5-2019

A Data-driven Approach for Detecting Stress in Plants Using Hyperspectral Imagery

Suraj Gampa
University of Nebraska-Lincoln, suraj.gampa@huskers.unl.edu

A DATA-DRIVEN APPROACH FOR DETECTING STRESS IN PLANTS USING HYPERSPECTRAL IMAGERY

by

Suraj Gampa

A THESIS

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfilment of Requirements

For the Degree of Master of Science

Major: Computer Science

Under the Supervision of Professors Ashok Samal and Sruti Das Choudhury

Lincoln, Nebraska

May, 2019

A DATA-DRIVEN APPROACH FOR DETECTING STRESS IN PLANTS USING HYPERSPECTRAL IMAGERY

Suraj Gampa, M.S.

University of Nebraska, 2019

Advisers: Ashok Samal and Sruti Das Choudhury

A phenotype is an observable characteristic of an individual and is a function of its genotype and its growth environment. Individuals with different genotypes are impacted differently by exposure to the same environment. Therefore, phenotypes are often used to understand morphological and physiological changes in plants as a function of genotype and biotic and abiotic stress conditions. Phenotypes that measure the level of stress can help mitigate the adverse impacts on the growth cycle of the plant. Image-based plant phenotyping has the potential for early stress detection by means of computing responsive phenotypes in a non-intrusive manner. A large number of plants grown and imaged under a controlled environment in a high-throughput plant phenotyping (HTPP) system, are increasingly becoming accessible to research communities. They can be useful to compute novel phenotypes for early stress detection.

In early stages of stress induction, plants manifest responses in terms of physiological changes rather than morphological, making it difficult to detect using visible spectrum cameras which use only three wide spectral bands in the 380nm - 740 nm range. In contrast, hyperspectral imaging can capture a broad range of wavelengths (350nm - 2500nm) with narrow spectral bands (5nm). Hyperspectral imagery (HSI), therefore, provides rich spectral information which can help identify and track even small changes in plant physiology in response to stress.

In this research, a data-driven approach has been developed to identify regions in plants that manifest abnormal reflectance patterns after stress induction. Reflectance patterns of age-matched unstressed plants are first characterized. The normal and stressed reflectance patterns are used to train a classifier that can predict if a point in the plant is stressed or not. Stress maps of a plant can be generated from its hyperspectral image and can be used to track the temporal propagation of stress. These stress maps are used to compute novel phenotypes that represent the level of stress in a plant and the stress trajectory over time. The data-driven approach is validated using a dataset of sorghum plants exposed to drought stress in a LemnaTec Scanalyzer 3D HTPP system.

DEDICATION

Dedicated to my parents and sister.

ACKNOWLEDGMENTS

Firstly, I would like to extend my sincere gratitude to my adviser and mentor Dr. Ashok Samal for his encouragement and ideas to approach a problem from a new dimension. I am thankful for his constant belief and motivation in this research work and for inspiring me to work towards the goal. His free hand in letting me try new possibilities and approaches while constantly reminding me of the bigger picture has made this research an experiential learning process for me.

I am also grateful to my co-advisor Dr. Sruti Das Choudhury for sparking interest in me to pursue this research. Her constant support through rough trails has helped me successfully reach the completion.

I would also like to use this opportunity to thank my committee member Dr. Vinodchandran N. Variyam for taking time from his busy schedule to review my work and offer constructive opinions.

And lastly, I would like to thank my parents and sister for their unconditional love and encouragement.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Phenotyping is the process of identifying attributes of an organism that can be quantified. In plants, phenotypes can broadly be divided into two types - morphological and physiological. Morphological phenotypes are derived from structure of the plant like leaf length, stem angle, etc. Physiological phenotypes are impacted by the underlying biological processes that occur within the plant like photosynthesis, water absorption, etc. In plants, stress can manifest in the form of physiological changes that can be quantified by phenotyping like photosynthesis rate, water usage efficiency, etc. Stresses in plants can be categorized into two types - biotic and abiotic. Biotic stresses are caused by organisms like parasites, insects and fungi while abiotic stresses are caused by factors like sunlight, temperature and soil. Understanding the impact of stress on growth of plants can help develop plants which can tolerate stress. Stress based phenotypes can help track temporal propagation of stress and study its relationship with different genotypes.

Phenotypes can be computed in two ways - intrusively and non-intrusively. Intrusive approaches involve using hand-held devices for measuring phenotypes directly from the plant. Non-intrusive phenotyping is an indirect approach of computing phenotypes from images of the plant. While intrusive approaches are accurate in measuring stress-based phenotypes, they are not scalable approach and hence are not suitable for large-scale phenotype computation (high-throughput plant phenotyping). Imaging is the most common non-intrusive way of measuring phenotypes in a high-throughput plant phenotyping system. There are many different imaging modalities including infrared thermography, fluorescent imaging and thermal imaging. Due to their ability to capture images at fine resolution, hyperspectral imaging can be particularly helpful in computing stress based phenotypes. A brief description of hyperspectral imaging and its potential for plant phenotyping is described in the following sections.

1.1 Hyperspectral Imaging

Hyperspectral imagery (HSI) gives reflectance across a wide range of the electromagnetic spectrum. While visible spectrum (RGB) cameras capture an image in the visible range of 400-700 nm, hyperspectral cameras can capture imagery beyond visible spectrum typically in the range of 350 - 2500 nm. It is a collection of images, each of which are acquired at narrow wavelength intervals across a continuous spectrum. For ease of management, these images are arranged in an increasing order of their wavelengths in the form of a cube called hyperspectral cube (HSC). Essentially, an output of a hyperspectral camera is a HSC of size (l × b × n) where l and b are the length and breadth of an image capturing spatial information and n is the number of spectral bands which carry the spectral information captured by the camera. Each pixel location on this HSC is a (1 x n) array that can be visually represented as a spectral reflectance curve. These reflectance curves can represent variations in reflectance values with respect to spectral bands. It can help identify spectral signatures.

Initially, hyperspectral imaging was used in remote sensing and found applications in weather forecasting, environmental study, mineral exploration and land use mapping. Later, it has been used to capture images in close range and found applications in the areas of biology, food safety and control, etc. In biology, HSI has helped in identification of specific chemical compositions and in the detection of onset of certain biological processes. In plants, HSI is being used to calculate vegetation indices that can, for example, estimate nitrogen content, canopy water content, evaluate light use efficiency, etc.

1.2 Plant Phenotyping using Hyperspectral Imagery

A phenotype is an attribute of an organism that can be quantified based on the morphological and physiological changes it undergoes due to its genotype and the environment around it [34]. They can be used to understand the underlying plant growth processes. They also play an important role in studying the responses of organisms to different conditions. In plants, they help estimate the effects of biotic and abiotic stresses. Some phenotypes that can be measured in plants are leaf area, stem diameter, plant height, area of convex hull, photosynthesis rate, nitrogen content, salt stress, canopy water content, etc. While plants belonging to different genotypes may manifest same phenotype when the environmental factor is constant, plants kept under different environmental conditions and belonging to same genotype may manifest different phenotypes.

Image based plant phenotyping is a non-intrusive way of measuring plant traits by using spectral information captured using visible, near-infrared (NIR), fluorescent or hyperspectral range cameras. Visible range cameras are widely used for measuring morphological phenotypes whereas NIR, fluorescent and hyperspectral range cameras are competent to measure physiological properties of the plants. The advantage of hyperspectral imagery over other spectral images lies in its ability to capture spectral reflectances in narrower bands over a wide range of wavelengths which can help increase the probability of identification of spectral signatures unique to a certain compound or a biological process. These spectral signatures can act as attributes to quantify physiological changes of a plant and define existing or new plant phenotypes. Some existing phenotypes that can be computed using HSI are chlorophyll content, water content, greenness, photosynthetic activity, etc. [1, 2, 3]. It has also been used to detect diseases in plants by way of quantifying the differences between reflectance data of normal plants and diseased plants. Given its ability to capture reflectance data over a broad range of wavelengths, HSI has the potential to detect and predict important events in a plant's life cycle like early-stage stress detection, prediction of flowering day, etc., as well as estimate the productivity of a crop and various factors effecting it.

1.3 Research Challenges

One of the main challenges of using HSI is the curse of dimensionality. A large number of spectral bands increases the spectral dimensionality of the dataset and thereby adds a new set of challenges. Also, high correlation in the reflectance data of nearby spectral bands due to closely spaced wavelength intervals can introduce data redundancy. Analysis of close-range hyperspectral imagery on plants must address irregular illumination effects caused by uneven leaf surfaces and overall canopy structure. Since they can significantly alter the reflectance values, they cause misclassification.

Development of algorithms for computing stress based phenotypes has many challenges as listed below:

- **Data collection:** A large dataset that includes multiple stressed and unstressed plants imaged over period of time in a uniform environment must be compiled before any algorithm development can start. This can be often difficult without a large greenhouse facility.

- **Lack of ground truth:** While we can record when the stress is introduced in a plant, it is unclear when and how different parts of the plant will be impacted and manifest in the captured images. Development of a classifier therefore must be preceded by identifying normal and outlier behavior.

- **Non-uniformity in plant's growth patterns:** Plants are complex organisms and hence different plants of same genotype growing under same conditions exhibit different responses. Furthermore, different parts of a plant (stressed on unstressed) may also have different reflectance properties. Characterization of the exact nature of the normal reflectance must therefore be a part of any stress identification algorithm.

1.4 Contributions

The aim of this study is to identify unique traits in reflectance patterns of a stressed plant when compared to a normal plant and build a classifier that can detect if a plant is under drought stress in its early stages. This thesis has the following novel contributions:

- A data-driven approach to identify classes in stressed and unstressed plants and generate ground truth by segmenting stressed responses from unstressed.

- A classification model that can predict stress in a plant based on reflectance.

- A novel stress phenotype that can quantify stress in a plant.

These techniques enable a non-intrusive approach for early-stage stress detection in plants. The stress phenotype will help identify genotypes which have better drought tolerant characteristics. As we will be able to contain the issue in early stages, it can also help increase the productivity of the plant.

1.5 Thesis Overview

This section discusses the topics covered in this thesis. Chapter 2 discusses related work in the field of computer vision applied to hyperspectral imaging. Chapter 3 defines the problem and the proposed methodology to solve it. Chapter 4 discusses the details of the experiment conducted to obtain the dataset and shows the results of applying our methodology on it. Chapter 5 concludes the report and discusses future work.

Chapter 2

Related Work

Employing hyperspectral data to compute physiological phenotypes that could detect and quantify stresses in plants has been an active research area in plant science and agronomy. However, HSI analysis often comes at a cost of dealing with large, high dimensional datasets. This problem can be mitigated using dimensionality reduction techniques that can convert a high dimensional dataset into a low dimensional dataset with minimal information loss. Section 2.1 focuses on related work in dimensionality reduction for close-range HSI of plants. Classifier models, when trained on needful HSI data, can help predict and quantify stress activity in plants. In this thesis, we focus on close range HSI of individual plants. Section 2.2 focuses on related work in classification of single plant images.

2.1 Dimensionality Reduction

Dimensionality reduction can help mitigate the effects of curse of dimensionality in HSI by extracting useful features that have higher inter-class variability. Some predominantly used techniques described below.

Principal Component Analysis: It is a data transformation technique that aims at reducing the dimensionality of the original data by calculating a new set of features called principal components (PCs). PCs are latent features that are linear combinations of the original features in the dataset. They are orthogonal to each other which make them an independent set of features. Qin et al. [5] used principal component analysis (PCA) on HSI of citrus fruit to extract useful features. Based on visual inspection of the score images, the first five principal components provided meaningful information that could differentiate normal grape fruits from the ones infected with citrus canker disease.

Kernel PCA: It is an extension of PCA where the data is projected onto a higher dimensional space defined by the kernel before performing dimensionality reduction. It is a non-linear dimensionality reduction technique typically used for data that is linearly inseparable. Here, the PC's are non-linear combinations of the original features where the non-linearity is defined by the kernel. Some widely used kernels for this operation are Radial Basis Function, Polynomial and Sigmoid.[14]

Independent Component Analysis: This technique assumes that a random variable in a given data object is a linear mixture of independent components (IC's) such that the IC's are statistically independent from each other and each IC has a non-gaussian distribution [15]. It has proven to be effective in performing source separation and has found applications in medical signal processing and image processing.

Autoencoders: This introduces a deep learning approach for extracting high level features from data in an unsupervised manner [20]. It works in an encoder-decoder model where input x is encoded into a lower dimensional latent feature space h and reconstructed back. The reconstruction loss computed from decoder output is used as a metric to fine tune the feature space. This model is resilient to noise present in the input data and can work even on unseen data.

While the aforementioned techniques are used for feature transformation in HSI, we have to understand that the transformed features no longer preserve the original spectral information which may be useful to deduce certain conclusions about the spectral bands like their biophysical meaning, etc [4]. The following statistical techniques can help identify features that contribute most towards inter-class variability and those which are most likely independent of class.

Feature Selection using Fischer's statistics: It is a filtered feature selection process that aims to enhance the class separability between two predefined classes [17] [4]. The Fischer's statistics (F-value) for a given feature is the ratio of between class variability and within class variability for that feature. If the F-value of a feature is larger than other features and its corresponding p-value is less than desired significance level, we can say that that feature offers better class separability compared to other features. Asaari et al. [4] used F-value criterion to rank individual spectral bands on HSI of maize plant. The top-scoring spectral bands that could clearly discriminate between healthy and drought stressed plants were selected. This technique, unlike PCA where new features are extracted from original features, preserves the original features that could later be interpreted for its properties.

Tree based feature selection: A Decision tree classifier is a white-box classification model [18]. The inverted tree look-alike structure is build up of decision nodes and tree nodes. The decision node represents a test on a feature of the dataset while the leaf node is the classification result. This classifier makes it possible to visualize the hierarchical decision making process in the tree. As a result, it is possible to ascertain the significance of features. However, it might require a certain degree of hyperparameter tuning to give optimal results.

2.2 Classification

Classification can help detect and quantify stress levels in plants. In HSI, it is widely used for remote sensing data [19] [20]. This thesis is focused on classification of individual pixels (spectral curves) for close-range HSI of individual plants. Some widely used techniques are:

Support Vector Machines (SVM): SVMs' [21] are widely applied to classification problems and nonlinear regressions [22]. It aims to find the best separating hyperplane between classes by means of maintaining the largest distance from the data points. This is achieved by adding two equidistant, parallel, imaginary hyperplanes to the separating hyperplane and trying to widen the margin between these two imaginary hyperplanes. The wider the margin, the higher is the generalization ability of the classifier. In HSI classification, SVM technique have helped in significant reduction of classification complexity and improved classification accuracy [23]. SVMs have also been proved as a valid and effective alternative to conventional pattern recognition approaches which employ a combination of feature reduction and classification methods to classify hyperspectral remote sensing data [24]. Behmann, et al. [12] have been able to detect drought stress in barley plants using a Linear Ordinal SVM classifier with an accuracy of 70%. Linear Ordinal SVM has proved to be a compact model which could be applied to a high-throughput phenotyping system under limited resources. Rumpf, et al. [10] used spectral vegetation indices calculated from hyperspectral data as features to classify the normal plants from plants affected by sugar beet diseases. They have used support vector machine (SVM) as a classifier to detect early onset of the disease and discriminate between three other diseases.

Artificial Neural Network (ANN): Artificial Neural Networks are a class of supervised and unsupervised machine learning algorithms which take inspiration from the biological nervous system. While kernel-based SVMs and decision trees employ a two-layered model, ANNs' use multiple layers of processing to extract more abstract features that are capable of achieving higher levels of accuracy in classification [25]. In HSI classification where training datasets can be large and high-dimensional with high degrees of spectral-spatial diversity, ANNs' can automate the process of feature construction by building high-level features from low-level ones. Baranowski, et al. [11] were able to achieve 90.5% prediction accuracy using a back-propagation neural network for predicting responses of Oilseed Rape to a fungal species of Genus Alternaria. Rojas, et al. [26] were able to quantify different levels of physical perturbation on the mushroom pilei plant using an ANN.

Chapter 3

Approach

3.1 Problem Definition

The problem addressed in this research is defined as follows. Given a dataset, derived from a stress based experiment that includes (a) a set of plants P defined as: $P = \{p_1, p_2, \ldots p_i, \ldots, p_n\}$, where plant p_i is given by $p_i = \{p_{i1}, p_{i2}, \ldots, p_{ij}, \ldots, p_{im}\}$ and $p_{i,j}$ is the hyperspectral cube of plant p_i at time j and (b) a set of stressed plants $P_s \subset P$, develop (a) a classifier χ which maps a hyperspectral cube of a plant, p to a stress image s, i.e. $s = \chi(p)$ and (b) novel stress phenotypes algorithm, ψ which maps the stress image s to a set of phenotypes Σ, i.e. $\Sigma = \psi(s)$.

3.2 Overview

The following methodology is divided into six modules as follows:

- **Preprocessing:** As illustrated in Figure 3.3, this step involves removing the most noisy bands from the spectrum, performing image segmentation on p_{ij} to separate plant part from the background and generating spectral curves SC_{ij} (refer Algorithm 1).

- **Denoising:** Denoising the spectral information involves eliminating the illumination effects from the universal set of spectral curves S (refer Algorithm 1), which otherwise adversely effect the analysis results.

- **Grouping spectral classes:** Spectral curves from denoised set N are divided into normal class SC^n and stress class SC^s using a list of stressed plants P_s (refer Algorithm 1). Each of the SC^n and SC^s are grouped into κ spectral classes and stored in C^n and C^s respectively. The spectral classes identify different physiological processes that occur in a plant.

- **Spectral Feature Selection:** As referred to in Algorithm 3 this process helps in identifying and selecting spectral bands that manifest large variability between stressed and unstressed plants.

- **Ground truth generation and labeling:** Here, the spectral curves which show stress behavior are segregated from normal and labeled likewise (refer Algorithm 4).

- **Stress classification modeling:** It involves developing a Hyperspectral Imagery Stress Classifier (HISC) (refer Algorithm 3) that can predict if a given spectral curve is stressed or unstressed, as illustrated in Figure 3.1.

- **Phenotype computation:** The output from the HISC for a given plant can be used to plot a stress map and compute novel stress phenotypes (refer Algorithm 4).

3.3 Preprocessing

The raw image data obtained from a hyperspectral camera can contain blacked-out images and non-plant objects like pot, soil, etc. As illustrated in Figure 3.3, the

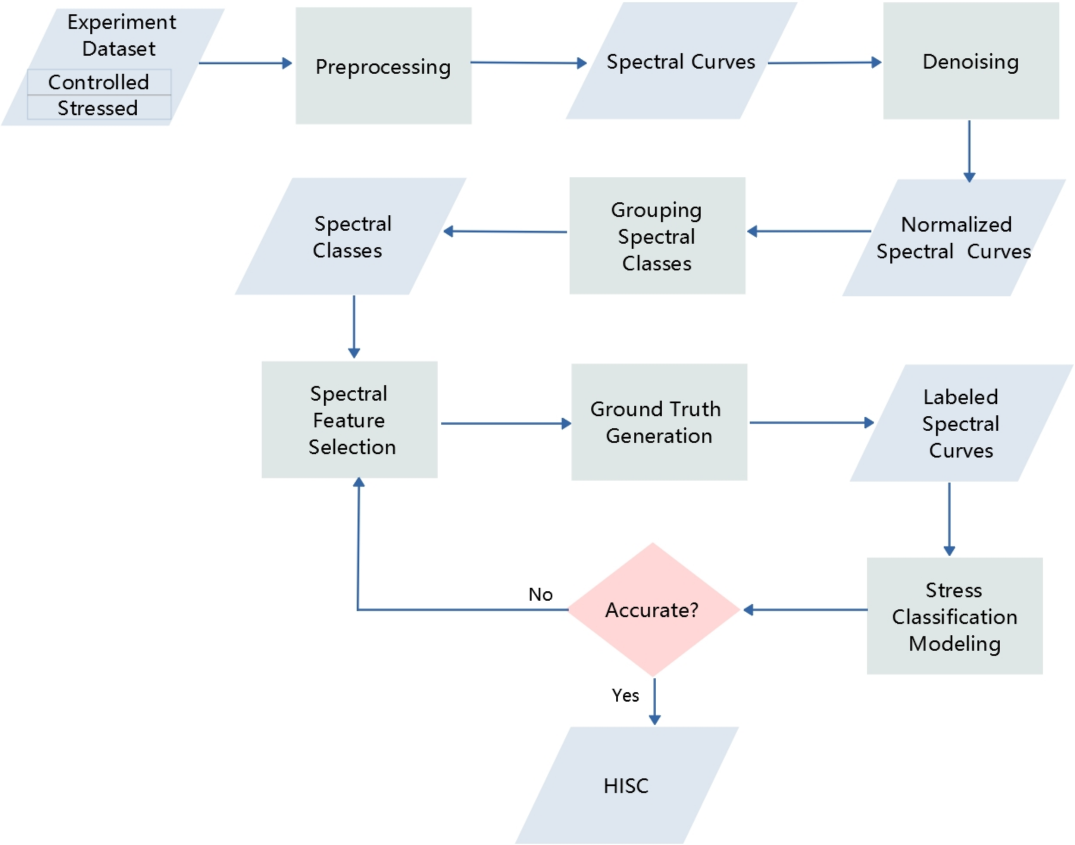


Figure 3.1: Overview of the hyperspectral imagery based stress classifier development

Algorithm 1 Stress Phenotype Computation Algorithm

Input: $P = \{P_{11}, P_{12}, ..., P_{1,n}, P_{2,1}, ..., P_{m,n}\}, P_S, A_{opt}, t_{dist}$

Output: Phenotypes

```
$S = \phi$
for i = 1...m do
    for j = 1...n do
        $S_{ij} \leftarrow Segment(P_{ij})$
        $SC_{ij} \leftarrow GenerateSpectralCurves(S_{ij})$
        $S = S \cup SC_{ij}$
$N \leftarrow Denoise(S)$
$\{SC^n, SC^s\} \leftarrow DivideSpectralCurves(N, L_S)$
$\{C^s, C^n\} \leftarrow Group(SC^n, SC^s)$
$HISC \leftarrow DevelopClassifier(C^s, C^n, A_{opt}, t_{dist})$
```

process of removing noisy bands, image segmentation and generating spectral curves is primary for data analysis.

3.3.1 Removal of Noisy Bands

The hyperspectral camera in a HTPP system is prone to capture extremely noisy images, especially in the initial spectral bands. It is not possible to extract the underlying spectral information from these images. If present, they can adversely effect classification and analysis results. So it is important to eliminate these spectral bands from our study.

3.3.2 Plant Segmentation

The HSI of sorghum plants can consist of non-plant objects like pot, soil, etc., which need to be separated from the plant part. To obtain spectral information corresponding only to the plant part, we perform plant segmentation. For close-range HSI of plants, one of the widely used segmentation techniques is Normalized Difference Vegetation Index (NDVI) based segmentation. Previously, it has been used in [8] and [9] for generating a plant mask which is superimposed on the HSI cube to extract plant part.

The NDVI based segmentation utilizes the disparity in absorption and reflectance patterns of plants to segregate pixels corresponding to the plant part. For spectrum in the HSI range (546 nm - 1700 nm), chlorophyll present in a plant absorbs a major portion of the visible part of spectrum (400 - 700 nm) but reflects about 50% of the infrared part of spectrum (700 - 1100 nm). In contrast, the non-plant parts absorb and reflect both parts of the spectrum uniformly as shown in Figure 3.2. The NDVI based segmentation utilizes this disparity in reflectances to segregate plant pixels. The NDVI is an index that measures photosynthetic activity by utilizing the disparity in absorption of visible and infrared spectra by the chlorophyll pigment. It is given by the formula:

$$NDVI = \frac{P_{\lambda_{NIR}} - P_{\lambda_{VIS}}}{P_{\lambda_{NIR}} + P_{\lambda_{VIS}}} \tag{3.1}$$

where $P_{\lambda_{NIR}}$ and $P_{\lambda_{VIS}}$ are the proportions of radiations reflected by a plant for wavelengths λ_{NIR} in near infrared and λ_{VIS} in visible range respectively. The range of NDVI is between -1 and 1. Higher values of NDVI would mean the plant is rich in chlorophyll content and lower values mean otherwise.

For segmentation of a HSC P_{ij}, a given pair of λ_{NIR} and λ_{VIS} values are selected. A mask of the plant part is created by subtracting the selected λ_{NIR} band image from the selected λ_{VIS} band image. This fades out the background part and exposes only the plant part. A binary version of this mask converts the values of background pixels to zero and plant pixels to one, which makes it easy to superimpose on a HSC. So, a segmented HSC S_{ij} is generated by binarizing the mask using a given threshold and then superimposing it on all the band images in P_{ij}. The values of λ_{NIR} and λ_{VIS} need to be fine tuned for a higher NDVI value. The higher the NDVI value, the better is the segmentation result.

3.3.3 Spectral Curves Generation

A spectral curve is a one dimensional vector quantity that represents a given pixel along its spectral dimension. In this thesis, it forms the input for pixel-wise data analysis. It can be graphically represented as shown in Figure 4.3. Spectral curves can help represent three-dimensional HSCs in a two-dimensional tabular format where rows represent the pixels and columns represent the reflection coefficients across each of the spectral bands. Each spectral curve is also tagged with its respective plant ID, time stamp and pixel location. For a given segmented HSC S_{ij}, the corresponding

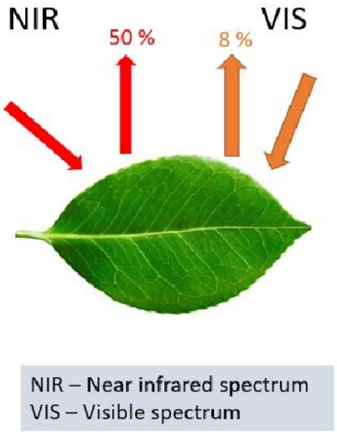


Figure 3.2: Difference in absorption of visible and near infrared wavelength by plants

set of spectral curves is given by SC_{ij}. For better management and portability of the dataset, the spectral curves SC_{ij} corresponding to all plants P are merged into a single set S given by Equation 3.2.

$$S = \{SC_{11}, SC_{12}, \ldots, SC_{1j}, SC_{2j}, \ldots, SC_{ij}\} \tag{3.2}$$

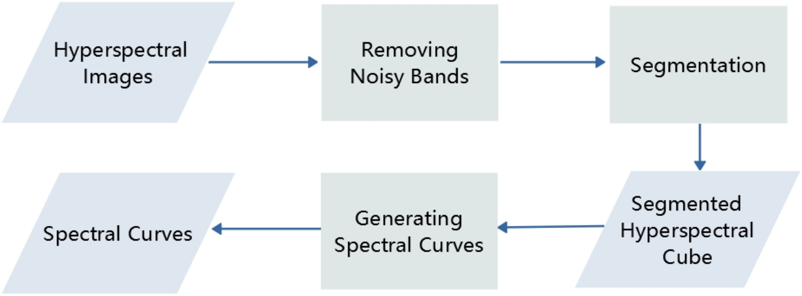


Figure 3.3: Preprocessing

3.4 Denoising Spectral Information

In a high-throughput plant phenotyping(HTPP) chamber, uneven illumination effects is a phenomenon that can be caused due to the morphological structure of the plant, mainly due to the leaves as shown in Figure 4.4. There are two ways that light gets reflected by leaves - one is reflections caused when leaf acts as a Lambertian surface (i.e. reflected rays scatter uniformly in all directions) and the second is specular reflections (i.e. incident light is reflected in a single direction) [4]. While the energy received by the hyperspectral camera due to leaf acting as a lambertian surface depends on factors like the cosine of the angle between light source and leaf and distance between leaf and camera, specular reflections are largely caused by the texture of the leaf and may vary from plant to plant [4]. Both these effects are wavelength independent and add as a scalar factor to the reflectance values. But their presence can pose undue advantage to certain spectral bands during analysis and classification. A viable approach to correct these illumination effects is using Standard Normal Variate (SNV) technique which is given by the equation,

$$SNV = \frac{X - \mu}{\sigma} \tag{3.3}$$

where X is the reflection coefficient for a pixel at a given spectral band and μ and σ are respectively the mean and standard deviation for the set of reflection coefficients for that band.

SNV transformation is performed for each spectral band independently. For an input S to this process, the resultant denoised dataset is given by N.

3.5 Grouping Spectral Classes

There can be several physiological processes that occur in plants during its life cycle which manifest uniquely in HSI. Also, these processes may respond differently to stress. It may be beneficial to divide the HSI data into distinct classes, each representing a unique process, and analyze them individually for stress patterns. In a data-driven approach, the optimal value of the number of clusters (κ) can be calculated using elbow method described in Section 3.5.1. Later, the closest distance to centroid based κ-means clustering technique is used to divide the spectral data into κ spectral classes.

3.5.1 Determination of Optimal Number of Clusters

There are different ways to determine optimal κ. Some widely used methods are elbow method, dendrogram, etc. In this thesis, we use compute it over the cumulative dataset of spectral curves S using elbow method. Elbow method helps in visually illustrating the optimal value of κ (κ_{opt}) by plotting the total Error Sum of Squares(SSE) values for different κ. The value of κ where we find a visual elbow is considered the optimal clustering number.

For a given cluster of data points A, SSE_A can be defined as the sum of the squares of differences between each data point in A and the mean of the cluster $\overline{x_A}$, given by the Equation 3.4.

$$SSE_A = \sum_{i \in A} (x_i - \overline{x_A})^2 \tag{3.4}$$

For a dataset X containing κ clusters of data points, the SSE over all the κ

clusters, SSE_k, is the sum of SSE's of each of the clusters, given by the Equation 3.5.

$$SSE_k = \sum_{j \in Y} SSE_j \ , \ where \ Y \ = \{1...\kappa\} \tag{3.5}$$

As the value of κ increases, the average distance of each point to its respective cluster center decreases resulting in lower SSE_k values. Typically, we notice a steady decrease in SSE for every increase in κ value. However, the impact of κ on SSE is not uniform. The elbow point is where an increase in κ does not lead to a substantial decrease in value of SSE i.e. the curve starts to flatten. The value of κ at this point is considered the optimal clustering number.

However, for large volume of high-dimensional data, the values of SSE's can be huge and may cause a hurdle in estimating the elbow on graph. To mitigate this problem, we use a modified cost function $\overline{SSE_k}$ derived from [4], given by Equation 3.6.

$$\overline{SSE_k} = |\log(SSE_k) \ - \ \log(SSE_{k-1})| \tag{3.6}$$

The elbow point on a plot of $\overline{SSE_k}$ values for different κ gives κ_{opt} for dataset S.

3.5.2 Clustering

In a data-driven approach, clustering may help identify different classes of physiological processes that occur in plants. For this, we use the closest distance to center based κ-means clustering. It is a technique that is used to partition N data points into κ clusters such that the distance from each data point to the center of cluster is minimum. The run-time complexity of this algorithm is $O(N^2)$ [35]. In HSI classification, κ-means algorithm (Algorithm 2) [30] has been used for data labelling and quantifying similarity. A specific value of κ can be decided based on a prior knowledge about the data or using techniques like elbow method (Section 3.5.1), dendrogram, etc., to examine the underlying data structure. In this thesis, κ_{opt} has been derived from Section 3.5.1 and clustering has been performed on spectral curves belonging to the normal plants (SC^n) and stressed plants (SC^s) separately, as shown in algorithm 2 [40]. The output of clustering is stored in cluster sets C^n and C^s corresponding to normal and stressed plants respectively.

Algorithm 2 κ-means algorithm for Spectral Curves

```
Input: $\{SC^i, \kappa_{opt}\}$
Output: $C^i = \{C^i_1, C^i_2, ..., C^i_\kappa\}$
Select $\kappa$ spectral curves as the initial centroids.
repeat
    Form $\kappa$ clusters by assigning all spectral curves to the closest centroid.
    Recompute the centroid of each cluster.
until the centroids do not change.
return Set of clusters ($C^i$).
```

To explore the abstract level behavior of each cluster, we plot the mean spectral curves graph which is a graphical representation of means calculated for each of the clusters across all the bands. They can be used to visualize the differences between clusters and mark unique spectral signatures. They can also be used to compare and contrast corresponding cluster behavior between a stressed and an unstressed plant. A sample of the mean spectral curves for a single, unstressed plant is shown in Figure 4.6. In this thesis, mean spectral curves are used as a reference to pair corresponding clusters in C^n and C^s in a hierarchical order and assign each pair a cluster pair index (CPI).

3.6 Spectral Feature Selection

Given the high dimensionality of the spectral data, there is a need to select only those spectral features (bands) which offer best inter-class separability while reducing the curse of dimensionality and overfitting during classification [7]. An effort has been made to find the contribution of each band towards differentiating unstressed from stressed plants using the Fisher's statistics (F-value) criterion [4]. For each cluster pair CPI (Section 3.5.2), F-value is computed for each of the spectral bands. The K highest scoring F-values whose p-values are less than a 0.05 significance level (95 % and above confidence level) are selected as K best features of the data set. The higher the F-value, greater is the between group variability for that spectral band which translates to that feature being able to better differentiate between the normal and stressed classes. The highest scoring feature offers the best inter class separability. In this thesis however, small K may not produce desired accuracies while classification as we consider stress manifesting as a complex function of interdependent features. As referred to in Algorithm 3, K acts as a tuning parameter to achieve optimal classification accuracy (A_{opt}) for stressed and unstressed spectral curves, X_n and X_s respectively. The value of K is incremented until the accuracy of classifier reaches A_{opt} (Algorithm 3).

3.7 Ground Truth Generation

As we assume that not all the spectral curves in plants that are subjected to stress actually manifest stress behavior, a methodology has been proposed, as shown in Figure 3.4, to identify the stressed spectral curves in the stressed class of data and label them accordingly. After selecting the best spectral features (Section 3.6), an effort as been made to identify the stressed cluster pairs (3.7.1) which noticeably manifest stress. To quantify the divergence of each spectral curve from normal, a normal curve (3.7.2) is generated and taken as a reference to compute the distances to each of the spectral curves. To identify spectral curves which manifest stressed behaviour, a statistical data binning approach [13] is used to plot histograms and identify spectral curves which manifest abnormal behavior. A detailed description of these methodologies are discussed in the following sections.

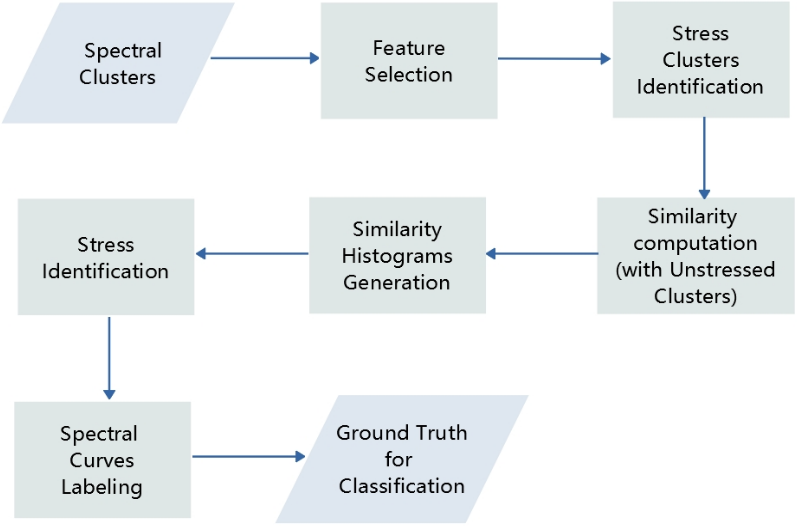

Spectral Clusters
Feature Selection
Stress Clusters Identification
Similarity computation (with Unstressed Clusters)
Similarity Histograms Generation
Stress Identification
Spectral Curves Labeling
Ground Truth for Classification

Figure 3.4: Generating ground truth

3.7.1 Stress Clusters Identification

From the notion that the physiological processes in a plant respond in varying magnitudes to stress, we assume that not all clusters manifest same degree of stress. This procedure aims to identify clusters which are noticeably effected by stress. We use SAM as a comparison metric due to it ability to use spectral direction and not magnitude as metric to measure similarity between spectra [29] [36]. As described in Algorithm 3, a normal cluster (C_j^n) and its corresponding stressed cluster (C_j^s) for a given CPI are taken and K best features are selected using the spectral feature selection process described in Section 3.6. Then the SAM value is computed between the mean spectral curves of the normal and stressed class for those K best features.

If this SAM value is greater than t_{dist}, then the stressed cluster (C_j^s) in cluster pair is manifesting noticeable stress behavior. This cluster pair is labeled stressed cluster pair (SCP) and its index j is added to the $Indices^s$ list. This process is carried out for all corresponding cluster pairs in stressed and unstressed classes.

A SCP is a cluster pair containing a cluster from normal class paired with its corresponding cluster in stress class which noticeably manifests stress. There can be more than one SCP for a given data set. If there are more than one SCP's, the K best features of the SCP with highest t_{dist} value is considered the universal set of best features and applied to the whole dataset for further analysis.

3.7.2 Similarity Computation and Histogram Generation

We were able to identify SCPs'. However, not all pixels in the stressed clusters corresponding to SCPs' show stressed behavior. There is a need to segregate stressed pixels from normal pixels. For this, each SCP is considered individually and a mean spectral curve of its normal cluster, i.e., a normal curve $(SC_{mean}^{n_j})$ is generated, as described in Algorithm 3. Spectral angles (SAM) are computed between this normal curve and each spectral curve in the normal and stressed clusters, (C_j^n) and (C_j^s) and stored in d_n and d_s respectively. A statistical data binning approach [13] is used to plot two histograms, h_n and h_s, for d_n and d_s respectively. To compare these histograms on a same scale, we normalize the bin heights by approximating it to a probability density function. This process is carried out for all SCPs'. At the end of this process, a pair of histograms, h_n and h_s, are generated for each SCP. They help with segregating stressed pixels from normal and labeling them accordingly.

3.7.3 Stress Identification and Spectral Curves Labeling

The histogram pairs generated for each SCP create a means to visually identify pixel groups in stressed classes which show unique behavior. For a given SCP, the histograms are compared with each other to decide on a threshold of SAM (t_{SAM}). However, the binning size is unknown, which makes it difficult to decide on the t_{SAM}. The normal density curves plotted for d_n and d_s may not always fit the shape of histogram correctly. To estimate an unknown probability density function for d_n and d_s, we go for kernel density estimation (KDE) [38] where a kernel function (gaussian, in our case) is fitted to each data point and summed over all data points. This produces a smooth density estimate curve for each of d_n and d_s. Now, when these two KDE curves are plotted together,t_{SAM} is selected as the point where the two curves diverge.

Each SCP may have a different t_{SAM} value. For a given SCP, all pixels in its stressed cluster whose SAM values are greater than t_{SAM} are labeled as stressed pixels (x_s) and all other points in both normal and stressed cluster of that SCP are labeled normal or unstressed (x_n). This process is implemented for all SCPs'. The x_n and x_s pixel sets pertaining to all SCP's are grouped into X_n and X_s respectively which correspond to all normal and stressed pixels. Also, all pixels pertaining to clusters in both stressed and unstressed classes which are not part of SCP (non-SCP) are labeled normal as well and added to X_n.

3.8 Stress Classification Modeling

Although t_{SAM} helps in differentiating between normal and stressed spectral curves in the stressed class, there may be quite a number of spectral curves even in the normal class whose SAM values are greater than t_{SAM}, as seen in Figure 4.9, which requires us to build a classifier that can differentiate between these two classes. Also,

Algorithm 3 *DevelopClassifier*

```
Input: $C^s, C^n, A_{opt}, t_{dist}$
Output: Hyperspectral Imagery Stress Classifier(HSISC)
$K$ = Number of features
$q = |C^s| = |C^n|$
for i = 1...$K$ do
    for j = 1...q do
        if $Distance(C^s_j, C^n_j) > t_{dist}$ then
            $Indices^s = \cup \{j\}$
            $SC^{n_j}_{mean} \leftarrow GenerateNormal(C^n_j, Indices^s)$
    for j in $Indices_s$ do
        $d_n \leftarrow GenerateDistances(C^n_j, SC^{n_j}_{mean})$
        $d_s \leftarrow GenerateDistances(C^s_j, SC^{n_j}_{mean})$
        $\{h_n, h_s\} \leftarrow GenerateSimilarityHistograms(d_n, d_s)$
        $\{x_n, x_s\} \leftarrow GroundTruth(h_s, h_n)$
        $X_n = \cup x_n$
        $X_s = \cup x_s$
    $Classifier_i \leftarrow BuildClassifier(X_n, X_s)$
    if $Accuracy(Classifier_i) >= A_{opt}$ then
        $HISC \leftarrow Classifier_i$
        return $HISC$
```

the classifier needs to be able to differentiate between spectral curves belonging to SCP and non-SCP classes. To learn these complex patterns and underlying hidden features, we use a supervised artificial neural network (ANN), also called a multi-layer perceptron based classifier.

An ANN is a collection of artificial neurons (also called perceptrons) structured with an input layer, hidden layer and an output layer. Each layer can have any number of nodes and there can be multiple hidden layers. Each perceptron in the hidden layer takes a weighted sum of all inputs from training data set, passes them through an activation function and fires an output. Each perceptron in the output layer in turn takes a weighted sum of outputs from each perceptron in the hidden layer, passes them through an activation function and gives a final output. These output labels are compared with their corresponding ground truth labels and an error is calculated.

Some widely used activation functions are Softmax, Sigmoid, Tanh and Rectifier Linear Unit (ReLU) [28]. In this thesis, we use a Proximity Rectifier Linear Unit (ReLU) as an activation function for perceptrons in the hidden layers and sigmoid in the output layer. The PReLU activation function [31] [32] supposedly improves model fitting with reduced risk of overfitting the model. The sigmoid activation function in the output is widely used for estimating labels in a binary classification model.

The output error can be reduced by fine tuning the weight vectors in the network. This is achieved using the back propagation algorithm [27] where the error is propagated back into the network and used to tune the weights connecting output layer to hidden layer, weights connecting sequential pairs of hidden layers (in case of a multi-layer network) and weights connecting input layer to the hidden layer. This back propagation mechanism is achieved using the gradient descent algorithm [27]. The process of fine tuning the weights is carried out iteratively until their values stabilize. Each iteration is called an epoch. The number of epoch that a model needs to be trained on is an important hyper parameter that needs to be defined while model building. We use early-stopping criteria with a given patience factor to decide on an optimal epoch number.

The classifier (C_i) is trained on two classes of spectral curves: Stressed (X_s) and Unstressed (X_n). To estimate the accuracy of the prediction model and also make sure that the model generalizes well on independent datasets, we perform K-fold cross-validation while training [39]. At the end of training, we calculate certain evaluation metrics like accuracy, precision and recall using the test dataset and compare it with A_{opt}. If accuracy is less than A_{opt}, we increment the size of feature set (K) and repeat the process. The qualifying classifier is named Hyperspectral Imagery Stress Classifier (HISC).

3.9 Stress Phenotype Computation

The HISC can perform point wise prediction of whether a given pixel (spectral curve) belongs to stressed or unstressed class. For HSI of a given plant (P_{ij}), the output of HISC can be used to generate a stress map on the plant image with S points marked as stressed and the remaining U marked unstressed. To measure the spatial extent of stress, we propose a novel phenotype called Plant Stress Index (PSI) that is given by the equation 3.7. It is the ratio of pixels marked stressed (S) and the total number of pixels ($S + U$). The range of PSI is in between 0 and 1.

$$PSI = \frac{S}{S+U} \tag{3.7}$$

For computing phenotypes of P_{ij}, as referred to in Algorithm 4, we need to perform plant segmentation, generate spectral curves S from segmented plant part S_{ij} and denoise the data, as shown in Figure 3.5. The denoised data N is now sent to HISC where the K best features are selected and then passed on to the classifier part. The output of classifier is used to generate a stress map and compute phenotypes.

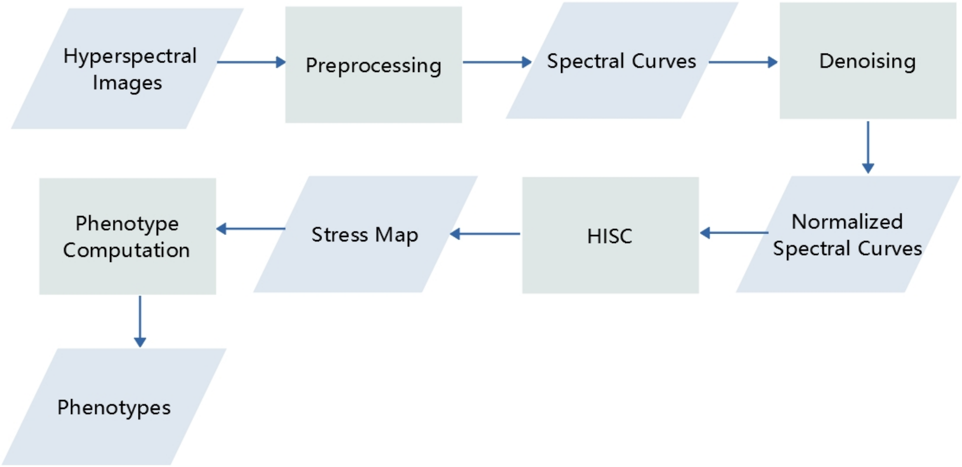


Figure 3.5: Stress phenotype computation

Algorithm 4 Stress Phenotype Computation Algorithm

Input: P_{ij}

Output: Phenotypes

```
$S_{ij} \leftarrow Segment(P_{ij})$
$S \leftarrow GenerateSpectralCurves(S_{ij})$
$N \leftarrow Denoise(S)$
$StressMap \leftarrow HISC(N)$
$Phenotypes \leftarrow ComputePhenotypes(StressMap)$
```

Chapter 4

Implementation and Results

In this chapter, we discuss the results for the methodology described in chapter 3.

4.1 Dataset Description

The HSI dataset P used for evaluating our algorithm was of sorghum plant. The experiment was conducted at Lemnatec Scanalyzer 3D high-throughput plant phenotyping (HTPP) facility in Nebraska Innovation Campus using a hyperspectral camera that was able to capture spectral information for wavelength bands between 546 nm and 1700 nm with a 4.7 nm wavelength interval (243 bands). The dataset comprised of ten plants belonging to the same genotype and have been imaged for a span of eight days. Of the ten plants, five were being grown under unstressed conditions and the other four were subjected to drought stress conditions where drought was induced on day one. The images were captured from the side view at an angle of 90 degrees.

4.2 Preprocessing

4.2.1 Removal of Noisy Bands

Through visual inspection, the eight spectral bands from wavelengths 546nm to 583nm were found to be noisy. Samples of noisy band images are shown in Fig.4.1., with most parts of the image either completely dark or white. The gray-scale images pertaining to these bands were removed.

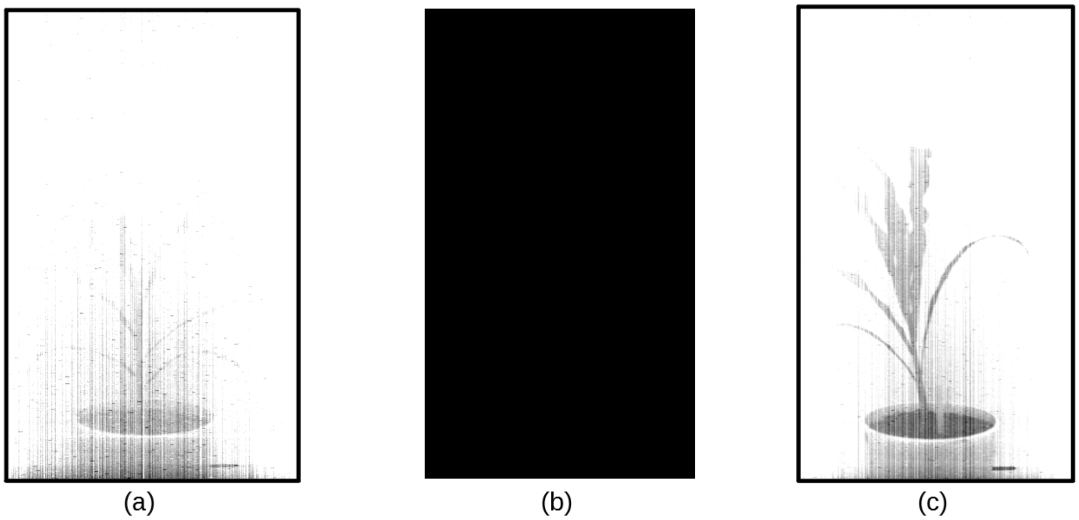


Figure 4.1: Noisy bands

4.2.2 Plant Segmentation

In this study, we used spectral images of plants at $\lambda_{VIS} = 678nm$, as shown in Fig.4.2 (a), and $\lambda_{NIR} = 800nm$, as shown in Fig.4.2 (b), to perform segmentation. Each of them is multiplied by two to increase the intensity of their pixels as shown in images Fig.4.2 (c and d). These brightened images are subtracted (Fig.4.2 (e)) and then subjected to binarization (Fig.4.2 (f)) with a threshold of 0.25. This creates a mask of the plant part. In some plants, there were bits and pieces of background

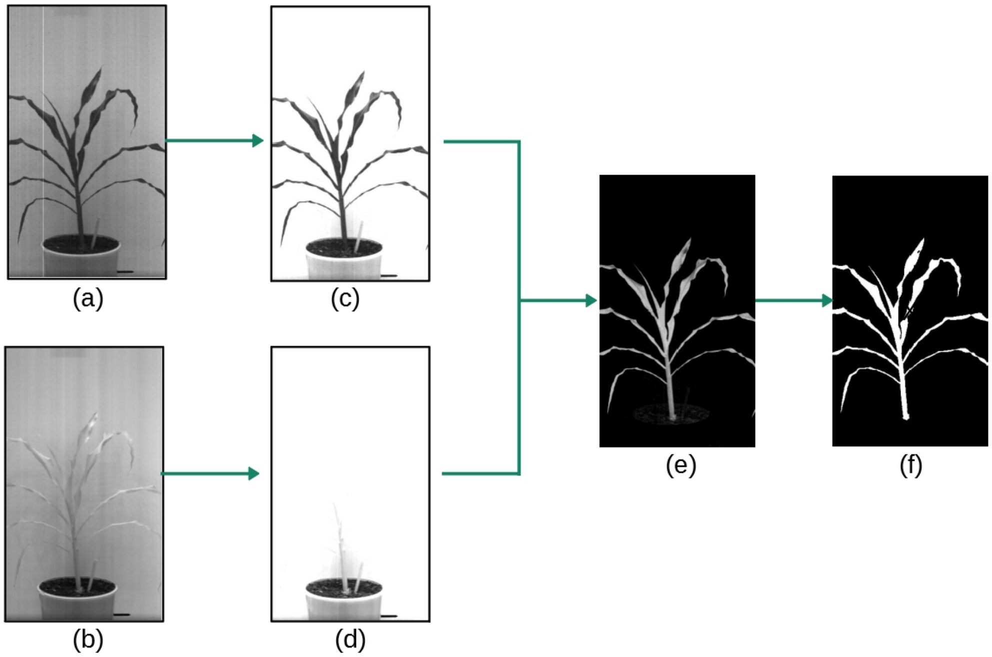


Figure 4.2: NDVI based plant segmentation

left after binarization. To eliminate this kind of noise, parts in the image that were disconnected from plant with a pixel count less than 500 were removed. The resultant mask is then mapped onto all the images across the HSC, P_{ij}, giving a segmented HSC, S_{ij}.

4.2.3 Spectral Curves Generation

For better manageability of data, spectral curves corresponding to all HSCs' are stored in a single set S. For each HSC, a pixel belonging to the plant part was scanned along the spectral dimension and appended to a 2-D array SC_{ij} along with its tag of attributes like pixel location, plant ID (i) and time stamp(j) for future reference. After scanning all the pixels in a HSC, its corresponding 2-D array is appended to S. The final S contained 1,632,478 spectral curves corresponding to all plants for all treatments across all days, each with its own tag.

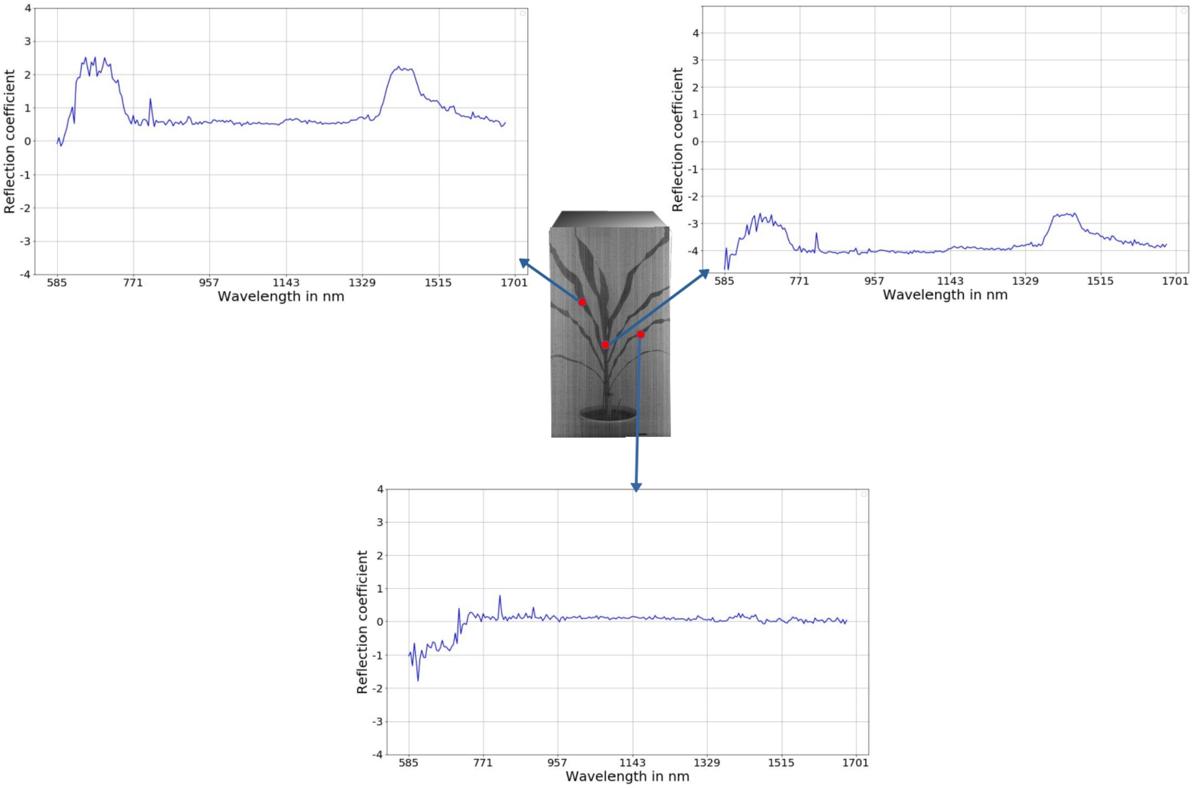


Figure 4.3: Spectral curves at different locations on a plant

4.3 Denoising Spectral Information

To evenly reduce the effects of illumination and shadowing on HSI, like the one shown in Figure 4.4, we perform SNV transformation on S. The mean (μ) and standard deviation (σ) are computed for each of the spectral bands. These μ and σ values must be used as parameters while preprocessing in Algorithm 4 for all future transformations of HSI data.

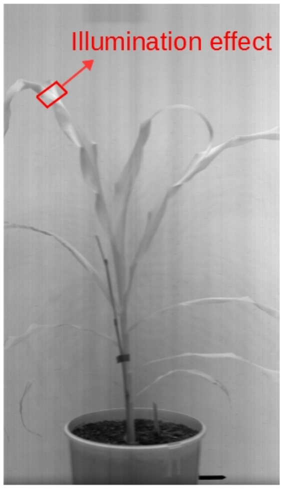


Figure 4.4: Illustration of uneven illumination effect

4.4 Grouping Spectral Classes

4.4.1 Determination of Optimal Number of Clusters

To determine the optimal κ, we employ the elbow method for dataset S. As shown in Figure 4.5, there is a consistent decrease in $\overline{SSE_k}$ values from $\kappa = 2$ to $\kappa = 5$ after which the curve starts to flatten out. However, there is a probable elbow point even at $\kappa = 7$. Considering the higher bend angle at 5, we go with $\kappa = 5$.

4.4.2 Clustering

We perform clustering individually on normal and stressed plants using κ-means algorithm. For this, S is regrouped into SC^n and SC^s corresponding to normal and stress classes respectively using P_s. The number of spectral curves in SC^n corresponded to 866,575 and those in SC^s corresponded to 765,903. To visualize the behavior of the clusters on an abstract level, we plot the mean spectral curves graph for normal and stress classes as shown in Figures 4.6 and 4.7. The proportion of points belonging to

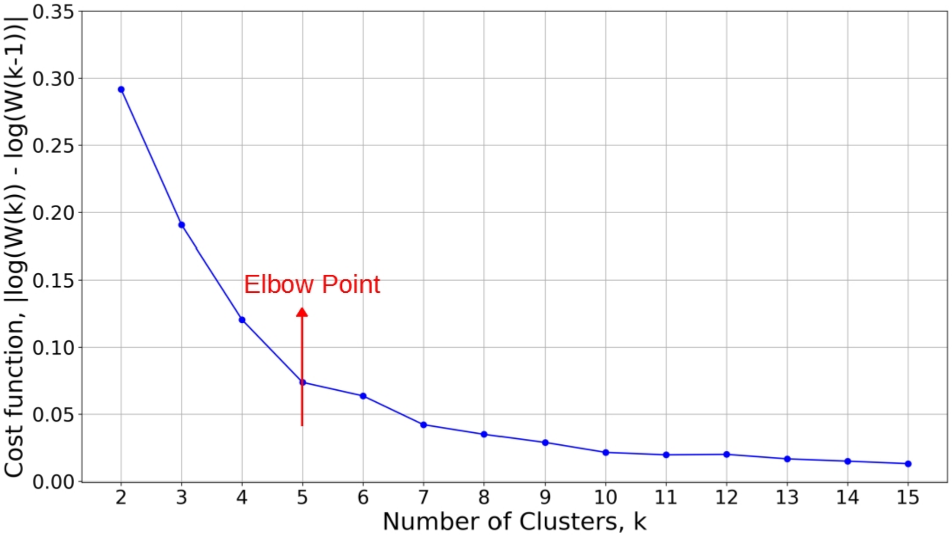


Figure 4.5: Elbow method for optimal cluster number determination

a specific cluster is shown in the legend of the graph next to the cluster number.

4.5 Spectral Feature Selection

To select the K spectral bands that offered best inter-class variability (between normal and stress), we opt for F-value criterion. For this, we group corresponding clusters into cluster pairs with the help of Figures 4.6 and 4.7 in a hierarchical manner and assign an index called cluster pair index (CPI) to each pair as shown in Table 4.1. For a given CPI, we select K best spectral bands. As we follow an incremental approach in the number of spectral features used for analysis, we obtain an optimal K, K_{opt} at 30.

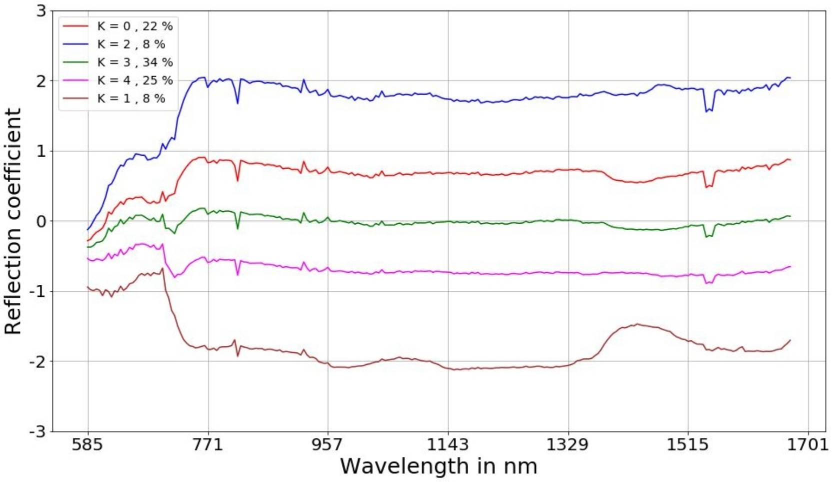


Figure 4.6: Mean spectral curves of each cluster group in the normal class

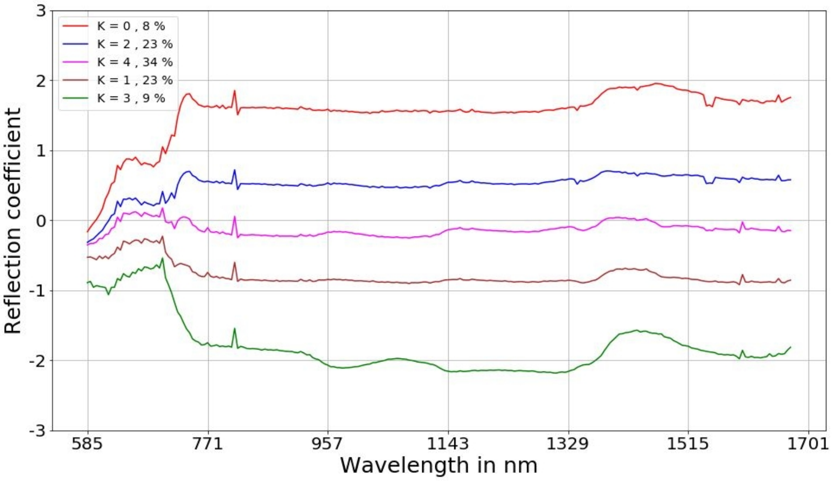


Figure 4.7: Mean spectral curves of each cluster group in the stress class

Table 4.1: Corresponding normal-stress cluster pairs

C_i^n	C_i^n (%)	C_i^s	C_i^s (%)	CPI
0	22	2	23	1
1	8	3	9	2
2	8	0	8	3
3	34	4	34	4
4	25	1	23	5

4.6 Ground Truth Generation

4.6.1 Stress Clusters Identification

To identify clusters that show stress behavior, we measure the spectral variability between C_j^n and C_j^s for each CPI. Then, we compute the mean spectral curves of C_j^n and C_j^s, as shown in Figure 4.8 (P = 50), and measure the spectral angle (SAM) between them, as shown in Table 4.2. If this distance is greater than t_{dist}, we label the CPI as a SCP. For this experiment, the t_{dist} was set at 2 based on which cluster pair with CPI equal to 3 was labeled as SCP, as shown in Table 4.2.

4.6.2 Similarity Computation and Histogram Generation

To quantify the stress factor for each point, we measure the spectral angle between that point and a normal curve, as shown in Fig.4.8. The spectral angles are computed for all spectral curves in normal and stress cluster of a given SCP individually. The spectral angles of normal cluster are stored in d_n and stress cluster in d_s. The normalized histograms h_n and h_s are generated individually for d_n and d_s as shown in Figs. 4.9 and 4.10.

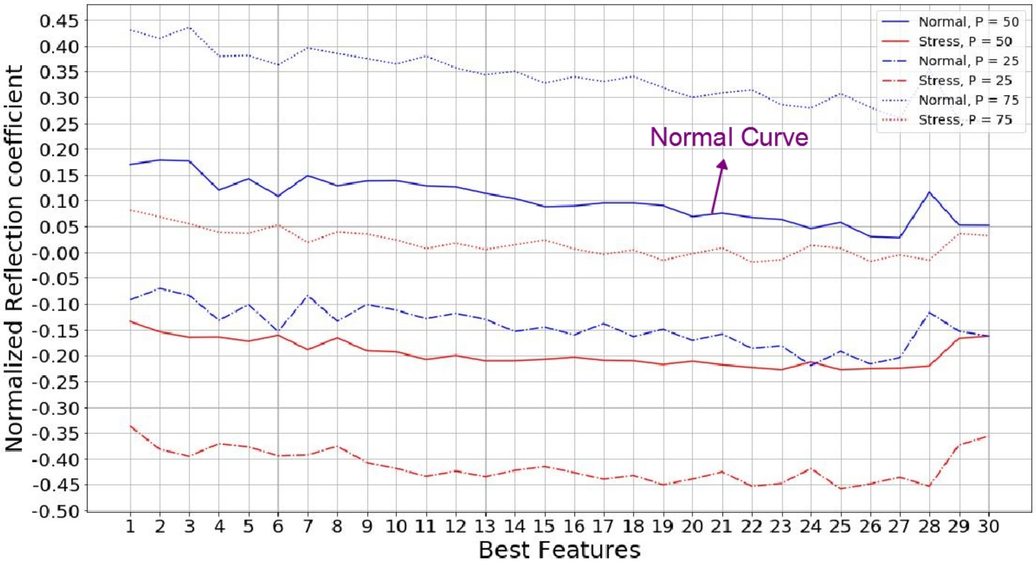


Figure 4.8: Spectral curves of a SCP for different percentile of normal and stress clusters

Table 4.2: Distance between the cluster pairs

CPI	Distance (SAM)	Label
1	0.017	non-SCP
2	0.041	non-SCP
3	2.666	SCP
4	0.058	non-SCP
5	0.091	non-SCP

4.6.3 Stress Identification and Spectral Curves Labeling

The histogram h_n has larger proportion of pixels closer to the normal curve (SAM range between 0.15 and 2) and smaller proportion far away from normal (SAM range between 2 and 3). However, histogram h_s has smaller proportion of pixels closer to normal and larger proportion far away. To compute a pin pointed value of threshold t_{SAM}, we plot the kernel density estimation curves for d_n and d_s. The point were these two curves diverge, t_{SAM}, is plotted at 2.1 as shown in Fig. 4.11. Based on

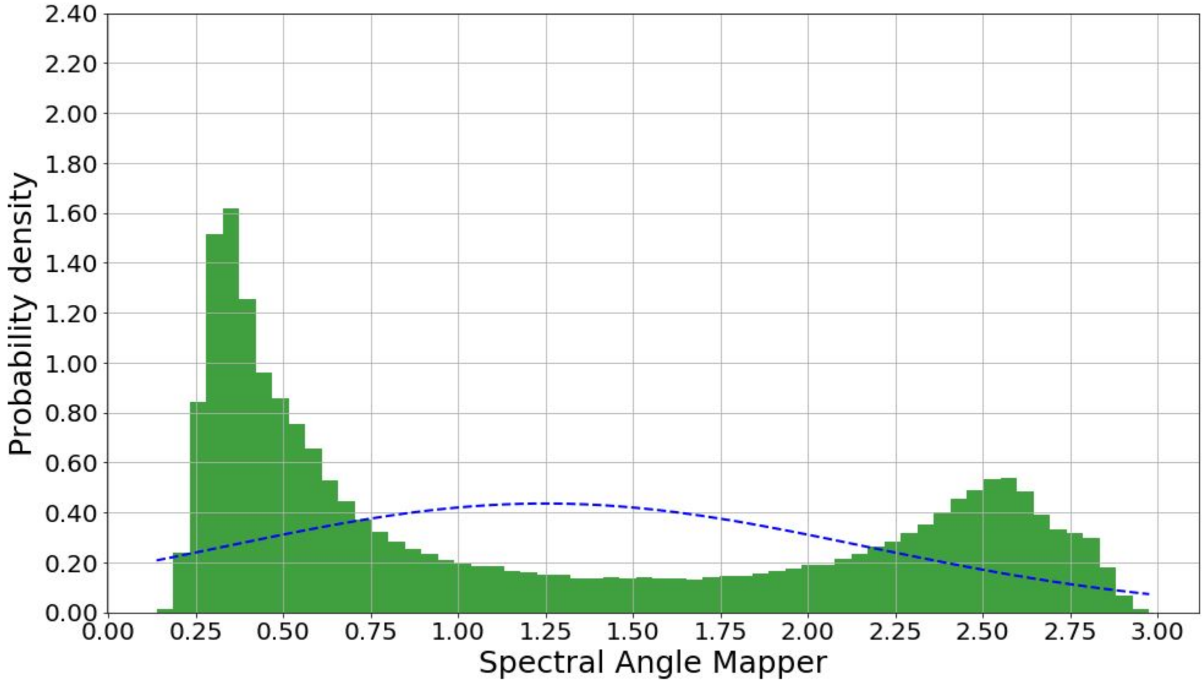


Figure 4.9: Normalized probability density with histogram for the normal class

t_{SAM}, the number of spectral curves labeled normal were 1,485,095 and those labeled stressed were 147,383.

4.7 Stress Classification Modeling

For point-wise stress classification, ANN is employed. As shown in Figure 4.12, it is designed using 2 hidden layers, each housing 15 perceptrons and trained on labeled data from Section 4.6.3. The input data is divided into 60% training and 40 % testing. For validation at each epoch, 20% of the training data is used. The classifier has been trained for 30 epochs based on early stopping criteria [33]. To validate the network, a plot of validation errors as shown in Fig. is used. To test the classifier, we compute a confusion matrix as shown in Table 4.7 where the rows rows represent actual class and columns represent predicted class. Further, we compute accuracy, sensitivity and

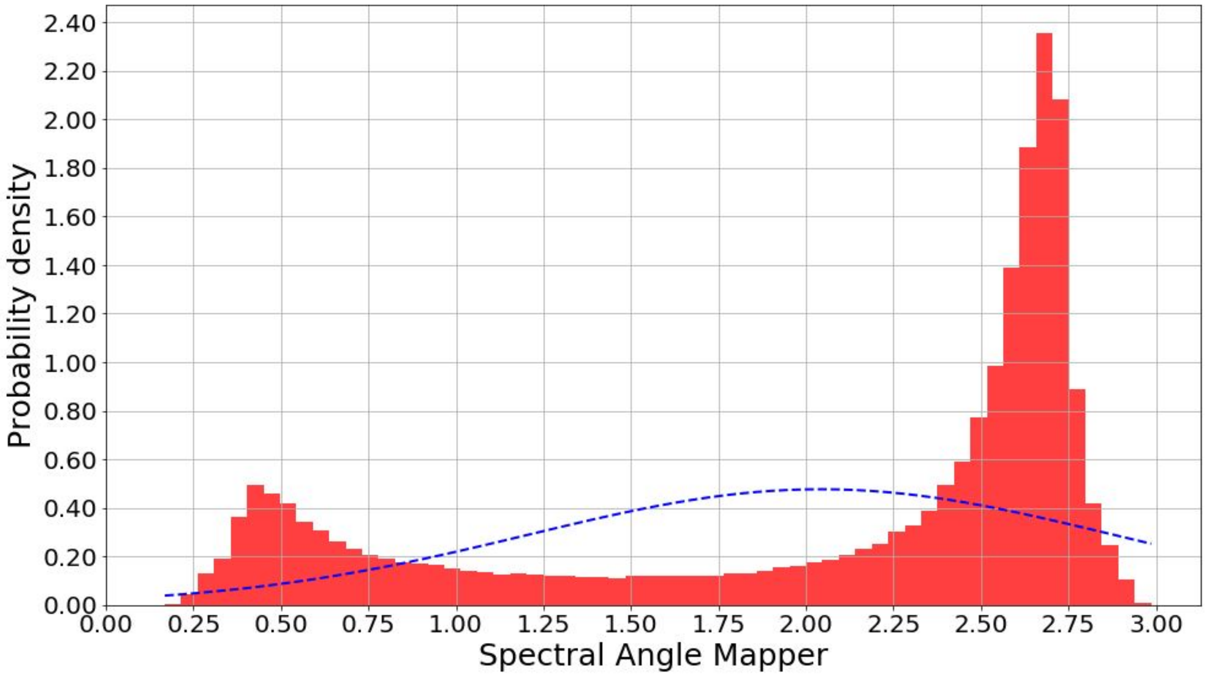


Figure 4.10: Normalized probability density with histogram for the stress class

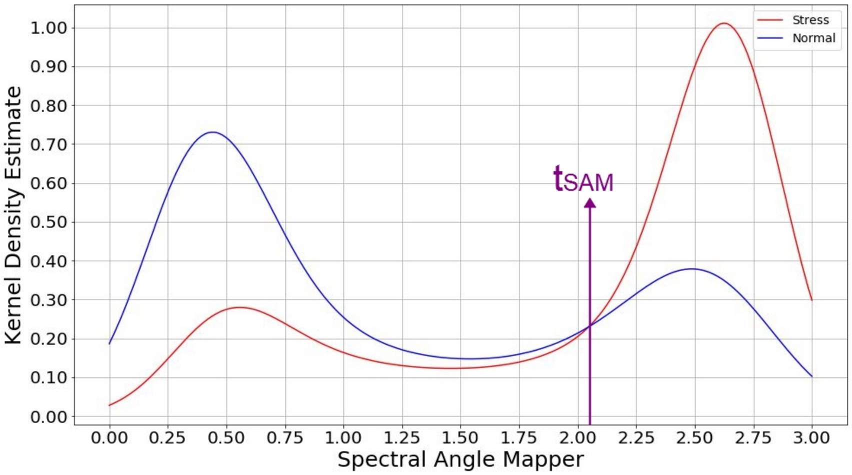


Figure 4.11: Kernel density estimate for spectral angles of SCP

Table 4.3: Confusion matrix for classification results on the test data set

Label	Normal	Stress
Normal	582457	11564
Stress	9519	49452

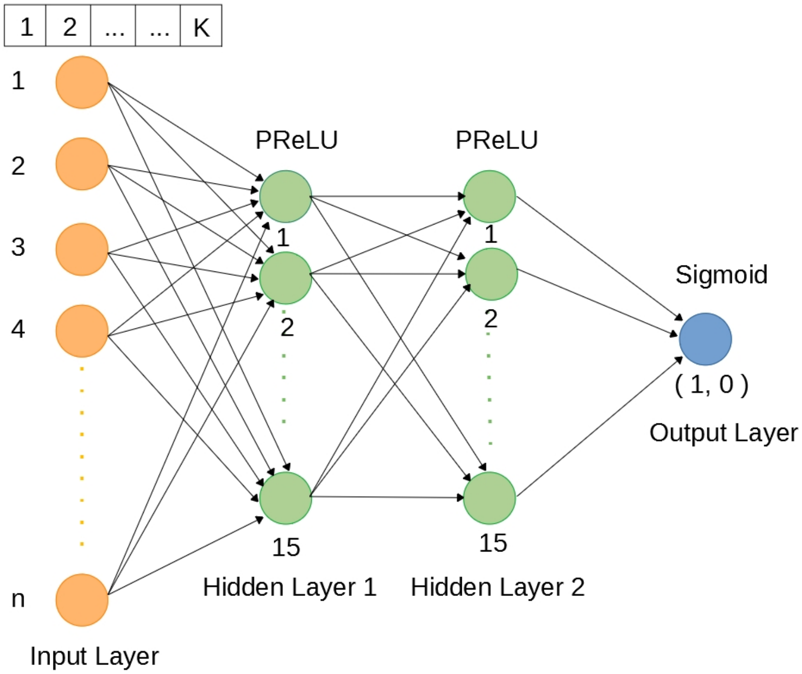


Figure 4.12: Architecture of artificial neural network for the HISC

specificity metrics to evaluate the performance of the HISC. For this experiment, we have set A_{opt} at an accuracy of 95%. For these parameters, optimal K, in reference to Section 4.5, is found at $K = 30$. The accuracy for HISC was found to be 96.7 % with a sensitivity of 98 % and specificity of 81 %.

4.8 Stress Maps

To compute and visualize the stress maps, we consider the HSI of an age-matched normal and a stressed plant from the validation set which are not used in the process of stress classification modeling. They undergo preprocessing, spectral curve generation and denoising using the same parameters used for training data in reference to 4. The label obtained for each point is mapped to its corresponding location on the plant image as shown in Figures 4.13 (a to h) and 4.14 (a to h) for each of the eight days. The points labeled stressed are marked red and unstressed are marked green.

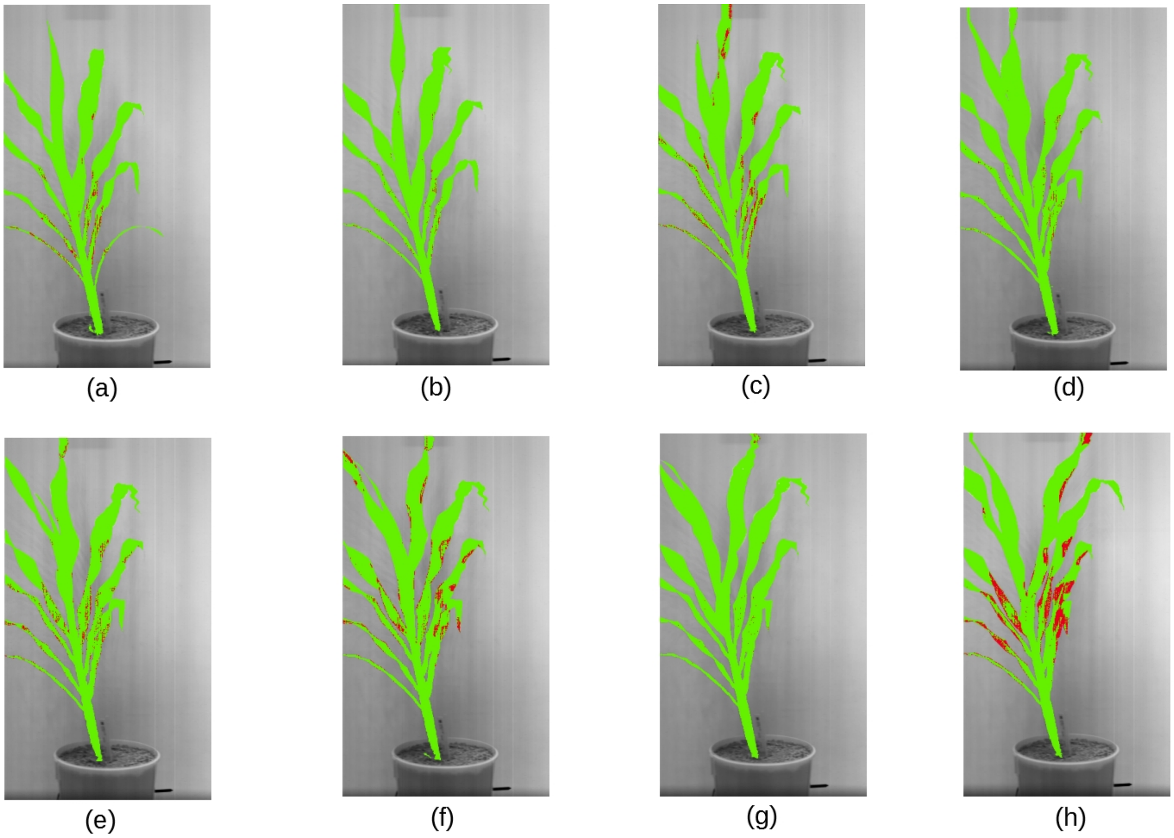
As we observe in Figure 4.14 (a to h), the spatial extent of stress is clearly visible in the stressed plant from the initial days as compared to the normal plant (Figure 4.13 (a to h)) where the normalcy is consistent. However, even for the normal plant, there are some pixels which are marked stressed, especially in the last day (Figure 4.13(h)). This can be mitigated by increasing the specificity of ANN from the current 81% (Section 4.7 through building a better classification model.(Table 4.7).

4.9 Phenotype Computation

The S and U values required for computing PSI is derived from the stress map. The PSI values for the stressed plant increases overtime as shown in Figure 4.15 when compared to a normal plant. The plot of means of the PSI values (Figure 4.15) shows that there is a clear difference between normal and stressed plants.

4.10 Evaluation

The PSI values computed for normal and stressed plants (Section 4.9), shows that HISC is able to clearly differentiate a stressed plant from early stages of stress induce-



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 4.13: Stress maps of normal plant, Day 1 to Day 8

ment. It is able to quantify the stress level for a whole plant in terms of the spatial extent of stress. Also, PSI may be considered as a viable phenotype which is able to detect reasonably early differences in stressed plants. Looking at Figure 4.15, there are fairly clear differences between two classes from day 3 though the early days of stress (day 1 and day 2) appear to look like aberrations.

Having in-situ measurements of stress and ground truth would have lead to the complete picture of stress manifestation. Nevertheless, this data driven approach too shows promising results as it can be seen that normal plants in general have consistently low PSI compared to stress plants.

From Figure 4.15, we observe that the PSI does not follow a uniform trend for

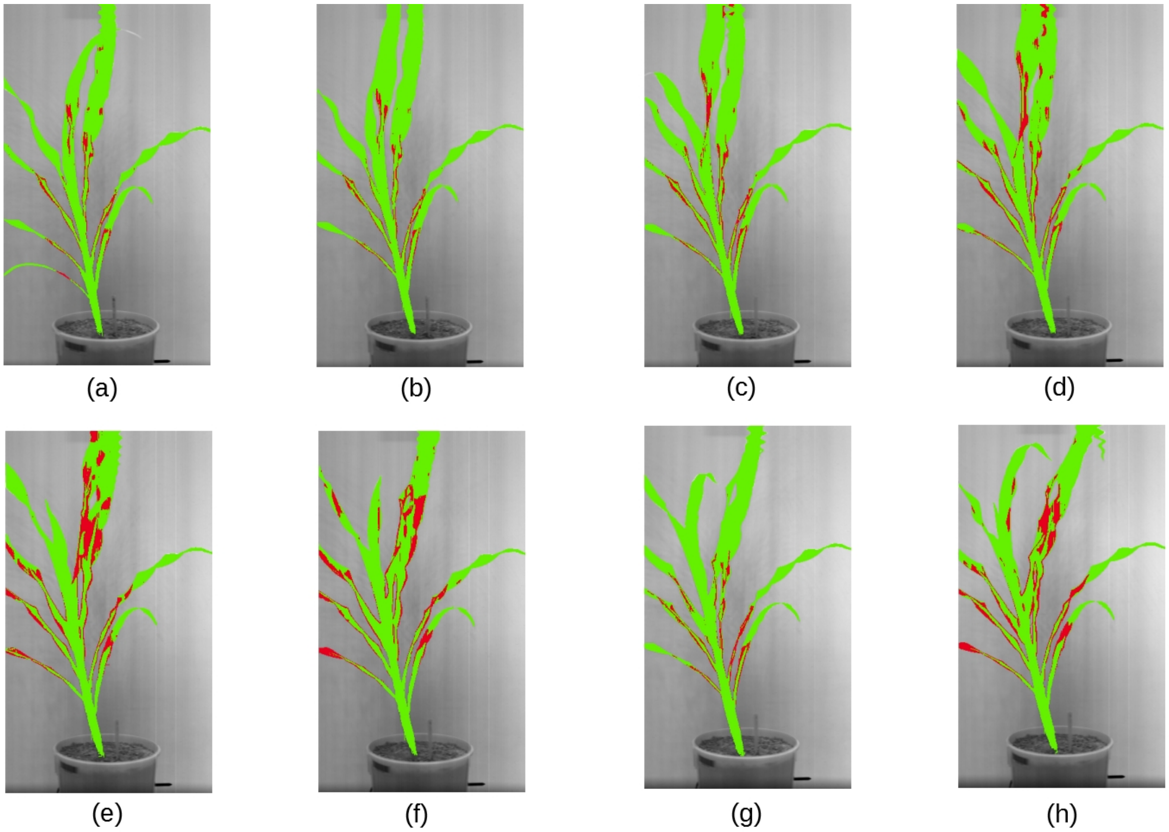


Figure 4.14: Stress maps of stressed plant, Day 1 to Day 8

either for normal or stressed plants. This may be due to the fact that the HSI was imaged only from the side view and it is difficult to capture all the plant pixels from one view. Since stress is manifesting only in the leaves part and leaves twist, turn and bend from day to day, the pixels that appear on a particular day may not reappear in the consecutive days. So the pixels manifesting stress behavior for a given day may not be imaged for other days thereby not giving a complete view of the spatial extent of stress.

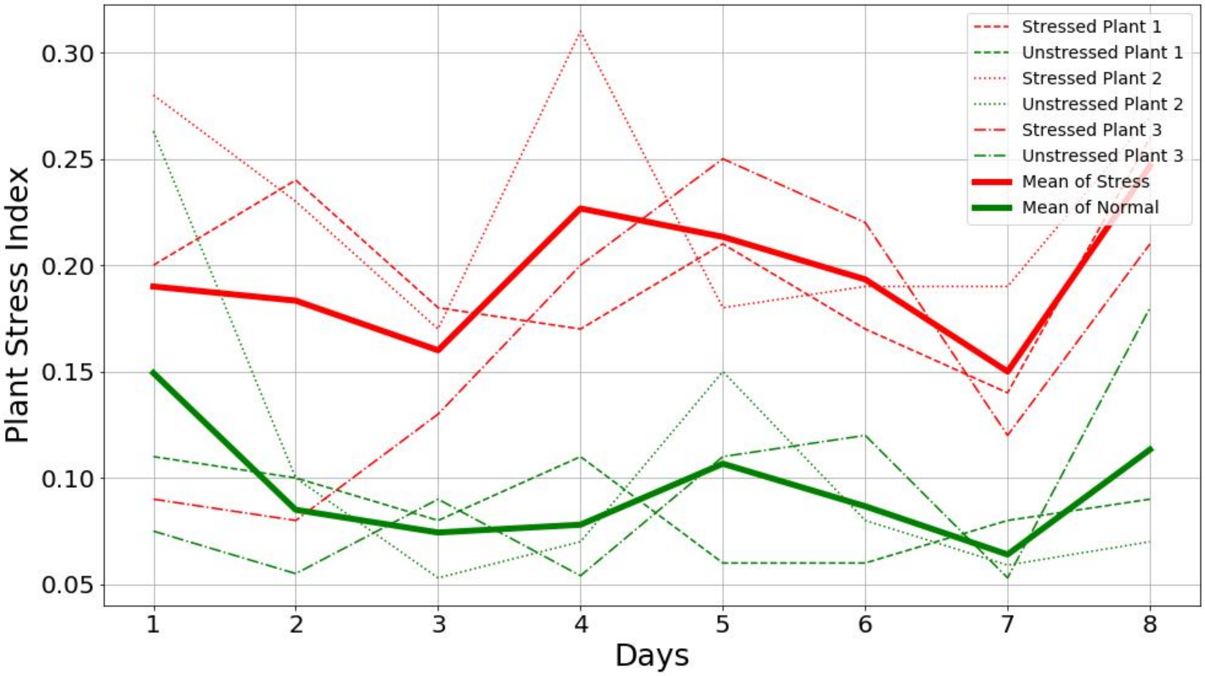


Figure 4.15: PSI values from Day 1 to Day 8

Chapter 5

Summary and Future Work

Physiological phenotyping is important to understand the impact of stress on a plant's growth. In this thesis, we focused on understanding stress using HSI. We have developed a data-driven approach to characterize normal behavior of a spectral curve as a cluster process and used it as a reference to generate ground truth. This ground truth was used to build a classifier that can classify a given spectral curve as stressed or unstressed.

Using the dataset derived from a HTPP system, results show that we can (a) Identify spectral curves in stressed plants that manifest stress behavior. (b) Classify a given spectral curve as stressed or unstressed. (c) Quantify stress level in a plant by defining a stress phenotype.

This work has not considered some interesting dimensions. One of them is that the ground truth is unknown. Perhaps some physical measurements that can locate stress parts more accurately can be used to build the ground truth for stress. Imaging the HSI of plants from multiple views and angles may help in computing better holistic stress phenotypes. Also, we considered this a two class classification problem although it may have had multiple classes (levels) of stress, which gives scope to multiclass stress classification. We could extend this work to other plant species and analyze the results. It would also be interesting to apply this methodology on plants which are in stress recovery stage.

Plotting the spectral curves at pixel level for high resolution HSI will help in tracking the behavior of the plant at a finer spatial level. It may also help visualize the transitions that a plant undergoes when subjected to stress. Our classification dataset was imbalanced with the number of spectral curves marked unstressed outnumbering those that are marked stress by ten times. Using larger datasets can help mitigate this problem. Modeling this problem as an outlier detection problem rather than a classification model may be an interesting approach.

Bibliography

[1] Claudio, H.C.; Cheng, Y.; Fuentes, D.A.; Gamon, J.A.; Luo, H.; Oechel, W.; Qiu, H.-L.; Rahman, A.F.; Sims, D.A. Monitoring drought effects on vegetation water content and fluxes in chaparral with the 970 nm water band index. Remote Sens. Environ. 2006, 103, 304–311. [Google Scholar]

[2] Mistele, B.; Schmidhalter, U. Spectral measurements of the total aerial n and biomass dry weight in maize using a quadrilateral-view optic. Field Crops Res. 2008, 106, 94–103. [Google Scholar]

[3] Schlemmer, M.R.; Francis, D.D.; Shanahan, J.; Schepers, J.S. Remotely measuring chlorophyll content in corn leaves with differing nitrogen levels and relative water content. Agron. J. 2005, 97, 106–112. [Google Scholar]

[4] M. S. M. Asaari et al., ”Close-range hyperspectral image analysis for the early detection of stress responses in individual plants in a high-throughput phenotyping platform”, ISPRS J. Photogramm. Remote Sens., vol. 138, pp. 121-138, Apr. 2018.

[5] Qin, Jianwei & F. Burks, Thomas & Kim, moon seok & Chao, Kuanglin& Ritenour, Mark. (2008). Citrus canker detection using hyperspectral reflectance imaging and PCA-based image classification method. Sensing and Instrumentation for Food Quality and Safety. 2. 168-177. 10.1007/s11694-008-9043-3.

[6] Khouj Y, Dawson J, Coad J, Vona-Davis L. 2018. Hyperspectral Imaging and K-Means Classification for Histologic Evaluation of Ductal Carcinoma In Situ.DOI: 10.3389/fonc.2018.00017

[7] Wikipedia contributors. Feature selection [Internet]. Wikipedia, The Free Encyclopedia; 2019 Mar 5, 06:28 UTC [cited 2019 Mar 11]. Available from: https://en.wikipedia.org/w/index.php?title=Feature_selection&oldid=886273579.

[8] V. Dworak, J. Selbeck, K.-H. Dammer, M. Hoffmann, A. A. Zarezadeh, and C. Bobda, "Strategy for the development of a smart NDVI camera system for outdoor plant detection and agricultural embedded systems.," Sensors (Basel, Switzerland), vol. 13, pp. 1523– 38, jan 2013.

[9] Y. Ge, G. Bai, V. Stoerger, and J. C. Schnable, "Temporal dynamics of maize plant growth, water use, and leaf water content using automated high throughput RGB and hyperspectral imaging," Computers and Electronics in Agriculture, vol. 127, pp. 625–632, 2016.

[10] T. Rumpf, et al. Early detection and classification of plant diseases with support vector machines based on hyperspectral reflectance. Comput. Electron. Agric., 74 (2010), pp. 91-99.

[11] P. Baranowski, et al. Hyperspectral and thermal imaging of oilseed rape (Brassica napus) response to fungal species of the genus Alternaria PLoS ONE, 10 (2015), p. e0122913.

[12] J. Behmann, et al. Ordinal classification for efficient plant stress prediction in hyperspectral data. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci., XL-7 (2014), pp. 29-36

[13] Wikipedia contributors. (2018, October 27). Data binning. In Wikipedia, The Free Encyclopedia. Retrieved 08:01, March 25, 2019, from https://en.wikipedia.org/w/index.php?title=Data_binning&oldid=866001459

[14] Arti Patle and Deepak Singh Chouhan. (2013). SVM kernel functions for classification, International Conference on Advances in Technology and Engineering. DOI: 10.1109/ICAdTE.2013.6524743.

[15] K. N. Leach, ”A survey paper on independent component analysis,” Proceedings of the Thirty-Fourth Southeastern Symposium on System Theory (Cat. No.02EX540), Huntsville, AL, USA, 2002, pp. 239-242. doi: 10.1109/SSST.2002.1027042

[16] Zhouhan Lin, Yushi Chen, Xing Zhao and Gang Wang, ”Spectral-spatial classification of hyperspectral image using autoencoders,” 2013 9th International Conference on Information, Communications & Signal Processing, Tainan, 2013, pp. 1-5. doi: 10.1109/ICICS.2013.6782778

[17] Grunauer, A., Vincze, M., 2015. Using dimension reduction to improve the classification of high-dimensional data. In: 39th Annual Workshop of the Austrian Association for Pattern Recognition (OAGM 2015). Available from: ¡arXiv:1505.01065v1¿.

[18] P. H. Swain and H. Hauska, ”The decision tree classifier: Design and potential,” in IEEE Transactions on Geoscience Electronics, vol. 15, no. 3, pp. 142-147, July 1977.

[19] Y. Chen, H. Jiang, C. Li, X. Jia and P. Ghamisi, ”Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Net-

works," in IEEE Transactions on Geoscience and Remote Sensing, vol. 54, no. 10, pp. 6232-6251, Oct. 2016. doi: 10.1109/TGRS.2016.2584107

[20] Zhouhan Lin, Yushi Chen, Xing Zhao and Gang Wang, "Spectral-spatial classification of hyperspectral image using autoencoders," 2013 9th International Conference on Information, Communications & Signal Processing, Tainan, 2013, pp. 1-5. doi: 10.1109/ICICS.2013.6782778

[21] Corinna Cortes, Vladimir Vapnik. (1995). Support-Vector Networks.

[22] F. Chu, G. Jin, and L. Wang. Cancer Diagnosis and Protein Secondary StructurePrediction Using Support Vector Machines, StudFuzz 177, 343–363 (2005)

[23] T A Moughal, Hyperspectral image classification using Support Vector Machine.(2013). 6th Vacuum and Surface Sciences Conference of Asia and Australia(VASSCAA-6). DOI:10.1088/1742-6596/439/1/012042.

[24] F. Melgani, L. Bruzzone. (2004). Classification of hyperspectral remote sensing images with support vector machines. IEEE Transactions on Geoscience and Remote Sensing (Volume: 42, Issue: 8, Aug. 2004). DOI: 10.1109/TGRS.2004.831865.

[25] Yushi Chen, Zhouhan Lin, Xing Zhao, Gang Wang, Yanfeng Gu. (2014). DeepLearning-Based Classification of Hyperspectral Data, IEEE Journal of SelectedTopics in Applied Earth Observations and Remote Sensing (Volume: 7, Issue:6). DOI: 10.1109/JSTARS.2014.2329330.

[26] Rojas-Moraleda, R., Valous, N.A., Gowen, A. et al. Neural Comput & Applic (2017) 28(Suppl 1): 969. https://doi.org/10.1007/s00521-016-2376-7.

[27] R. Rojas. (1996). Neural Networks - A Systematic Introduction, pp 151 - 184, Springer-Verlag, Berlin.

[28] Cenk Bircano glu, Nafi Arıca. (2018). A comparison of activation functions in artificial neural networks, 2018 26th Signal Processing and Communications Applications Conference(SIU), Izmir, Turkey. DOI:10.1109/SIU.2018.8404724.

[29] H. Yao, Z. Hruska, R. Kincaid, A. Ononye, R. L. Brown and T. E. Cleveland, ”Spectral Angle Mapper classification of fluorescence hyperspectral image for aflatoxin contaminated corn,” 2010 2nd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, Reykjavik, 2010, pp. 1-4. doi: 10.1109/WHISPERS.2010.5594920

[30] Dibya Jyoti Bora1, Dr. Anil Kumar Gupta. (2014). A Comparative study Between Fuzzy Clustering Algorithm and Hard Clustering Algorithm, 2014 International Journal of Computer Trends and Technology (IJCTT) – volume 10 number 2 – Apr 2014.

[31] K. He, X. Zhang, S. Ren, J. Sun, ”Delving deep into rectifiers: Surpassing human-level performance on imagenet classification”, 2015.

[32] S. Mei, J. Ji, Q. Bi, J. Hou, Q. Du and W. Li, ”Integrating spectral and spatial information into deep convolutional Neural Networks for hyperspectral classification,” 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, 2016, pp. 5067-5070. doi: 10.1109/IGARSS.2016.7730321

[33] Y. Shao, G. N. Taff and S. J. Walsh, ”Comparison of Early Stopping Criteria for Neural-Network-Based Subpixel Classification,” in IEEE Geoscience and Remote Sensing Letters, vol. 8, no. 1, pp. 113-117, Jan. 2011. doi: 10.1109/LGRS.2010.2052782

[34] H. Scharr, T. Pridmore and S. A. Tsaftaris, ”Computer Vision Problems in Plant Phenotyping, CVPPP 2017: Introduction to the CVPPP 2017 Workshop Papers,” 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, 2017, pp. 2020-2021. doi: 10.1109/ICCVW.2017.236

[35] M. K. Pakhira, ”A Linear Time-Complexity k-Means Algorithm Using Cluster Shifting,” 2014 International Conference on Computational Intelligence and Communication Networks, Bhopal, 2014, pp. 1047-1051. doi: 10.1109/CICN.2014.220

[36] A. Panda and D. Pradhan, ”Hyperspectral image processing for target detection using Spectral Angle Mapping,” 2015 International Conference on Industrial Instrumentation and Control (ICIC), Pune, 2015, pp. 1098-1103. doi: 10.1109/IIC.2015.7150911

[37] J. W. Herring and David, “Measuring Vegetation (NDVI EVI): Feature Articles,” 2000.

[38] JooSeuk Kim and C. Scott, ”Robust kernel density estimation,” 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, 2008, pp. 3381-3384. doi: 10.1109/ICASSP.2008.4518376

[39] Wikipedia contributors. ”Cross-validation (statistics).” Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 4 Apr. 2019. Web. 6 Apr. 2019.

[40] Dibya Jyoti Bora, Dr. Anil Kumar Gupta, ”A Comparative study Between Fuzzy Clustering Algorithm and Hard Clustering Algorithm,” International Journal of Computer Trends and Technology (IJCTT) – volume 10 number 2 – Apr 2014, Page 108.