ORNL/TM-2024/3279

Hyperkube: A Kubernetes Based System For The Automation Of Processing And Analysis Of Hyperspectral Data Obtained From Multiple Hyperspectral Imaging Systems

Hong-Jun Yoon
Daniel Hopp
Kellen Leland
Ryan Prout
Paul Bryant
Stanton Martin

February 2024

OAK RIDGE National Laboratory

ORNL IS MANAGED BY UT-BATTELLE LLC FOR THE US DEPARTMENT OF ENERGY

ORNL/TM-2024/3279

Center for Bioenergy Innovation
&
Laboratory Directed Research and Development

HYPERKUBE: A KUBERNETES BASED SYSTEM FOR THE AUTOMATION OF PROCESSING AND ANALYSIS OF HYPERSPECTRAL DATA OBTAINED FROM MULTIPLE HYPERSPECTRAL IMAGING SYSTEMS

Authors
Hong-Jun Yoon, Daniel Hopp, Kellen Leland, Ryan Prout, Paul Bryant, Stanton Martin

February 2024

Prepared by
OAK RIDGE NATIONAL LABORATORY
Oak Ridge, TN 37831
managed by
UT-BATTELLE LLC
for the
US DEPARTMENT OF ENERGY
under contract DE-AC05-00OR22725

Contents

ABSTRACT

Hyperspectral imagery is an emerging field of technology that has enormous potential for remote and proximal sensing in numerous areas of research. The plant phenotyping community is applying this technology to advance the throughput and accuracy of plant phenotypes based on airborne and lab-based hyperspectral imaging technology. Here we report an automated processing and analysis pipeline for four different hyperspectral imaging platforms, discuss the data issues involved, and present a strategy for computing and data architecture to handle hyperspectral data.

Introduction

Hyperspectral imaging represents a relatively new modality by which scientists can observe various optical phenomena of interest. Instrument vendors such as Specim, Hyspex, Headwall, Surface Optics, and others offer a broad array of sensors which can be tuned to specific applications in forestry and agriculture. In 2018, Specim technologies debuted what it claimed was the world's first mobile hyperspectral camera, the Specim IQ. The Specim IQ has the capability to collect imagery in the 400 to 1000 nm range.1 Oak Ridge National Laboratory acquired two Specim IQ instruments in 2020, a Headwall instrument in 2021, and two instruments VNIR and SWIR 1700 series from Photon Systems Instruments (PSI) in 2022. These instruments were deployed in 2023 and we have generated an automated pipeline to manage the data streams, and integrate the analyses with other instrumentation at the Advanced Plant Phenotyping Laboratory (APPL) located at Oak Ridge National Laboratory. The APPL facility is an example of a dedicated plant phenotyping facility that utilizes multi modal imaging to determine structure, performance, and tolerance to limitations of an individual plant or group of plants in a greenhouse environment.2 The capabilities of APPL are combined with field and biochemical studies to provide a holistic picture of plant physiological traits under various conditions of interest. The primary contribution of this study is the design and implementation of a robust and efficient scientific software pipeline. This pipeline effectively captures hyperspectral data from APPL and aims to seamlessly manage the data stream to ensure smooth data flow. Furthermore, the pipeline incorporates generating data products and conducting data analysis using traditional as well as the latest deep learning-based approaches, thus validating the accuracy and integrity of the captured data.

BACKGROUND

Multispectral imaging operates on the foundational principle that each substance possesses a distinctive spectral signature, which serves as a unique identifier and provides valuable insights into its constituent elements and surface properties.3 By analyzing the spectrum of a single pixel within a multispectral image, precise information about the material can be obtained. Over the past few decades, significant advancements in multispectral image sensing technologies have paved the way for capturing images that span an extensive spectral range. These cutting-edge techniques allow for the simultaneous acquisition of several hundred spectral bands, encompassing the entirety of the observational scene in a single scan. The utilization of hyperspectral imaging techniques further enhances the spectral resolution, enabling a comprehensive examination of land surfaces and the discrimination of different materials present within the observed scene.4

Hyperspectral images exhibit a unique combination of spatial and spectral resolutions, both of which are crucial in extracting detailed information.5 Spatial resolution quantifies the geometric arrangement and relationship of individual image pixels, whereas spectral resolution determines the level of variation within each pixel as a function of wavelength. The hyperspectral data is often represented in the form of a three-dimensional hyperspectral data cube, where each axis represents spatial coordinates and spectral information. The spectral resolution is typically characterized by the number of spectral bands captured by the sensor and the breadth of the spectrum measured. Hyperspectral sensors offer the ability to measure and capture a wide range of electromagnetic energy across the designated wavelength range, enabling precise observation and analysis of distinct surface features and changes exhibited by various materials.

In the field of hyperspectral imaging, reflectance plays a crucial role as a metric for understanding the complex interaction between incident light and the surface of a material.6 Reflectance is precisely quantified as the ratio of reflected energy to incident energy, evaluated as a function of wavelength. Researchers can gain critical insights into the optical properties and compositional intricacies of a material by examining its reflectance behavior. Reflectance values within a designated electromagnetic spectrum range can be leveraged to infer valuable information about a material's spectral characteristics and optical behavior. When all light energy directed towards an object at a specific wavelength is reflected back to the imaging sensor, the reflectance value reaches its peak at 100%. Conversely, when the material absorbs all incident light at a specific wavelength, the reflectance value descends to its nadir at 0%. Comparative analysis can be performed by plotting the reflectance values of distinctive materials present on the surface of an object to generate spectral signatures or spectral response curves.4 By meticulously examining and comparing these spectral signatures, researchers can gain insights into the composition and optical behavior of the respective materials. The spectral resolution of the image sensor used in hyperspectral imaging is crucial for capturing the intricate details within the spectral signatures, facilitating comprehensive classification and discrimination of materials based on their spectral characteristics.

Deep learning is a novel machine learning approach that has demonstrated impressive results in various image processing applications.7 Recently, this approach has been extended to the detection and classification of spectral and spatio-spectral signatures in hyperspectral images.8,9 The high dimensionality of hyperspectral data, coupled with limited labeled training data, makes deep learning an attractive method for comprehensive hyperspectral data analysis. Deep learning models, which rely on artificial neural networks, are capable of automatically learning complex feature representations from raw data, eliminating the need for manual feature engineering.10 With their layered architectures and interconnected neurons, deep learning models can effectively navigate the intricate interactions within hyperspectral images and accurately identify and classify spectral signatures.11 Furthermore, the scarcity of labeled training data in the hyperspectral domain enhances the appeal of deep learning by leveraging unlabeled data for unsupervised pre-training. By fine-tuning with limited annotated samples, deep learning models achieve exceptional generalization and classification performance. The integration of deep learning techniques unleashes the potential of hyperspectral imaging, enabling enhanced interpretation, detection, and classification of spectral and spatio-spectral signatures with unmatched accuracy and efficiency.

RESULTS

The automated pipeline can produce three primary data products from each instrument:

1. A csv file that contains statistics for the pixels of interest across all hyperspectral bands. These files are located in the signatures directory for each project.
2. A masked file that contains only pixels of interest from the instrument (e.g. Plant files). These files are stored in parquet format. They are stored in the parquet directory for each project
3. A Portable Network Graphics (PNG) representation of the masked imagery. These files are stored in the "masked_png" directory.

A post analysis tool is used to create graphical representations of the parquet files and comma-separated values (CSV) files in the form of a reflectance plot. The output of this tools is represented in the figures below

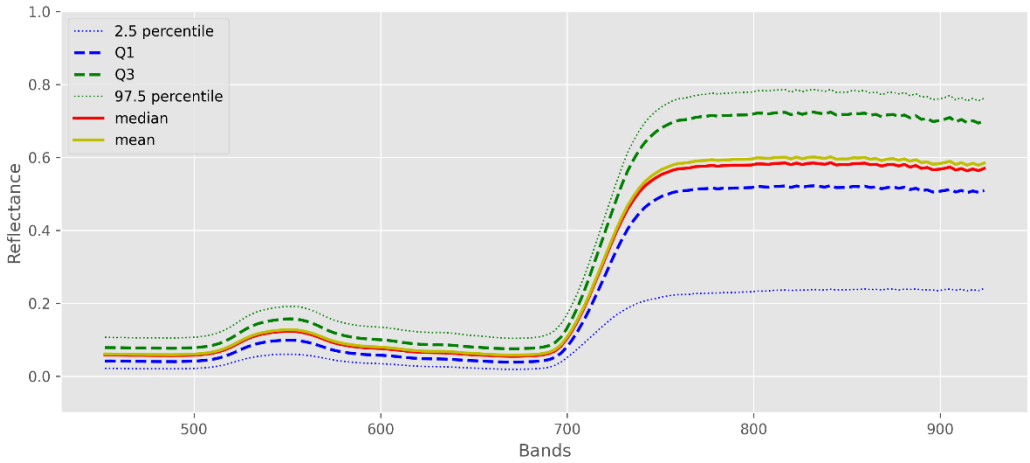

Figure 1: Reflectance plot of masked pixel statistics across all bands for the Specim IQ instrument. This is a one dimensional representation of the hyperspectral data from a single instrument

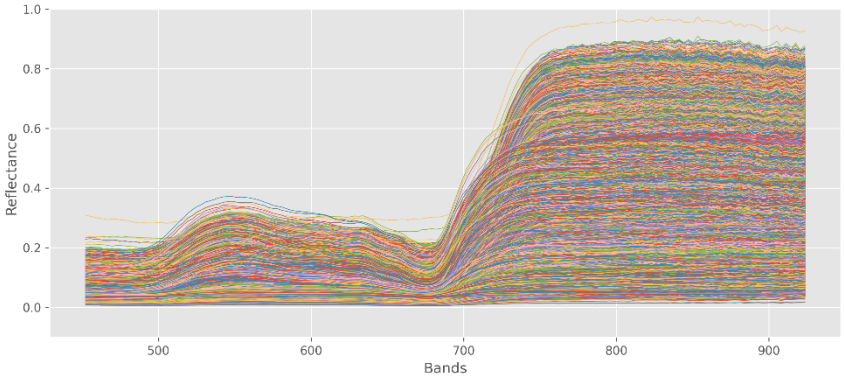

Figure 2: Reflectance plot of all masked pixels from the corresponding parquet file for Figure 1. This is a two dimensional representation of the hyperspectral data.

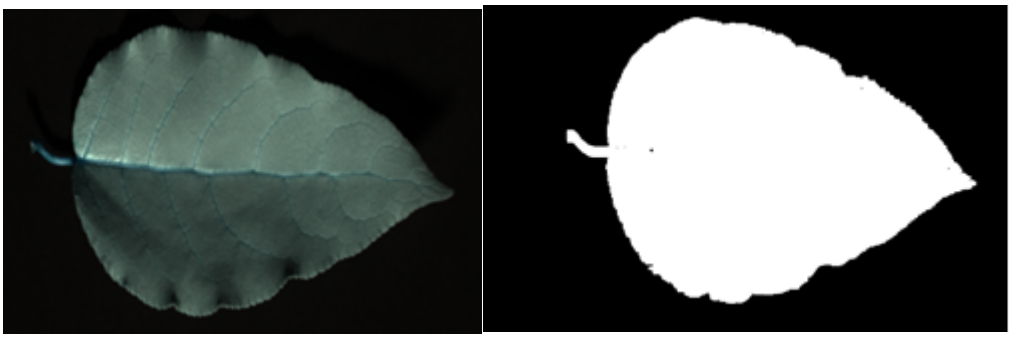

Figure 3: A PNG reconstruction of the masked imagery in RGB color space. This is used for visual quality control.

Data streams from multiple instrumentation can be combined to reveal the full spectra of variation between 400 and 2500 nm. This is done by merging data products from instruments with different native spectral ranges. In this case, the Specim IQ with a native range of 400 to 1000 nm was combined with a Headwall native range of 1000 to 2500 nm. A careful visual inspection of the data revealed that the useable data for this dataset range from ~452 to ~923 nm for the Specim IQ data and ~895 to ~2500 nm for the Headwall instrument. An example combined data plot is shown in Figure 4.

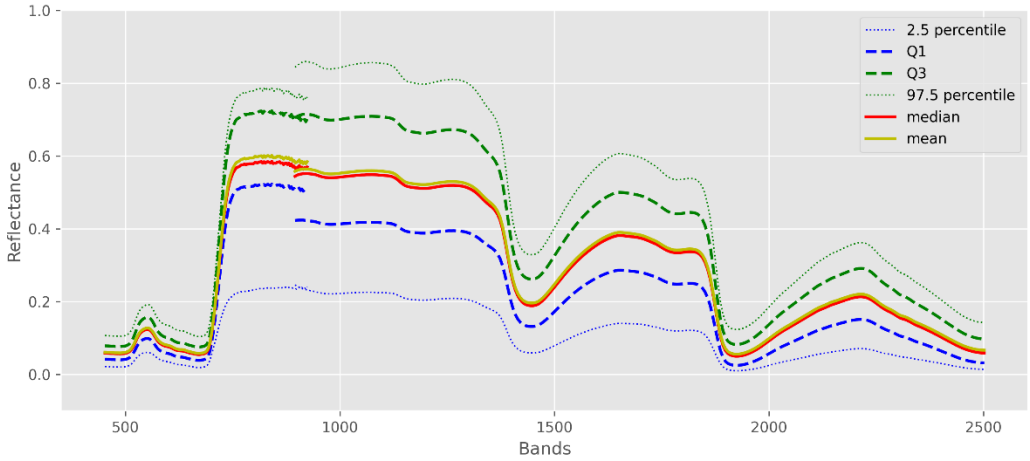

Figure 4: One-dimensional combined plot of Specim IQ and Headwall data. The gap in the two plots represents the ends of the spectra for each instrument where data is trimmed for quality control.

The different data products created above can be used for various analyses, based on the experimental design. We tested the process using an experiment in which cuttings of greenhouse grown *Populus Trichocarpa* were cultivated under different regiments of nitrogen application. A DOI has been assigned to this code via the DOE CODE repository. It is available at:

https://doi.org/10.11578/dc.20231101.1

We devised different analysis pipelines to process the one-dimensional data products. In the simplest scenario, we regarded three treatments as categorical variables. We applied a classification analysis to the entirety of the data and assessed the performance of the algorithm to correctly classify the data based on its treatment.

Given an input signature $\bar{x} = H(x)$ from a hyperspectral image x, where $H(\cdot)$ is a translation into a one-dimensional hyperspectral signature, the classification can be expressed as

$$y_k = f_c(\bar{x}),$$

where $f_c(\cdot)$ is a classification function, $k = \{0,\cdots,K\}$ is the k-th output, and K represents the number of classes in the problem. This particular case was a three-class problem, namely, no treatment, 0.5, and 1; hence, $K = 3$. The implementation is written in Python and utilizes the scikit-learn package.[12] The source code can be accessed at classification.py.

In the second scenario, we consider treatments as continuous variables and perform regression analysis to estimate their numerical values. The analysis can be expressed as

$$y = f_r(\bar{x}),$$

where $f_r(\cdot)$ is a regression function. The implementation with scikit-learn can be found at regression.py.

The aforementioned methods are classical statistical approaches that have been widely used in various scientific applications for many years.[13]

A second approach is to apply modern artificial intelligence (AI) and machine learning (ML) techniques to both classification and regression problems.[14] It is achieved by using a one-dimensional convolutional neural network (CNN), which employs a convolutional layer that contains a set of one-dimensional convolution filters. These filters capture latent representations of the features present in the hyperspectral signatures. The extracted features are then transferred to the fully connected layer, where the final decision is made. Specifically, the output layer is activated using the softmax function for classification, while for regression the softmax function, while for regression, the sigmoid function is utilized as the activation function. We can then compare results of both classification and estimation routines from both the classical and machine learning approaches as indicated in Figure 5.

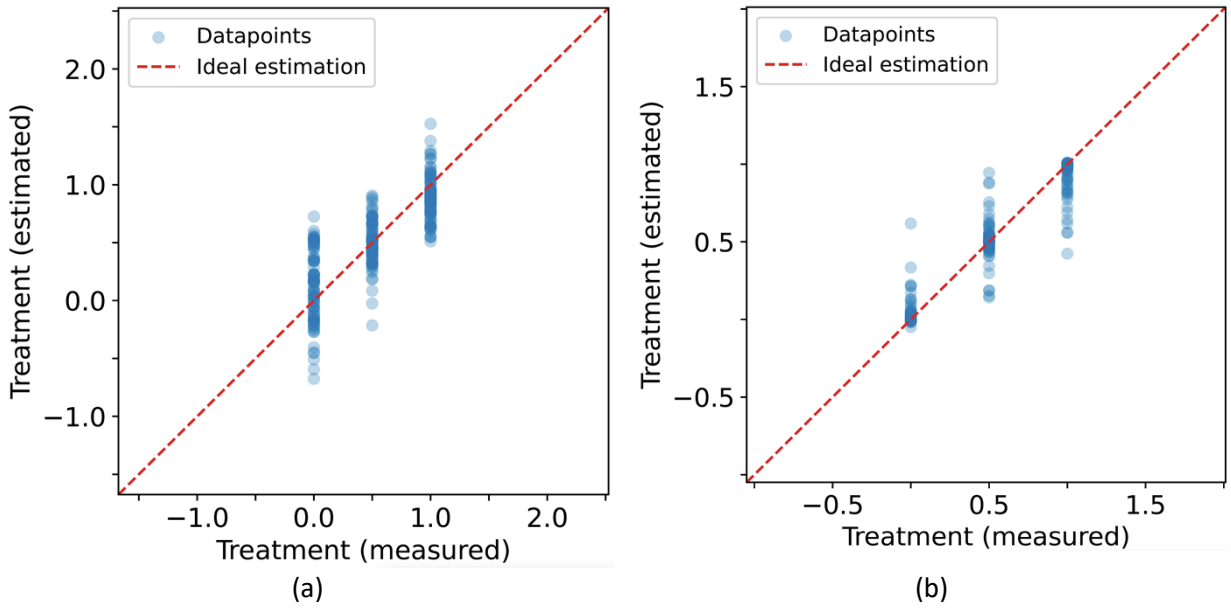

Figure 5: Plots of estimating nitrogen treatment levels using hyperspectral signatures based on (a) partial least squares regression algorithms and (b) one-dimensional convolutional neural network algorithm.

The above graph shows the comparison of traditional statistical methodology based on partial least square regression (PLSR) compared with the CNN. The R^2 value for PLSR is 0.56, while the R^2 value for CNN is 0.88, indicating that the CNN-based regression model significantly outperforms the PLSR model in terms of accuracy. The source code can be accessed at hyperspectral_CNN.py.
When dealing with parquet files, we encounter multidimensional data instead of one-dimensional data. The second dimension corresponds to the spatial location of each pixel within the hypercube. In the field of hyperspectral image processing, the most widely used analytical approach for 2-dimensional data is spectral unmixing. This involves computing the fractional contribution of elementary spectra, also known as endmembers. These endmembers are the vertices of a convex polytope that covers the image data points in high-dimensional spaces. The image model is a linear combination of the endmembers, with positive coefficients that add up to one.

One intuitive method for extracting endmembers in hyperspectral image processing is to analyze the spectra as linear combinations of non-negative components. This approach utilizes a Principal Component Analysis (PCA) dimensional reduction based on the sample spectral correlation matrix of the image. The implementation of this method using Python and scikit-learn can be found in endmember.py.[15]

DISCUSSION

Hyperspectral imaging is an emerging technology that shows great potential for rapid plant phenotyping. However, to realize this potential, a number serious of challenges need to be overcome. This includes capturing and managing the raw data streams, providing effective quality control, generating data products suitable for analysis, and applying appropriate statistical techniques to the data pipeline. In this work, we have developed and implemented a software pipeline to capture data, manage the data stream, perform quality control, and produce data products suitable for further analysis. Future work on the analytical methods will be able to leverage this software solution to create high throughput phenotyping workflows based on hyperspectral imagery. The system allows for numerous different analytical algorithms to be applied to simultaneously to the data products via the Kubernetes pods. This capability allows for a dramatic reduction in analysis time. Additionally, the data products created from this pipeline are more robust than those typically generated. For instance, the traditional hyperspectral signature as defined in database such as Aster, EcoStress, the Vegetation Spectral library, and others consists of a mean only. Our method not only includes the mean, but also other summary statistics so end users can see the full range of values that may be generated by the plant tissue of interest. Future work will be required to partition the variance found in these signatures to more specific tissue subtypes. For this task the end member data product may be interrogated to classify pixels according to leaf tissue subtypes, such as leaf, stem, veins, etc. The classification may also pick up plant tissue defects such as necrotic lesions or insect damage. The features can be counted, characterized, and added as a phenotype for downstream statistics. For example, the diameter and size of necrotic lesions could be informative in determining host resistance to a pathogen.

METHODS

Architecture

We chose to use the Kubernetes platform as the underlying architecture for HyperKube. Kubernetes is a modern, cloud native infrastructure that facilitates modular code development in the form of microservices. Each component of HyperKube is discretized into a Kubernetes pod, defined by a yaml file. The Kubernetes infrastructure orchestrates each pod so microservices are called automatically in response to events.

The Kubernetes cluster itself is comprised of several namespaces used for development and production. Each namespace contains its own hyperspectral imaging and database pods. The hyperspectral imaging pod consists of Python and its required packages including asyncio, cProfile, cv2, email.mime, matplotlib, numpy, os, pandas, PIL, pstats, psycopg2, pysptools, rasterio, shutil, and spectral.

Pods are usually created with a Docker container build instruction, Dockerfile, Docker run command, and virtual operating system environment instruction files. These files are saved in web repository, and it uses an inherent pipeline to launch a deployment into a Kubernetes namespace. The pipeline is

triggered manually or when a change is applied to a repository file. Deployments are controlled by Argo CD, an automated continuous delivery tool for Kubernetes.
The Python files within its pod queries a PostgreSQL database pod that is in turn connected to a MySQL database pod through a foreign data wrapper. The script iterates through a list of hyperspectral files, processes them, and outputs the data products indicated above.
The initial hyperspectral image file and its file path metadata are generated at the APPL user facility at ORNL. The image is stored on a local hard drive and the file path metadata is recorded in a local MySQL database. Subsequently, the images are transferred to the Themis storage enclave, which is accessible at the Oak Ridge Leadership Computing Facility (OLCF), and undergo data integrity verification through checksums. A copy of the APPL greenhouse database is maintained within the MySQL pod.

Workflow

The hyperspectral workflow consists of a series of discrete events:

1. Data acquisition
2. A data discovery event
3. A data ingestion event
4. Data masking and segmentation
5. Data product generation
6. Data analyses

The workflow begins with a data acquisition event. Data is acquired by one or more hyperspectral imagers and the raw (level 0) data is deposited into an appropriate directory structure. Data discovery is implemented as an automated response to the deposition event. When a directory for a particular instrument receives a raw data file, it triggers an instrument specific data ingestion process. The header file for the hyperspectral data is read and quality controlled to ensure it is indeed from the instrument that is expected. Data is then read into an in-memory array where it can be called by the data masking process. Since our pipeline is specific to plant species with green leaves, we use a dual parameter strategy: First, a color-based segmentation analysis is utilized where background pixels that are non-plant are masked. Subsequently, a shape-based segmentation analysis is performed for quality control purposes to ensure that anomalous plant pixels are not erroneously discarded. The pixels that are identified as “plant” pixels are then passed to the data product generation routines. There are four data products produced automatically: A masked pixels file, a masked PNG image of the data, a signatures file, and an endmembers file. The masked pixels file consists of a three-dimensional array of all the original pixels in the dataset that are classified as plant material. The mask is implemented as a color test where RGB pixel values are first translated into HSV color space. The green ranges, obtained observationally for each instrument, are used in the masking procedure. The procedure itself is implemented using the open-cv (cv2) python package. The algorithm starts with the full image, and systematically erodes away the non-plant of pixels via the native erode function in the open-cv software package. This is followed by a dilation routine where the image is magnified. Following magnification the erosion routine is called again. This cycle continues until the desired mask is achieved. Once the mask is

achieved, it is saved as a parquet file. A portable network graphics (png) representation of the masked image is also produced. The masked data is used as input to generate the remaining data products. The signature file contains summary statistics for the reflectance values for each band over all masked pixels in the image. The summary statistics include the mean, the standard deviation and boundaries for the 2.5th percentile, the 25th percentile, the 75th percentile and the 97.5th percentile. For cases where there are two hyperspectral instruments with different spectral ranges, the two signature files are merged. This creates a continuous signature representing the entire spectral range of the two instruments. A gap in the spectra represents areas of non-overlap between the two instruments.

ACKNOWLEDGEMENTS

We thank Dave Weston and Philip Bingham for their valuable insight and advice for this project. Dave Weston conceived, designed and executed the nitrogen stress experiment with assistance from Madhavi Martin, Hunter Andrews, and Sara Jawdy. We appreciate their willingness to share the data with us.

REFERENCES

1. Specim Corporation, “AisaKESTREL Hyperspectral Imaging System,” http://www.specim.fi/hyperspectral-remotesensing/
2. Rijad Sarić, Viet D. Nguyen, Timothy Burge, Oliver Berkowitz, Martin Trtílek, James Whelan, Mathew G. Lewsey, Edhem Čustović, Applications of hyperspectral imaging in plant phenotyping,Trends in Plant Science,Volume 27, Issue 3,2022,Pages 301-315,ISSN 1360-1385, https://doi.org/10.1016/j.tplants.2021.12.003.
3. Chang, Chein-I. *Hyperspectral data processing: algorithm design and analysis*. John Wiley & Sons, 2013.
4. Khan, Muhammad Jaleed, et al. "Modern trends in hyperspectral image analysis: A review." *Ieee Access* 6 (2018): 14118-14129.
5. Landgrebe, David. "Information extraction principles and methods for multispectral and hyperspectral image data." *Information processing for remote sensing*. 1999. 3-37.

6. ElMasry, Gamal, and Da-Wen Sun. "Principles of hyperspectral imaging technology." *Hyperspectral imaging for food quality analysis and control*. Academic Press, 2010. 3-43.
7. Dargan, Shaveta, et al. "A survey of deep learning and its applications: a new paradigm to machine learning." *Archives of Computational Methods in Engineering* 27 (2020): 1071-1092.
8. Fabelo, Himar, et al. "Spatio-spectral classification of hyperspectral images for brain cancer detection during surgical operations." *PloS one* 13.3 (2018): e0193721.
9. Petersson, Henrik, David Gustafsson, and David Bergstrom. "Hyperspectral image analysis using deep learning—A review." *2016 sixth international conference on image processing theory, tools and applications (IPTA)*. IEEE, 2016.
10. Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.
11. Ahmad, Muhammad, et al. "Hyperspectral image classification—Traditional to deep models: A survey for future prospects." *IEEE journal of selected topics in applied earth observations and remote sensing* 15 (2021): 968-999.
12. Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." *the Journal of machine Learning research* 12 (2011): 2825-2830.
13. Stanton, Jeffrey M. "Galton, Pearson, and the peas: A brief history of linear regression for statistics instructors." *Journal of Statistics Education* 9.3 (2001).
14. Hsieh, Tien-Heng, and Jean-Fu Kiang. "Comparison of CNN algorithms on hyperspectral image classification in agricultural lands." *Sensors* 20.6 (2020): 1734.
15. Chang, Chein-I., and Antonio Plaza. "A fast iterative algorithm for implementation of pixel purity index." *IEEE Geoscience and Remote Sensing Letters* 3.1 (2006): 63-67.