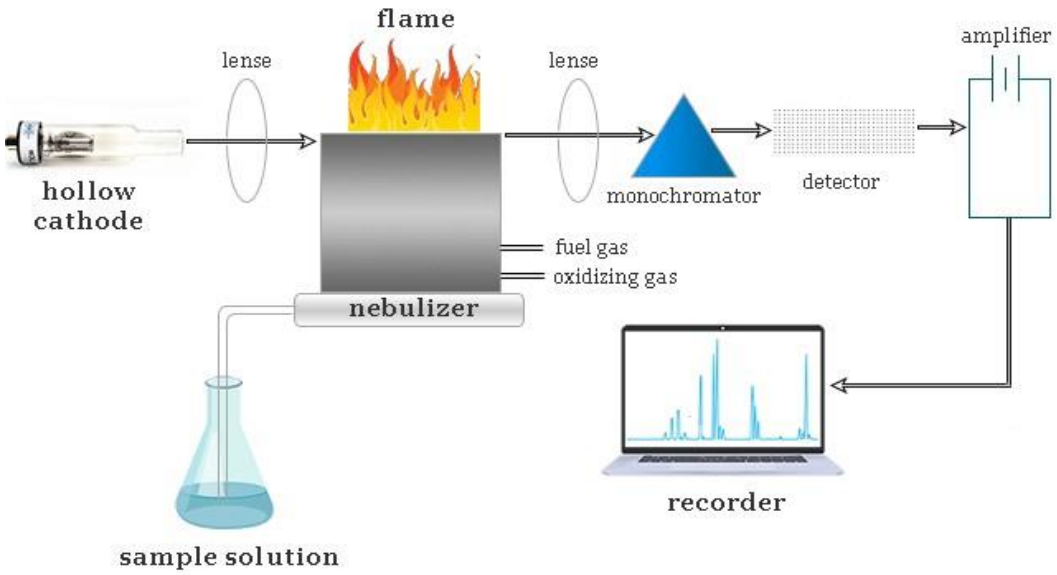


Figure 1 Atomic absorption spectroscopy instrumentation [23]

As for ICP-AES and ICP-MS, soil samples are first dried at 105°C for 3 h [24]. Around 0.2 g of samples are accurately weighted to then be placed in a pressure vessel. After adding 4 ml of concentrated nitric acid and 0.5 ml of concentrated hydrofluoric acid, the samples were acid digested using microwave power [24]. The samples were then diluted with 100 ml of water after cooling. For ICP-AES analyses, a Jobin -Yvon 38 spectrometer was used. As for ICP-MS, 20 µg of Rh per liter was added as an internal standard, and an ELAN 600 instrument was used for samples analysis [24].

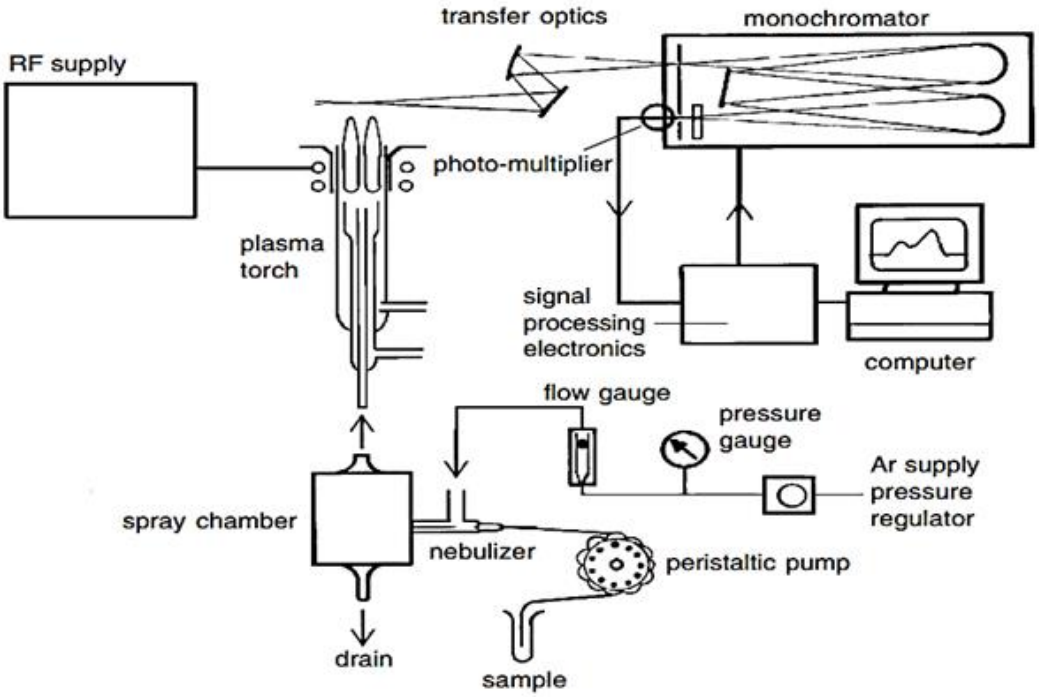


Figure 2 Inductively coupled plasma-atomic emission spectroscopy [24]

D. Hyper Spectroscopy

An informative way to work with materials, identify them or define their properties is to study how light interacts with them. This study of light interaction with material is called spectroscopy [25]. Spectroscopy examines how light behaves in the target and recognizes materials based on their different spectral signatures. These spectral signatures can be identified from the spectrum of the material. The spectrum describes the amount of light in different wavelengths and shows how much light is emitted, reflected, and transmitted from the target [26].

Remote sensing-based-reflectance spectroscopy, referred to as hyperspectral imaging, allows for rapid and inexpensive quantitative mapping by remotely capturing the electromagnetic radiation reflected by the target [25, 26]. Hyperspectral imaging has found applications in detecting petroleum hydrocarbons, landmine detection, and detecting heavy metals in soils and vegetation [27-29]. A successful quantification based on hyperspectral images could allow the quantification through airborne mapping using new hyperspectral sensor techniques and machine learning algorithms, and overcome the qualitative mapping of heavy metals in soils [30].

The vision of the human eye is based on the three basic colors; red, green, and blue, which is considered a small range on the electromagnetic spectrum, ranging from 400 nm to 700 nm. Hyperspectral imaging (HSI), on the other hand, is one of the optical analysis techniques that achieve data in the form of hundreds of adjoining spectral bands [31]. Hyperspectral imaging is a technique that collects spatial data (x and y coordinates) and spectral data (λ) of the sample image to form a multidimensional hypercube [32]. Figure 3 further explains how hyperspectral cubes are formed from hyperspectral images, where each image is a combination of pixels.

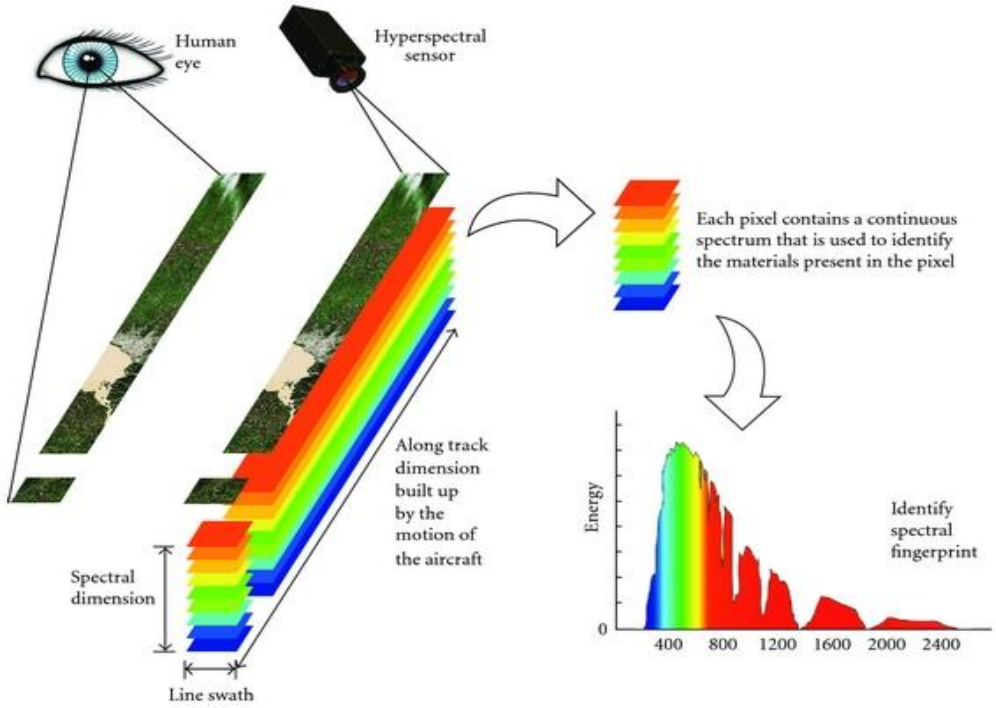


Figure 3 Hyperspectral Remote Sensing [31]

After a constant spectrum for imagery pixels is obtained using collective data, imagery spectrums are analyzed with field reflectance or laboratory spectra, allowing for the detection and mapping of materials, objects, chemical components, etc. in an area of interest [31, 32]. For example, for plant hyperspectral imaging, the typical spectrum ranges from ultraviolet up to short-wave infrared (from UV ~ 250 nm to SWIR~ 250 nm). To increase the coverage of the spectrum, several sensors are usually combined, where each captures a sub-range [33].

The spectral responses of heavy metals in soils and food crops are generally measured using a hyperspectral ASD FieldSpec3 or ASD FieldSpec4 meter over the full

visible-mid-infrared- (MIR) region ranging from 350 nm to 2500 nm at a 1 nm resolution [34-37]. The measurement can be taken either in situ or indoors. Measurements taken in situ undergo minimal disturbance, however a variety of factors such as solar radiation and ambient conditions may influence the measured reflectance [5]. By taking the sample to the laboratory where the spectral reflectance can be measured in a controlled environment, this problem can be avoided [5]. The collected soil and vegetation must be homogenized to fine powders before any measurement to remove lumps and aggregates [38]. Several measurements are taken for each sample and the average is used as the final result.

These measurements are crucial as they form the foundation of hyperspectral sensing of heavy metals as a close correlation between hyperspectral reflectance of heavy metals and heavy metal concentration has been reported [39, 40]. The high degree of correlation reflects the potential of using hyperspectral sensing methods to determine the type and degree of contamination damage [5].

The soil act as a basin for multiple metals to co-exist at the same time. To isolate the spectral response of a given metal, the plant undergoes a series of predefined amounts of metals added to its pot [41]. For example, the reflectance of Salicornia Virginica showed sensitivity to early stress levels for Cd after being treated with two metals [42]. Nevertheless, different metals have their unique spectral peculiarity over different wavelengths. A decrease in the spectral response near 540 nm and NIR 750-1400 nm bands are associated with a higher Cd concentration [5]. At the absorption depth of 500 nm, all metals are correlated positively due to their intrinsic absorption at low concentrations [43].

A variety of variables can affect the spectral behavior of heavy metals in plants, of which the internal structure, chlorophyll, and water content are the most important [5]. Soil organic matter, total iron, and clay minerals have a critical impact on the soil's spectral response. The signal of the soil constituents such as heavy metals is overwhelmed by a reduced reflectance response if the soil organic content exceeds 2% [5]. In addition, since metals co-exist together in the soil, interference of spectral responses between different metals occurs. Thus, it may not be possible to extract the unique response of a particular metal of interest [5].

There is no universal answer for which spectral band(s) is the most effective for heavy metals' detection in soils as every metal is sensitive over a specific wavelength. For example, wavebands centering around 838, 1930, and 2148 nm are considered the best for studying soil Pb) content [44]. Bands centering around 460, 1400, 1900, and 2200 nm are sensitive to As and Cu in a mining area [45]. The absorption peaks at 700 nm and 870 nm are the best for detecting Fe and Fe_2O_3, while 1140-1200 nm bands are best for studying soil mercury (Hg) [23, 46]. The original spectra after averaging are shown in Figure 4.

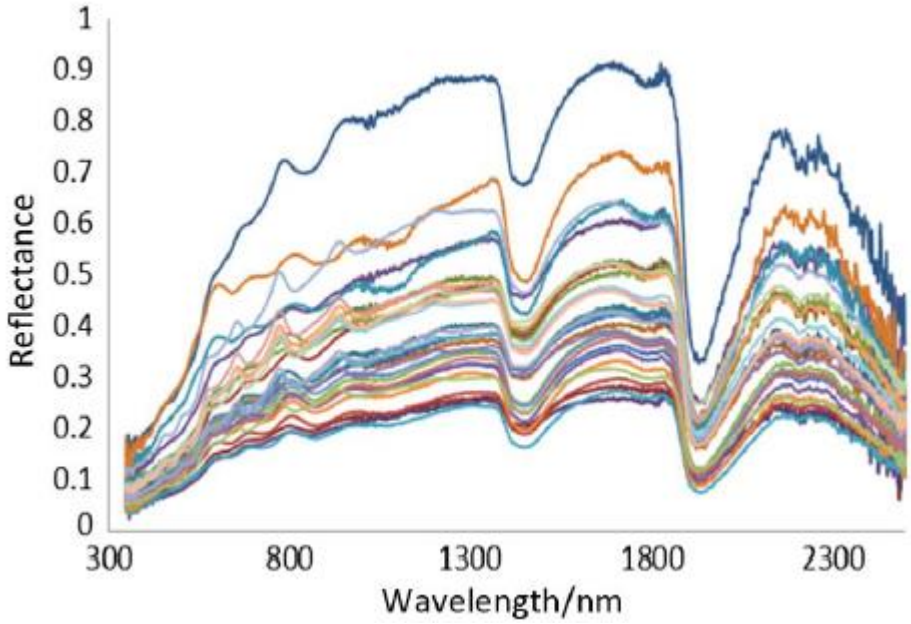


Figure 4 Original spectra of 30 soil samples [36]

Figure 5 shows the spectral changes of the 39 soil samples with different heavy metal contaminations examined by Zhang et al [47]. The overall trend and shape of the spectral curve are similar, with variations in the numerical reflectivity of different heavy metal concentrations. The spectral reflectance follows a rising trend with gradual stability of the near-infrared band as wavelengths approach the visible wavelength band [47]. The spectral reflectance of the spectral feature bands were primarily detected at 590-623, 810-830, 970-1025, 1389-1417, 1613-1641, 1725-1753, 1921-1949, and 2229-2257 nm [47]. Excluding the 900 nm band, molecular metal elements with hydroxyl groups, such as Fe-OH, AI-OH, and Mg-OH are assigned to the near-infrared region [47]. The peak at about 1000 nm is clearly observable, which is the absorption band of hydroxide Fe^{3+}. The

vibration of CO_3^{2-} groups in soil carbonates generated the absorption peak near 2455 nm [48].

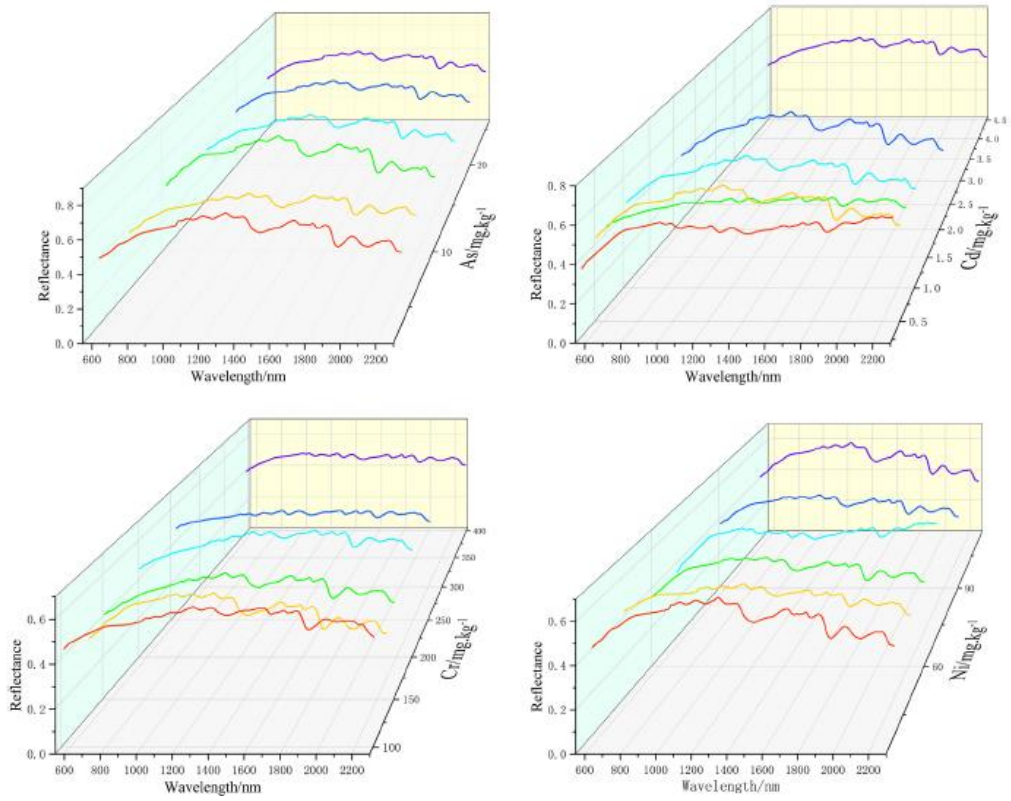


Figure 5 Spectral reflectance of soils with different heavy metal concentrations [39, 47]

Spectral reflectance of vegetation has also been used intensively and primarily to detect and determine heavy metal contamination. For the detection of Lead (Pb), Zinc (Zn), Copper (Cu), and Arsenic (As) in paddy leaves, the most sensitive bands lay around 460, 560, 660, and 1100 nm [5]. The reflectance at 690-1300 is best for monitoring Cadmium (Cd) in Rapa Chinensis leaves, with the optimal wavelength being 782 nm [49]. Spectral sensitive bands to Pb contamination are mainly distributed over 450, 550, 670, 760, and 1240 nm [5]. Heavy metal-contaminated plants are best studied over the following bands:

554, 631, and 557 nm [5]. The reflectance of plant leaves and their metal content are negatively but linearly correlated at 1240 nm [42]. Therefore, a wavelength that is optimal for the detection of metals in soils may not necessarily be the best for detecting the same metals in vegetation. As illustrated in Figure 6, the spectral reflectance of wheat and rapeseed plants after treatment of Cu and Ni is negatively correlated with the metal content [5].

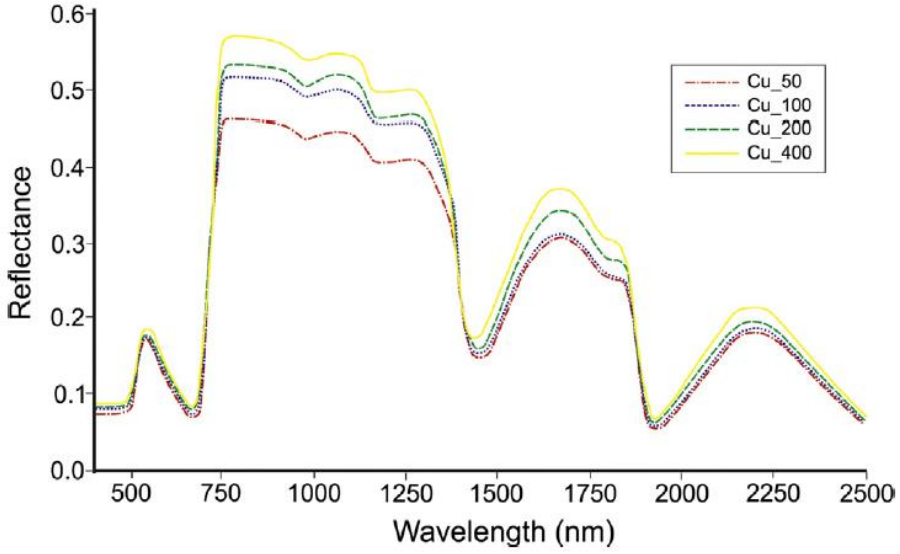


Figure 6 Spectral reflectance of wheat plants treated with four levels of Cu [5]

Indicating the most useful and sensitive spectral bands is achieved through calculating the correlation between the optimized spectral reflectance (Figure 7) or their transformed indices and the in situ measured heavy metal content. Estimation models are constructed based on the wavelengths that result in the highest correlation being candidate variables [5]. The correlation can also be calculated for the derivative spectra instead of the

raw spectral reflectance. Any parameter, that is closely correlated with the metal concentration measured in situ, can be an explanatory variable in the estimation model needed to be constructed.

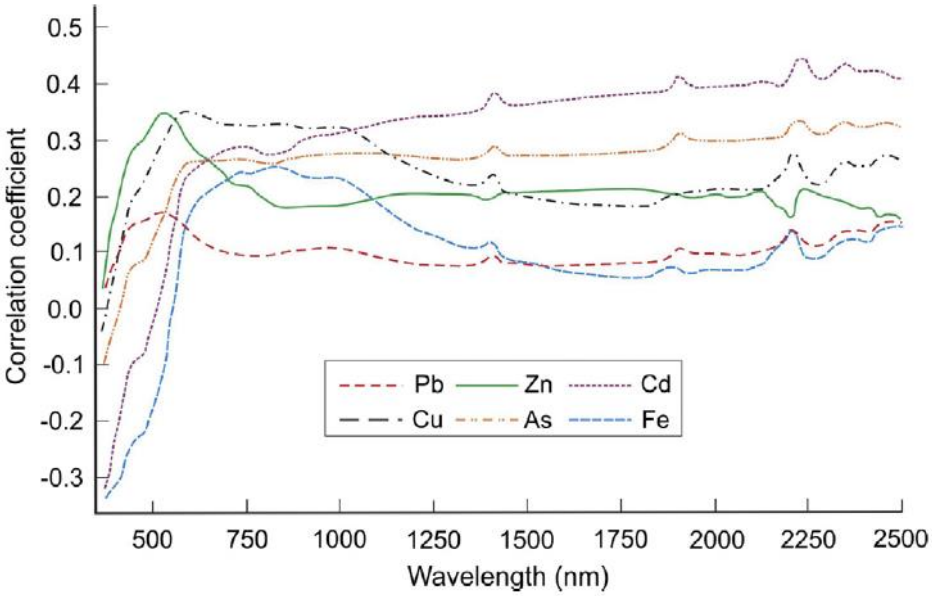


Figure 7 Correlation coefficient between spectral response and six levels of metal [5]

Compared to bare soils, detecting heavy metals in food crops and vegetation is significantly more difficult as vegetation influences the captured reflectance. For that reason, vegetation is generally used as a proxy, mostly through changed chlorophyll structure and content, abnormal growth, and other altered conditions [5]. For example, the close correlation between rice leaf chlorophyll content and Cu, As, Cr, and Cd content in soil suggests that rice leaves are the perfect surrogate for heavy metal contamination studying, since the increase of rice leaf's C/N ratio is due to the increase of soil Cd concentration [5]. Different responses in vegetation are provoked by various metals, where some responses are conductive to pigment synthesis while others are destructing the

chlorophyll structure, resulting in a high NIR reflectance [50]. Metals that have no role in plant growth, such as Pb, impact the water absorption of plants and thus the normal synthesis of pigments.

As previously mentioned, not all metals are spectrally responsive to radiation. Compared to water and organic matter, heavy metals are microscopic components that may have an extremely low response or not be spectrally detectable. If the soil's metal concentration is less than 4.0 mg g^{-1}, it is impossible to detect the spectral response of some metals [26]. In addition, vis-NIR and mid-IR radiation are not absorbable by pure metals. Nevertheless, the co-variation between spectrally indistinctive and spectrally active metals, such as Fe oxides and clay, allows for the possibility of sensing non-detectable metals. For example, Ni, Cu, As, Hg, Zn, Cr, Cd, and Pb in soils are correlated with organic matter [5]. A closer correlation results in higher estimation accuracy. However, the established relationships are influenced by seasonal influences, increased levels of spectral noise, and species heterogeneity. Thus, each correlation is unique for the site in which data are collected. Also, this method is not feasible for metals that establish no correlation with other spectrally responsive metals, such as between Fe and As [51].

On the other hand, the use of hyperspectral imaging as a remote sensing tool is accompanied by challenges, especially with dealing with data. The application of HIS relies on sophisticated and complex data analysis methods [51]. These difficulties appear as a result of the high dimensionality and size of the hyperspectral data cube. In addition to that, the measurement process associates to it spectral mixing (linear and nonlinear) and degradation mechanisms, such as noise and atmospheric effects Also, limitations in the

application of the learning methods and finding an optimal solution are considered the main challenges in this field [52].

CHAPTER II

MACHINE LEARNING APPLICATION FOR SOILS CONTAMINATION DETECTION BASED ON HYPERSPECTRAL DATA

A. Introduction

Machine learning methods have been very successful in hyperspectral analysis tasks by automatically finding the correlation between the reflectance spectrum and the desired data, and being robust against external noises and disturbances [52]. In comparison with traditional methods such as AAS and ICP, studies have shown that machine learning techniques can perform better results [53]. Machine learning algorithms can better handle external disturbances and spectral and ground truth variability when compared to traditional methods [54]. Moreover, machine learning techniques allow for high dimensionality data handling using band selection and dimensionality reduction techniques [55]. For those reasons, in addition to the ability to produce more accurate results, machine learning algorithms allow for the construction of models capable of identifying highly sensitive bands and relating their reflectance to the soil heavy metal concentrations in the field of hyperspectral imaging.

Obtained hyperspectral data are disturbed by noises, so they undergo a series of optimization techniques, such as spectral derivatives and vegetation indexing, to enhance their signal features and suppress noises. It was noticed that at a sufficiently high metal content, only a slight disparity existed in the spectral responses of certain metals. However, not all metals have their own unique spectral reflectance signature. Detecting them relies on

their coexistence with other spectrally responsive metals or organic matter in soils or food crops, as their presence influences the other metals' spectral response. The heavy metal concentration in soils is usually low, so it might not have obvious spectral characteristics [22]. For this reason, it is necessary to preprocess the data to reflect the weak information. Several preprocessing methods are widely used, such as SG, FD, SD, SNV, and CR [22]. These methods can smooth the spectra, suppress the noise in data acquisition, and eliminate errors caused by the instrument, thus enhancing the weak spectral data related to heavy metals [22].

Preprocessing the data is usually followed by developing an estimation model from relevant explanatory variables. Various analytical methods and machine learning algorithms have been investigated to relate the metal concentration in soils and food crops to their hyperspectral responses from hyperspectral imaging. The most common mathematical methods used for the prediction of different soil and vegetation constituents include MLR, PLSR, and PCR. As for machine learning algorithms, SVM, RF, and ANN are the ones widely used. However, researchers are always investigating new learning algorithms that might offer better generalization performance and higher accuracy [22]. Therefore, hyperspectral imaging for the detection of heavy metal contamination in soils and food crops offers a fascinating option to deal with the related environmental concerns.

B. Preprocessing

Since the environment and instruments produce irrelevant information, spectral noises, and overlapping peaks, optimization of the measured spectral data is required. In addition, as mentioned before, some heavy metals are present in low concentrations, thus almost impossible to be spectrally detected. Therefore, it is necessary to implement some basic preprocessing steps, along with several spectral pretreatments such as transformations and their derivatives.

Signal preprocessing is divided into two parts; Spectral Derivatives and Vegetation Indexing. The most common preprocessing methods in spectral derivatives are SG, FD, SD, SNV, and CR. One can possibly choose a single preprocessing method or a combination of techniques. However, this decision will influence the outcome predicted concentration of the heavy metal from the estimation model. So, it is crucial to choose the appropriate method of preprocessing the data. To achieve a successful heavy metal concentration retrieval, it is also essential to choose the bands that are most sensitive to heavy metal content. This selection is critical specifically in hyperspectral data due to the immense redundancy among hundreds of available bands [5].

1. Scatter Corrections

Under scatter corrections, the most commonly used techniques are MSC, SNV, and normalization. These methods reduce the physical variability between samples due to scatter and adjust for baseline shifts [56].

a. Multiplicative Scatter Correction (MSC)

The most widely preprocessing technique for NIR (closely followed by derivation and SNV) is Multiplicative Scatter Correction. MSC removes artifacts or imperfections from the data matrix before modeling the data [56]. Implementing MSC is as follows:

i. Estimation of the correction coefficients (additive and multiplicative contributions)

$$\boldsymbol{x}_{org} = b_0 + b_{ref,1}.\boldsymbol{x}_{ref} + \boldsymbol{e} \tag{1}$$

ii. Correcting the recorded spectrum

$$\boldsymbol{x}_{corr} = \frac{\boldsymbol{x}_{org} - b_0}{b_{ref,1}} = \boldsymbol{x}_{ref} + \frac{\boldsymbol{e}}{b_{ref,1}} \tag{2}$$

where $\boldsymbol{x_{org}}$ is one original sample spectra measured by the NIR instrument, $\boldsymbol{x_{ref}}$ is a reference spectrum used for the preprocessing of the entire datasheet, $\boldsymbol{e}$ is the un-modeled part of $\boldsymbol{x_{org}}$, $\boldsymbol{x_{corr}}$ are the corrected spectra, and b_0 and $b_{ref,1}$ are scalar parameters.

b. Standard Normal Variate (SNV)

The second most common method for scatter correction of NIR/NIT data is Standard Normal Variate preprocessing [56]. Due to the similarities between MSC and SNV, the format of SNV is as follows:

$$\boldsymbol{x_{corr}} = \frac{\boldsymbol{x_{org}} - \boldsymbol{a_0}}{\boldsymbol{a_1}} \tag{3}$$

where a_0 is the average value of the sample spectrum to be corrected and a_1 is the standard deviation of the sample spectrum.

2. Spectral Derivatives

a. Savitzky-Golay (SG)

Spectral derivatives techniques allow for the removal of overlapping spectra and multiplicative effects in the spectra. Savitzky-Golay method smoothes the spectra after reducing random noise. It also eliminates the influence of numerical jumps of adjacent bands [57]. Savitzky and Golay (1964) used simplified least-square fit convolution for smoothing and computing derivatives of a set of consecutive values (a spectrum) [57]. The general equation can be represented as follows:

$$Y_j^{*} = \frac{\sum_{i=-m}^{m} C_i Y_{j+1}}{N} \tag{4}$$

Where Y is the original spectrum, Y^{*} is the smoothed spectrum, C_i is the coefficient for the ith spectral value of the filter (smoothing window), and N is the number of convoluting integers. The index j is the running index of the original ordinate data table [57]. Deciding on the number of points used to calculate the polynomial (the window) and the degree of polynomial fitted is a significant task. The derivative with the highest order is dependent on the degree of the polynomial used in the fitting. For instance, a second-order polynomial can be utilized to approximate the second-order derivative. Ren et al [45] smoothed all the spectral data by the Savitzky-Golay filter with a second order polynomial fit and window size of 11. The filter reserved general trends such as heights and widths

with peak values while eliminating noise. Figure 8 shows both raw and smoothed spectra from 1950 to 2500 nm [45].

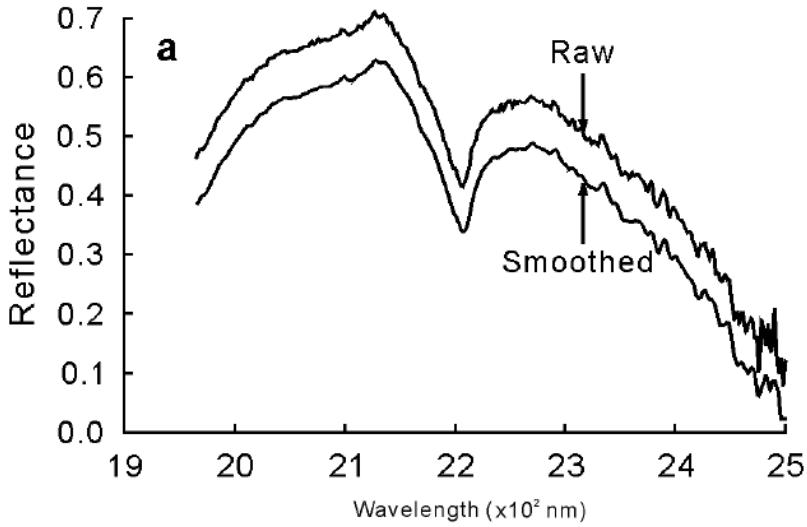


Figure 8 Raw reflectance spectra and smoothed spectra by the Savitzky-Golay filter [45]

b. <u>First Derivative (FD) & Second Derivative (SD)</u>

The first & second derivative is a technique suggested and elaborated on by Norris and Williams in 1984 (also called Norris-Williams derivation) [56]. This basic method is developed to avoid noise inflation in finite differences and calculate the derivative of NIR/NIT spectra. The derivation is achieved by performing a two-step procedure [56]:

i. <u>Smooth the spectra while averaging over a given number of points:</u>

$$x_{smooth,i} = \frac{\sum_{j=-m}^{m} x_{org,i+j}}{2m+1} \tag{5}$$

where m is the number of points in the smoothing window centered around the current measurement i.

ii. Calculate the first derivative by taking the difference between two smoothed values having a gap size greater than zero. For the second derivative, take twice the smoothed value at point i and the smoothed value at gap distance on either side:

$$x_i' = x_{smooth,i+gap} - x_{smooth,i-gap} \tag{6}$$

$$x_i'' = x_{smooth,i-gap} - 2.x_{smooth,i}x_{smooth,i+gap} \tag{7}$$

c. Continuum Removal (CR)

The CR method is highly effective in highlighting the absorption, emission, and reflection characteristics of a spectral curve. It quantifies the absorption of materials at a certain wavelength presuming that no other heavy metal has stronger absorption features over this specific wavelength [58]. The continuum is approximated by joining two local reflectance maxima placed on both shoulders (λ_{min} and λ_{max}) of the peak absorption wavelength $\boldsymbol{\lambda_{peak}}$ by a straight line [58]. Figure 9 shows the VNIR/SWIR spectrum recorded by the ASD portable spectroradiometer corresponding to a soil sample that has clay and $CaCO_3$ and the VNIR/SWIR HYMAP spectrum for the location of this soil sample.

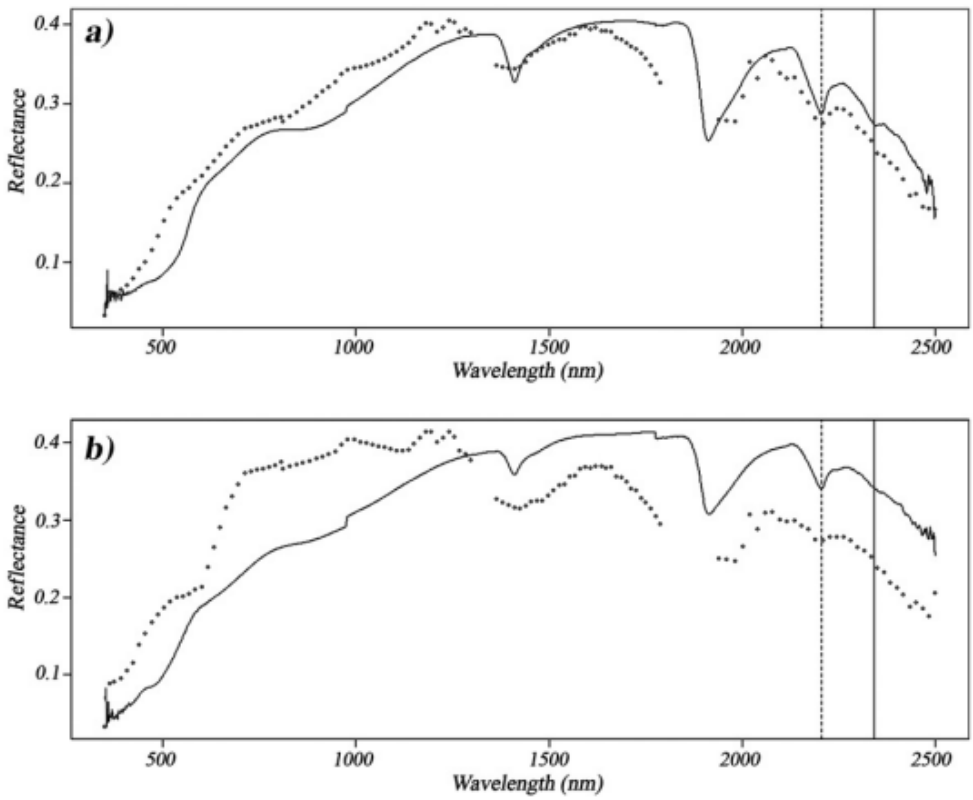


Figure 9 VNIR/SWIR spectrum recorded by the ASD pro FR portable spectroradiometer corresponding to soil samples with different clay content [58]

Yousefi et al compared the effect of five spectral preprocessing methods on Cu-Ni modeling accuracy via the PLSR method. Table 2 displays the applicability of using a series of preprocessing techniques on two heavy metals and compares the results (accuracy) of these methods [59].

Table 2 The effect of different preprocessing methods on modeling accuracy by partial least squares regression model [59]

Pre-processing Method	Cu				Ni			
	Calibration		Validation		Calibration		Validation	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
Non-Pre-processing	0.826	0.004	0.684	0.0088	0.841	0.0015	0.882	0.0012
Savitzky - Golay filter	0.938	0.0026	0.825	0.0046	0.898	0.0011	0.905	0.0012
Standard Normal Variate (SNV)	0.97	0.0018	0.396	0.0084	0.887	0.0012	0.778	0.0017
First Derivative + Savitzky - Golay filter	0.946	0.0025	0.694	0.0056	0.949	0.0007	0.747	0.0022
Second Derivative + Savitzky - Golay filter	0.573	0.007	0.302	0.01	0.979	0.0005	NA	0.0038

C. Vegetation Indexing

Although spectral preprocessing eliminates some of the noises, however, these methods can't differentiate the slight spectral differences in metal-induced vegetation conditions. The optimization of this disparity is fulfilled through vegetation indexing [5]. Vegetation indexing is essential especially when plants are being studied. This approach suppresses noises coming from environmental conditions and enhances signals related to vegetation [5]. Due to their sensitivity to structural changes, vegetation indices play the role of an indicator of the level of metal accumulation in food crops and vegetation. Ratio Vegetation Index (RVI), Normalized Difference Vegetation Index (NDVI), Modified

Chlorophyll Absorption Reflectance Index (MCARI), and Soil Adjusted Vegetation Index (SAVI) are the most common indices proposed.

All vegetation indices utilize at least two wavebands, with each having a close correlation with the dependent variable. Depending on the metal of interest and its level of concentration, the same index can contain bands of different wavelengths [5]. Each index has its particular strengths in heavy metal detection. For example, Ren et al [45] used R510 and R810 for Pb and As, R660 and R870 for Cu, and R510 and R870 for Zn, to derive normalized vegetation indices. Various spectral indices are also effective for the assessment of heavy metal contamination in agricultural soils. For instance, the ratio vegetation index is used for chlorophyll detection, and in particular, the ratio R725/R675 is considered an influential index for observing heavy metal toxicity due to its relationship with several metals [5].

D. Feature Extraction and Dimensionality Reduction

Feature extraction is the process of transforming the data from a high-dimensional space to a lower dimensional space while conserving the information in the data as much as possible. For the aim of reducing the computational time and avoiding the problem of the high spectral resolution of the image in comparison to the low number of training samples, the feature extraction technique is used in hyperspectral imaging analysis [28]. Several feature extraction methods are examined, where some are linear, and others are nonlinear.

Principle Component Transformation (PCT) and Linear Discriminant Analysis (LDA) are well-known feature extraction algorithms that will be discussed further.

1. Principle Component Analysis (PCA)

Principal Component Analysis, sometimes referred to as principal component transformation or Karhunen-Loeve transformation, is a method based on the minimization of representation error that reduces the dimension of the space. This method is based on choosing the bands that are most sensitive based on the eigenvalue decomposition of the covariance matrix of the hyperspectral image [60]. As a first step of implementing PCA, one must calculate the covariance matrix of the image matrix. This is followed by the calculation of the eigenvalues of the covariance matrix. Finally, after extracting the eigenvectors, the image matrix is projected onto a new subspace constituted by the k orthogonal eigenvectors corresponding to the highest eigenvalues. This is performed through:

$$Y = W^T x \tag{8}$$

where Y is the new $\boldsymbol{k}$ x 1-dimensional sample in the new subspace, W is the transformational matrix of k orthogonal eigenvectors, and x is a d x 1-dimensional vector representing one image pixel.

In theory, the PCT transformation influences the hyperspectral image classification. However, the comprehensive influence of the classification does not affect the general class patterns, thus the dominating classification result remains true.

2. Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis, also named Discriminant Analysis Feature Extraction (DAFE), is a statistical-based technique used for dimensionality reduction and feature extraction. By maximizing class discrimination and minimizing the within-class distance, LDA evaluates an optimal transformation, thus achieving maximum class discrimination [61]. The first step in computing the LDA method is calculating the within-class distance, total scatter, and between-class matrices. Then, by applying the eigenvector decomposition on the scatter matrix, define and compute the transformation matrix.

One major disadvantage of this method is that it requires the scatter matrix data to be nonsingular. Some other drawbacks include a large number of training samples to ensure the reliability of estimation between the between-class and within-class of the scatter matrix, the maximum number of features extracted must be equal to the of classes minus one, and the requirement for more training samples for hyperspectral images to calculate the class statistical parameters at full dimension [28, 62]. It is also a time-consuming technique compared to other methods.

E. Estimation Models

In hyperspectral sensing of heavy metals in soils and food crops, it is essential to develop an estimation model that predicts the heavy metal contamination. The model has the metal concentration as the dependent variable while the spectral signatures or its

transformed indices (or plant growth variables) are the explanatory variables [63]. There is no restriction on the number of variables included in the model, as long as a higher estimation accuracy is achieved (e.g., a high R^2 value).

The empirical relationship between the dependent and explanatory variables (between soil metal content and spectral responses) is achieved through several statistical methods and machine learning algorithms. The most common statistical methods are Multivariate Linear Regression (MLR) and Partial Least Square Regression (PLSR). As for machine learning algorithms, the most efficient and accurate algorithms used are Random Forest (RF), Support Vector Machines (SVM), and Artificial Neural Networks (ANN).

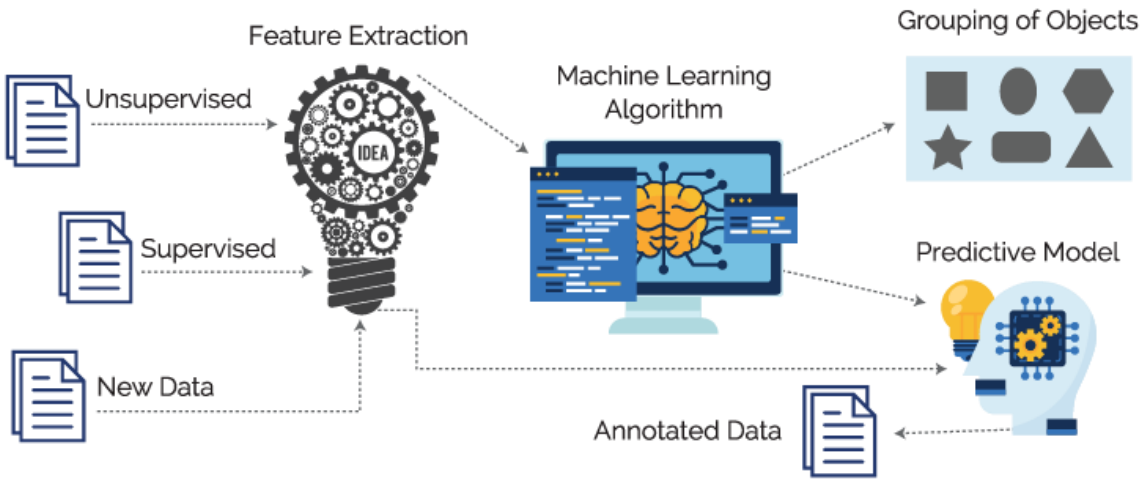


Figure 10 Machine learning model infrastructure [52]

1. *Multivariate Linear Regression (MLR) & Partial Least Square Regression (PLSR)*

Multivariate Linear Regression (MLR) is a statistical method that allows for the prediction of an outcome or response variable based on several explanatory or independent variables. Multivariate analysis is usually used to develop the estimation model from the measured spectral reflectance. Multivariate statistics can detect pedogenic elements such as As and Mn, and anthropogenic elements such as Cr, Cu, Hg, Ni, Pb, and Zn [5].

A particular form of the MLR is the Partial Least Square Regression is a particular form of Multivariate Linear Regression. PLSR is based on the assumption that a linear combination of explanatory variables can estimate the dependent variables [5]. PLSR is considered a combination of correlation analysis, principal component analysis (PCA), and a sum of regression analysis. A score matrix $\boldsymbol{S}$ and a loading matrix $\boldsymbol{P^T}$ plus an error matrix $\boldsymbol{E}$, which both derived from PCA, constitute the input matrix $\boldsymbol{X}$ comprising m explanatory variables with a total of n samples.

$$X = SP^T + E \tag{9}$$

Similarly, the response or output matrix $\boldsymbol{Y}$ has p variables of n samples. Also, $\boldsymbol{Y}$ constitutes of a score matrix $\boldsymbol{U}$ and a loading matrix $\boldsymbol{Q^T}$plus the error matrix $\boldsymbol{F}$.

$$Y = UQ^T + F \tag{10}$$

where $\boldsymbol{S}$ and $\boldsymbol{P}$ have a dimension of n x k and m x k ($k, l =$ number of principal components for reconstructing $\boldsymbol{X}$ and $\boldsymbol{Y}$), respectively. $\boldsymbol{U}$ and $\boldsymbol{Q}$ are of dimensions n x 1 and p x 1 respectively. Both $\boldsymbol{E}$ (n x m) and $\boldsymbol{F}$ (n x p) are residual matrices.

PLSR is feasible for detecting soil metals using hyperspectral reflectance measurements due to the close relationship between surface soil metal concentration and reflectance spectroscopy, as it maps soil heavy metal abundance [35]. PLSR regression coefficients can point out the critical wavelengths for the prediction of heavy metal contamination [26]. A relationship between the spectrally active clay (and soil organic matter) and their reflectance spectra in the VNIR region allows for the detection of Cd and Zn [34]. It also allows for Pb level retrieval from hyperspectral data. Zhou et al [5] found that the PLSR estimation model ($R^2 = 0.984$) is the most accurate compared to the third-order polynomial ($R^2 = 0.511$), linear ($R^2 = 0.488$), logarithm ($R^2 = 0.460$), and power models ($R^2 = 0.379$).

Several comparative studies have been done on the performance of MLR and PLSR regarding their ability to develop estimation models. In general, both methods were found able to increase the estimation accuracy. However, the metal under study plays an essential role in the pace of improvement [5]. For example, MLR is better for Ni detection, but PLSR is better for the other six metals. Nevertheless, although MLR models are too dependent on training samples, they are more accurate than singular models [5]. Besides, even though both PLS and the narrow band vegetation index MSAV12mm are highly accurate when studying heavy metal contamination in soils, MSAV12mm outperformed the best PLS model [26]. PLSR was also found ill-adapted to unknown samples [5].

2. Random Forest (RF)

The Random Forest algorithm is a predictive modeling algorithm based on classification and regression trees and the bagging learning strategy. In RF, a decision tree is randomly generated from a fixed-size subset of all attributes giving the advantage of a reduced computational cost [22]. To generate a fixed number of subset training samples, random sampling is repeated K times. A classification or regression tree is generated by randomly selecting subsamples with their corresponding attributes, where all the generated trees form a forest [22]. Finally, based on the scores of the class voting from all trees, the results are obtained.

The trained forest $\hat{F}_{RF}^{K}(x)$ with K trees can be given as:

$$\hat{F}_{RF}^{K}(x) = \frac{1}{K}\sum_{K=1}^{K} T(x_s) \tag{11}$$

where $T(x)$ is a single tree, x is all the training samples, and x_s is each tree's training sample data obtained with the bootstrap sampling method.

Random Forest is one of the popular machine algorithms because of its robustness, ease of use, and relatively high accuracy. RF is sometimes preferred over other machine learning methods due to its interpretation of results [22]. Its essence lies in the enhancement of the decision tree algorithm and its ability to handle many input variables. Nevertheless, even when missing data are present, the tree structure can deal with missing sample attributes, and high accuracy is still maintained [64]. For all these reasons, RF is found preferable for the processing of hyperspectral data. With more trees being used, we can improve the attribute subset to cover all of the property space and prevent overfitting [22].

Tan et al [22] compared the accuracies of different estimation models in Table 3. Based on the coefficient of determination (R^2) and the root-mean square error ($RSME$), we can see that RF achieved higher accuracies compared to PLS and SVM in the prediction of the presented types of heavy metals. RF achieved the highest accuracy in predicting Cr, As, and Pb, followed by SVM and PLS respectively. Furthermore, a better generalization performance can be drawn from the RF model as it fully uses the information contained in the variables [22].

Table 3 Model performance statistics for selected bands [22, 26]

Metal	**Evaluation**	**PLS**	**SVM**	**RF**
Zn	**R^2**	0.8730	0.8851	0.8929
	RMSE	6.9413	7.0360	6.7740
Cr	**R^2**	0.8681	0.9005	0.9110
	RMSE	5.9399	4.9732	4.5683
As	**R^2**	0.9431	0.9720	0.9912
	RMSE	1.3683	0.8760	0.5327
Pb	**R^2**	0.9071	0.9622	0.9576
	RMSE	2.6327	1.5919	1.1694

3. Support Vector Machines (SVM)

Support Vector Machines (SVM) is a powerful supervised machine learning algorithm used for two-group classification and regression. It was first proposed by Vapnik in 1994 for classification and regression, then used in the classification of hyperspectral

images by Gualtieri and Cromp [65]. This method works by finding the best separation, the separating hyperplane, between two classes based on significant training data, which is referred to as support vectors. Although there are infinite hyperplanes that separate the data, SVM finds the optimal separating hyperplane, which is the plane that maximizes the distance between the support vectors (maximizes the margin). If the classes are linearly separable, the SVM classifier is represented by the function:

$$f(x) = w.x + b \tag{12}$$

where w is a vector $\in R^d$, and b is a real bias $\in$ R.

To maximize the distance and by introducing the Lagrangian formalism, the optimization problem is transformed to the dual problem:

$$\text{Maximize: } \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \left(x_i . x_j\right) \tag{13}$$

$$\text{subject to: } \sum_{i=1}^{N} \alpha_i y_i = 0 \tag{14}$$

$$\alpha_i \geq 0 \tag{15}$$

Where (x_i, y_i) are the training samples, y_i is a class label equal to ∓ 1 which indicates the class of the pixel and x_i is a d-dimensional vector that represents the spectrum of the pixel in d wavelengths in the case of hyperspectral images, and α_i are the Lagrangian multipliers that can be estimated using quadratic programming.

If the data are not linearly separable, which is the common case, a suitable kernel function is used to project the input features into a higher dimensional space called the feature space.

In this higher dimensional space, the data become linearly separable, thus linearly classified. So, the inner product $(x_i . x_j)$ used for maximization is replaced with the kernel function $k(x_i . x_j)$. Table 3 also presents the accuracy of the SVM model while predicting several metals.

4. Artificial Neural Networks (ANN)

Artificial Neural Networks (ANN) is a sophisticated machine learning algorithm that mimics the human brain. It receives an input (similar to the external stimulus received by humans) and decodes it into information or electrical signals that travel through neurons. It finally reaches the central nervous system, the brain, where the brain processes this information and initiates a command such as moving muscles, etc. Neural networks function similarly. ANN allows the computer to produce an output closely resembling the input that has already been trained to the computer [5]. In neural networks, units are connected in a way that ensures the unidirectional flow of information. As illustrated in Figure 11, They pass from the input units, through the units located in the hidden layers, to reach the units on the output layer.

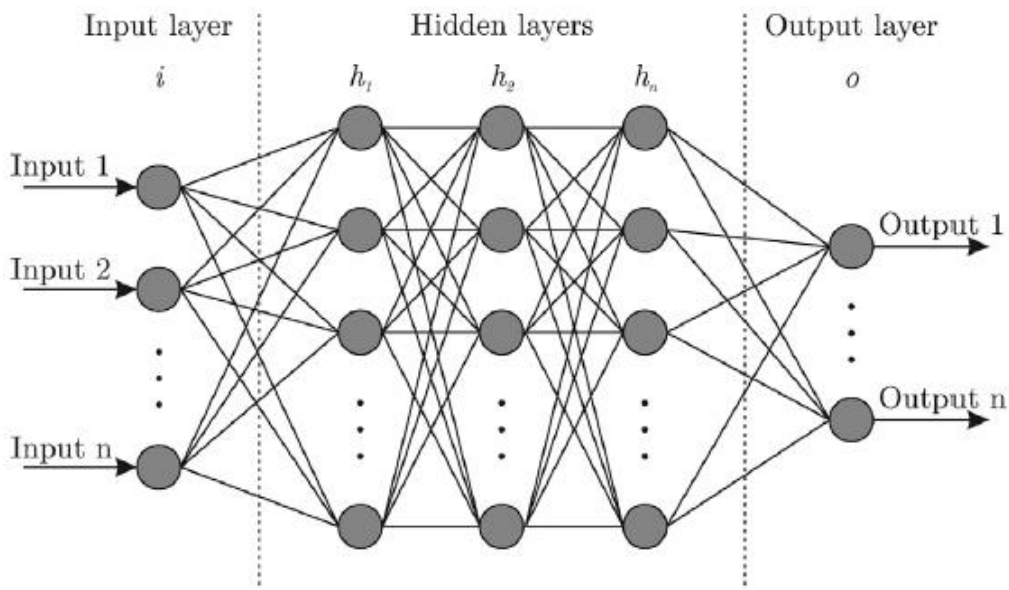


Figure 11 Artificial neural network architecture [66]

The nodes in the input layer represent the independent variables. REP, modified NDVI, NDVI, and MTCI, soil pH value, wavelet coefficients, the fractal dimension of reflectance with wavelet transform (FDWT), and the influencing factors of total Cu concentration are all variables that have been used as inputs [5]. As for the output node, it contains the heavy metal contamination level or its surrogate indicators, such as chlorophyll. Tan et al [36] achieved, using ANN, a $R^2 = 0.94$ in detecting As abundance.

A critical disadvantage of ANN is the very time-consuming process of properly configuring its structure to achieve optimum performance. Li et al [5] found that the 4-11-7-1 network configuration with logsig transfer function achieves a 100% classification accuracy for each metal contamination level while doing ground-sample spectral analysis.

Also, a two hidden layer model results in a high fitting degree with a prediction accuracy of 85.4% [67].

CHAPTER III

PROBLEM STATEMENT & OBJECTIVES

The traditional methods of estimating the spatial distribution of heavy metals are inefficient, time-consuming, and costly processes. So, studying the use of VNIR hyperspectral imaging (HSI) technology, combined with machine learning algorithms, to detect heavy metals in contaminated soils is crucial. For that reason, this work aims to:

1) Explore several preprocessing and dimensionality reduction techniques and study their effect on the prediction results. The data handling and dimensionality reduction techniques are the following: SG, MSC, SNV, PCA, and LDA.
2) Investigate several machine learning algorithms, namely RF, SVM, and KNN, and evaluate their performance based on multiple evaluation metrics.
3) Use the above-mentioned to accurately detect and predict heavy metal contamination in soil.

In the following two chapters, we will present the use of UV VIS spectral measurements to identify certain contaminants and demonstrate the concept. However, to apply the ML algorithms, and due to the lack of available data, we will use the available three datasets: two vegetation datasets and one paint condition assessment dataset. These datasets will be investigated upon to be an alternative and to prove the ability of the built models to accurately detect contamination levels in soils and vegetation.

CHAPTER IV

METHODOLOGY & SPECTRAL MEASUREMENTS

The methodology of work, which is the design and implementation for estimating the heavy metal contamination in soils, is illustrated in Figure 12. The framework can be divided into two main bodies: model establishment and model evaluation. In the first process, data preprocessing, feature extraction, and algorithm optimization are performed on training samples and spectral data. In the evaluation process, the same preprocessing and feature extraction methods are used as in the model establishment, but this time on test data. This is followed by evaluating the performance by comparing the results with the true concentration values. After evaluating all the models using the testing sub-data set, prediction accuracies will be compared, and the most accurate model will be chosen.

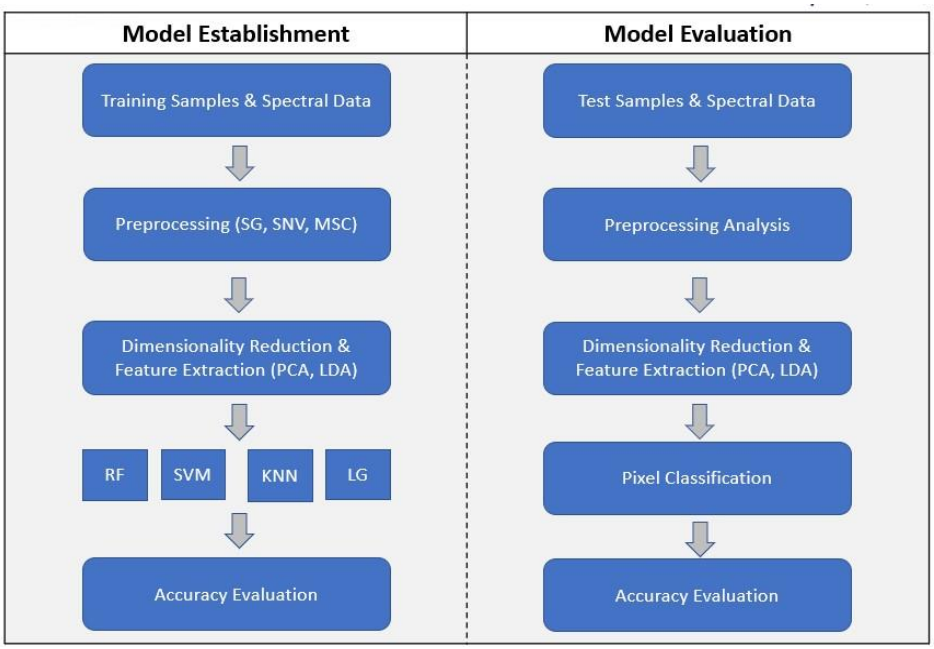


Figure 12 Proposed framework for heavy metal estimation

Applying the above-mentioned methodology to various soil samples from different locations will provide us with the needed knowledge and data to minimize the effects of soil contamination in Lebanon, and hopefully put it to an end in the future.

A. Soil Preparation & Sampling

Soil samples were collected from an area in Mount Lebanon, Lebanon where no agricultural and/or agricultural activities took place. Surface soil samples are collected at a depth ranging from 0-5 cm, 0-10 cm, or 0-20 cm from different locations in the same area, then mixed. The fresh soil samples are air dried until all the moisture content is evaporated, then placed in sealed plastic bags and taken to the laboratory. Soil samples are then ground using a mortar and pestle and passed through a 2mm sieve to remove coarse materials, lumps, stones, and other debris.

B. Soil Contamination Process

In this study, Cr and Zn were the metals being tested for. Each soil sample was weighted to be 50 g. Treatment of the soil samples consisted of one uncontaminated sample and four levels with 40 mg/kg, 80 mg/kg, 100 mg/kg, 120 mg/kg and 50 mg/kg, 70 mg/kg, 80 mg/kg, and 100 mg/kg of Cr and Zn respectively. The artificially contaminated soil samples were then air dried for 48 to 72 hours until completely dry. Also, multiple soil samples were contaminated with both Cr and Zn together at the same time with the above-mentioned concentration levels in increasing order.

C. UV VIS-Measurements

After contaminating the soil samples as mentioned above, UV VIS spectral measurements were taken using the JASCO V-570. The below figures show the spectral reflectance of contaminated soils with Cr, Zn, and both combined over the wavelength range of 400 nm – 2100 nm.

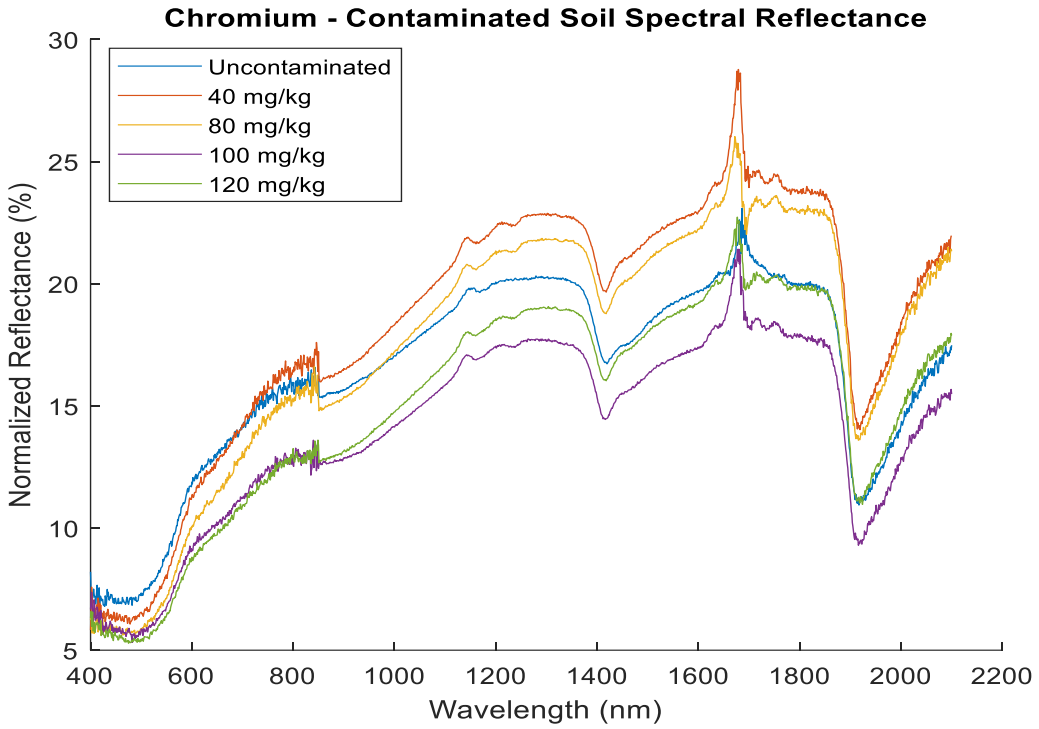


Figure 13– Spectral reflectance of soil contaminated with multiple levels of Cr

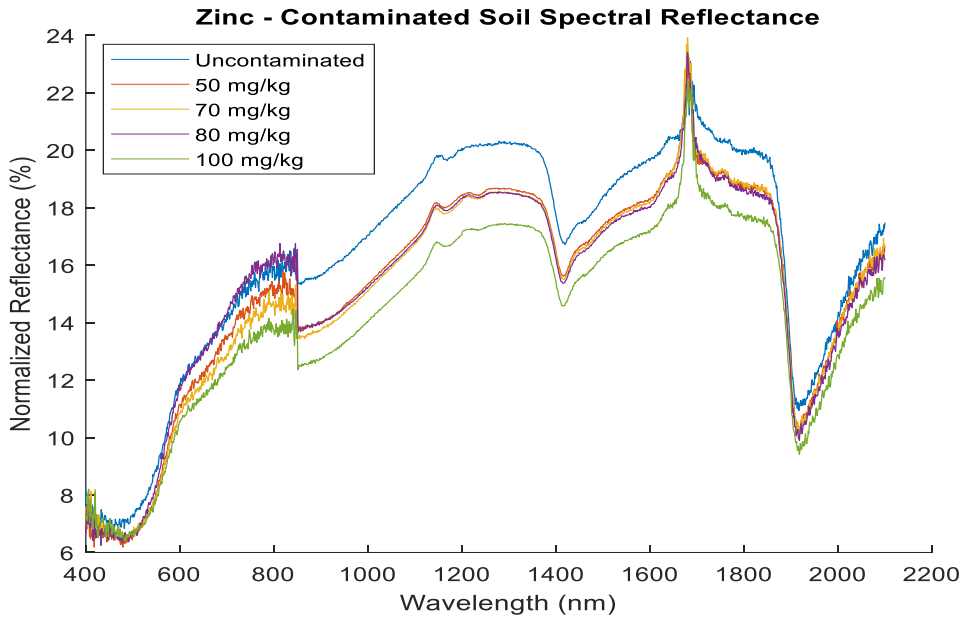


Figure 14 Spectral reflectance of soil contaminated with multiple levels of Zn

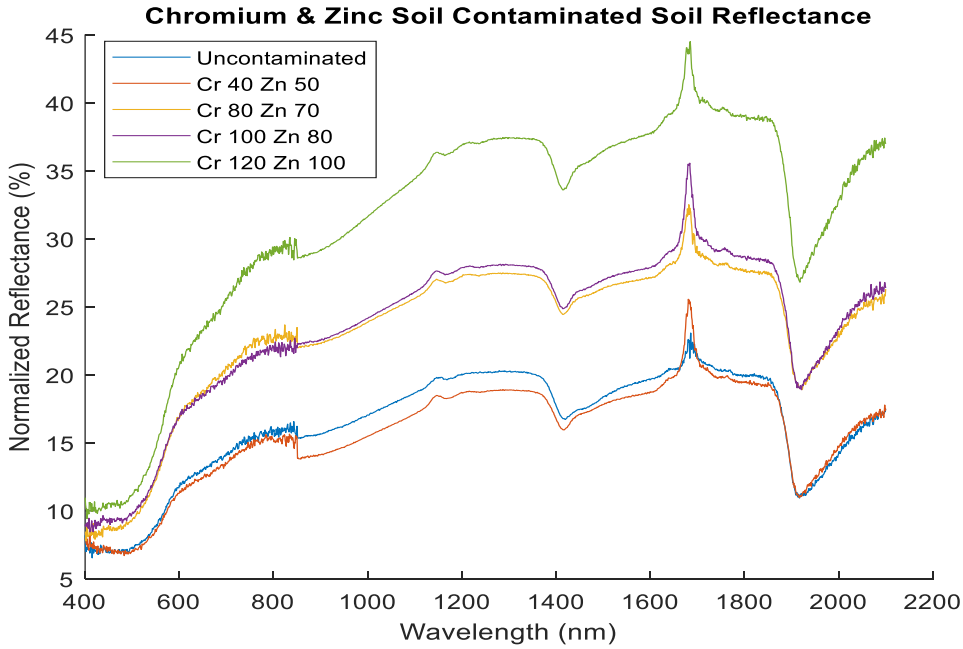


Figure 15 Spectral reflectance of soil contaminated with multiple levels of Cr and Zn

The results shown in the UV – VIS Measurements section constitute the high potential for the detection and prediction of Cr and Zn in contaminated soils. Figure 13, Figure 14, and Figure 15 showed distinctive features and spectral peaks at certain ranges along the measured spectrum. It is seen that there are distinguishing spectral trends for both Cr and Zn in the ranges of 1170 nm – 1450 nm and 1600 nm – 1800 nm. These trends will allow for key observations ofCr and Zn, and the prediction of their concentration levels in the tested soil. Although other wavelengths can be useful and provide insights about the tested-for-metal, these wavelengths might be neglected for reducing the data size and computational time and speed purposes. Furthermore, the controlled laboratory indoor conditions, where the measurements are being taken, allow for deviation from the true results desired in the complex reality. Therefore, the measured spectral response can present guidelines for choosing the best spectral wavelengths for identifying a particular metal.

D. Preprocessing & Machine Learning Algorithms

Since the environment and instruments produce irrelevant information, spectral noises, and overlapping peaks, optimization of the measured spectral data is required. In addition, as mentioned before, some heavy metals are present in low concentrations, thus almost impossible to be spectrally detected. Therefore, it is necessary to implement some basic preprocessing steps, along with several spectral pretreatments such as transformations and their derivatives.

The most common preprocessing methods in spectral derivatives are the SG, SNV, MSC, FD, SD, and continuum removal CR. One can possibly choose a single preprocessing method or a combination of techniques. However, this decision will influence the outcome predicted concentration of the heavy metal from the estimation model. So, it is crucial to choose the appropriate method of preprocessing the data. To achieve a successful heavy metal concentration retrieval, it is also essential to choose the bands that are most sensitive to heavy metal content. This selection is critical specifically in hyperspectral data due to the immense redundancy among hundreds of available bands.

In this study, data handling will start applying smoothing techniques such as SG, SNV, and MSC. For feature extraction and dimensionality reduction, PCA and LDA methods will then be implemented on each of the preprocessed data. Finally, the data will be split into training and testing sub-datasets, where RF, SVM, and KNN will be used for the classification and prediction purposes of each label in the working dataset. Models are then compared to each other, and the model with the highest prediction accuracy will be chosen based on several evaluation metrics compared.

E. Testing on Various Datasets

Studying soil contamination and predicting its concentration levels requires conducting tests on large datasets, collected both in situ and in the lab, to derive consistent results and conclusions. However, due to the unavailability of the required hyperspectral camera and the lack of ability to gather and compile huge datasets on the studied

contaminated soil, different datasets were used for testing and evaluation. Two datasets; Sorghum Plant and Salinas A-Scene, are vegetation-related datasets, that provide a high resemblance to the contaminated soil data. Also, a Paint Condition Assessment dataset is used to further prove the robustness of the classification models, especially since it contains long and short-range images. This will be presented in the following chapter.

CHAPTER V

APPLICATION OF MACHINE LEARNING TO HYPERSPECTRAL IMAGING

In this chapter, three different datasets. Several data handling and preprocessing, in addition to machine learning algorithms will be tested and their performance will be compared. The robustness and the high accuracy of the models on these datasets will prove the capacity of using machine learning and artificial intelligence with HSI on soil-contaminated datasets for the prediction of heavy metal contamination levels.

A. Sorghum Plant Dataset

The Sorghum Plant dataset is a sorghum pixel classification dataset provided by the project titled “Sorghum and Maize Segmentation using Hyperspectral Imagery” created on the Zooniverse crowdsourcing science platform (https://www.zooniverse.org/projects/alejandropages/sorghumand-maize-segmentation-using-hyperspectral-imagery). The dataset is a selection of ten pixels per class per image for the four classes (background, leaf, stalk, and panicle as shown in Figure 16) and a total of 7560 classified pixels (189 images x 4 classes x 10 pixels per class per image) saved as a NumPy array (.npy).

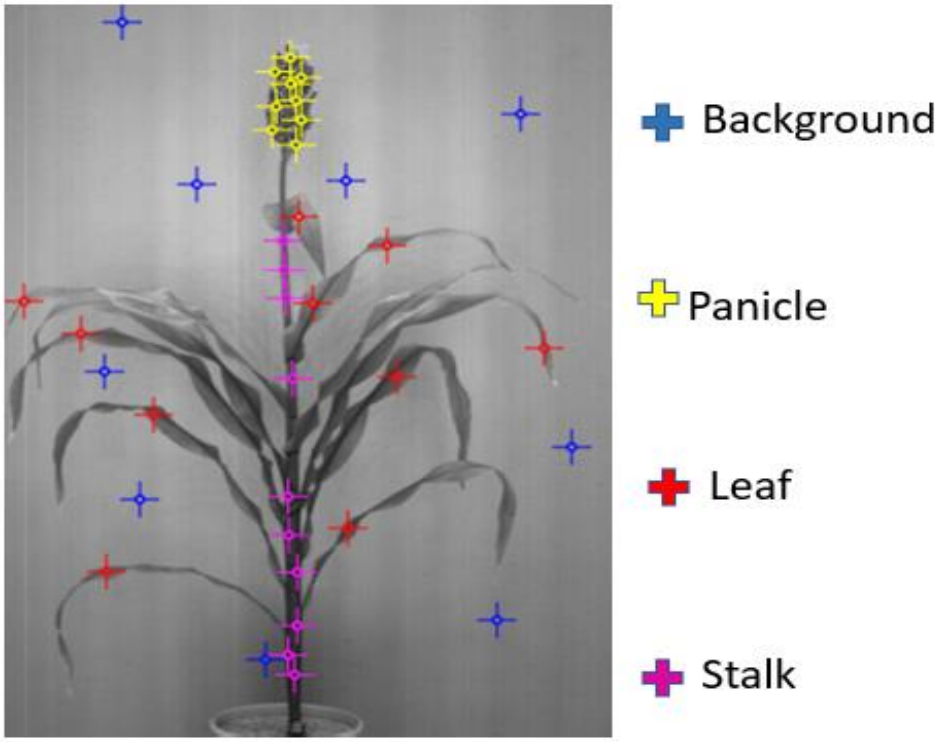


Figure 16 Sorghum plant hyperspectral image [68]

Figure 17 illustrates the spectral reflectance of the sorghum plant dataset against the wavelength. By exploring the data, it is seen that the data has irrelevant information, spectral noises, and overlapping peaks along the wavelengths, thus requiring preprocessing. Also, it is visible that the data has numerous features that produce no valuable information in our study. For that reason, dimensionality reduction and feature extraction are also required.

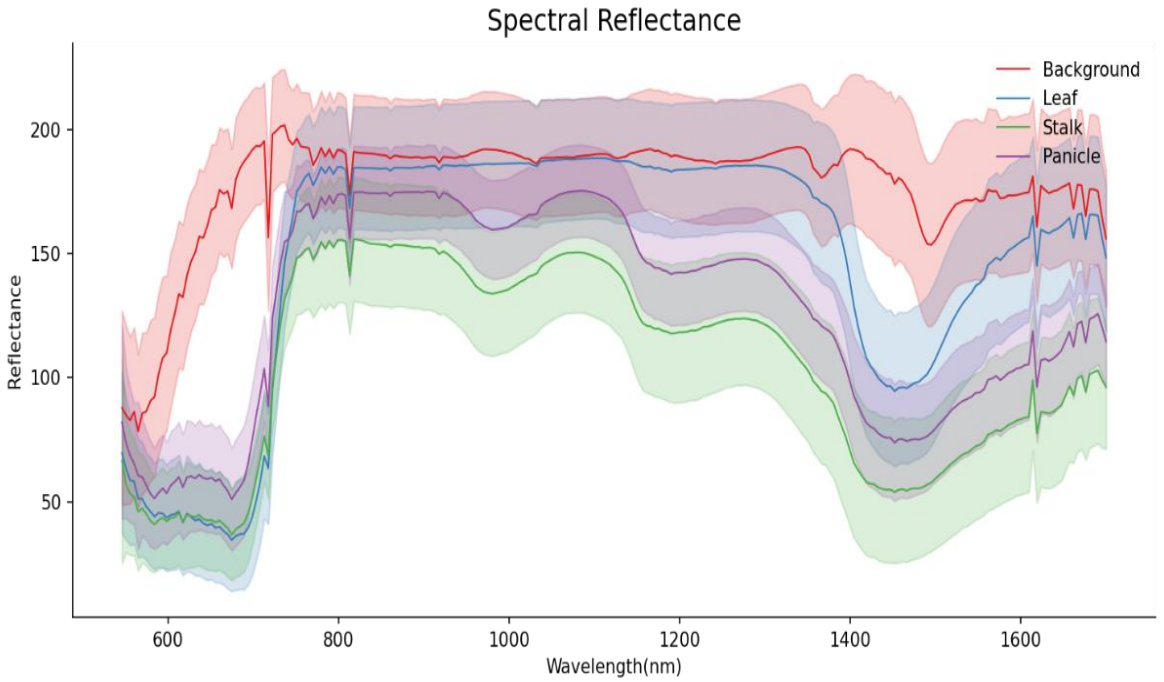


Figure 17 Spectral reflectance of sorghum plant dataset

After preprocessing the data and extracting the valuable features, a set of machine learning algorithms for classification purposes was implemented. RF, SVM, and KNN were explored in the combination of the datasets, and their detection accuracies are presented in

Table *4*.

	RF	**RF+PCA**	**RF+LDA**	**SVM**	**SVM+PCA**	**SVM+LDA**	**KNN**	**KNN+PCA**	**KNN+LDA**
Not Normalized	0.962	0.848	0.981	0.972	0.856	0.981	0.946	0.84	0.98
SG	0.970	0.947	0.973	0.972	0.853	0.975	0.956	0.848	0.978
MSC	0.971	0.785	0.973	0.953	0.750	0.973	0.956	0.777	0.971
SNV	0.970	0.846	0.977	0.966	0.837	0.980	0.959	0.835	0.979

	RF	**RF+PCA**	**RF+LDA**	**SVM**	**SVM+PCA**	**SVM+LDA**	**KNN**	**KNN+PCA**	**KNN+LDA**
Not Normalized	0.962	0.848	0.981	0.972	0.856	0.981	0.946	0.84	0.98
MSC	0.971	0.785	0.973	0.953	0.750	0.973	0.956	0.777	0.971
SNV	0.970	0.846	0.977	0.966	0.837	0.980	0.959	0.835	0.979

Table 4 Classification detection accuracy results of sorghum dataset

After testing the above-mentioned approach, it is seen that RF and SVM, after performing LDA, both produced the highest accuracy of 98.1%. SVM with LDA and SVN was also detected to perform well under SNV and LDA preprocessing with an accuracy of 98%. Figure 18 and Figure 19 show the confusion matrix for both the classification using RF and SVM after performing LDA on the dataset respectively.

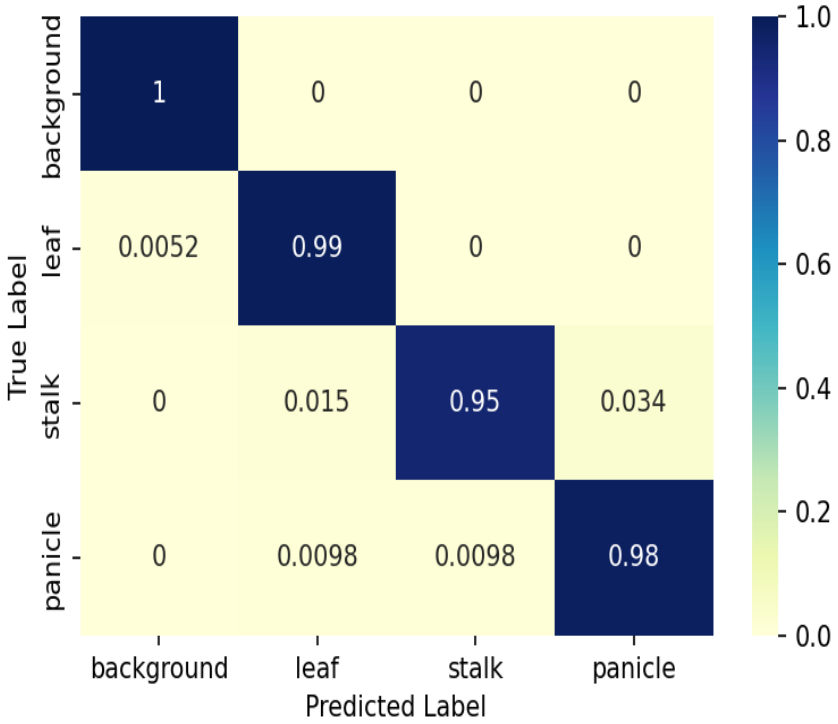


Figure 18 Confusion matrix for RF without normalization and using LDA

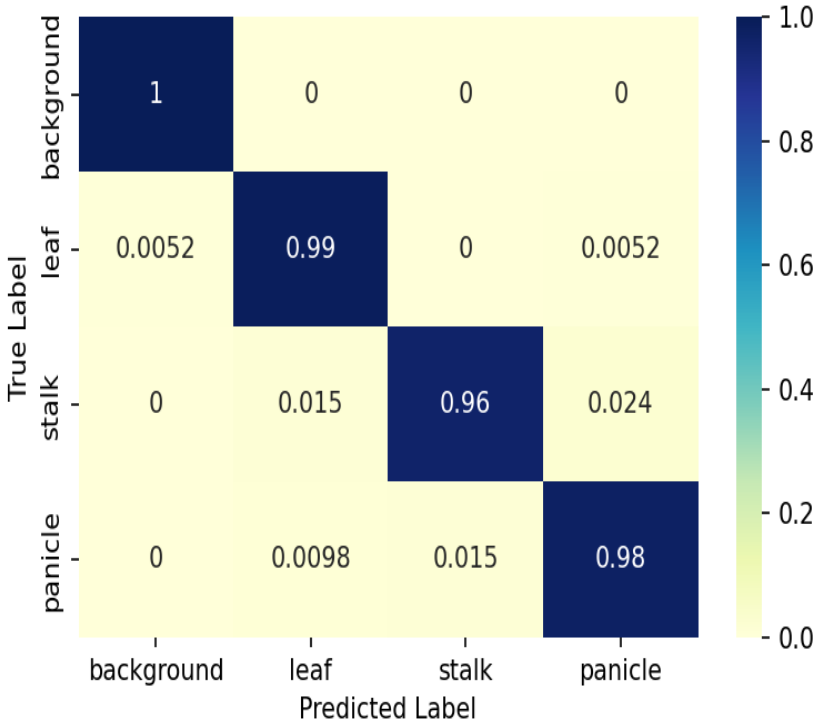


Figure 19 Confusion matrix for SVM without normalization and using LD

Moreover, Figure 20 and Figure 21 compare the classification results of the mentioned algorithms respectively. It is seen that both algorithms performed extremely well, with minimal errors on the boundaries of the sorghum plant.

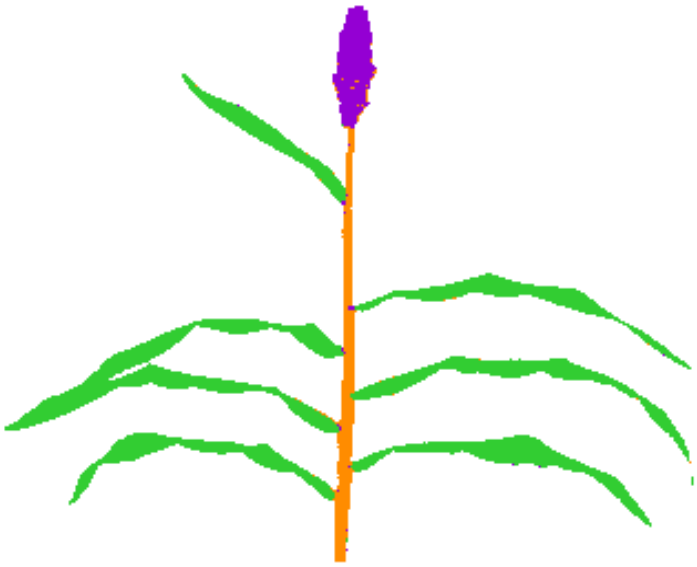


Figure 20 Pixel prediction using RF without normalization and using LDA

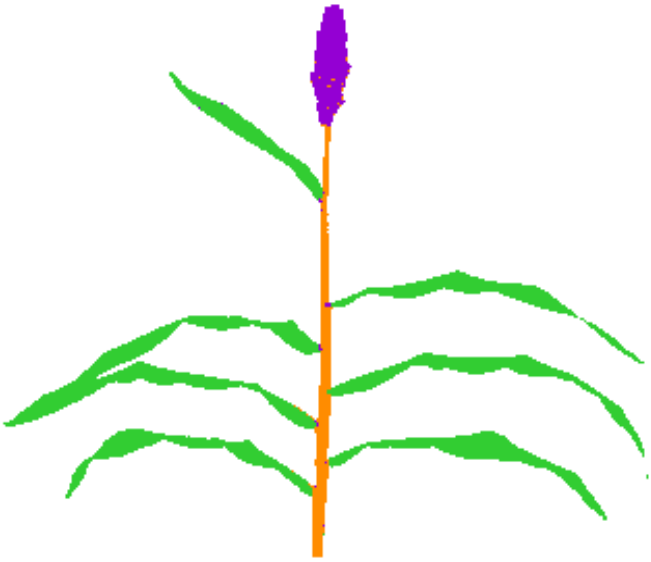


Figure 21 Pixel prediction using SVM without normalization and using LDA

A set of 9 classification algorithms (combinations between preprocessing and classification) were evaluated for their ability to correctly classify hyperspectral pixels. The average classification accuracy of the 6 algorithms (disregarding the combinations with PCA), estimated from fivefold cross-validation, exceeded 95% as shown in

Table *4*. This is because when PCA was applied as a data reduction technique,

	RF	**RF+PCA**	**RF+LDA**	**SVM**	**SVM+PCA**	**SVM+LDA**	**KNN**	**KNN+PCA**	**KNN+LDA**
Not Normalized	0.962	0.848	0.981	0.972	0.856	0.981	0.946	0.84	0.98
Normalized	0.968	0.845	0.973	0.973	0.857	0.975	0.953	0.841	0.978
SG	0.970	0.947	0.973	0.972	0.853	0.975	0.956	0.848	0.978
MSC	0.971	0.785	0.973	0.953	0.750	0.973	0.956	0.777	0.971
SNV	0.970	0.846	0.977	0.966	0.837	0.980	0.959	0.835	0.979

crucial signatures at certain specific wavelengths were removed, resulting in poor prediction. As expected, given the distinct reflectance patterns observed in Figure 17, all the methods have a very high classification accuracy in the prediction of background pixels. Also, all methods showed quite an accurate classification (> 97%) for leaf pixels. Random Forests and Support Vector Machines, before normalization, had the highest accuracy for leaf (99%), stalk (95%), and panicle (98%), respectively, although the overall differences were quite small.

Qualitatively, classification accuracy was high, with common errors in small areas and pixels in the leaves' edges where the model was misclassified as stalk. The lignified and thick midribs of the sorghum plant leaves may have resulted in the production of similar reflectance patterns to the stalk rather than to the remainder of the leaf blade. The semantic segmentation of sorghum hyperspectral images at a pixel level allows for the automated estimation of a variety of plant traits.

B. Salinas A-Scene Dataset

The Salinas A-Scene dataset is a small subscene of the Salinas image, denoted as Salinas A, shown in Figure 22. This scene was collected by the 224- band AVIRIS sensor over Salinas Valley, California, and is characterized by high spatial resolution (3.7-meter pixels) and saved as a MATLAB file (.mat). Salinas A dataset is provided by Grupo de Inteligencia Computacional (GIC) of the University of Basque Country (http://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes#Salinas-A_scene). It compromises 86 x 83 pixels and includes six classes; Broccoli Green Weeds, Corn Senesced Green Weeds, Lettuce Romaine 4wk, Lettuce Romaine 5wk, Lettuce Romaine 5wk, Lettuce Romaine 6wk, and Lettuce Romaine 7wk.

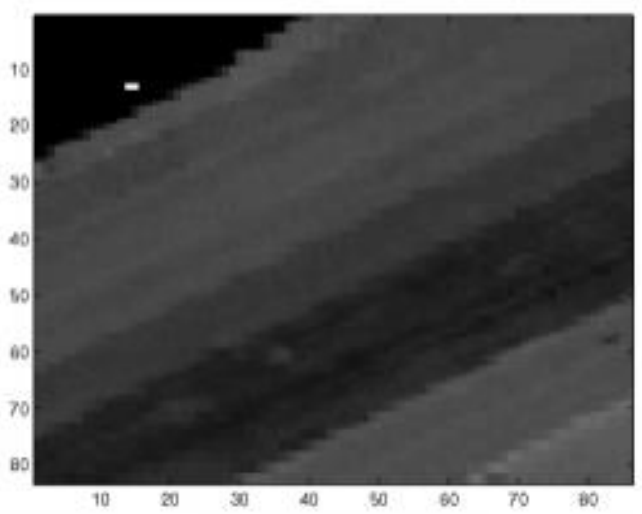


Figure 22 Salinas A Scene hyperspectral image [69]

A similar approach was followed for the Salinas A-Scene dataset. The results of the study are shown in Table 5.

Table 5 Classification detection accuracy results of Salinas A-Scene dataset

	RF	RF+PCA	RF+LDA	SVM	SVM+PCA	SVM+LDA	KNN	KNN+PCA	KNN+LDA
Not Normalized	0.894	0.878	0.809	0.754	0.757	0.733	0.868	0.844	0.731
Normalized	0.9018	0.870	0.804	0.737	0.741	0.746	0.874	0.829	0.738
SG	0.900	0.868	0.804	0.737	0.741	0.743	0.876	0.827	0.730
MSC	0.9011	0.866	0.813	0.744	0.756	0.766	0.885	0.812	0.756
SNV	0.900	0.867	0.826	0.741	0.761	0.781	0.882	0.827	0.760

As for the Salinas A-Scene dataset, the RF algorithm, after normalization, resulted in the highest classification accuracy of 90.18%. Also, RF was classified accurately after performing SG, MSC, and SNV preprocessing with 90%, 90.11%, and 90% detection accuracy respectively. Figure 23 presents the confusion matrix for the RF algorithm after normalization.

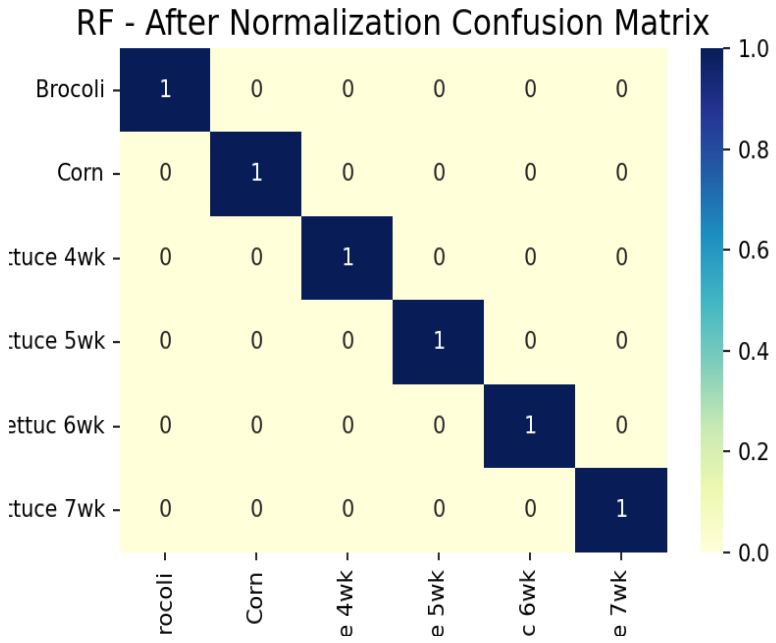


Figure 23 Confusion matrix for RF with normalization

Furthermore, Figure *24* and Figure *25* compare the classification of the Salinas A-Scene with the RF-Normalized algorithm against the ground truth labels.

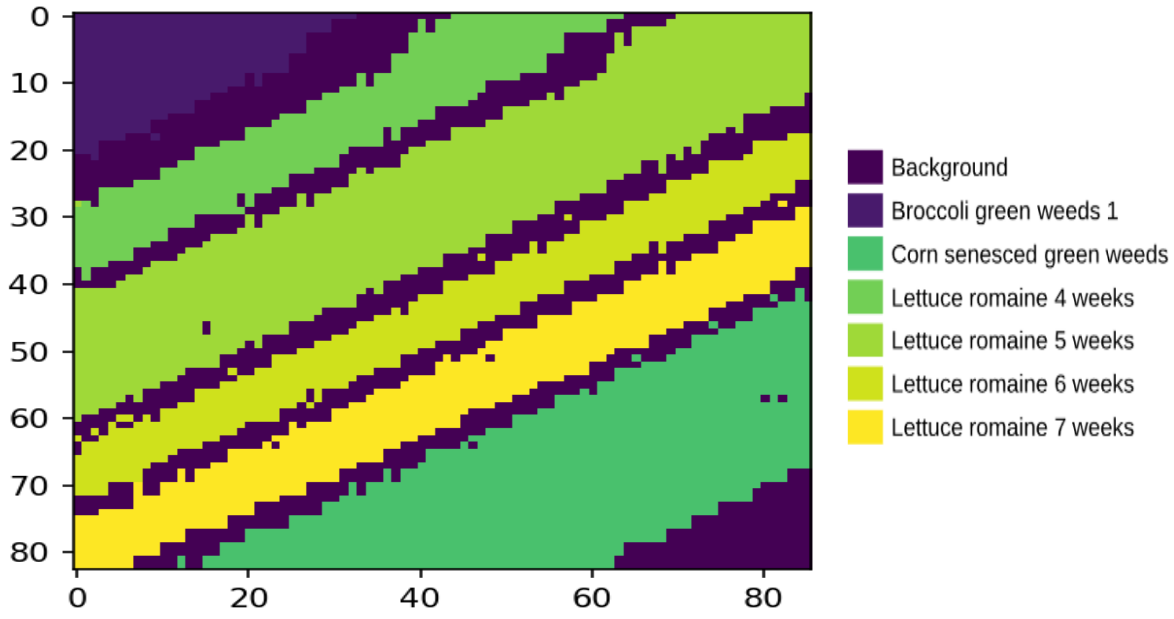


Figure 24 Pixel prediction using RF with normalization

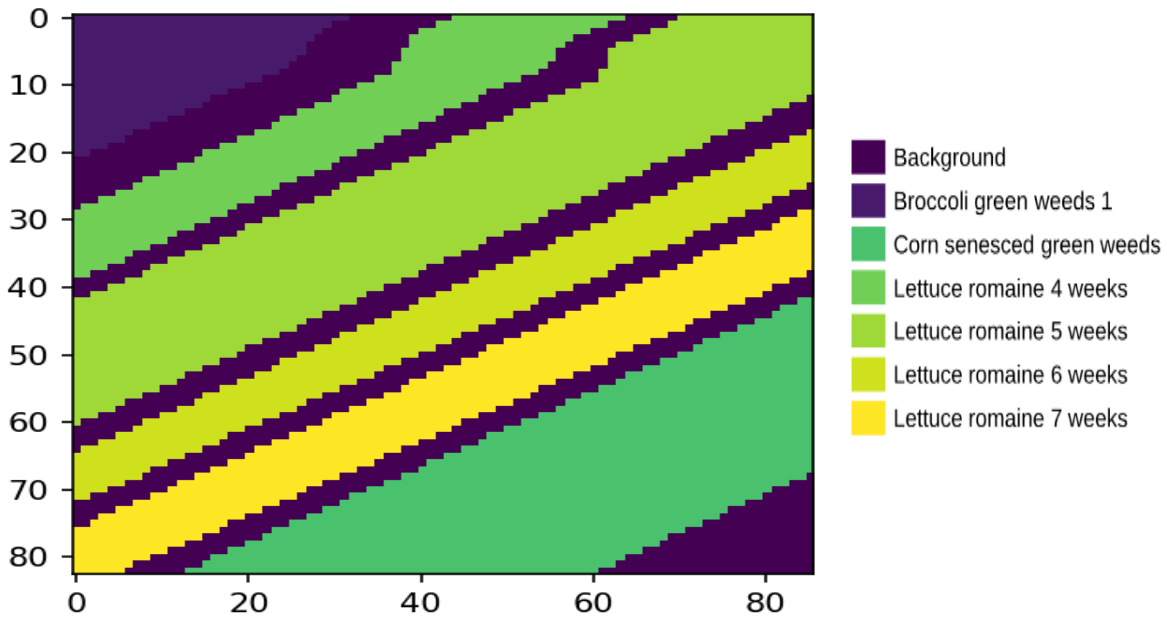


Figure 25 Ground truth

It is seen that the RF algorithm classified the image with high accuracy. Some errors and misclassification can be detected around the boundaries of each label.

Furthermore, the hyperspectral data distribution of the Salinas A-Scene does not follow the trend of being separable in parallel hyperplanes. That is why SVM showed lower classification accuracies compared to RF and KNN. Also, with 7 classes to classify, Random Forests outperforms KNN in a multiclass dataset. Moreover, since the data is not completely raw and has been preprocessed by the source, applying further preprocessing techniques such as PCA, LDA, SNV, and SG will result in removing some of the important features and signatures and would cause the data to lose some of its distinctive properties, resulting in poorer performances compared to non-preprocessed data by the above-mentioned techniques.

C. Paint Condition Assessment of Sydney Harbor Bridge

The hyperspectral images of this dataset were acquired of steel surfaces located at long (mid-range) and short distances on the Sydney Harbour Bridge with an Acousto-Optics Tunable filter (AOTF) hyperspectral camera (consisting of 21 bands in the visible spectrum), the structure is subjected to several levels of deterioration and rust [70]. The chosen locations ensured the presence of the four different levels of paint failures described below. For labeling the images, several images were randomly selected and labeled by a human bridge inspector. The expert, who was present at the collection site to minimize mislabeling, assigned labels to the corresponding condition ratings based on his observations of the actual surface.

The labeling process was based on the four-level rating system which is popular in Australia and other countries for evaluating and assessing civil steel structures [70]. The images are taken over a wavelength range of 450-650 nm with a 10 nm incremental step between successive bands. In our case, the inputs (features) for the models are the number of bands in the hyperspectral images, which are 21 bands for both the short and long-range images.

The images contain five labels:

1. Label 1: the protective coating is undamaged – Level 1.

2. Label 2: white or red rusting manifests on the protective coating in a small area – Level

3. Label 3: white rusting in the area between 2 – 5% of the total surface area – Level 3.

4. Label 4: damaged protective coating, white and red rusting covered greater than 5% of the total surface area – Level 4.

5. Label 5: the shadow – Level 5.

As for extracting material reflectance, it is assumed that the scene is uniformly illuminated by a single light source and the irradiance arriving at the camera sensor is proportional to the scene radiance. The matter of separating the material reflectance from the illumination power spectrum given the irradiance image is closely related to the large body of works in computational color constancy. Therefore, a simple and widely adopted method known as the Grey-World Method was leveraged. This method relies on the hypothesis that the spatial average of surface reflectance in a scene is achromatic, i.e., the illuminant spectrum can be estimated by taking the average of the sensor responses in the image.

Different models were trained and then compared based on the pre-mentioned metrics and approach. The comparison between the models took place after tuning the hyperparameters of each algorithm, resulting in the highest accuracy for the models. The results for long-range and short-range images are discussed below.

After comparing the classifiers, it is observed in Table 6 that Decision Trees outperformed the other models. In terms of F1 score, Decision Trees and Support Vector Machines resulted in quite similar F1 scores of 0.92 and 0.89 respectively, however lower scores (0.65 and 0.74) resulted from Naïve Bayes and Logistic Regression. On the other

hand, the Decision Trees classifier resulted in high accuracy of 0.98, which surpassed Support Vector Machines and Logistic Regression's 0.93 model accuracies. As for Naïve Bayes, it poorly classified the image and resulted in a 0.67 classification accuracy, which is considered far lower than the performance of the other models.

Table 6 Model metric results for long-range images

Model Metric	Decision Tree	Support Vector Machines	Naïve Bayes	Logistic Regression
MSE	**0.05**	**0.16**	**1.5**	**0.19**
F1	**0.92**	**0.89**	**0.65**	**0.74**
Precision	**0.92**	**0.65**	**0.77**	**0.96**
Recall	**0.92**	**0.67**	**0.6**	**0.69**
Accuracy	**0.98**	**0.93**	**0.67**	**0.93**

As for short-range images, it is seen in Table 7 that the Decision Tree classifier resulted in the highest classification accuracies and evaluation metrics, with the lowest RMSE for short-range images. Decision Trees classifier resulted in a relatively high F1 score of 0.97 compared to 0.79, 0.66, and 0.74 for Support Vector Machines, Naïve Bayes, and Logistic Regression respectively. As for classification accuracies, Decision Trees' accuracy of 0.98 slightly exceeded Support Vector Machines' 0.96 and Logistic Regression's 0.95 classification accuracies but performed much better than Naïve Bayes with 0.83 classification accuracy.

Table 7 Model metric results for short-range images

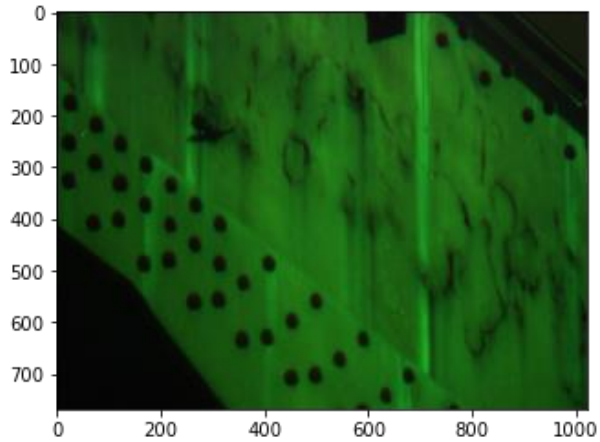
Model Metric	Decision Tree	Support Vector Machines	Naïve Bayes	Logistic Regression
MSE	**0.05**	**0.16**	**1.5**	**0.19**
F1	**0.92**	**0.89**	**0.65**	**0.74**
Precision	**0.92**	**0.65**	**0.77**	**0.96**

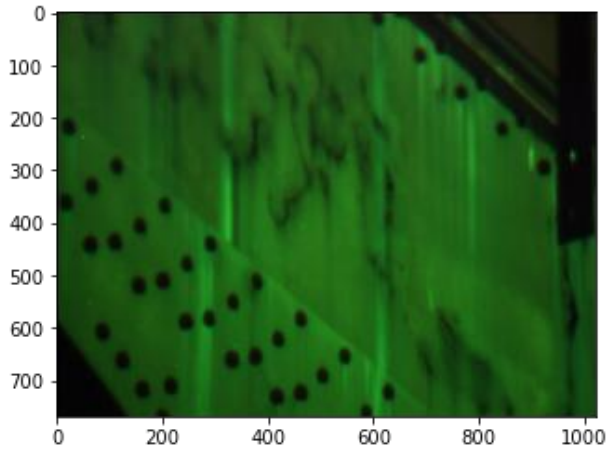
Recall	**0.92**	**0.67**	**0.6**	**0.69**
Accuracy	**0.98**	**0.93**	**0.67**	**0.93**

Since Decision Trees can be easily prone to overfitting, Grid Search algorithm was performed using the testing dataset to identify the best hyperparameters that results in the highest accuracy on both the training and testing dataset. The hyperparameters for the Decision Trees model, that resulted in the optimal performance, are as follows:

The Criterion is chosen to be “Entropy Function”, and Maximum Depth is selected to at 1000. The strategy used to choose the split at each node is the “Best” splitter.

Figure 26, Figure 27, Figure 28, and Figure 29 visualize the classification results as a heat map (Pixel–wise classification) for long- and short-range images respectively. For both image sets, the background class was accurately classified. In addition, Figure 26 and Figure 27show that Level 1 pixels in the long-range image were also classified well. However, it is also seen that there is some confusion when predicting Levels 1 and 2, in addition to predicting Levels 3 and 4.





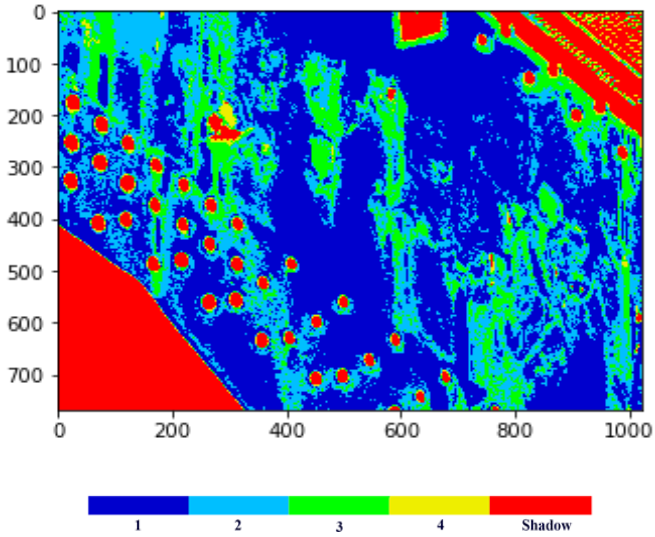


Figure 26 Prediction result for long-range image using Decision Trees

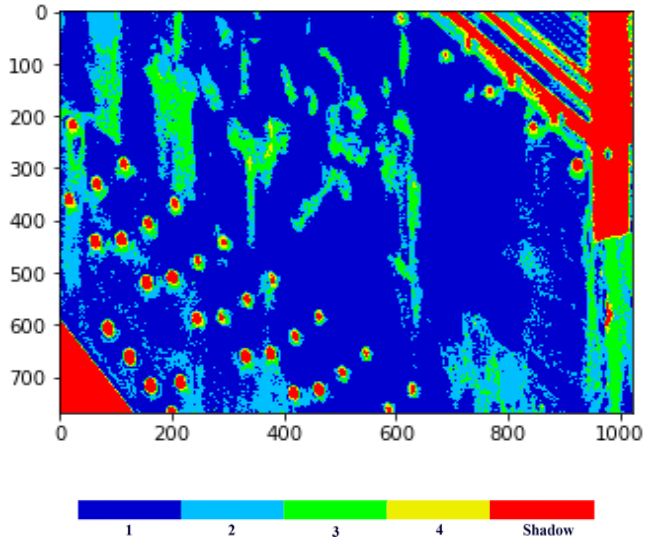
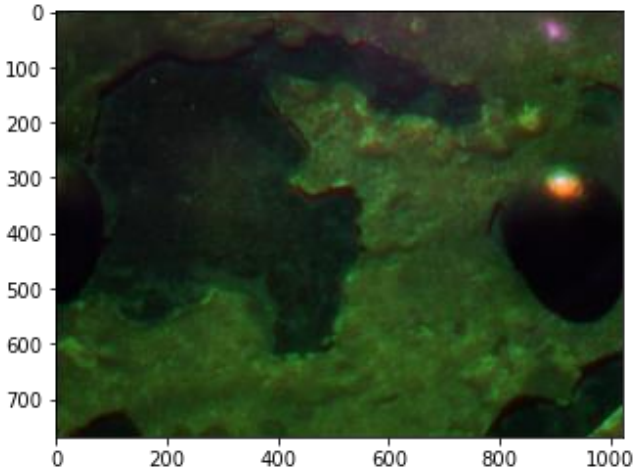
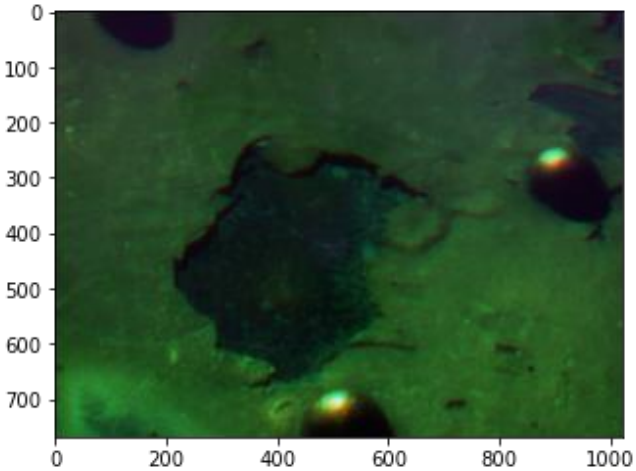


Figure 27 Prediction result for long-range image using Decision Trees

On the other hand, it was seen in Figure 28 and Figure 29 that the Decision Trees model accurately classified Levels 1 and 4. Nevertheless, some misclassifications between Levels 3 and 4 are observed.





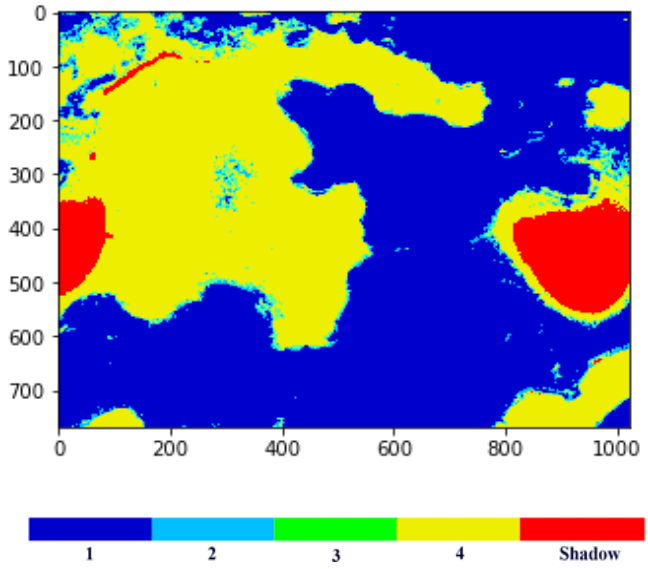


Figure 28 Prediction result for short-range image using Decision Trees

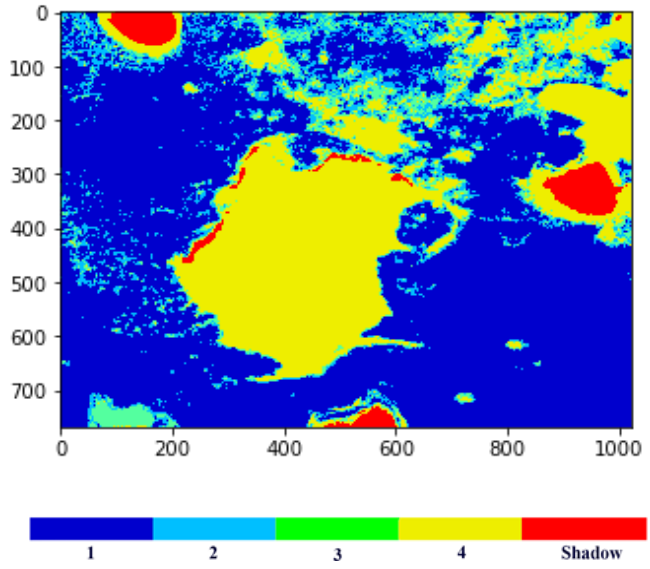


Figure 29 Prediction result for short-range image using Decision Trees

The misclassifications in both image sets can be explained by noting that the reflectance spectra for these levels in the training set exhibit similar trends and variations over the wavelength range. Regardless of this confusion, the Decision Tree classifier achieved an acceptable overall classification accuracy of 0.98 for both long- and short-range images, which is comparably higher than Huynh et al's Multiclass Support Vector Machines model that resulted in overall classification rates of 85.54% and 75.19% for long- and short-range images respectively for the same set of images.

The overall observation is that the Decision Tree classifier distinguishes with a high level of accuracy the regions with intact coating from the corroded areas, i.e., Level 1 (undamaged the protective coating) from Level 2, 3, or 4 (white and red rusting). Furthermore, the background class is accurately classified in both long-range and short-range images. On the other hand, the classifier appears to misclassify and confuse surfaces

with similar ratings, such as Levels 1 and 2 and Levels 3 and 4. The Decision Trees outperformed the other models due to the ability to handle categorical data, or possibly due to the poor choice of SVM kernel and regularization hyperparameters that resulted in lower SVM accuracy. On the other hand, Naïve Bayes is a linear classifier and assumes that all predictors are independent, which is not generally the case, leading to lower accuracy. Logistic Regression is also a linear classifier that assumes a linear relationship between the input and output.

Overall classification accuracy of 0.98 was obtained for both short- and long-range data. The background class, which are the areas that are not considered under the four paint levels, was accurately classified from the different paint levels. Levels 1 and 2 were also reasonably classified for long-range images. However, some misclassifications are found when detecting Levels 1 and 2, in addition to Levels 3 and 4. As for short-range images, Levels 1 and 4 were accurately classified. Nonetheless, some confusion was present when distinguishing Levels 3 and 4 from each other.

The high overall classification accuracy presented using the Decision Tree classifier has proven the potential use of HSI for assessing the paint condition on structures. Future works might include developing a model that is robust to external factors such as daylight conditions and provide a real-time classification that can be used in situ. Also, other techniques such as spectral unmixing can be considered. Finally, a longer wavelength can also be investigated for the identification of iron-oxides, such as hematite and goethite.

CHAPTER VI

CONCLUDING REMARKS AND FUTURE WORK

Hyperspectral imaging for the detection of heavy metals offers a compelling and viable option to consider in dealing with heavy metal soil contamination in Lebanon which has been raising environmental concerns. This study aimed at discussing the methodology and results of various papers that have dealt with and reported on the use of hyperspectral imaging for the detection of heavy metals in soils in the lab or situ using machine learning algorithms and statistical methods. In addition, it compares the accuracies of different machine learning models for the prediction of heavy metal content in the soil after optimizing the spectra and smoothing it.

The significance of accurately detecting and classifying areas of interest using hyperspectral remote sensing was investigated in three case studies, Sorghum Plant, Salinas A-Scene, and paint condition assessment of Sydney Harbor Bridge datasets. After processing the data using SG, SNV, and MSC, multiple classifiers (RF, SVM, and KNN) were evaluated. Overall, all models showed acceptable classifications, but the RF model performed best in the multiclass and non-separable in a parallel plane dataset, with a classification accuracy of 98 %, 90 %, and 98 % for each case study respectively. In all three studies, machine learning techniques using hyperspectral imaging have proved to be a rapid, accurate, and robust approach that delivers better accuracy compared to traditional approaches for optical remote sensing data processing. The case studies presented prove the ability of these preprocessing, data reduction, and prediction techniques to be a practical option for the detection of heavy metal contamination in Lebanon and relate their reflectance to the soil's heavy metal concentrations.

Following this approach and proving the impressive capabilities of machine learning models, after applying the required preprocessing techniques, can classify hyperspectral images related to vegetation and paint assessment, it proves the ability of these models to accurately detect and predict contamination levels in soils. This work can be extended by evaluating real-time classification and prediction that may be of use when the application requires airborne remote sensing. The rapidly evolving multidisciplinary area of science and engineering, allows for other techniques to be investigated in data handling and preprocessing, such as spectral unmixing, in addition to exploring the effect of the use of deep learning instead of machine learning as a classification and prediction

tool. Moreover, especially after achieving promising results, implementing these techniques on a contaminated soil dataset, with soil being contaminated with different heavy metals, different concentrations, and in different atmospheric conditions, is considered an important extension of this work.

REFERENCES

1. Su, C., L. Jiang, and W. Zhang, *A review on heavy metal contamination in the soil worldwide: Situation, impact and remediation techniques.* Environ. Skept. Crit., 2014. **3**: p. 24-38.
2. Han, F.X., et al., *Industrial age anthropogenic inputs of heavy metals into the pedosphere.* Naturwissenschaften, 2002. **89**(11): p. 497-504.
3. M.R.G. Sayyed, M.H.S., *Variations in the heavy metal accumulations within the surface soils from*

   the Chitgar industrial area of Tehran. Proceedings of the International Academy of Ecology and Environmental Science, 2011. **1**(1): p. 12.
4. Zhang, W., F. Jiang, and J. Ou, *Global pesticide consumption and pollution: With China as a focus.* Proceedings of the International Academy of Ecology and Environmental Sciences, 2011. **1**: p. 125-144.
5. Wang, F.H., J. Gao, and Y. Zha, *Hyperspectral sensing of heavy metals in soil and vegetation: Feasibility and challenges.* Isprs Journal of Photogrammetry and Remote Sensing, 2018. **136**: p. 73-84.
6. Jyothi, N.R., *Heavy Metal Sources and Their Effects on Human Health*. IntechOpen.
7. Lee, D.H., Y.G. Zo, and S.J. Kim, *Nonradioactive method to study genetic profiles of natural bacterial communities by PCR-single-strand-conformation polymorphism.* Appl Environ Microbiol, 1996. **62**(9): p. 3112-20.

8. Barajas Aceves, M., et al., *Soil microbial biomass and organic C in a gradient of zinc concentrations in soils around a mine spoil tip.* Soil Biology and Biochemistry, 1999. **31**(6): p. 867-876.
9. Chander, K., P.C. Brookes, and S.A. Harding, *Microbial biomass dynamics following addition of metal-enriched sewage sludges to a sandy loam.* Soil Biology and Biochemistry, 1995. **27**(11): p. 1409-1421.
10. Qin, T., Y. Wu, and H. Wang, *Effect of cadmium, lead and their interactions on the physiological and biochemical characteristics of Brassica chinensis.* Acta Ecologica Sinica, 1994. **14**(1): p. 46-50.
11. Yabe, J., M. Ishizuka, and T. Umemura, *Current levels of heavy metal pollution in Africa.* J Vet Med Sci, 2010. **72**(10): p. 1257-63.
12. Borjac, J., et al., *Quantitative Analysis of Heavy Metals and Organic Compounds in Soil from Deir Kanoun Ras El Ain Dump, Lebanon.* TheScientificWorldJournal, 2020. **2020**: p. 8151676-8151676.
13. Lasota, J. and E. Błońska, *Polycyclic Aromatic Hydrocarbons Content in Contaminated Forest Soils with Different Humus Types.* Water, Air, & Soil Pollution, 2018. **229**.
14. Pignatello, J., B. Katz, and H. Li, *Sources, Interactions, and Ecological Impacts of Organic Contaminants in Water, Soil, and Sediment: An Introduction to the Special Series.* Journal of environmental quality, 2010. **39**: p. 1133-8.
15. Balkhair, K., *Microbial contamination of vegetable crop and soil profile in arid regions under controlled application of domestic wastewater.* Saudi Journal of Biological Sciences, 2015. **23**.
16. Shellito, K., *The Economic Effect of Refugee Crises on Host Countries and Implications for the Lebanese Case.* 2016.
17. Ltd, M.A.P. *ENVIRONMENTAL IMPACT ASSESSMENT REPORT SOLID WASTE TREATMENT FACILITY IN AIN BAAL, CAZA OF TYRE, SOUTH LEBANON*. 2005.
18. Chbib, C., et al., *Distribution of Organochlorine Pesticides and Heavy Metals in Lebanese Agricultural Soil: Case Study—Plain of Akkar.* International Journal of Environmental Research, 2018. **12**.
19. Darwish, T., et al., *Preliminary contamination hazard assessment of land resources in Central Beka plain of Lebanon.* Lebanese Science Journal, 2008. **9**.
20. Brigden, K., Stringer, R. & Santillo, D, *Heavy metal and radionuclide contamination of fertilizer products and phosphogypsum waste produced by The Lebanese Chemical Company, Lebanon, 2002.* Greenpeace Research Laboratories, Department of Biological Sciences, University of Exeter, Exeter EX4 4PS, UK, 2002: p. 16.
21. Wood, J.M., *Biological cycles for toxic elements in the environment.* Science, 1974. **183**(4129): p. 1049-52.
22. Tan, K., et al., *Random forest–based estimation of heavy metal concentration in agricultural soils with hyperspectral sensor data.* Environmental Monitoring and Assessment, 2019. **191**.
23. Kemper, T. and S. Sommer, *Estimate of heavy metal contamination in soils after a mining accident using reflectance spectroscopy.* Environmental Science & Technology, 2002. **36**(12): p. 2742-2747.
24. Moor, C., T. Lymberopoulou, and V.J. Dietrich, *Determination of Heavy Metals in Soils, Sediments and Geological Materials by ICP-AES and ICP-MS.* Microchimica Acta, 2001. **136**(3): p. 123-128.

25. von Steiger, B., et al., *Mapping heavy metals in polluted soil by disjunctive kriging.* Environmental Pollution, 1996. **94**(2): p. 205-215.
26. Wu, Y., et al., *Possibilities of reflectance spectroscopy for the assessment of contaminant elements in suburban soils.* Applied Geochemistry, 2005. **20**(6): p. 1051-1059.
27. Scafutto, R.D.P.M., C.R. de Souza Filho, and W.J. de Oliveira, *Hyperspectral remote sensing detection of petroleum hydrocarbons in mixtures with mineral substrates: Implications for onshore exploration and monitoring.* ISPRS Journal of Photogrammetry and Remote Sensing, 2017. **128**: p. 146-157.
28. Makki, I., et al., *A survey of landmine detection using hyperspectral imaging.* ISPRS Journal of Photogrammetry and Remote Sensing, 2017. **124**: p. 40-53.
29. Liu, M., et al., *Wavelet-based detection of crop zinc stress assessment using hyperspectral reflectance.* Computers & Geosciences, 2011. **37**(9): p. 1254-1263.
30. Swayze, G.A., et al., *Using Imaging Spectroscopy To Map Acidic Mine Waste.* Environmental Science & Technology, 2000. **34**(1): p. 47-54.
31. Mateen, M., et al., *The role of Hyperspectral Imaging: A literature review.* International journal of advanced computer science & applications, 2018. **9**(8): p. 51-62.
32. González-Cabrera, M., et al., *Multisensor hyperspectral imaging approach for the microchemical analysis of ultramarine blue pigments.* Scientific Reports, 2022. **12**(1): p. 707.
33. Lowe, A., N. Harrison, and A.P. French, *Hyperspectral image analysis techniques for the detection and classification of the early onset of plant disease and stress.* Plant Methods, 2017. **13**(1): p. 80.
34. Kooistra, L., et al., *Exploring field vegetation reflectance as an indicator of soil contamination in river floodplains.* Environmental Pollution, 2004. **127**(2): p. 281-290.
35. Pandit, C.M., G.M. Filippelli, and L. Li, *Estimation of heavy-metal contamination in soil using reflectance spectroscopy and partial least-squares regression.* International Journal of Remote Sensing, 2010. **31**(15): p. 4111-4123.
36. Tan, K., et al., *Estimation of Arsenic Contamination in Reclaimed Agricultural Soils Using Reflectance Spectroscopy and ANFIS Model.* Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2014. **7**(6): p. 2540-2546.
37. Mohamed, E.S., et al., *Near infrared spectroscopy techniques for soil contamination assessment in the Nile Delta.* Eurasian Soil Science, 2016. **49**(6): p. 632-639.
38. Kokaly, R.F. and R.N. Clark, *Spectroscopic Determination of Leaf Biochemistry Using Band-Depth Analysis of Absorption Features and Stepwise Multiple Linear Regression.* Remote Sensing of Environment, 1999. **67**(3): p. 267-287.
39. Siebielec, G., et al., *Near- and Mid-Infrared Diffuse Reflectance Spectroscopy for Measuring Soil Metal Content.* Journal of Environmental Quality, 2004. **33**(6): p. 2056-2069.
40. Choe, E., et al., *Qualitative analysis and mapping of heavy metals in an abandoned Au–Ag mine area using NIR spectroscopy.* Environmental Geology, 2009. **58**(3): p. 477-482.
41. Dunagan, S.C., M.S. Gilmore, and J.C. Varekamp, *Effects of mercury on visible/near-infrared reflectance spectra of mustard spinach plants (Brassica rapa P.).* Environmental Pollution, 2007. **148**(1): p. 301-311.
42. Rosso, P.H., et al., *Reflectance properties and physiological responses of Salicornia virginica to heavy metal and petroleum contamination.* Environmental Pollution, 2005. **137**(2): p. 241-252.

43. Wu, Y., et al., *A Mechanism Study of Reflectance Spectroscopy for Investigating Heavy Metals in Soils.* Soil Science Society of America Journal, 2007. **71**(3): p. 918-926.
44. Xia, Z., W. Jianting, and Z. Dong. *Band selection method for retrieving soil lead content with hyperspectral remote sensing data*. in *Proc.SPIE*. 2010.
45. Ren, H.-Y., et al., *Estimation of As and Cu Contamination in Agricultural Soils Around a Mining Area by Reflectance Spectroscopy: A Case Study.* Pedosphere, 2009. **19**(6): p. 719-726.
46. Evangelou, M.W.H., M. Ebel, and A. Schaeffer, *Chelate assisted phytoextraction of heavy metals from soil. Effect, mechanism, toxicity, and fate of chelating agents.* Chemosphere, 2007. **68**(6): p. 989-1003.
47. Zhang, S., et al., *Hyperspectral inversion of heavy metal content in reclaimed soil from a mining wasteland based on different spectral transformation and modeling methods.* Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2019. **211**: p. 393-400.
48. Genshan, J. and X. Binbin, *REFLECTANCE OF SOIL CLAY MINERALS AND ITS APPLICATION IN PEDOLOGY.* ACTA PEDOLOGICA SINICA, 1987. **24**(1): p. 67-76.
49. WANG Ting, Z.C., GU YanWen, MA WenChao, LIU Yuan, WEI Hong, *Hyperspectral Estimation of Cadmium Content in Tumorous Stem Mustard Based on the Wavelet-Fractal Analysis.* Scientia Agricultura Sinica, 2018. **51**(1): p. 71-81.
50. Zhou, E., et al., *Field Tests to Determine Static and Dynamic Response to Traffic Loads of Fiber-Reinforced Polyester No-Name Creek Bridge.* Transportation Research Record, 2007. **2028**(1): p. 231-237.
51. Norra, S., et al., *Impact of irrigation with As rich groundwater on soil and crops: A geochemical case study in West Bengal Delta Plain, India.* Applied Geochemistry, 2005. **20**(10): p. 1890-1906.
52. Gewali, U., S. Monteiro, and E. Saber, *Machine learning based hyperspectral image analysis: A survey*. 2018.
53. Darvishzadeh, R., et al., *Mapping grassland leaf area index with airborne hyperspectral imagery: A comparison study of statistical approaches and inversion of radiative transfer models.* ISPRS Journal of Photogrammetry and Remote Sensing, 2011. **66**(6): p. 894-906.
54. Schneider, S., et al., *Gaussian Processes with OAD Covariance Function for Hyperspectral Data Classification*. Vol. 1. 2010. 393-400.
55. Heming, L. and Q. Li, *Hyperspectral Imagery Classification Using Sparse Representations of Convolutional Neural Network Features.* Remote Sensing, 2016. **8**: p. 99.
56. Rinnan, Å., F.v.d. Berg, and S.B. Engelsen, *Review of the most common pre-processing techniques for near-infrared spectra.* TrAC Trends in Analytical Chemistry, 2009. **28**(10): p. 1201-1222.
57. Savitzky, A. and M.J.E. Golay, *Smoothing and Differentiation of Data by Simplified Least Squares Procedures.* Analytical Chemistry, 1964. **36**(8): p. 1627-1639.
58. Gomez, C., P. Lagacherie, and G. Coulouma, *Continuum removal versus PLSR method for clay and calcium carbonate content estimation from laboratory and airborne hyperspectral measurements.* Geoderma, 2008. **148**(2): p. 141-148.
59. Yousefi, G., M. Homaee, and A.A. Norouzi, *Estimating soil heavy metals concentration at large scale using visible and near-infrared reflectance spectroscopy.* Environmental Monitoring and Assessment, 2018. **190**(9): p. 513.

60. Rodarmel, C. and J. Shan, *Principal Component Analysis for Hyperspectral Image Classification.* Surv Land inf Syst, 2002. **62**.
61. Boulgouris, N.V., N.P. Konstantinos, and M.-T. Evangelia, *Discriminant Analysis for Dimensionality Reduction: An Overview of Recent Developments*, in *Biometrics: Theory, Methods, and Applications*. 2010, IEEE. p. 1-19.
62. Hsu, P.-H., *Feature extraction of hyperspectral images using wavelet and matching pursuit.* ISPRS Journal of Photogrammetry and Remote Sensing, 2007. **62**(2): p. 78-92.
63. Rumiana, K. and G. Georgi. *Assessing Cd-induced stress from plant spectral response*. in *Proc.SPIE*. 2014.
64. Carranza, E.J. and A. Laborte, *Random forest predictive modeling of mineral prospectivity with small number of prospects and data with missing values in Abra (Philippines).* Computers & Geosciences, 2014. **74**.
65. Gualtieri, J.A. and F.C. Robert. *Support vector machines for hyperspectral remote sensing classification*. in *Proc.SPIE*. 1999.
66. Bre, F., J. Gimenez, and V. Fachinotti, *Prediction of wind pressure coefficients on building surfaces using Artificial Neural Networks.* Energy and Buildings, 2017. **158**.
67. Huiping, L. and L. Xiangnan. *Hyperspectral analysis of leaf copper accumulation in agronomic crop based on artificial neural network*. in *2008 International Workshop on Earth Observation and Remote Sensing Applications*. 2008.
68. ZOONIVERSE, *Sorghum and Maize Segmentation Using Hyperspectral Imagery*. ZOONIVERSE.
69. (GIC), G.D.I.C., *Hyperspectral Remote Sensing Scenes* 2021, Grupo De Inteligencia Computacional (GIC).
70. Huynh, C.P., et al., *Multi-class support vector machines for paint condition assessment on the Sydney Harbour Bridge using hyperspectral imaging.* Structural Monitoring and Maintenance, 2015. **2**: p. 181-197.