Rapid peanut phenotyping and water quality monitoring using remote sensing and machine learning techniques

by

Kamand Bagherian

A thesis submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Auburn, Alabama
December 10, 2022

Keywords: Remote sensing, machine learning, precision agriculture, high-throughput plant phenotyping, hyperspectral imaging, satellite imaging, harmful algal blooms

Approved by

Yin Bao, Chair, Assistant Professor of Biosystems Engineering
William David Batchelor, Professor of Biosystems Engineering
Alvaro Sanz-Saez, Assistant Professor of Crop, Soil and Environmental Sciences

Abstract

As aerial and sensing technologies have been developed, so have their applications in plant phenotyping and water quality monitoring. This is especially true with respect to the mass manufacture of easily portable quadcopters and hexacopters and publicly available satellite imagery. These platforms enable researchers to acquire data at a rapid pace, eliminating the need for manual and labor-intensive measurements. This paradigm shift constitutes a dramatic reduction in the cycle time of hypothesis testing, and ultimately enables us to glean more insights into the nature of reality, faster. The research in this thesis exploits these remote sensing technologies, coupled with machine learning techniques, to propose new solutions to rapid peanut phenotyping for breeding drought tolerance and water quality monitoring for the prediction of harmful algal blooms.

Direct measurement of the agronomical and physiological traits of peanuts is labor-intensive and time-consuming, and these traits hold invaluable information for breeders who need to select peanut genotypes with high-yielding and resilient characteristics. As part of this study, UAV-based hyperspectral imaging and machine learning (ML) techniques were used to predict three agronomic traits (biomass, pod count, and yield) as well as two physiological traits (photosynthesis and stomatal conductance) in peanut plants under drought stress.

An evaluation of two different approaches was conducted. Using 80 narrow-band vegetation indices as input features, the first approach employed an ensemble model of K-nearest neighbors, support vector regression, random forest, and multi-layer perceptron (MLP) to predict the agronomic and physiological traits. Second, the mean and standard deviation of canopy spectral reflectance were calculated per band, resulting in a total of 400 features that were used to train an

end-to-end deep learning (DL) model for the prediction of the same traits; biomass, pod count, pod yield, photosynthetic rate and stomatal conductance. This model consisted of several one-dimensional convolutional layers, followed by an MLP regressor. Agronomic traits predicted by feature learning and deep learning (R^2 = 0.45-0.73; sMAPE = 24-51%) outperformed those predicted by traditional machine learning and feature engineering (R^2 = 0.44-0.61, sMAPE = 27-59%). While the ensemble model did not match the DL model's performance in predicting agronomic traits, it was slightly better in predicting physiological traits, achieving R^2s in the range of 0.35-0.57 and sMAPEs in the range of 37-70%, while the DL model achieved R^2s between 0.36 and 0.52 and sMAPEs between 47 and 64%. It was demonstrated that using advanced remote sensing tools such as UAV-based hyperspectral imaging, coupled with machine learning, could enable peanut breeders to screen genotypes quickly for improved yield and drought tolerance.

Another problem addressed in this thesis was predicting chlorophyll-a (chl-a) concentrations and detecting harmful algal blooms (HABs) as chl-a concentration is often used as an indicator of algal blooms. Traditionally, collected water samples are required for lab-based cell taxonomy in order to measure chlorophyll-a concentrations. Using satellite images, it is possible to monitor inland water bodies extensively and rapidly. MODIS images were used in this study to predict chl-a concentrations and HAB events in Lake Okeechobee, the second largest freshwater lake in the United States. These images were acquired using Google Earth Engine (GEE) and processed in batches automatically for the period of 2011-2020. Ten years of time-series reflectance data were extracted from these images and several additional features were appended to it including cloud cover, chl-a estimations using the OCx algorithm, temperature data, and the sine transform of timestamps. These complex time-series data were

trained on a Long-short term memory (LSTM) model, a recurrent neural network (RNN) with the ability to learn long-term dependencies. The dataset was structured such that each day with a chl-a measurement was linked to same day reflectance data, as well as several days of reflectance data preceding the measurement day. Twelve variations of training sets were generated using different numbers of days of study before event dates, to study the effect of the time period on the result, and also to determine the optimum number of days we need to look back in time to detect HABs. The time variations ranged from 3 to 25 days before each chl-a measurement, and the results showed that a time period of fifteen days with a resolution of 4 days before each event, had the best performance with a root mean square error (RSME) of 11.95 µg/L, mean absolute error (MAE) of 8.55 µg/L and coefficient of determination (R^2) of 0.43. It was shown that satellite imagery and additional environmental features, together with a recurrent neural network such as LSTM, have the potential to detect HABs and estimate chl-a concentrations in Lake Okeechobee.

Acknowledgments

I would like to express my deepest gratitude to my advisor, Dr. Yin Bao, for his continuous support, guidance and kindness in every step of my Master's studies in the past two years. I am thankful for the opportunity to work under his supervision and gain multi-disciplinary research experience. I would also like to thank my committee members, Dr. Alvaro Sanz-Saez and Dr. Batchelor for their support and helpful feedback.

I had the opportunity to work with many amazing people through collaborations on numerous projects conducted by Dr. Bao and different departments and teams. I would like to thank Dr. Alvaro Sanz-Saez and his students once again, for their help with research and assistance with in-situ measurements associated with the work described in Chapter 2. I would also like to thank Dr. Stephanie Rogers and her team for their support with the research presented in Chapter 3.

Graduate school wouldn't have been as amazing as it was without my wonderful friends and colleagues, Rafael Bidese, Mary Beth Cassity and Gabi Itokazu. I would like to express my gratitude in particular for Rafael's unending assistance with respect to much of this research, especially for his company during the many challenging days of field work.

I would also like to thank my family for their never-ending support and encouragement, even when they're thousands of miles away. And lastly, I would like to thank my husband for his love, support, patience and the sacrifices he made to help me during my graduate studies at Auburn University.

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

As interest in automation and rapid monitoring grows, time-consuming and labor intensive tasks in biological applications–particularly in agriculture–are increasingly utilizing remote sensing technologies. Remote sensing platforms offer many advantages over traditional methods of monitoring and phenotyping. These advantages include key differences in the amount of data that can be obtained over a given amount of time, and also the fact that remote sensing is inherently non-destructive to the subject matter.

One application of remote sensing in agriculture is monitoring crop phenotypes in a non-destructive and efficient manner (Araus and Cairns, 2014; Araus, Kefauver, Zaman-Allah, Olsen, and Cairns, 2018). Satellite imaging is one of the high-throughput plant phenotyping techniques that has been used in the past few years (Chawade et al., 2019). This technique has been utilized for different phenotyping applications, such as estimating leaf area index (Kaplan et al., 2021; Wei et al., 2017), above-ground biomass (Han et al., 2017; Sibanda, Mutanga, Rouget, and Kumar, 2017) and yield prediction (Peralta, Assefa, Du, Barden, and Ciampitti, 2016; Schwalbert et al., 2018). Despite having the potential to be used for phenotyping agronomic traits, satellite images have a limited spatial resolution and cannot be used for assessing physiological characteristics at the plot level.

UAV (unmanned-aerial vehicle)-based images have several advantages over satellite images that make them preferable in several agricultural applications. UAVs give researchers the flexibility to choose different types of sensors to be mounted on the UAV according to the problem, and they can provide significantly better resolution in all aspects; a better temporal resolution can be achieved since flight missions can be scheduled based on the needs of the project. Additionally, the possibility of mounting different sensors on the platform, gives the

users and researchers the flexibility of choosing the sensor of their choice which can provide better spatial and spectral resolutions (Araus and Cairns, 2014; Araus et al., 2018). A growing number of high-resolution cameras with high-resolution sensors have become available to consumers and researchers. These cameras can provide valuable information regarding plant phenotypes. RGB, multispectral and hyperspectral cameras are three of the most commonly used sensors in this field. By combining two or more bands collected by these sensors, vegetation indices (VI) can be calculated, highlighting vegetation properties, plant health, and stress. Thus, UAV imagery can therefore provide quick insights into plant health over large areas. The most common and low-cost type of sensor on UAVs is RGB cameras that are often used to assess the physiological and agronomic traits of plants. For instance, Choudhary, Biswal, Saha, and Chatterjee (2021) evaluated the nitrogen status of wheat using aerial RGB images. Similarly, barley biomass was estimated with the same type of images in a study conducted by Bendig et al. (2014). However, RGB sensors lack the near-infrared (NIR) band which is critical when making decisions about plant health and phenotypes. Plants have strong reflectivity in the NIR region, since healthy vegetation absorbs light in the blue and red wavelengths for photosynthesis and creates chlorophyll which is highly reflective in near infrared. Therefore, most research studies take this band into account, by utilizing multi-spectral or hyperspectral sensors. The NIR band is often combined with other bands such as red or red-edge, providing metrics for plant health. This combination of bands are called vegetation indices and are commonly used for estimating crop attributes such as above-ground biomass, leaf area index (LAI), water stress, yield and chlorophyll content (Maresma, Ariza, Martínez, Lloveras, and Martínez-Casasnovas, 2016; Qi et al., 2021; Romero, Luo, Su, and Fuentes, 2018; Su et al., 2019). Both multispectral and hyperspectral sensors provide images from these wavelengths, and therefore vegetation indices

can be created from their output imagery. However, there are several key differences between multi and hyperspectral cameras; hyperspectral cameras are more complicated and costly, and resulting images from these cameras require large amounts of storage, but they provide the reflectance in over two hundred narrowband continuous wavelengths, whereas multispectral images are provided in few broadband wavelengths. The continuous reflectance from hyperspectral sensors gives us the spectral signature of the crop, which holds determining information that allows a deeper analysis of plant characteristics in high-throughput phenotyping applications.

Fenghua et al. (2017) used a hyperspectral camera mounted on a UAV for phenotyping LAI (leaf area index), leaf chlorophyll content (Cab), canopy water content (Cw), and dry matter content (Cdm) of rice. Yield and biomass prediction are also a common use of these sensors (X. Feng et al., 2020; Moghimi, Yang, and Anderson, 2020). Hyperspectral sensors have also been used for assessing photosynthetic attributes in several studies. Kanning et al. (2018) estimated chlorophyll content and LAI from UAV-based hyperspectral data. To the best of the authors' knowledge, there is no previous work estimating peanut photosynthetic rate and stomatal conductance from aerial hyperspectral imagery. A more common method for assessing these traits in the available literature is via spectrometers based on leaf contact. Buchaillot et al. (2022) estimated peanut and soybean photosynthetic traits such as mid-day photosynthesis, maximum rubisco capacity (Vcmax) and maximum RuBP regeneration capacity (Jmax) using leaf spectral reflectance obtained by a handheld spectrometer (Field Spec Hi-Res 4, Malven Analytics). Qi et al. (2020) employed the same device for measuring peanut leaves chlorophyll content. However, handheld spectrometers do not allow high-throughput screening and are time-consuming, as the

reflectance of each plant needs to be assessed individually. By contrast, aerial hyperspectral imagery can cover a larger area more effectively by looking at several plants simultaneously.

Choosing the right remote sensing platform, including the sensor (RGB, thermal, multispectral or hyperspectral) and the platform (drone or satellite) involves a tradeoff between spectral, temporal and spatial resolution. Even though UAV-based hyperspectral images provide the highest spectral resolution, and atmospheric conditions do not interfere with the imaging- unlike satellite imagery- this method is not suitable for monitoring areas larger than a few acres due to the battery and flight time limit of these vehicles. Therefore, satellite images are preferable for monitoring water quality in large lakes. In the modern era, the entire globe has been imaged at a multitude of different wavelengths, and given that much of this data is freely accessible, it is easier than ever to leverage when working on research questions that do not require a very high degree of resolution.

Satellite multispectral images can assist water quality monitoring, since similar to plants, chlorophyll are found in algae and highly reflective in near infrared and their reflectance in different bands can provide insights into chlorophyll concentrations. Chlorophyll-a (chl-a) is commonly used as an indicator of harmful algal blooms (HAB), and chl-a concentrations higher than 10 μg/L classify as HABs (World Health Organization [WHO] (2022). Numerous studies have been done for both estimating chl-a concentrations and detecting/predicting HABs, and satellites of different types have been used in their research. Landsat is one of the most commonly used satellite products for monitoring in-land algal blooms (Khan et al., 2021). However, it is limited by its 16-day temporal resolution. These long revisit intervals limit the utility of Landsat for mapping algal blooms' temporal variability. Sentinel-2 is another satellite with high spatial resolution which is used for monitoring freshwater regions. This satellite was

launched in 2015 and therefore developing a model with a limited time range can be limiting, especially since inland water bodies are not sampled as frequently as marine waters and have less field observations (Ventura et al., 2022). Choosing the right type of satellite involves a tradeoff between the range of availability, temporal and spatial resolution. MODerate resolution imaging spectroradiometer (MODIS) is one of the satellites offering an archive of long-term image series of daily global coverage. The high temporal resolution of this satellite increases the probability of getting cloud-free images in the areas of interest, and its long observation record (since 1999) allows a deeper analysis of temporal dynamic blooms in inland waters.

A number of previous studies have demonstrated that MODIS products can be used to estimate chlorophyll-a levels in large inland water bodies. By studying thirteen lakes in Brazil with water surface areas ranging from 1.85 to 441 km2, Ventura et al. (2022) explored the potential of using MODIS imagery to estimate chl-a on lakes of different sizes. The results showed that the three biggest lakes with the highest frequency of field sampling showed the best results, with $R^2 > 0.5$. Zhang et al. (2011) used the reflectance from MODIS band 2 (near infrared) and an empirical model to make predictions on chl-a in Lake Taihu. Another study by Li et al. (2019) also explores chl-a predictions in Lake Taihu using a classification-based MODIS land-band algorithm. A study on Lake Okeechobee demonstrated the potential of using MODIS imagery for estimating chl-a, using three different models; a genetic programming (GP) model, an artificial neural network (ANN) model and a multiple linear regression (MLR) model (Chang, Yang, Daranpob, Jin, and James, 2011).

With the advent of UAVs and public availability of satellite data, and their proven importance in crop and water quality monitoring, it has become clear for most researchers what platforms and sensors are most suitable for the problem that's being addressed, and their focus

has shifted from data collection to different methods of data analysis. As sensor data becomes more complex, more sophisticated methods are needed to correlate the underlying patterns between sensor data and the observable characteristics of the subject. Machine learning and deep learning models have shown their capability of extracting information from complex, high-dimensional and time series data. Both studies in this thesis utilize data that has been acquired using one of the aforementioned platforms and propose new solutions using machine learning and deep learning techniques to achieve the best result. They are individually discussed and compared to traditional methods in the following chapters and a summary of their objective is as follows.

Objective 1. Phenotyping Agronomic and Physiological Traits in Peanut Using UAV-Based Hyperspectral Imaging and Machine Learning

This study is provided in the second chapter of this thesis, assesses the feasibility of predicting pod yield, pod count, biomass, photosynthetic rate, and stomatal conductance in peanuts using UAV-based hyperspectral imaging and machine learning methods. There were two approaches compared, each representing a machine learning paradigm (i.e., feature engineering and feature learning). The first approach utilized vegetation indices as the input features to an ensemble model of conventional machine learning models, while the second approach employed a deep one-dimensional (1-D) convolutional neural network (CNN) that took the average and standard deviation of peanut canopy reflectance as input features. In addition, the importance of the vegetation indices utilized for the ensemble model and the wavelengths in the deep learning model were evaluated using permutation importance scores.

Objective 2. Detecting harmful algal blooms in Lake Okeechobee using MODIS satellite imagery and long-short term memory (LSTM)

The second study presented in chapter three, investigates the possibility of detecting harmful algal blooms and estimating chl-a concentration in Lake Okeechobee using MODIS images from 2011 to 2020 and several additional features that were appended to the dataset. A recurrent neural network, long-short term memory (LSTM), was employed and trained on this time-series data. In addition, three machine learning models were trained on the same dataset and their performance was compared to LSTM's performance. These models were trained on both single-time and time-series inputs to assess if temporal features have any effect on predictions. Another experiment was testing different ranges of data points, from 3 to 25 days, preceding the day events were recorded. The results were compared to find the optimal time period for evaluating HABs and chl-a predictions, and finally, a feature importance analysis was conducted to find out which features contributed most to the model's performance.

References

Araus, J. L., & Cairns, J. E. (2014). Field high-throughput phenotyping: The new crop breeding frontier. Trends in Plant Science, 19(1), 52–61.

Araus, J. L., Kefauver, S. C., Zaman-Allah, M., Olsen, M. S., & Cairns, J. E. (2018). Translating High-Throughput Phenotyping into Genetic Gain. Trends in Plant Science, 23(5), 451–466.

Bendig, J., Bolten, A., Bennertz, S., Broscheit, J., Eichfuss, S., & Bareth, G. (2014). Estimating biomass of barley using Crop Surface Models (CSMs) derived from UAV-based RGB imaging. Remote Sensing, 6(11), 10395–10412.

Buchaillot, Ma. L., Soba, D., Shu, T., Liu, J., Aranjuelo, I., Araus, J. L., Sanz-Saez, A. (2022). Estimating peanut and soybean photosynthetic traits using leaf spectral reflectance and advance regression models. Planta, 255(4), 93.

Chang, N. B., Yang, Y. J., Daranpob, A., Jin, K. R., & James, T. (2011). Spatiotemporal pattern validation of chlorophyll-a concentrations in Lake Okeechobee, Florida, using a comparative MODIS image mining approach. 33(7), 2233–2260.

Chawade, A., Van Ham, J., Blomquist, H., Bagge, O., Alexandersson, E., & Ortiz, R. (2019). High-Throughput Field-Phenotyping Tools for Plant Breeding and Precision Agriculture. Agronomy, 9

Choudhary, S. S., Biswal, S., Saha, R., & Chatterjee, C. (2021). A non-destructive approach for assessment of nitrogen status of wheat crop using unmanned aerial vehicle equipped with RGB camera. Arabian Journal of Geosciences, 14(17), 1739.

Feng, X., Zhan, Y., Wang, Q., Yang, X., Yu, C., Wang, H., He, Y. (2020). Hyperspectral imaging combined with machine learning as a tool to obtain high-throughput plant salt-stress phenotyping. The Plant Journal, 101(6), 1448–1461.

Fenghua, Y., Tongyu, X., Wen, D., Hang, M., Guosheng, Z., & Chunling, C. (2017). Radiative transfer models (RTMs) for field phenotyping inversion of rice based on UAV hyperspectral remote sensing. International Journal of Agricultural and Biological Engineering, 10(4), 150–157.

Han, J., Wei, C., Chen, Y., Liu, W., Song, P., Zhang, D., Huang, J. (2017). Mapping Above-Ground Biomass of Winter Oilseed Rape Using High Spatial Resolution Satellite Data at Parcel Scale under Waterlogging Conditions. Remote Sensing, 9(3), 238.

Kanning, M., Kühling, I., Trautz, D., & Jarmer, T. (2018). High-Resolution UAV-Based Hyperspectral Imagery for LAI and Chlorophyll Estimations from Wheat for Yield Prediction. Remote Sensing, 10(12), 2000.

Kaplan, G., Fine, L., Lukyanov, V., Manivasagam, V. S., Malachi, N., Tanny, J., & Rozenstein, O. (2021). 66. Estimating processing tomato water consumption, leaf area index and height using Sentinel-2 and Venμs imagery. Precision Agriculture ’21, 551–557. The Netherlands: Wageningen Academic Publishers.

Khan, R. M., Salehi, B., Mahdianpari, M., Mohammadimanesh, F., Mountrakis, G., & Quackenbush, L. J. (2021). A meta-analysis on harmful algal bloom (HAB) detection and monitoring: A remote sensing perspective. Remote Sensing, 13(21).

Li, J., Gao, M., Feng, L., Zhao, H., Shen, Q., Zhang, F., Zhang, B. (2019). Estimation of chlorophyll-a concentrations in a highly turbid eutrophic lake using a classification-based MODIS land-band algorithm. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12(10), 3769–3783.

Maresma, Á., Ariza, M., Martínez, E., Lloveras, J., & Martínez-Casasnovas, J. (2016). Analysis of vegetation indices to determine nitrogen application and yield prediction in maize (Zea mays L.) from a Standard UAV Service. Remote Sensing, 8(12), 973.

Moghimi, A., Yang, C., & Anderson, J. A. (2020). Aerial hyperspectral imagery and deep neural networks for high-throughput yield phenotyping in wheat. Computers and Electronics in Agriculture, 172, 105299.

Peralta, N., Assefa, Y., Du, J., Barden, C., & Ciampitti, I. (2016). Mid-season high-resolution satellite imagery for forecasting site-specific corn yield. Remote Sensing, 8(10), 848.

Qi, H., Wu, Z., Zhang, L., Li, J., Zhou, J., Jun, Z., & Zhu, B. (2021). Monitoring of peanut leaves chlorophyll content based on drone-based multispectral image feature extraction. Computers and Electronics in Agriculture, 187, 106292.

Qi, H., Zhu, B., Kong, L., Yang, W., Zou, J., Lan, Y., & Zhang, L. (2020). Hyperspectral inversion model of chlorophyll content in peanut leaves. Applied Sciences 2020, Vol. 10, Page 2259, 10(7), 2259.

Romero, M., Luo, Y., Su, B., & Fuentes, S. (2018). Vineyard water status estimation using multispectral imagery from an UAV platform and machine learning algorithms for irrigation scheduling management. Computers and Electronics in Agriculture, 147, 109–117.

Schwalbert, R. A., Amado, T. J. C., Nieto, L., Varela, S., Corassa, G. M., Horbe, T. A. N., Ciampitti, I. A. (2018). Forecasting maize yield at field scale based on high-resolution satellite imagery. Biosystems Engineering, 171, 179–192.

Sibanda, M., Mutanga, O., Rouget, M., & Kumar, L. (2017). Estimating biomass of native grass grown under complex management treatments using WorldView-3 spectral derivatives. Remote Sensing, 9(1), 55.

Su, W., Zhang, M., Bian, D., Liu, Z., Huang, J., Wang, W., Guo, H. (2019). Phenotyping of Corn plants using unmanned aerial vehicle (uav) images. Remote Sensing, 11(17).

Ventura, D. L. T., Martinez, J. M., de Attayde, J. L., Martins, E. S. P. R., Brandini, N., & Moreira, L. S. (2022). Long-term series of vhlorophyll-a concentration in Brazilian Semiarid Lakes from Modis imagery. Water (Switzerland), 14(3).

Wei, C., Huang, J., Mansaray, L. R., Li, Z., Liu, W., Han, J., Thenkabail, P. S. (2017). Remote sensing estimation and Mapping of Winter Oilseed Rape LAI from High Spatial Resolution Satellite Data Based on a Hybrid Method. Remote Sensing, 9, 488.

World Health Organization (WHO). (2022).

Zhang, Y., Lin, S., Qian, X., Wang, Q., Qian, Y., Liu, J., & Ge, Y. (2011). Temporal and spatial variability of chlorophyll a concentration in Lake Taihu using MODIS time-series data. Hydrobiologia, 661(1), 235–250.

Chapter 2. Phenotyping Agronomic and Physiological Traits in Peanut Using UAV-Based Hyperspectral Imaging and Machine Learning

2.1 Abstract

Agronomic and physiological traits in peanut are important to breeders for selecting high-yielding and resilient genotypes. However, direct measurement of these traits is labor-intensive and time-consuming. This study assessed the feasibility of using UAV-based hyperspectral imaging and machine learning (ML) techniques to predict three agronomic traits (biomass, pod count, and yield) and two physiological traits (photosynthesis and stomatal conductance) in peanut under drought stress. Two different approaches were evaluated. The first approach employed eighty narrow-band vegetation indices as input features for an ensemble model that included K-nearest neighbors, support vector regression, random forest, and multi-layer perceptron (MLP). The second approach utilized mean and standard deviation of canopy spectral reflectance per band. The resultant 400 features were used to train a deep learning (DL) model consisting of one-dimensional convolutional layers followed by a MLP regressor. Predictions of the agronomic traits obtained using feature learning and DL (R^2 = 0.45-0.73; sMAPE = 24-51%) outperformed those obtained using feature engineering and conventional ML models (R^2 = 0.44-0.61, sMAPE = 27-59%). In contrast, the ensemble model had a slightly better performance in predicting physiological traits (R^2 = 0.35-0.57; sMAPE = 37-70%) compared to the results obtained from the DL model (R^2 = 0.36-0.52; sMAPE = 47-64%). The results showed that the combination of UAV-based hyperspectral imaging and machine learning techniques have the

potential to assist breeders in rapid screening of genotypes for improved yield and drought tolerance in peanuts.

2.2 Introduction

Peanut is one of the most important cash crops in the United States, valued at over one billion U.S. dollars. Over 3 million tons of peanuts were harvested in 2020 from approximately 1.6 million acres in the United States (USDA-NASS 2020). Peanuts are grown in many Southern states in the U.S. and around the world. For this reason, peanut breeding programs aim to develop cultivars that have desirable and improved traits that can be adapted to their respective environments. Moreover, as droughts are becoming more frequent, severe, and widespread, drought-tolerant cultivars need to be developed for regions affected by drought (NASA, 2021). In a breeding program, a breeder may need to measure multiple traits for hundreds to thousands of peanut genotypes at multiple field locations every year. Typical agronomic traits in peanut include biomass, pod count, and pod yield, which quantify how a peanut plants convert energy and nutrients into different yield components. Measuring the three agronomic traits is normally done manually, which involves drying, weighing, counting, and shelling. Physiological traits such as photosynthesis rate and stomatal conductance can indicate whether a plant is under drought stress (Buezo et al., 2019; Zhang et al., 2022). These traits are measured by using a portable infrared gas analyzer that detects the plant's CO2 fixation and the water liberated through stomata. Both procedures are labor-intensive and time-consuming, especially at large scales (Baslam et al., 2020).

High-throughput plant phenotyping (HTPP) offers solutions to alleviate the phenotyping bottleneck in breeding programs. Remote sensing techniques have made it possible to monitor crop phenotypes in a non-destructive and efficient manner and are thus a valuable tool for

estimating agronomic traits. The use of unmanned aerial vehicles (UAVs) for precision agriculture has recently gained significant attention because of their greater flexibility in mission scheduling, and the possibility of mounting different high-resolution sensors on the platform (Araus and Cairns, 2014; Araus et al., 2018). High-resolution sensors such as RGB, multispectral, and hyperspectral cameras have become available to researchers and consumers and can provide valuable information regarding plant phenotypes. The images collected by these sensors can be used to calculate vegetation indices (VI), which are mathematical combinations of two or more bands to highlight vegetation properties, plant health, or stress. Therefore, UAV imagery can rapidly reveal information about the health of plants in a large area. RGB cameras are the most accessible and common type of sensor utilized on UAVs. They are often used to assess plant physiological and agronomical traits. Examples include a study by Choudhary et al. (2021), where vegetation indices obtained from an RGB camera were used to assess the nitrogen status of wheat. Bendig et al. (2014) also used UAV-based RGB imagery to estimate the biomass of barley. Due to strong reflectivity of plant canopy at near-infrared (NIR) wavelengths, multispectral sensors incorporating NIR channels are becoming more popular. NIR gives information about the cellular structure within leaves and when combined with a band like red or red-edge, it gives VIs such as NDVI (Normalized Difference Vegetation Index) and NDRE (Normalized Difference Red Edge Index), which provide measurements for overall plant health. These VIs have been applied in numerous studies to estimate above-ground biomass, leaf area index (LAI), water stress, and yield prediction (Maresma et al., 2016; Romero et al., 2018; Su et al., 2019). Qi et al also used vegetation indices from a multispectral camera to monitor chlorophyll content in peanut leaves (Qi et al., 2021).

Hyperspectral cameras are more complicated and costly compared to RGB and multispectral sensors, and the resulting images from these cameras require large amounts of storage. Despite their complexity, they provide invaluable information about crops' reflectance in hundreds of narrow spectral bands. This allows a more advanced analysis of plant characteristics in high-throughput phenotyping applications. Fenghua et al. (2017) used a hyperspectral camera mounted on a UAV for phenotyping LAI, leaf chlorophyll content (Cab), canopy water content (Cw), and dry matter content (Cdm) of rice. Yield and biomass predictions are also a common use of these sensors (Feng et al., 2020; Moghimi et al., 2020). Hyperspectral sensors have also been used for assessing photosynthetic attributes in several studies. Kanning et al. (2018) estimated the chlorophyll content and LAI from UAV-based hyperspectral data. To the best of our knowledge, there is no previous work on estimating peanut photosynthetic rate and stomatal conductance from aerial hyperspectral imagery. A more common method for assessing these traits in the literature is to measure individual leaves using a spectrometer. Buchaillot et al. (2022) estimated peanut and soybean photosynthetic traits such as mid-day photosynthesis, maximum rubisco capacity (Vcmax) and maximum RuBP regeneration capacity (Jmax) using leaf spectral reflectance obtained by a handheld spectrometer (Field Spec Hi-Res 4, Malven Analytics). Qi et al. (2020) employed the same device for measuring peanut leaves chlorophyll content. However, handheld spectrometers do not allow high-throughput screening and are time consuming, as the reflectance of individual plants needs to be assessed manually. Aerial hyperspectral imaging on the other hand, can capture a large area of plants in a far more efficient and automated manner.

As UAVs and sophisticated cameras become more viable, compact and affordable, the focus of HTPP has shifted from data collection to data analytics. A common approach for

analyzing the data is correlating extracted vegetation indices with crop phenotypes or diseases using statistical methods. Patrick et al. assessed the correlation of several vegetation indices such as NDRE, green difference vegetation index (GDVI) with tomato spot wilt disease in peanut using linear regression. Balota and Oakes (2017) compared vegetation indices derived from the ground and aerial sensor data with leaf wilting, pod yield, and crop value in peanut based on Pearson correlation results. However, as sensor data becomes more complex, such as hyperspectral data, more advanced methods are required to determine the underlying patterns between sensor data and phenotypes.

Machine Learning (ML) and Deep Learning (DL) models have been shown to be highly capable of extracting information from complex and high-dimensional data and for that reason they have become a popular data analytics method for HTPP. A common approach is feeding the extracted vegetation indices as the input to a ML model such as K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), etc. (Eugenio et al., 2020; Maimaitijiang et al., 2017; Qi et al., 2021; Sankaran et al., 2021; Wang et al., 2021). Feng et al. (2020) et al. showed that ensemble models are more powerful than individual ML models. They developed an ensemble model by combining ML models and trained the model on narrow-band vegetation indices derived from aerial hyperspectral imagery for alfalfa yield prediction. Instead of using predetermined wavelengths for the VIs, Feng et al. (2020) used ANOVA, multilayer perception, and reduced sampling to identify the most significant wavelengths, which were then utilized for the construction of new VIs that are able to detect bacterial wilt. Another approach is using the average spectrum at the plot level with no dimension reduction and training a DL model to learn the most significant bands of the spectrum. DeepRWC was developed by Rehman et al. (2020), an end-to-end DL model to predict the relative water content (RWC) of plants directly from

mean spectral reflectance. Moghimi et al. (2020) implemented a deep neural network consisting of fully connected layers for high-throughput yield phenotyping in wheat. Both mean and standard deviation in addition to the area of leaves and spikes were used as the input.

There are numerous studies on assessing remote sensing and ML techniques for phenotyping crops such as barley, alfalfa, and rice. However, there is limited research on high-throughput phenotyping of agronomic and physiological traits in peanut. This study evaluated the feasibility of predicting pod yield, pod count, biomass, photosynthetic rate, and stomatal conductance in peanut using UAV-based hyperspectral imaging and ML methods. Two approaches representing two ML paradigms (i.e., feature engineering and feature learning) were compared. The first approach utilized vegetation indices as the input features to an ensemble model of conventional ML models, while the second approach employed a deep one-dimensional (1-D) convolutional neural network (CNN) that took the average and standard deviation of peanut canopy reflectance as input features. Another objective of this study was to find the best day for data collection to get the best predictions for the agronomic traits of drought-stressed peanuts. There were three data collections between the start of drought and harvest, fourteen, eighteen and twenty nine days after drought. Both the deep learning and the ensemble machine learning model were trained on the data collected on each day, and their performances were evaluated. Finally, the importance of the features utilized in both models were evaluated using permutation importance scores.

2.3 Materials and Methods

2.3.1 Experimental Design

The field experiment was conducted at the U.S. Department of Agriculture - Agricultural Research Service National Peanut Research Laboratory in Dawson, Georgia, USA

(31.759875793753956, -84.43488756104786). The field was divided into four blocks and each block was equipped with an automatic rainout shelter (Blankenship, Mitchell, Layton, Cole, and Sanders, 1989). Each metal shelter covers a ground area of 5.5 m ×12.2 m and automatically closes when a rain detector (Agrowtek IR Digital Rain Sensor, Agroetek, Brookfield, Wisconsin) is triggered. Each shelter was planted as a common garden experiment and was further divided into 16 rows and 4 columns, resulting in 64 individual plant plots per shelter. Two of the four rainout shelters were employed to impose drought treatments while the others were maintained under well irrigated conditions. Each plot is 0.3 m × 0.9 m in dimensions, with a 0.15-m row spacing. A single peanut plant was grown in each plot following a generalized randomized block design. The plant materials were PI502120, AU-NPL 17, Ga-Green, AP-3, x587, C76-16, AT3085RO, Line 8, and TifRunner parent cultivars as well as the F1 population of crossing of Tifrunner with the other parent lines. Each parent cultivar and F1 descendant was replicated 3 times per shelter. PI502120, AU-NPL 17, and Line 8 are of high drought tolerance; C76-16, TifRunner, and x587 are of moderate drought tolerance and AP-3, Ga-Green, and AT3085RO are drought sensitive (Q. Zhang et al., 2022). A set of Water Mark soil moisture sensors (Irrometer, Riverside, CA, USA) were placed in the center of the field under each shelter at depths of 0.1 m and 0.2 m. Irrigation was triggered when the soil water potential was under -60 KPa before the drought was imposed. During the drought, the irrigated shelters followed the same regime but the drought shelters did not received any water. The simulated drought was imposed on July 26, 2021 and terminated with rewatering after six weeks. The peanuts were harvested on September 23, 2021. During this period, UAV-based hyperspectral images were collected on August 9th, August 13th, and August 24th, which are 14, 18, and 29 days after drought (DAD), respectively. Biomass, pod yield, and pod count were measured after harvest.

Photosynthesis and stomatal conductance were measured on the same image collection dates using four LI-6400 systems (LI-COR Biosciences, Lincoln, NE, USA) at midday (11:00 to 13:30). Measurements were performed on fully expanded young leaves corresponding with the second/third leaf from the top of the main stem. The LI-6400 chambers were set to display the same environmental conditions (i.e., light, relative humidity, temperature) as the atmospheric condition varied between measurement days. A summary of the statistics of the measured ground-truth data is shown in Table 2.1, and an aerial image of the field, taken on August 5th, is presented in Figure 2.1. Moreover, the distribution of spectral reflectance across 256 plots and three data collections is shown in Figure 2.2.

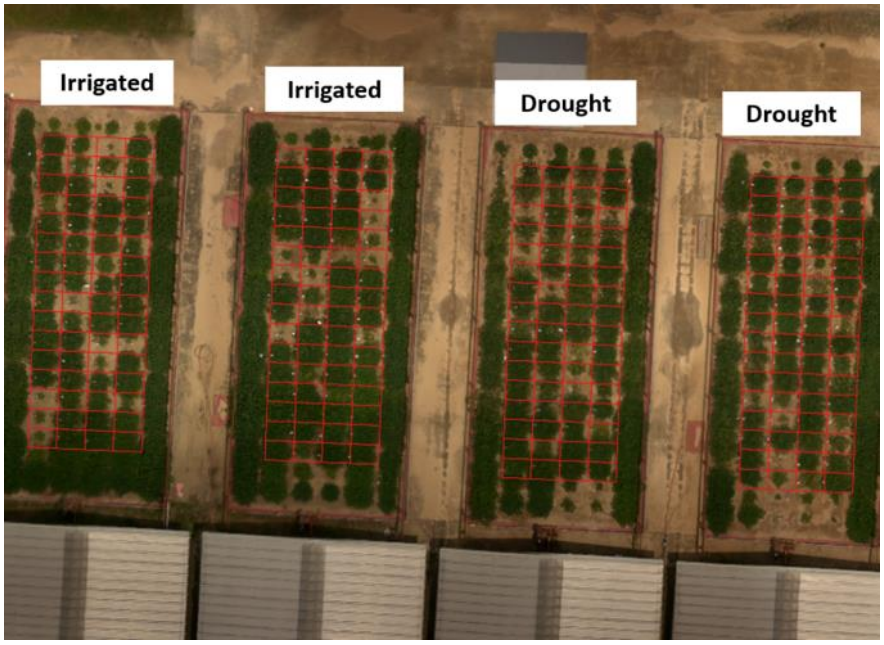

Figure 2.1 Aerial image of the experimental field with four rainout shelters open. A grid is overlaid on the image to indicate plot boundaries.

Table 2.1 Summary of the statistics of the measured agronomic and physiological traits.

Trait	Min	Max	Mean	Standard Deviation
Biomass (g/plant)	1.11	409.52	156.14	82.14
Yield (g/plant)	0.00	326.7	122.00	80.01
Pod Count	0.00	342.00	120.85	70.37
Photosynthetic rate – 14 DAD ($\mu mol.m^{-2}s^{-1}$)	-0.25	47.63	19.89	11.61
Photosynthetic rate – 18 DAD	-3.71	40.83	14.90	10.80
Photosynthetic rate – 29 DAD ($\mu mol.m^{-2}s^{-1}$)	0.37	42.66	16.45	10.48
Stomatal conductance – 14 DAD ($mmol.m^{-2}s^{-1}$)	-0.05	1.08	0.30	0.25
Stomatal conductance – 18 DAD ($mmol.m^{-2}s^{-1}$)	-0.12	1.02	0.19	0.21
Stomatal conductance – 29 DAD ($mmol.m^{-2}s^{-1}$)	-0.04	2.18	0.29	0.31

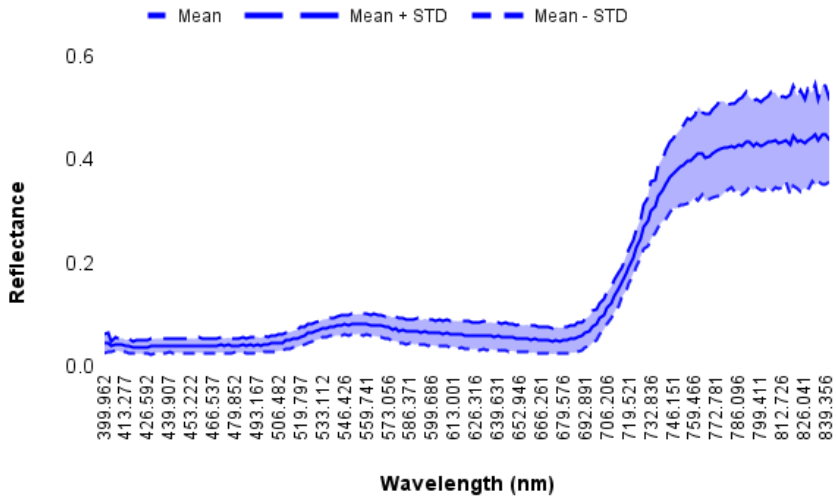

Figure 2.2 Average spectral reflectance across 256 plots and three data collections ± the standard deviation.

2.3.2 UAV-Based Hyperspectral Imaging and Data Preprocessing

The UAV platform used was a Matrice 600 Pro hexcopter (Shenzhen DJI Sciences and Technologies Ltd., China). Flight missions were planned using UgCS (SPH Engineering, Latvia) with 1% forward overlap and 40% side overlap. The camera was faced nadir during the flight and was stabilized using a Ronin-MX gimbal (Shenzhen DJI Sciences and Technologies Ltd., China) on the UAV. A push-broom visible-near-infrared (VNIR) hyperspectral camera (Nano-Hyperspec, Headwall Photonics, Inc., MA, USA) was used for the data collection. This camera covers a spectral range of 400-1000nm with a spectral resolution of 2 nm. Each line scanned by this sensor contains 640 pixels with a pixel pitch of 7.4 µm. Exposure time was adjusted using a white PVC panel so that its reflectance covered about 75% of the maximum reflectance the camera can capture. 700 frames were acquired per image cube. After the acquisition of the hyperspectral images, the raw files were radiometrically calibrated using the dark reference collected on the same day before the flight. The dark reference is a single image cube acquired with the lens cap on, with the same exposure settings as the other image cubes. The resulting radiance cubes were then calibrated to reflectance using a 3m by 3m calibration tarp with three regions of 56%, 32%, and 11% reflectivity, respectively. This panel was placed in the field on a flat surface for every data collection. Following the conversion to reflectance, all the images were geometrically corrected. The described post-processing steps were performed using SpectralView, a software provided by Headwall Photonics, Inc. Subsequently, a hyperspectral orthomosaic was created from the orthorectified images, a grid was overlaid on the orthomosaic in QGIS 3.18.2 (QGIS.org, 2022), and individual plots were extracted from the map (Figure 2.1). Soil pixels were removed from each plot Image using a normalized difference vegetation index (NDVI) threshold of 0.2. NDVI is defined in Equation 2.1 where R(x) denotes the reflectance at

wavelength x. This threshold was determined empirically, similar to previous studies (Liang et al., 2015; Moghimi et al., 2020). Additionally, noisy spectral bands above 844 nm were removed. The general workflow for the procedure of the explained procedure is shown in Figure 2.3.

$$NDVI = \frac{R(804) - R(693)}{R(804) + R(693)} \quad (2.1)$$

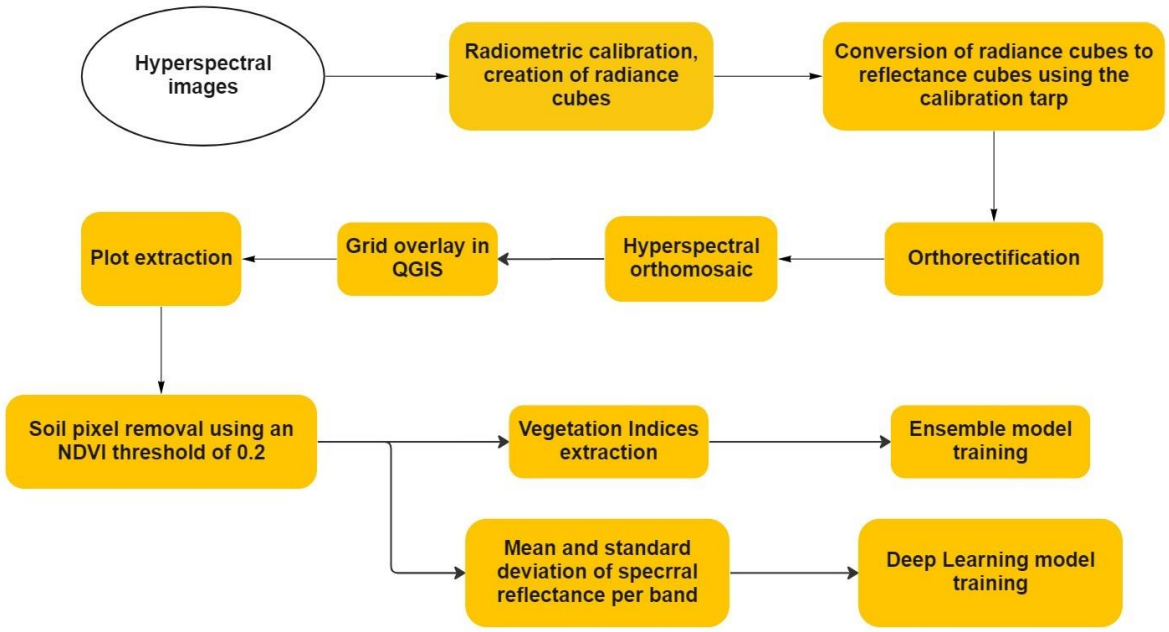

Figure 2.3 Hyperspectral imaging and processing workflow.

2.3.3 Machine Learning (ML) Models

Two methods were implemented in this study. The first method employed an Ensemble ML model consisting of four ML models including K-nearest neighbors (KNN), support vector regression (SVR), random forest (RF), and a multi-layer perceptron (MLP) regressor. The inputs to this model were eighty narrow-band VIs. The second method was a DL model with a 1-D CNN followed by a MLP regressor. The inputs to this model were the mean and standard deviation of spectral reflectance per band for each plot. The ensemble model and the DL model

were both trained for 2000 epochs and their hyperparameters were tuned with grid search. Cross validation with a total of 10 folds was performed to ensure the robustness of the model. There were a total of 256 plants, therefore the initial dataset had 256 data points. After removing several data points related to diseased plants, the dataset had 248 data points. Breaking the dataset to 10 folds, each fold had 228 data points for model training and 28 for model testing.

2.3.3.1 Ensemble Model

The hyperspectral orthomosaic has 270 continuous spectral bands and adjacent bands are normally correlated. Instead of using all of the original bands, eighty different VIs (Table 2.2) were computed at the plot level based on the work by Feng et al. (2020). These vegetation indices included twelve narrow-band NDVIs and nineteen simple ratio indices (SRIs). NDVI and SRI were examined more precisely due to their capability of characterizing canopy vigor, biomass, and photosynthetic rate.

Table 2.2 The vegetation indices used for the ensemble model.

Name	Index	Formula
Normalized difference vegetation index	NDVI [471, 584]	(R584 − R471)/(R584 + R471)
	NDVI [521, 689]	(R689 − R521)/(R689 + R521)
	NDVI [550, 760]	(R760 − R550)/(R760 + R550)
	NDVI [667, 740]	(R740 − R667)/(R740 + R667)
	NDVI[670, 800]	(R800 − R670)/(R800 + R670)
	NDVI[705, 750]	(R750 − R705)/(R750 + R705)
	NDVI[710, 750]	(R750 − R710)/(R750 + R710)
	NDVI[710, 780]	(R780 − R710)/(R780 + R710)
	NDVI[717, 732]	(R732 − R717)/(R732 + R717)
	NDVI[717, 770]	(R770 − R717)/(R770 + R717)
	NDVI[720, 820]	(R820 − R720)/(R820 + R720)
	NDVI[734, 750]	(R750 − R735)/(R750 + R734)
Physiological reflectance index	PRI[528,567]	(R528 − R567)/(R528 + R567)
	PRI[531,570]	(R570 − R531)/(R531 + R570)
Normalized difference red edge	NDRE	(R790 − R720)/(R790 + R720)
Modified normalized difference vegetation index	mND	(R750 − R705)/(R750 + R705 − 2×R445)
Green normalized difference vegetation index	GNDVI	(R750 − R550)/(R750 + R550)

Renormalized difference vegetation index	RDVI	$(R800 - R670)/\sqrt{(R800 + R670)}$
Normalized difference cloud index	NDCI	$(R762 - R527)/(R762 + R527)$
Curvature index	CI	$R675 \times R690/R6832$
-	Datt1	$(R850 - R710)/(R850 - R680)$
	Datt2	$R850/R710$
	Datt3	$R754/R704$
Double Difference index	DD	$(R749 - R720) - (R701 - R672)$
Double peak canopy nitrogen index	DCNI	$R720 - R700/[(R700 - R670)(R720 - R670 + 0.03)]$
-	Gitelson1	$1/R700$
	Gitelson2	$(R750\text{-}R800/R695\text{-}R740) - 1$
-	Carte1	$R695/R760$
	Carte2	$R605/R760$
	Carte3	$R710/R760$
	Carte4	$R695/R670$
Simple ratio index	SRI[533,565]	$R565/R533$
	SRI[550,750]	$R750/R550$
	SRI[550,760]	$R760/R550$
	SRI[560,810]	$R810/R560$
	SRI[629,734]	$R734/R629$
	SRI[660,810]	$R810/R660$
	SRI[670,700]	$R700/R670$
	SRI[670,800]	$R800/R670$
	SRI[675,700]	$R675/R700$
	SRI[680,800]	$R800/R680$
	SRI[690,752]	$R752/R690$
	SRI[700,750]	$R750/R700$
	SRI[705,750]	$R750/R705$
	SRI[706,755]	$R706/R755$
	SRI[708,747]	$R747/R708$
	SRI[710,750]	$R750/R710$
	SRI[717,741]	$R741/R717$
	SRI[720,735]	$R735/R720$
	SRI[720,738]	$R738/R720$
Modified simple ratio index	mSRI[550,780]	$R780/R550\text{-}1$
	mSRI[710,780]	$R780/R710\text{-}1$
	mSRI[720,750]	$R750/R720\text{-}1$
	mSR705	$(R750 - R445)/(R705 - R445)$
	mSR	$(R750/R705 - 1)/(\sqrt{R750/R705 + 1})$
New vegetation index	NVI1	$(R777 - R747)/R673$
	NVI2	$R705/(R717 + R491)$
Enhanced vegetation index	EVI	$2.5(R800 - R670)/(R800 - 6R670 - 7.5R475 + 1)$

Transformed Chlorophyll absorption in reflectance index	TCARI1	3[(R700 − R670) − 0.2(R700 − R550)(R700/R670)]
	TCARI2	3[(R750 − R705) − 0.2(R750 − R550)(R750/R705)]
Modified chlorophyll absorption ratio index	MCARI1	[(R700 − R670) − 0.2(R700 − R550)](R700/R670)
	MCARI2	[(R750 − R705) − 0.2(R750 − R550)](R750/R705)
	MCARI3	[(R750 − R710) − 0.2(R750 − R550)](R750/R715)
Optimized soil-adjusted vegetation index	OSAVI1	(1 + 0.16)(R800 − R670)/(R800 + R670 + 0.16)
	OSAVI2	(1 + 0.16)(R750 − R705)/(R750 + R705 + 0.16)
Combined TCARI/OSAVI	TCARI/OSAVI1	TCARI1/OSAVI1
	TCARI/OSAVI2	TCARI2/OSAVI2
Combined MCARI/OSAVI	MCARI/OSAVI1	MCARI1/OSAVI1
	MCARI/OSAVI2	MCARI2/OSAVI2
Triangular greenness index	TGI	−0.5[190(R670-R550) − 120(R670 − R480)]
Modified triangular vegetation index	MTVI	1.2[1.2(R800 − R550) − 2.5(670 − R550)]
MERIS terrestrial chlorophyll index	MTCI1	(R750 − R710)/(R710 − R680)
	MTCI2	(R754 − R709)/(R709 − R681)
Spectral polygon vegetation index	SPVI	0.4 × [3.7(R800 − R670) − 1.2|R550 − R670|]
Red edge position index	REP1	700 + 45[(R670 + R780)/2 − R700]/(R740 − R700)
	REP2	700 + 40[(R670 + R780)/2 − R700]/(R740 − R700)
-	VOG1	R740/R720
	VOG2	(R734 − R747)/(R715 + R726)
	VOG3	(R734 − R747)/(R715 + R720)
Optimal vegetation index	Viopt	(1 + 0.45)(R8002 + 1)/(R670 + 0.45)

It has been shown in several studies that ensemble models outperform individual ML models and have more robust results due to their diverse nature and not depending on an individual model's results (L. Feng et al., 2020; Q. Zhang et al., 2022). Our ensemble model used a voting regressor to give a final prediction from four models: K-nearest neighbors (KNN), support vector regression (SVR), random forest (RF), and a multi-layer perceptron (MLP) regressor. All individual models were tuned using grid search and the top performing versions

were used in the ensemble model. KNN regression is a non-parametric supervised ML algorithm that works based on the assumption that similar samples exist in close proximity to K nearest samples in the feature space, where K is a hyperparameter that needs to be tuned for a specific dataset. After ranking samples based on their distance to the unknown (testing) sample, it estimates the response by taking the average of the responses of K nearest neighbors in the training set. K was tuned to 4 for the KNN model. SVR is a supervised ML model that transforms input data into another space using a kernel function. A linear kernel function was selected in our case. RF regression is a combination of regression trees and the final prediction value is the average of all trees. The MLP regressor was configured as one hidden layer with 100 neurons and was trained for 2000 epochs. Adam (Adaptive Moment Estimation) was chosen as the optimizer with a learning rate of 0.1, and exponential decay rates of 0.9 and 0.99 for the first and second moment estimates, respectively. The ensemble model and its estimators were implemented in Python 3.9.7 using the libraries scikit-learn (Pedregosa et al., 2011) and NumPy (Harris et al., 2020).

2.3.3.2 Deep Learning (DL) Model

With a small dataset with a total of 248 data points, the input data needed to be simplified to reduce the model complexity and the number of trainable parameters. Instead of 3-D hyperspectral cubes, mean and standard deviation of reflectance were used as the inputs to the deep learning model, inspired by the work by Moghimi et al. (2020). 1-D convolution was chosen as the convolution method for these 1-D inputs. The architecture of the DL model consisted of four 1-D convolution layers and three dense layers, each of which is followed by a batch normalization layer and a dropout layer. The activation function of all layers in this model

is a Leaky ReLu (rectified linear unit), with α = 0.3 in the Leaky ReLU function (Equation 2.2). Unlike ReLU, Leaky ReLU allows a small gradient when the unit is not active.

$$f(x) = \alpha * x \text{ if } x < 0 \quad \textbf{(2.2)}$$

Dropout was added after Leaky ReLU, with dropout rates of 0.5 for the convolution layers, 0.3 and 0.1 for the fully connected layers, as shown in Figure 2.2. The purpose of dropout was to avoid overfitting and Adam was chosen as the optimizer with a learning rate of 0.1, and exponential decay rates of 0.9 and 0.99 for the first and second moment estimates, respectively. The input to this model was the normalized average and standard deviation of canopy reflectance per plot as a 1-D feature vector. The implementation and training of this model was done using Python 3.9.7, TensorFlow 2.5.0, and Keras 2.5.0 on an NVIDIA GeForce RTX 2080 Max-Q Graphics Processing Unit. The architecture of this model is shown in Figure 2.4.

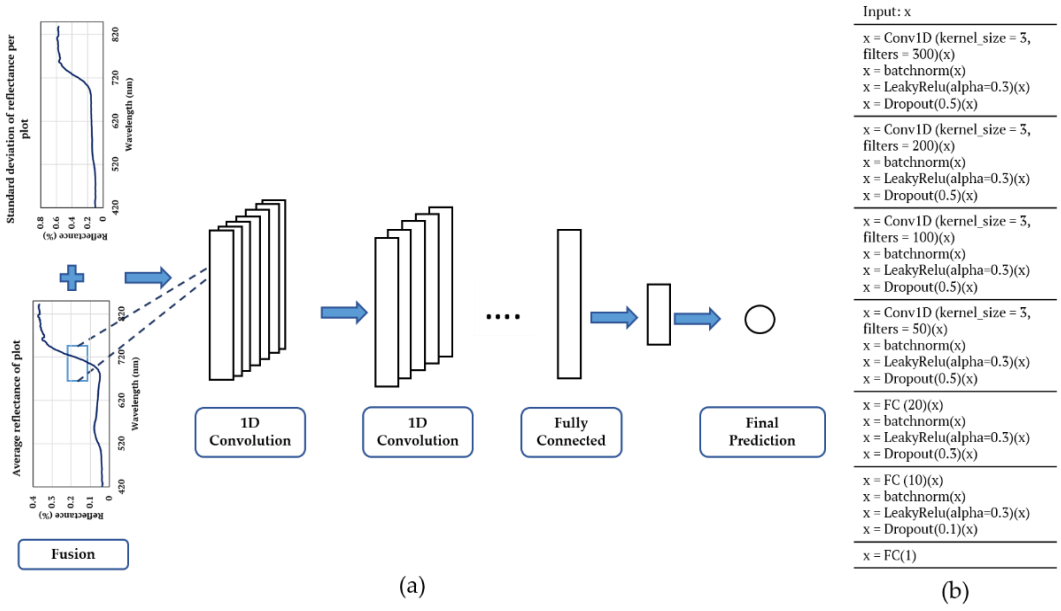

Figure 2.4 DL model architecture: (a) model schematic, (b) detailed DL model layers and their hyperparameters.

To evaluate the models, root mean square error (RMSE), coefficient of determination (R^2) and symmetric mean absolute percentage error (sMAPE) were used, and the equations of these metrics are shown in Equations (2.3)–(2.5), respectively.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad \textbf{(2.3)}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \textbf{(2.4)}$$

$$sMAPE = \frac{2}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i|} \quad \textbf{(2.5)}$$

2.4 Results

K-fold cross validation was used in both methods described in the ML models section, and the presented results are the median of the calculated metrics across all folds. The performance of the ensemble model and the DL model on testing data are shown in Tables 2.3-2.6 for both agronomic and physiological traits. Overall, both models have results close to each other, with DL having a slightly higher accuracy for prediction of agronomic traits, and the ensemble model performing marginally better on predictions of physiological traits. Averaging across the three data points, predictions of biomass had R^2 s of 0.60 and 0.51, RMSEs of 49.03 g·plant-1and 61.09 g·plant-1, and sMAPEs of 26.60% and 29% from the DL and ML model, respectively (Table 2.3, 2.4). Pod count estimations using the two models yielded the same average R^2 at 0.55, and slightly lower RMSE and sMAPE using the DL model, at 47.29 g·plant-12 and 42% compared to the ML model with RMSE and sMAPE values of 53.27 g·plant-1 and 53%. Yield predictions achieved an R^2 of 0.6, RMSE of 54.23 g·plant-1 and sMAPE of 50.50% from the DL approach. The same metrics using the ML method for yield predictions were 0.48 (R^2), 54.04 g·plant-1 (RMSE) and 38% (sMAPE) (Table 2.3, 2.4). ML and DL had close RMSE values for photosynthetic rate predictions at 8.33 µmol·m-2·s-1 and 8.54 µmol·m-2·s-1, respectively, and the same average sMAPE of 51%. ML yielded a slightly higher R^2 of 0.48,

compared to the R^2 from DL at 0.44. Stomatal conductance predictions produced average R^2s of 0.43 and 0.40, sMAPEs 63% and 79% using the DL and ML method, respectively. Both methods had an average RMSE of 0.19 mmol·m-2·s-1 (Table 2.5, 2.6).

The fold with the closest R^2 to the median R^2 of all folds was chosen for scatter plots shown in Figures 2.5-2.8. These plots present ground truth values versus predicted values for each trait. Eighteen days after drought (DAD) has the highest R^2 among most dates using DL, and the highest R^2 across traits corresponds to the predictions of biomass from data collected on this day, 18 DAD.

Table 2.3 Performance of the DL model on testing data for each agronomic trait, 14, 18 and 29 days after drought (DAD).

Metric	Biomass			Pod count			Yield		
	14 DAD	18 DAD	29 DAD	14 DAD	18 DAD	29 DAD	14 DAD	18 DAD	29 DAD
R^2	0.60	0.73	0.49	0.56	0.65	0.45	0.60	0.61	0.51
RMSE	54.18 ($\frac{g}{plant}$)	42.74 ($\frac{g}{plant}$)	50.18 ($\frac{g}{\text{plant}}$)	47.60	37.69	56.60	53.89 ($\frac{g}{\text{plant}}$)	54.57 ($\frac{g}{\text{plant}}$)	57.18 ($\frac{g}{\text{plant}}$)
sMAPE (%)	30	26	24	41	35	50	51	50	39

Table 2.4 Performance of the ensemble ML model on testing data for each agronomic trait, 14, 18 and 29.

Metric	Biomass			Pod count			Yield		
	14 DAD	18 DAD	29 DAD	14 DAD	18 DAD	29 DAD	14 DAD	18 DAD	29 DAD
R^2	0.48	0.61	0.44	0.60	0.52	0.54	0.59	0.54	0.50
RMSE	57.61 ($\frac{g}{plant}$)	64.29 ($\frac{g}{\text{plant}}$)	61.39 ($\frac{g}{\text{plant}}$)	44.06	60.11	55.65	54.11 ($\frac{g}{\text{plant}}$)	53.98 ($\frac{g}{\text{plant}}$)	58.48 ($\frac{g}{\text{plant}}$)
sMAPE (%)	32	28	27	50	59	50	40	38	54

Table 2.5 Performance of the DL model on testing data for each physiological trait, 14, 18, and 29 days after drought (DAD).

Metric	Photosynthetic rate			Stomatal Conductance		
	14 DAD	18 DAD	29 DAD	14 DAD	18 DAD	29 DAD
R^2	0.36	0.52	0.44	0.40	0.40	0.50
RMSE	9.94 $(\frac{\mu mol}{m^2.s})$	7.55 $(\frac{\mu mol}{m^2.s})$	8.13 $(\frac{\mu mol}{m^2.s})$	0.23 $(\frac{mmol}{m^2.s})$	0.16 $(\frac{mmol}{m^2.s})$	0.18 $(\frac{mmol}{m^2.s})$
sMAPE (%)	47	54	54	75	66	64

Table 2.6 Performance of the ensemble ML model on testing data for each physiological trait, 14, 18, and 29 days after drought (DAD).

Metric	Photosynthetic rate			Stomatal Conductance		
	14 DAD	18 DAD	29 DAD	14 DAD	18 DAD	29 DAD
R^2	0.41	0.56	0.48	0.35	0.52	0.57
RMSE	8.91 $(\frac{\mu mol}{m^2.s})$	7.24 $(\frac{\mu mol}{m^2.s})$	8.84 $(\frac{\mu mol}{m^2.s})$	0.21 $(\frac{mmol}{m^2.s})$	0.14 $(\frac{mmol}{m^2.s})$	0.18 $(\frac{mmol}{m^2.s})$
sMAPE (%)	37	55	61	56	70	63

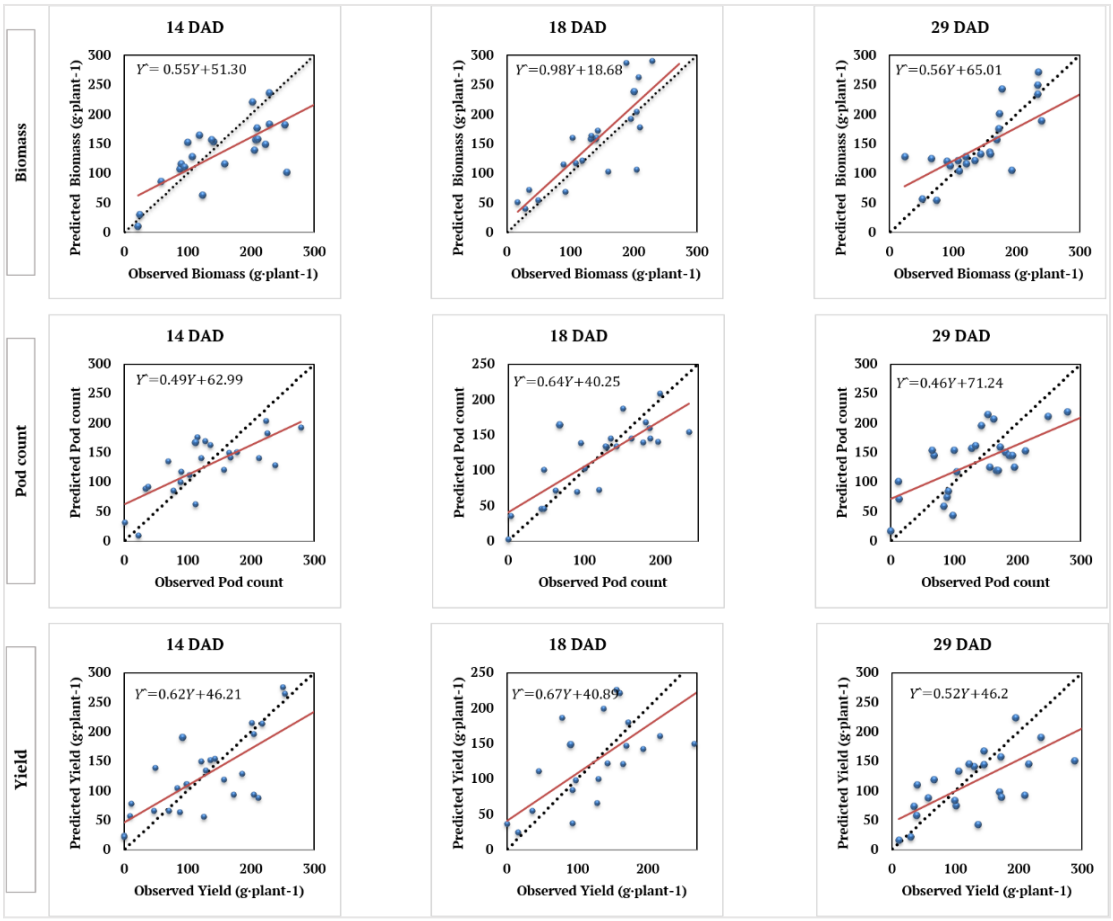

Figure 2.5 Performance of the DL model on testing data for predictions of biomass, pod count, and yield, 14, 18 and 29 days after drought (DAD). The shown data points for these scatter plots are the folds from the test dataset with the closest R^2 to the median values shown in Table 2.3.

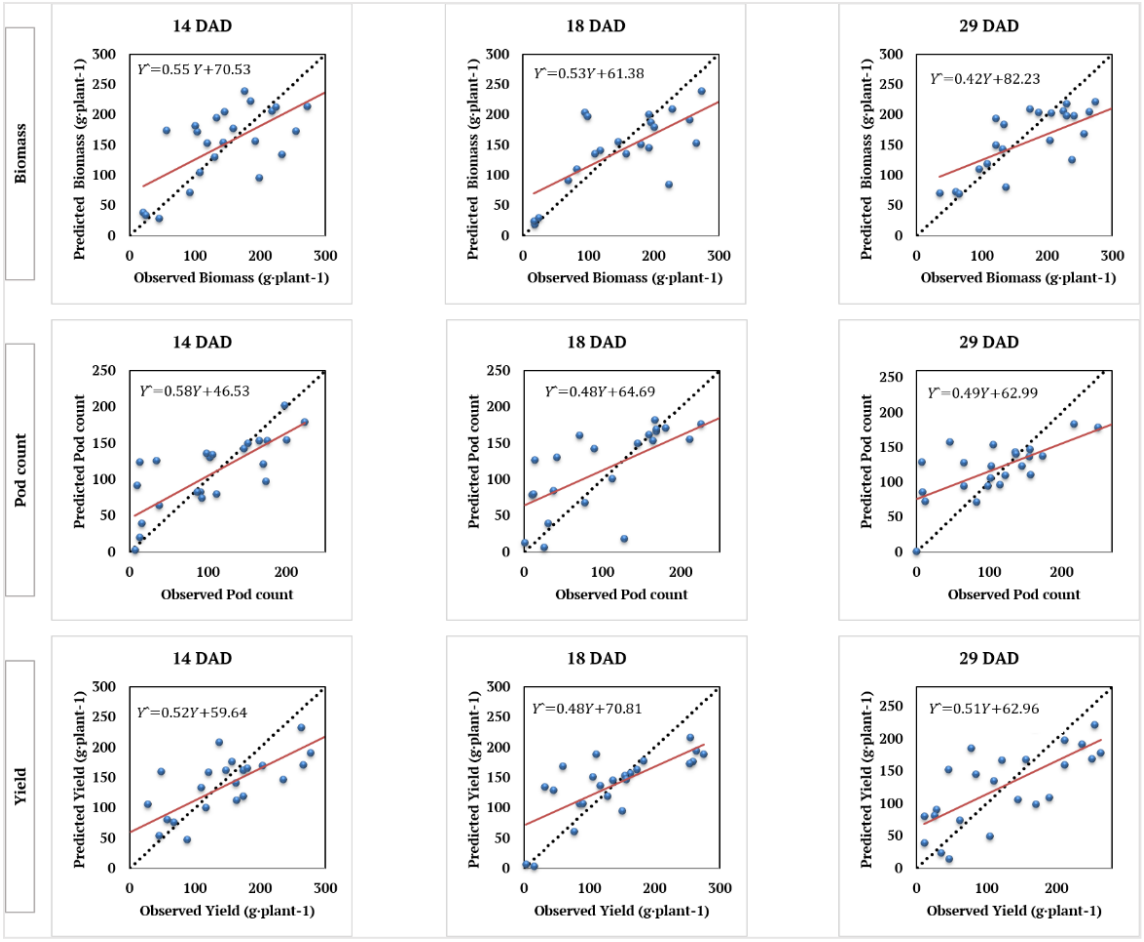

Figure 2.6 Performance of the ensemble model on testing data for predictions of biomass, pod count and yield, 14, 18 and 29 days after drought (DAD). The shown data points for these scatter plots are the fold from the test dataset with the closest R^2 to the median values shown in Table 2.4.

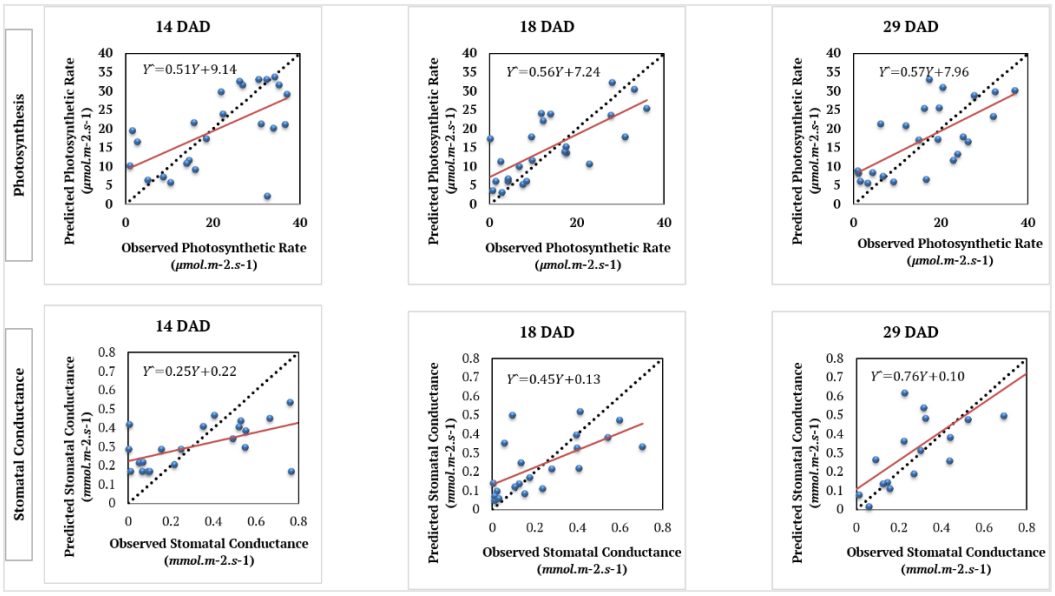

Figure 2.7 Performance of the DL model on testing data for predictions of photosynthetic rate and stomatal conductance, 14, 18 and 29 days after drought (DAD). The shown data points for these scatter plots are the fold from the test dataset with the closest R^2 to the median values shown in Table 2.5.

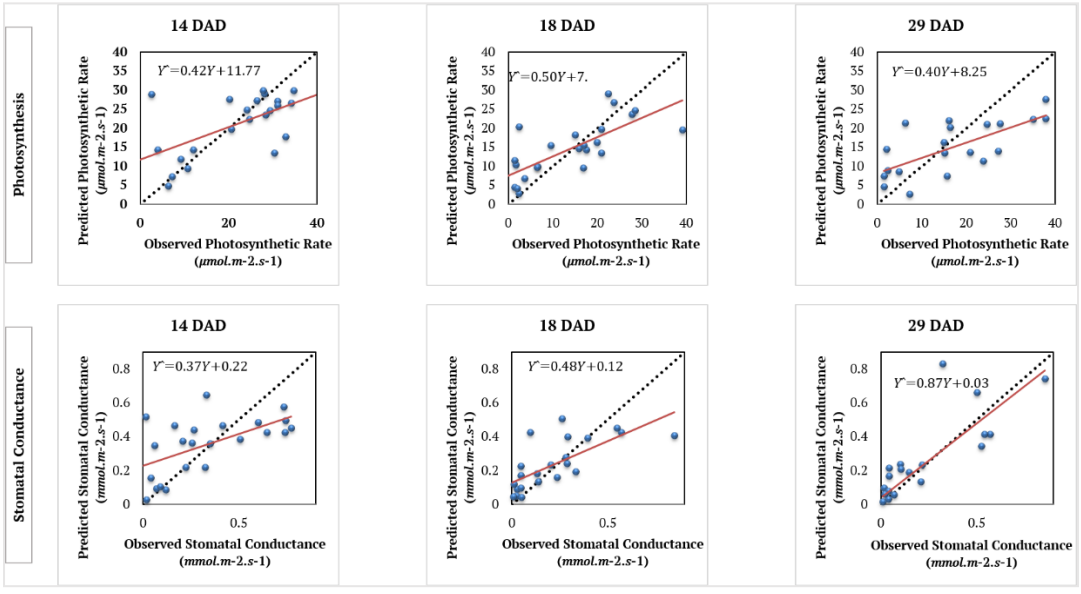

Figure 2.8 Performance of the ensemble model on testing data for predictions of photosynthetic rate and stomatal conductance, 14, 18 and 29 days after drought (DAD). The shown data points for these scatter plots are the fold from the test dataset with the closest R^2 to the median values shown in Table 2.6.

2.4.1 Feature Importance

2.4.1.1 The Ensemble Machine Learning Model

The importance of the studied eighty VIs on the trained ensemble model was evaluated using permutation importance (Altmann, Tolo¸si, Tolo¸si, Sander, and Lengauer, 2010). This algorithm provides insight into the importance of each data feature by assessing how much the model accuracy decreases when a feature is not available. This would be computationally intensive if performed during training, so instead it was performed during testing on the trained ensemble model. The model expects all features to be present during training and testing, so instead of removing each feature, they were replaced with random noise. This noise was drawn from the same distribution as the original values, by shuffling values for a feature and using other examples' feature values. The metric used in this algorithm was R^2, and the reported permutation importance score is the amount that R^2 decreased when a feature was not present. This algorithm was applied on the models used to report the results in Table 2.3-2.6, from the same fold, on the data collected 18 days after drought. This dataset (18 DAD) was chosen since it had a higher correlation with the ground truth data, and therefore the model is more capable of identifying the most significant VIs. The results of this analysis, the top 10 VIs for each model are shown in Figure 2.9. Gitelson1, SRI [710,750], and Gitelson2 were found to be the most important VIs for photosynthesis, stomatal conductance, and biomass, respectively, and Carte4 was the common most important VI for pod count and pod yield. There are several mutual top VIs across the traits; for example Gitelson2 and NVI1 were both among top 10 VIs in the photosynthesis and stomatal conductance models. Overall, Gitelson2, variations of NVI, MCARI, TCARI, and REP were among the most common top features across all models.

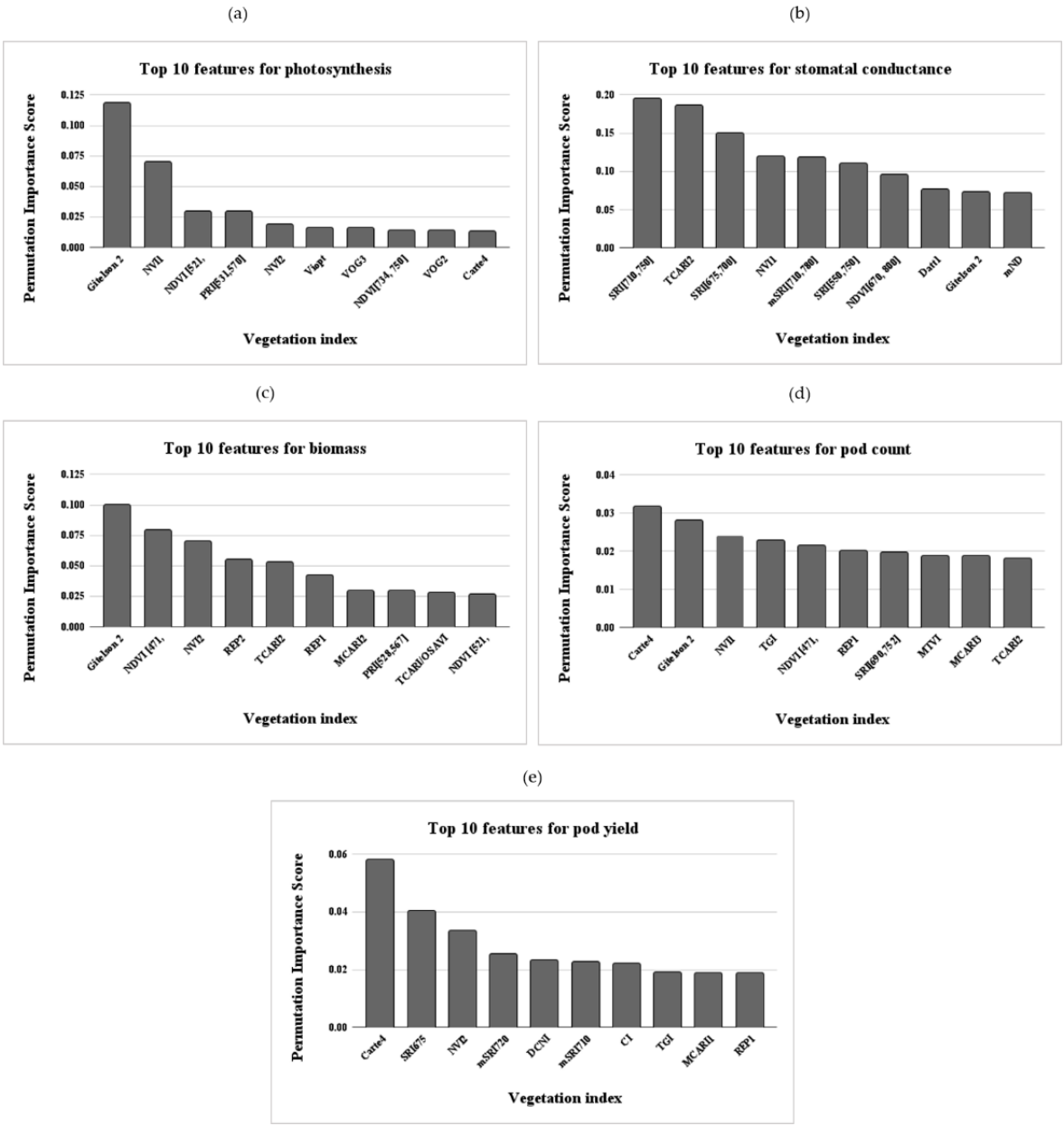

Figure 2.9 Top 10 vegetation indices used in the ensemble model trained for a) photosynthesis, b) stomatal conductance, c) biomass, d) pod count, and e) pod yield.

2.4.1.2 The Deep Learning (DL) Model

To identify the most important wavelengths in the DL model, the same approach applied for the ensemble model, permutation importance, was used. Overall, there were 200 features (wavelengths) from the average reflectance of each plot, and 200 features from the standard

deviation of reflectance per band per plot (400 features total). Since many features are adjacent to each other, the top 20 features for each model were selected, and among the adjacent bands within a ± 2 nm range, if there were any, the band with the highest permutation importance score was chosen. For example, in case of having 407, 409, 411 (nm) in top features, the one with the highest score was chosen as a representative in these charts. Therefore, the number of top features for each model is not the same. The model trained on the data from August 13th was chosen for this analysis, since the highest R^2 was achieved from training the model on this dataset. The retrieved top wavelengths are shown in Figure 2.10. There were top wavelengths for biomass in blue, green, red, red edge, and NIR, but the highest concentration is seen in the green region. Pod count and pod yield had relatively similar results, with top wavelengths in the blue and red-edge region. The top wavelengths in the photosynthesis model were mostly in green and red-edge, similar to stomatal conductance. Stomatal conductance also had some top features in the blue region. As discussed before, these features were chosen from plot-level average reflectance and the standard deviation (SD) of reflectance per plot. To explore the variation of plot-level mean and SD of reflectance across all data points, the mean and standard deviation of the reflectance profiles across 256 plots were calculated and are shown in Figure 2.11. As it can be seen in this figure, SD has a higher variation within plots.

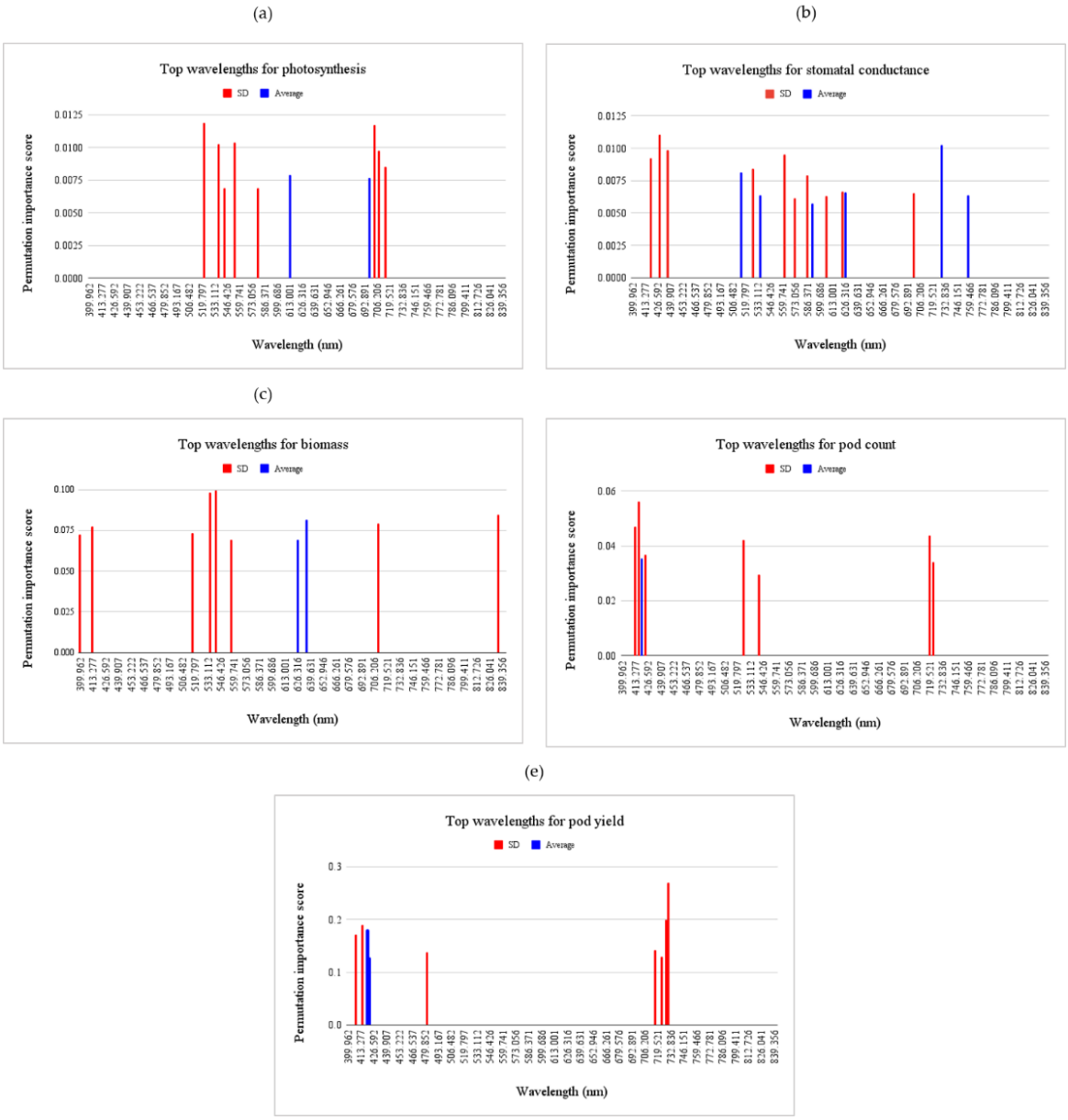

Figure 2.10 Top wavelengths used in the DL model trained for a) photosynthesis, b) stomatal conductance, c) biomass, d) pod count and e) pod yield.

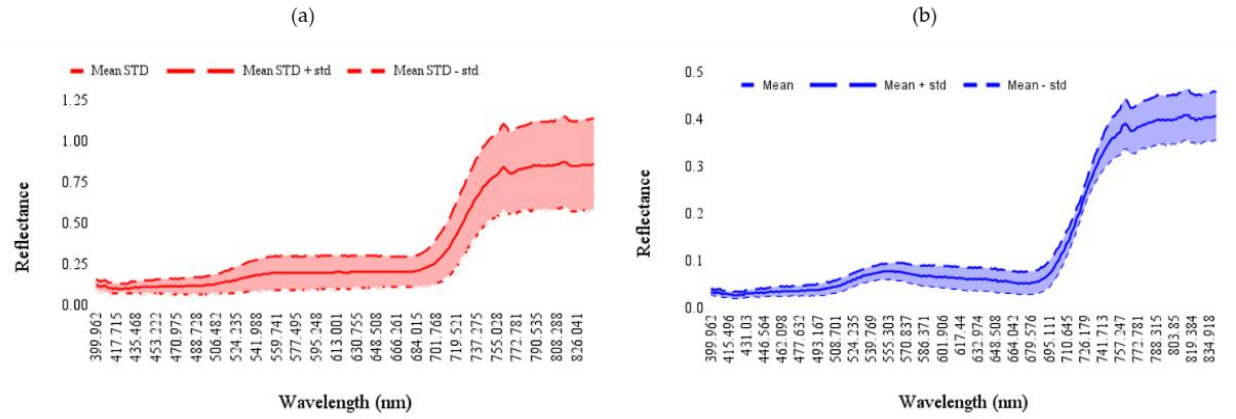

Figure 2.11 a) Mean of plot-level standard deviation across 256 plots ± the standard deviation (std) of plot-level standard deviation (STD) across 256 plots, b) Mean of plot-level average reflectance across 256 plots ± the standard deviation (std) of plot-level average reflectance across 256 plots.

2.5 Discussion

2.5.1 Ensemble ML Model vs DL Model

Both models performed well on our dataset; the highest R^2 achieved for above-ground biomass was 0.73 using the DL model (Table 2.3). Most of our results fall into the same range (0.64 to 0.89) achieved by Masjedi et al. (2020), where hyperspectral camera and LiDAR were used to estimate Sorghum biomass.

Pod count and yield predictions obtained via both methods estimated these complex traits with $R^2 > 0.5$, with the highest R^2 being 0.65 for pod count and 0.61 for pod yield. This R^2 value for pod yield is comparable to the highest R^2 achieved by Patrick et al. (2017) using NDRE ($R = 0.79$, $R^2 = 0.62$) even though this paper provides yield estimations for whole plots, whereas in this work the predictions correspond to individual plants. Averaging over a plot helps model estimations as it helps attenuate the signal to noise ratio. Using NDRE and correlating it with single-plant pod yield from this study, the R^2 dropped to 0.32, 0.44, and 0.34 for 14, 18, and 24 DAD respectively. This decrease in accuracy signifies the importance of using ML/DL models in predictions of complex traits.

The attained R^2 for yield in our study is also greater than the one reached by Balota and Oakes (2016) where the highest R^2 using color and RGB-derived indices were 0.39 and 0.26 respectively. Using a different approach, Bidese et al (2021) estimated the peanut yield of breeding lines with an R^2 of 0.59 by directly imaging infield pods with three RGB cameras after digging and before harvest and then using statistical models to predict yield based on image-derived pod counts. The R^2 in this study is slightly higher than the one obtained in the mentioned article, despite not directly observing the peanut pods.

Several studies have used hyperspectral imagery for yield predictions of other crops and shown the remote sensing method to be highly effective. An R^2 of 0.87 was obtained using an ensemble machine learning model and hyperspectral imaging for alfalfa yield predictions (L. Feng et al., 2020). Moghimi et al. (2020) used a deep neural network to predict yield in wheat using the same type of sensor and achieved R^2 in the range of 0.64-0.81. These studies have higher R^2 compared to those achieved in this study, which can be due to the nature of the crops. Peanuts are below-ground nuts and it is more challenging to predict their yield values, whereas yield properties of alfalfa and wheat are above-ground and can be directly seen from the sensor.

In this study, the top R^2 for estimations of photosynthetic rate and stomatal conductance were 0.56 and 0.57 respectively, using the ML model. The work by Buchaillot et al. (2022) showed a higher R^2 (0.62) for estimations of photosynthetic rate, using a handheld spectrometer and advanced regression models. This higher R^2 can be due to twofold. First, the same leaves were measured for physiological traits and reflectance in that study, whereas in this paper, the reflectance of the canopy of each plant was measured. Secondly, the plants grown in the mentioned paper are grown in a controlled environment, and the trained model might not perform well in the field due to variations in environmental conditions.

El-Hendawy et al. (2019) also used a portable spectroradiometer with a spectrum range of m 350 to 2500 nm to predict photosynthesis, transpiration and stomatal conductance properties of wheat. Using PLSR models, results showed moderate to high R^2 for predictions of photosynthesis (0.58 – 0.98) and stomatal conductance (0.44 – 0.92). In that paper, the measurements are also the reflectance of leaves (same as the work by Buchailot et al. (2022)), whereas in this paper the reflectance of whole plants were measured and averaged. This can be a possible reason behind greater R^2 achieved in that study. Another reason for better results could be the wider spectrum of the spectroradiometer, as the importance of SWIR bands was shown in the same study.

On average the DL model had a better performance for predicting biomass, pod yield, and pod count (the agronomic traits) and the ensemble ML model had a superior performance in predicting the physiological traits, photosynthesis rate and stomatal conductance. The reason for better performance of the two models for different traits could be in the choice and existence of relevant vegetation indices, and their capability of explaining the studied traits (Abdu, Mokji, and Sheikh, 2020). It is possible that the VIs included in this model, were better indicators of the physiological traits and not able to fully explain the variability in the agronomic traits. Therefore, the physiological phenotyping models had less input features and less trainable parameters, and therefore the training gave better results. Assuming the studied vegetation indices were not great indicators of pod count, biomass, and pod yield, the DL model was a more adaptive solution as the 1-D CNN could capture the detailed shape of the canopy reflectance of each plot, and learn which wavelengths are more important during training, unlike the ensemble model where the wavelengths chosen for the VIs were predetermined. Moreover, the standard deviation (SD) of canopy reflectance within each plot provided the model with information regarding the

distribution of pixels' reflectance per band per plot. When an additional DL model was trained without the SD in the input, R^2 decreased by 9% on average.

2.5.2 Interpretation of the Most Important Features

As shown in Figure 2.7, Gitelson2, variations of NVI, MCARI, TCARI and REP were among the most common top features across all the models. Most of these vegetation indices include the spectral response in a wavelength in each of the NIR, red-edge, green, and red regions, which was expected. Gitelson2 includes wavelengths from the NIR, red, and red-edge regions, and NVI includes wavelengths in the blue, red-edge and NIR region. MCARI and TCARI look at reflectance in the green and NIR region, which normally are used to estimate chlorophyll absorption. REP also includes wavelengths from red and NIR regions. These results confirmed the importance of reflectance in the mentioned regions of the electromagnetic spectrum for rapid plant phenotyping.

Despite the black-box nature of DL models, top features in the DL models were also found using permutation importance. Results from both analysis, show that green and red-edge (RE) are the most important regions of the spectrum for predicting biomass. The top VIs from the ML model including Gitelson2, NDVI [471,584], NVI2 and REP2 contain wavelengths from these ranges too, confirming the importance of these wavelengths. The range of 410-430 nm (blue) and 710-740 nm (RE) held the most important wavelengths for prediction of pod count and pod yield. The top VIs for these models also include wavelengths from the RE region (such as Carte4, SRI [675,700], NVI1 and NVI2) but there are not any indices including blue wavelengths. The green and RE ranges were shown to contain the most significant wavelengths for the prediction of photosynthesis and stomatal conductance. Most top VIs for these models also include RE in their formulas, such as NVI, Gitelson, NDVI [734,750], and VOG.

Considering all top wavelengths assessed in Section 2.3.1., plot-based SD delivered most of the important features. This could be due to the fact that SD has a higher variation between all plots, according to Figure 2.11.

2.5.3 Effect of Drought on Peanut Canopy Spectral Response

Hyperspectral data collected 18 DAD resulted in the best overall prediction accuracy. This could be due to the fact that this date was when the effect of drought was most severe. Since there were drought tolerant varieties among the peanut genotypes, some experienced wilting and recovered by August 24 (i.e., 29 DAD), therefore the effect of drought is not seen thoroughly in the plants' spectral response. Below are examples of the spectral responses of a drought tolerant genotype (Line-8) and a drought sensitive genotype (AP-3), 14, 18, and 29 DAD. As can be seen in Figure 2.12, the drought sensitive variety's reflectance in the NIR region decreased 18 DAD and stayed about the same until 29 DAD. However, the drought tolerant variety's reflectance in NIR was lowest 18 DAD, and surged after about 11 days. Since high reflectance in NIR is an indication of high plant vigor, the rise of spectral response in this region suggests the recovery of the drought tolerant genotype. The temporal changes of VNIR spectral response of peanut canopy may assist peanut breeders in quantifying recoverability from water stress in peanut.

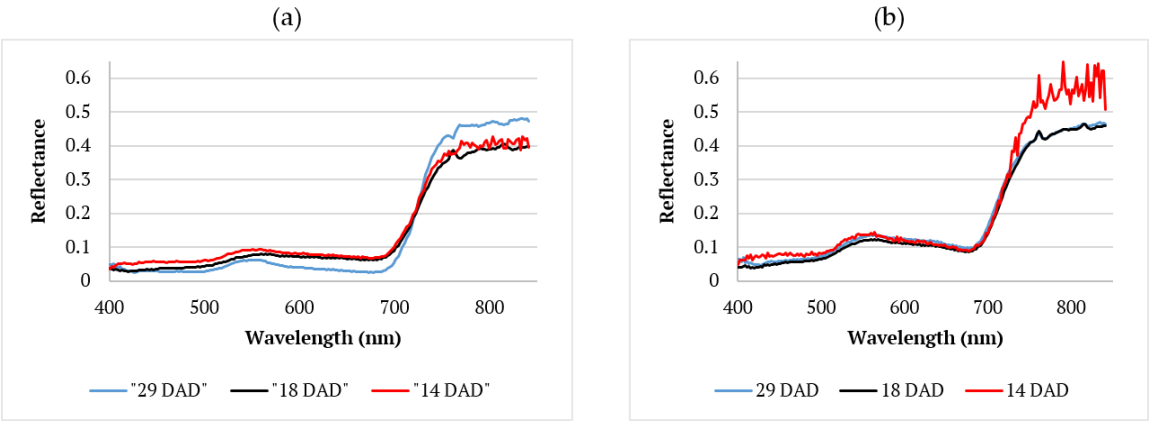

Figure 2.12 The spectral responses of a drought tolerant genotype, Line-8 (a), and a drought sensitive genotype, AP-3 (b).

2.6 Conclusions

In this study, the feasibility of using UAV-based hyperspectral imaging and ML for prediction of biomass, pod count, pod yield, photosynthesis rate, and stomatal conductance in peanut was evaluated. Two common approaches in this domain were compared: ML and feature engineering versus DL and feature learning. Both methods showed promising results; the DL model outperformed the ensemble ML model in predicting the agronomic traits and the ensemble ML model had a better performance in estimating the physiological traits. Moreover, data collected on 14, 18, and 29 days after the start of drought were tested on both models, and 18 days after drought was found to provide the most valuable information to achieve the highest accuracy. Additionally, the most important input features of both the ML and DL model were investigated, and the most effective detection wavelengths were in the visible, near infrared and red-edge region. This paper demonstrated the ability of both DL and ML models to extract valuable information from hyperspectral imagery for phenotyping the agronomic traits in peanuts 30 to 45 days before harvest, and estimate the physiological traits for same-day measurements. For future work, we will explore Recurrent Neural Networks (RNNs) such as long short-term memory (LSTM) and combine data from multiple dates to capture temporal features and

improve prediction accuracy. Another possible future direction can be training the deep learning model with a reduced number of features (the top wavelengths) and comparing the results.

2.7 References

Abdu, A. M., Mokji, M. M. M., & Sheikh, U. U. U. (2020). Machine learning for plant disease detection: An investigative comparison between support vector machine and deep learning. IAES International Journal of Artificial Intelligence (IJ-AI), 9(4), 670.

Altmann, A., Tolo¸si, L., Tolo¸si, T., Sander, O., & Lengauer, T. (2010). Data and text mining Permutation importance: A corrected feature importance measure. Bioinformatics, 26(10), 1340–1347.

Araus, J. L., & Cairns, J. E. (2014). Field high-throughput phenotyping: The new crop breeding frontier. Trends in Plant Science, 19(1), 52–61.

Araus, J. L., Kefauver, S. C., Zaman-Allah, M., Olsen, M. S., & Cairns, J. E. (2018). Translating high-throughput phenotyping into genetic gain. Trends in Plant Science, 23(5), 451–466.

Balota, M., & Oakes, J. (2016, May 17). Exploratory use of a UAV platform for variety selection in peanut (J. Valasek & J. A. Thomasson, Eds.).

Balota, M., & Oakes, J. (2017). UAV remote sensing for phenotyping drought tolerance in peanuts. In J. A. Thomasson, M. McKee, & R. J. Moorhead (Eds.), Proceedings of the SPIE (p.

Baslam, M., Mitsui, T., Hodges, M., Priesack, E., Herritt, M. T., Aranjuelo, I., & Sanz-Sáez, Á. (2020). Photosynthesis in a changing global climate: scaling up and scaling down in crops. Frontiers in Plant Science, 11.

Bendig, J., Bolten, A., Bennertz, S., Broscheit, J., Eichfuss, S., & Bareth, G. (2014). Estimating biomass of barley using Crop Surface Models (CSMs) derived from UAV-based RGB imaging. Remote Sensing, 6(11), 10395–10412.

Bidese Puhl, R., Bao, Y., Sanz-Saez, A., & Chen, C. (2021). Infield peanut pod counting using deep neural networks for yield estimation. 2021 ASABE Annual International Virtual Meeting, July 12-16, 2021. American Society of Agricultural and Biological Engineers.

Blankenship, P. D., Mitchell, B. W., Layton, R. C., Cole, R. J., & Sanders, T. H. (1989). A low-cost microcomputer system to monitor and control an environmental control plot facility. Computers and Electronics in Agriculture, 4(2), 149–155.

Buchaillot, Ma. L., Soba, D., Shu, T., Liu, J., Aranjuelo, I., Araus, J. L., Sanz-Saez, A. (2022). Estimating peanut and soybean photosynthetic traits using leaf spectral reflectance and advance regression models. Planta, 255(4), 93.

Buezo, J., Sanz-Saez, Á., Moran, J. F., Soba, D., Aranjuelo, I., & Esteban, R. (2019). Drought tolerance response of high-yielding soybean varieties to mild drought: Physiological and photochemical adjustments. Physiologia Plantarum, 166(1), 88–104.

Choudhary, S. S., Biswal, S., Saha, R., & Chatterjee, C. (2021). A non-destructive approach for assessment of nitrogen status of wheat crop using unmanned aerial vehicle equipped with RGB camera. Arabian Journal of Geosciences, 14(17), 1739.

El-Hendawy, S., Al-Suhaibani, N., Alotaibi, M., Hassan, W., Elsayed, S., Tahir, M. U., Schmidhalter, U. (2019). Estimating growth and photosynthetic properties of wheat grown in simulated saline field conditions using hyperspectral reflectance sensing and multivariate analysis. Scientific Reports, 9(1), 16473.

Eugenio, F. C., Grohs, M., Venancio, L. P., Schuh, M., Bottega, E. L., Ruoso, R., Fernandes, P. (2020). Estimation of soybean yield from machine learning techniques and multispectral RPAS imagery. Remote Sensing Applications: Society and Environment, 20, 100397.

Feng, L., Zhang, Z., Ma, Y., Du, Q., Williams, P., Drewry, J., & Luck, B. (2020). Alfalfa yield prediction using UAV-based hyperspectral imagery and ensemble learning. Remote Sensing, 12(12), 2028. https://doi.org/10.3390/rs12122028
Feng, X., Zhan, Y., Wang, Q., Yang, X., Yu, C., Wang, H., He, Y. (2020). Hyperspectral imaging combined with machine learning as a tool to obtain high-throughput plant salt-stress phenotyping. The Plant Journal, 101(6), 1448–1461.

Fenghua, Y., Tongyu, X., Wen, D., Hang, M., Guosheng, Z., & Chunling, C. (2017). Radiative transfer models (RTMs) for field phenotyping inversion of rice based on UAV hyperspectral remote sensing. International Journal of Agricultural and Biological Engineering, 10(4), 150–157.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585(7825), 357–362.

Kanning, M., Kühling, I., Trautz, D., & Jarmer, T. (2018). High-Resolution UAV-Based Hyperspectral Imagery for LAI and Chlorophyll Estimations from Wheat for Yield Prediction. Remote Sensing, 10(12), 2000.

Liang, L., Di, L., Zhang, L., Deng, M., Qin, Z., Zhao, S., & Lin, H. (2015). Estimation of crop LAI using hyperspectral vegetation indices and a hybrid inversion method. Remote Sensing of Environment, 165, 123–134.

Maimaitijiang, M., Ghulam, A., Sidike, P., Hartling, S., Maimaitiyiming, M., Peterson, K., Fritschi, F. (2017). Unmanned Aerial System (UAS)-based phenotyping of soybean using multi-sensor data fusion and extreme learning machine. ISPRS Journal of Photogrammetry and Remote Sensing, 134, 43–58.

Maresma, Á., Ariza, M., Martínez, E., Lloveras, J., & Martínez-Casasnovas, J. (2016). Analysis of vegetation indices to determine nitrogen application and yield prediction in maize (Zea mays L.) from a Standard UAV Service. Remote Sensing, 8(12), 973.

Masjedi, A., Crawford, M. M., Carpenter, N. R., & Tuinstra, M. R. (2020). Multi-temporal predictive modelling of sorghum biomass using UAV-based hyperspectral and LiDAR data. Remote Sensing, 12(21), 3587.

Moghimi, A., Yang, C., & Anderson, J. A. (2020). Aerial hyperspectral imagery and deep neural networks for high-throughput yield phenotyping in wheat. Computers and Electronics in Agriculture, 172, 105299.

NASA: Climate Change and Global Warming. (2021, September 27).

Patrick, A., Pelham, S., Culbreath, A., Holbrook, C. C., De Godoy, I. J., & Li, C. (2017). High throughput phenotyping of tomato spot wilt disease in peanuts using unmanned aerial systems and multispectral imaging. IEEE Instrumentation & Measurement Magazine, 20(3), 4–12.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Brucher, M. (2011). Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research, 12(85), 2825–2830.

Qi, H., Wu, Z., Zhang, L., Li, J., Zhou, J., Jun, Z., & Zhu, B. (2021). Monitoring of peanut leaves chlorophyll content based on drone-based multispectral image feature extraction. Computers and Electronics in Agriculture, 187, 106292.

Qi, H., Zhu, B., Kong, L., Yang, W., Zou, J., Lan, Y., & Zhang, L. (2020). Hyperspectral inversion model of chlorophyll content in peanut leaves. Applied Sciences 2020, Vol. 10, Page 2259, 10(7), 2259.

Rehman, T. U., Ma, D., Wang, L., Zhang, L., & Jin, J. (2020). Predictive spectral analysis using an end-to-end deep model from hyperspectral images for high-throughput plant phenotyping. Computers and Electronics in Agriculture, 177, 105713.

Romero, M., Luo, Y., Su, B., & Fuentes, S. (2018). Vineyard water status estimation using multispectral imagery from an UAV platform and machine learning algorithms for irrigation scheduling management. Computers and Electronics in Agriculture, 147, 109–117.

Sankaran, S., Marzougui, A., Hurst, J. P., Zhang, C., Schnable, J. C., & Shi, Y. (2021). Can high-resolution satellite multispectral imagery be used to phenotype canopy traits and yield potential in field conditions? Transactions of the ASABE, 64(3), 879–891.
Su, W., Zhang, M., Bian, D., Liu, Z., Huang, J., Wang, W., Guo, H. (2019). Phenotyping of corn plants using Unmanned Aerial Vehicle (UAV) images. Remote Sensing, 11(17), 2021.

United States Department of Agriculture, National Agricultural Statistics Service. (2020). Retrieved April 13, 2022, from https://www.nass.usda.gov/

Wang, J., Wu, B., Kohnen, M. V., Lin, D., Yang, C., Wang, X., Gu, L. (2021). Classification of rice yield using UAV-based hyperspectral imagery and lodging feature. Plant Phenomics, 2021.

Zhang, Q., Chen, C., Feng, Y., Batchelor, W., Dang, P., Lamb M, & Sanz-Saez, A. (2022). Tolerance to mid-season drought in peanut can be achieved by high water use efficiency or high efficient use of water. Crop Science.

Chapter 3. Detecting harmful algal blooms in Lake Okeechobee using MODIS satellite imagery and long-short term memory (LSTM)

3.1 Abstract

Harmful algal blooms (HABs) in inland water bodies are a global concern due to their negative impact on human and animal health. It is possible to detect HABs by monitoring chlorophyll-a (chl-a) concentration as an indicator of these events. Such monitoring requires extensive observations in terms of space and time, which can be achieved via remotely-sensed satellite images with high temporal resolution. To accomplish this, MODIS images from 2011 to 2020 were used to extract 10 years of time-series reflectance data. The dataset was expanded by adding several environmental features and derived products from the MODIS images, and a long-short term memory (LSTM) model was employed to examine and learn the complex data. In the structured dataset, each chl-a measurement was associated with reflectance data for the same day, as well as reflectance data for several days preceding the measurement. Different temporal windows were evaluated in this study to investigate the effect of window size on chl-a estimations. This method was evaluated on Lake Okeechobee in Florida and the results showed that fifteen days before HAB events with a temporal resolution of four days gives the highest prediction accuracy, with a root mean square error (RSME) of 11.95 µg/L, mean absolute error (MAE) of 8.55 µg/L and coefficient of determination (R^2) of 0.43. A recurrent neural network, such as an LSTM, together with satellite imagery was proven effective in capturing the temporal features of the spectral reflectance and environmental attributes preceding HAB events, to estimate the concentrations of chlorophyll-a and detect HABs in Lake Okeechobee.

3.2 Introduction

An algal bloom is a phenomenon in which the population of phytoplankton (algae) increases rapidly in a water body, such as a river, lake, or sea. Harmful algal bloom (HAB) is a type of algal bloom with the potential to harm human health or aquatic ecosystems. These HABs can be produced by microorganisms called cyanobacteria, also known as blue-green algae. Some cyanobacterial HABs (cHABS) can produce toxins, which pose threats to people, animals, aquatic ecosystems, the economy, drinking water supplies, and recreational activities. The death of marine organisms, human health risks due to the consumption of contaminated seafood and water, and the decline in watersports and tourism are examples of these threats. In the US alone, an annual economic loss of at least $82 million is estimated as a result of HABs (Hoagland and Scatasta, 2006). For all these reasons, cHABs are a global concern and need to be monitored.

Lake Okeechobee, the case study in this paper, is the second largest freshwater lake in the United States, with a surface area of about 1890 km^2 and despite its remarkable size, it's very shallow with an average depth of only 9 feet (SFWMD, 2022). This lake is a key source of water supply and is home to fish, wading birds, and other wildlife. The surrounding watersheds around Lake Okeechobee result in extensive amounts of nutrients from agricultural and urban activities, and the occurrence of HABs as a result of these nutrients can pose a threat to human and animal health.

Traditionally, water samples are collected for lab-based cell taxonomy in order to measure algae concentrations and evaluate HAB events. These manual measurements are labor-intensive, and extremely time-consuming, which makes this type of measurement limited spatially and temporally (Craig et al., 2006). In contrast, remote sensing methods, which have been used in the past decades, allow a much higher coverage of the regions of interest in less

time. Remote sensing-based HAB detection methods use Chlorophyll-a as an indicator of HABs; Landsat is one of the most commonly used satellite products for monitoring in-land algal blooms (Khan et al., 2021). However, it is limited by its 16-day temporal resolution. These long revisit intervals limit the utility of Landsat for mapping algal blooms' temporal variability. Sentinel-2 is another satellite with a high spatial resolution which is used for monitoring freshwater regions. This satellite was launched in 2015 and therefore developing a model with a limited time range can be limiting, especially since inland water bodies are not sampled as frequently as marine waters and have fewer field observations. Choosing the right type of satellite involves a tradeoff between the range of availability, temporal and spatial resolution. Moderate resolution imaging spectroradiometer (MODIS) is one of the satellites offering an archive of long-term image series of daily global coverage. The high temporal resolution of this satellite increases the probability of getting cloud-free images in the areas of interest, and its long observation record (since 1999) allows a deeper analysis of temporal dynamic blooms in inland waters.

Previous studies have shown the capability of MODIS products for estimating chlorophyll-a in large inland water bodies. Ventura et al. (2022) explored the potential of using MODIS imagery to estimate chl-a concentrations of lakes in different sizes by studying thirteen lakes in Brazil, with water surface areas ranging from 1.85 to 441 km^2. The results showed that the three biggest lakes with the highest frequency of field sampling showed the best results, with $R^2 > 0.5$. Zhang et al. (2011) used the reflectance from MODIS band 2 (near infrared) and an empirical model to make predictions on chl-a in Lake Taihu. Another study by Li et al. (2019) explored chl-a predictions in Lake Taihu using a classification-based MODIS land-band algorithm. A study on Lake Okeechobee demonstrated the potential of using MODIS imagery for estimating chl-a, using three different models; a genetic programming (GP) model, an artificial

neural network (ANN) model, and a multiple linear regression (MLR) model (Chang et al., 2011).

Common methods of estimating chl-a concentrations and detection of HABs are based on water-leaving reflectance and regression models (Liu, Ling, Wu, Su, and Cao, 2021; Ventura et al., 2022; Xu, Pu, Zhu, Luan, and Shi, 2021). However, a common problem with these models is that they cannot be used in other locations and often have to be calibrated in order to ensure cross-sensor and temporal consistency (Xu et al., 2021). Moreover, it is not possible to trace the pattern of algal bloom growth using regression models.

MODIS's high temporal resolution allows studying long and short term dependencies in temporal information. Therefore, choosing a model capable of effectively capturing time dependencies can be beneficial. Therefore, LSTM (Long Short-Term Memory) (Hochreiter and Schmidhuber, 1997) was selected as the model for this study. LSTM is a type of Recurrent Neural Network (RNN) capable of characterization of time-varying signals. Several studies have shown the capability of LSTMs for chl-a estimations/predictions in marine waters but there is limited work on using this powerful model for inland water bodies. Yussof et al. (2021) used LSTM on MODIS and GEBCO images to predict chl-a concentrations in the west coast of Sabah. In this study, the convolution neural network (CNN) and LSTM were employed and the results revealed that the LSTM model outperformed the CNN model in terms of accuracy (R^2 and root mean square error (RMSE). HABnet also showed that an LSTM-based network achieved the highest accuracy in predicting HABs in a classification problem (Hill, Kumar, Temimi, and Bull, 2020).

A possible explanation for the limited use of LSTM or more generally, machine learning models, on lakes might be that HABs are more prevalent in coastal regions, so there are more

ground-truth measurements of chl-a and therefore more resources are available to study and monitor them. The higher frequency of ground-truth measurements allows for a larger training dataset, which is essential for deep learning and machine learning models. In this chapter, the first main goal is to address this gap by making use of feature engineering, and simplifying the input data to the LSTM model so that the model trains more easily and therefore can be used for smaller lakes with fewer field measurements. The dataset was further enhanced with the addition of cloud cover, chl-a estimations based on the OCx algorithm, temperature data, and the sine transform of timestamps. LSTM was trained on the time-series data and its performance was compared to three traditional machine learning models; K-Nearest Neighbors (KNN), Support Vector Machine (SVM) and Random Forest (RF). These models were tested on single and time-series inputs and the effect of adding temporal features was evaluated. Furthermore, twelve window sizes before event days, which were the day chl-a values were measured, were assessed to investigate the number of days of data that are needed to make the most accurate chl-a estimations.

3.3 Materials and Methods

3.3.1 Study site

The areas of study at Lake Okeechobee were six stations across the lake with an average of 105 data points from each station. These stations are shown in Figure 3.1 and their coordinates, min, max, and average chl-a concentration between 2011 and 2020 are shown in Table 3.1. All stations combined for the same time period, the chl-a concentration has an average chl-a concentration of 20.56 µg/L. With a threshold of 10 µg/L for categorizing HAB/No HAB events, there were 357 HAB events and 191 No HAB events. The distribution of chl-a concentrations is shown in a histogram in Figure 3.2.

Table 3.1 Stations in Lake Okeechobee and their characteristics.

Station	Data points	Coordinates (Latitude, Longitude)	Min, Max Chl-a (µg/L)	Average chl-a (µg/L)
CLV10A	112	26.916078, -80.624663	0.0, 60.0	12.62
KISSR0.0	106	27.141301, -80.846	0.025, 49.6	14.42
L005	105	26.95673, -80.972385	2.61, 142.0	30.30
LZ2	104	27.189756 -80.82804	1.51, 117.0	20.75
LZ30	102	26.796971, -80.860095	1, 278.0	15.24
POLESOUT	104	27.038198, -80.918541	5.7, 110.0	30.54

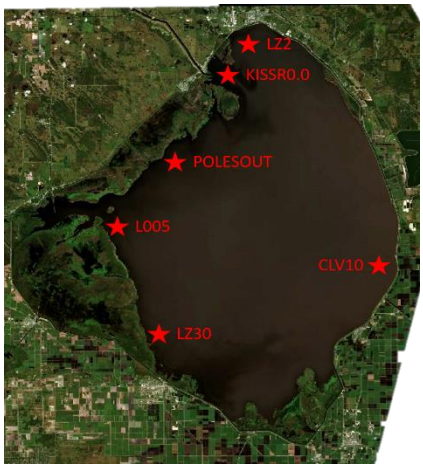

Figure 3.1 Study stations in Lake Okeechobee.

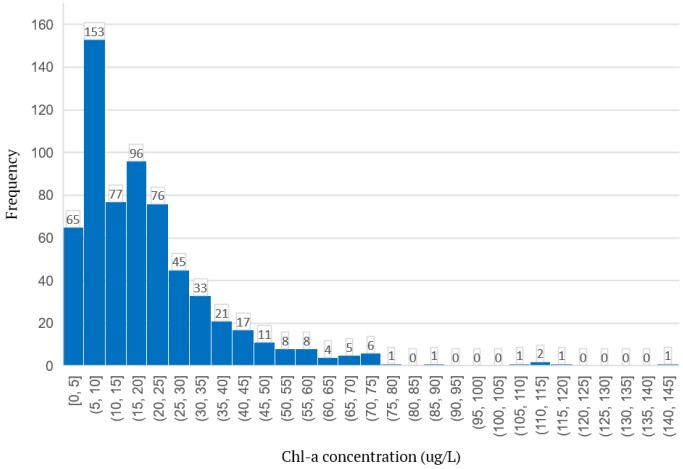

Figure 3.2 Histogram of chl-a concentrations across all stations between 2011 and 2020.

3.3.2 MODIS Images

The satellite images were acquired from version 6 of MODIS products, the MCD43A4 Nadir Bidirectional Reflectance Distribution Function (BRDF)-Adjusted Reflectance (NBAR) dataset. This dataset is produced daily using both Terra and Aqua MODIS with a resolution of 500 meters, and all images are atmospherically corrected and consistent in seven bands that are shown in Table 3.2.

Table 3.2 Corresponding wavelengths of MODIS bands 1–7.

Name	Wavelength
Band 1	620-670nm
Band 2	841-876nm
Band 3	459-479nm
Band 4	545-565nm
Band 5	1230-1250nm
Band 6	1628-1652nm
Band 7	2105-2155nm

The MCD43A4 dataset is publicly available on Google Earth Engine (GEE) for noncommercial purposes (Gorelick et al., 2017). A pipeline was developed in Python 3.9.0 to automate the image acquisition workflow using GEE. In the first step of the workflow, requests were sent using GEE's Python API, each image was clipped to the boundary of the lake and the correct scale (0.0001) was applied to the bands and they were saved to the hard drive. These images had a time range of 2011 through 2020, which matched the time period of ground truth measurements.

Saving the images and extracting the pixel values in a second step allowed more flexibility and repeatability for evaluating different approaches. Therefore, once a complete dataset was ready,

reflectance values were extracted from 6 by 6 windows surrounding each station. The maximum value from each band was selected as suggested by Yussof et al. (2021), since we're interested in extreme values and detecting HABs. This workflow is shown in Figure 3.3.

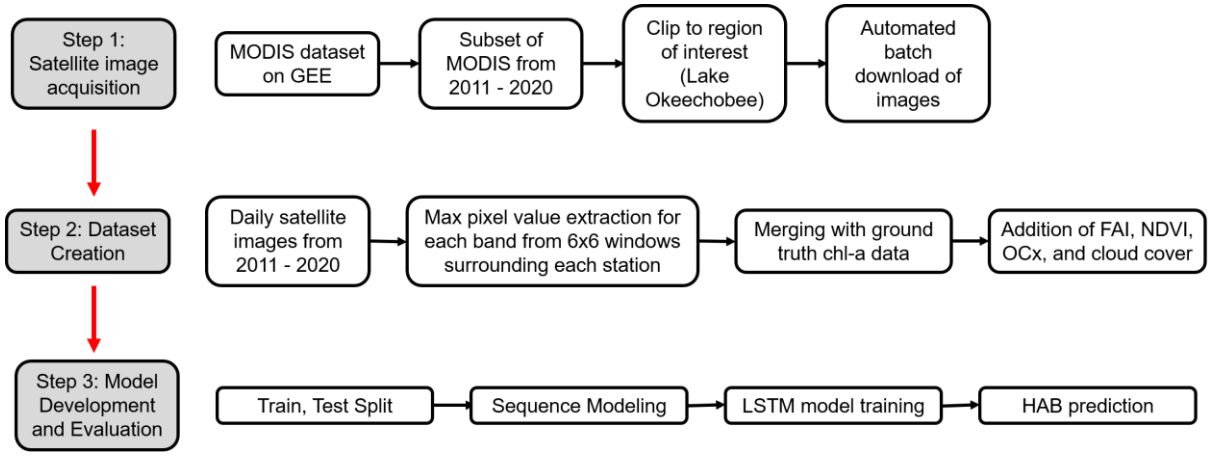

Figure 3.3 The overall MODIS image acquisition and processing workflow.

3.3.3 Additional features

Several additional features were calculated using the seven bands of MODIS images, and they were added to the features for training. The extracted features were mostly derived from the satellite images so that this methodology can also be applied to other lakes in future studies, and not rely on field measurements or data that might not be accessible for every lake. The first added feature was chl-a estimation using the OCx algorithm, which is a fourth-order polynomial equation between chl-a and a ratio of reflectance values of bands green and blue. This method was introduced by Hu, Lee, and Franz (2012), and it was recommended for chl-a retrievals above 0.2 µg/L. Since 99.8% of the chl-a data in the dataset used in this study is above 0.2 µg/L, OCx was selected as the preferred method. This algorithm is shown in Equations (3.1) and (3.2).

$$log_{10}(chl-a) = a_0 + \sum_{i=1}^{4} a_i \left(log_{10}\left(\frac{R_{rs}(\lambda_{blue})}{R_{rs}(\lambda_{green})}\right)\right)^i \quad (3.1)$$

$$a = [a_0, a_1, a_2, a_3, a_4] \quad (3.2)$$

Where λ_{blue} and λ_{green} are the instrument-specific wavelengths closest to 443, and 555 nm respectively. These wavelengths correspond to bands 3 and 4 in MODIS images, and R_{rs} is the reflectance value from these bands. Values a0 to a4 are 0.1464, -1.7953, 0.9718, -0.8319 and -0.8073 respectively.

The second added feature was the cloud cover derived from the satellite images. Light is one of the factors affecting the growth of algae and resulting algal blooms, and to account for the amount of light on each day, cloud cover was derived from the quality band provided by MODIS. Pixel values in the quality bands are either 0 or 1, where 0 means the pixel has good quality and is cloud-free and 1 means covered by clouds. With this information, the number of cloud-free pixels was calculated and the cloud percent cover was added to the model.

The third additional feature was the date, which in its original format as a string, is not a useful input to the model and doesn't carry any information. Therefore, each date was converted to seconds and its sine transform was calculated. This conversion provides signals of the time of the year and takes the seasonal changes of HABs into account. For this calculation, dates were converted to timestamps (seconds) and the sine wave was derived using the following equation.

$$Year\ sin\ =\ sin(timestamp\ *\ 2\ *\ \Pi/\ seconds_{year}) \quad \textbf{(3.3)}$$

$$seconds_{year}\ =\ 365.2425\ *\ 24\ *\ 60\ *\ 60 \quad \textbf{(3.4)}$$

Last, the dataset was expanded to include air temperature. It was decided to use air temperature over water temperature because water temperature depends on the depth and also the time of the day the sample was collected. By using air temperature, this complexity is eliminated and it is standardized across different stations of the lake and also can be applied to other lakes for future studies. Air temperature is directly correlated to water temperature, and therefore can be used as a proxy (O'Reilly et al., 2015). Minimum and maximum air temperatures in each day

were retrieved from NOAA Climate Data Database (NOAA, 2022) from the “Okeechobee 27.1 Nnw " meteorological station. This dataset is publicly available and is free to use.

3.3.4 In-situ Chl-a measurements

Chlorophyll-a data were collected from DBHYDRO, South Florida Water Management District’s corporate environmental database (www.sfwmd.gov/science-data/dbhydro, 2021) and the National Water Quality Council’s Water Quality Portal (www.waterqualitydata.us, 2022). Only discrete surface samples collected between 01/01/2011-12/31/2020 and analyzed via high performance liquid chromatography or via solvent extraction followed by fluorometry were considered. In-situ chl-a measurements were acquired for all station in Lake Okeechobee, as well as features such as date, depth, latitude and longitude. These additional parameters were later used for matching and merging with the satellite data.

3.3.5 LSTM model and training

3.3.5.1 Model development

Long-short term memory is a recurrent neural network (RNN) capable of handling long-term dependencies, hence a good choice for analyzing the behavior of algal blooms which have temporal patterns (Gianella, Burrows, Swan, Turner, and Davidson, 2021). LSTMs are able to remember information for long periods of time due to their special architecture. These networks are composed of a forget gate, a keep gate, and an output gate. The forget gate decides whether a current input (x_t) should be remembered and added to the cell state or discarded (C_{t-1}). This process, shown in Equation 3.5, is done by concatenating the current input by the previous hidden state (h_{t-1}), calculating the sigmoid of the concatenation, and multiplying it by the previous cell state (C_{t-1}). This, in practice, means that the model is deciding what features of day

t-1 have valuable information and which hidden units from that input should be ignored. The keep gate decides which values will be updated and what values need to be added to the final state using a hyperbolic tangent and a sigmoid layer (Equations (3.6), (3.7)). Finally, the old state is multiplied by the forget gate discarding the information it decided to forget using Equation 5, and added to the output of the keep gate to create the new cell state, C_t (Equation (3.8)). What the model outputs is a filtered version of the cell state shown in Equation (3.10).

$$f_t = \sigma(W_f \times x_t + U_h \times h_{t-1} + b_f) \quad \text{Forget gate} \quad \textbf{(3.5)}$$

$$i_t = \sigma(W_i \times x_t + U_i \times h_{t-1} + b_i) \quad \text{Input gate} \quad \textbf{(3.6)}$$

$$\hat{C}_t = tanh(W_c \times x_t + U_c \times h_{t-1} + b_c) \quad \text{Cell entrance} \quad \textbf{(3.7)}$$

$$C_t = f_t \times C_{t-1} + i_t \times \hat{C}_t \quad \text{New cell state} \quad \textbf{(3.8)}$$

$$o_t = \sigma(W_o \times x_t + U_o \times h_{t-1} + b_o) \quad \text{Output gate} \quad \textbf{(3.9)}$$

$$h_t = o_t \times tanh(C_{t-1}) \quad \textbf{(3.10)}$$

$$\sigma(x) = \frac{1}{1+e^x} \quad \text{Sigmoid function} \quad \textbf{(3.11)}$$

$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad \text{Tanh function} \quad \textbf{(3.12)}$$

Where W_f, W_i, W_o, and W_c are the weights connecting the input, x_t to the forget, input, output gates and the cell entrance respectively. Similarly, H_f, H_i, H_o, and H_c the weights connecting h_{t-1} to the same gates. b_f, b_i, b_o, and b_c are the bias terms for the mentioned gates and cell entrance.

The LSTM model implemented for this study is a bidirectional LSTM that consists of two LSTM networks as described above. The second LSTM reverses the flow of information flow. In this case, it means that if we're looking at a period of 5 days, in the first LSTM day 1 is the first time step to the model, and in the second LSTM, day 5 is the first time step. This means that the output layer can get information from past and future states simultaneously. Each LSTM model was chosen to have 60 units, resulting in a total of 120 units. Drop out with a rate of 0.5 was added to the layer to avoid overfitting and it was followed by a dense layer with one unit for the final prediction of chlorophyll estimation. This architecture was modified to a classification model by adding one unit and a sigmoid activation layer to the dense layer to get the probability

of HAB/No HAB events. A threshold of 10 µg/L was applied to chl-a values in the output for the classification model. The implementation and training of this model were done using Python 3.9.0, Tensor-Flow 2.7.0, and Keras 2.7.0 on an NVIDIA GeForce RTX 2080 Max-Q Graphics Processing Unit. Each dataset was split with a 90:10 train-to-test ratio and was trained for 150 epochs on regression models and 80 epochs on classification models. The number of epochs was chosen based on the performance of the models to achieve the best performance and avoid overfitting. Adam was chosen as the optimizer with a learning rate of 0.001, and exponential decay rates of 0.9 and 0.99 for the first and second moment estimates, respectively.

Finally, three commonly used machine learning models were employed to compare the performance of LSTM to non-recurrent models; K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Random Forest. Since these models can't be trained on three-dimensional data, the feature and time step dimensions were flattened into a 1-D vector. This experiment is referred to as "time-series input". Another issue is whether temporal features would improve predictions or not. To answer this question, the time step corresponding to the event day was chosen as a single input. For regression models, a KNN regressor, Support Vector Regression (SVR), and Random Forest (RF) regressor were trained, and a KNN classifier, SVM, and RF classifier were deployed for classification tasks. The same test-to-train ratio was applied to the datasets of all models, and hyperparameters were tuned for each of them using grid searches; K=3 was chosen as the number of neighbors for the KNN models and Manhattan distance was selected as the distance metric. Radial basis function (RBF) and $\varepsilon = 0.7$ were selected as the kernel and epsilon for SVM models, respectively, and the RF model had 3 estimators, a maximum depth of 6, and the ideal number of maximum features were picked by $\log_2$

(n_features) where n_features is the total number of features; that means the RF model takes a subset of $\log_2$(n_features) features to find the best split during training.

3.3.5.2 Dataset Structure

The extracted reflectance from each band, including all the additional features, were produced from the beginning of 2011 until the end of 2020. This data was merged with the chl-a data on the basis of date and station using the Pandas library in Python (McKinney, 2010). Therefore, the final dataset had twelve features including the maximum reflectance values from the seven bands of MODIS shown in Table 3.2, OCx, cloud cover, minimum temperature, maximum temperature, and date, for every day of the mentioned time period. Event days were retrieved for the days that the output values, chl-a (or HAB/No HAB), were available. Several variations of training sets were generated using different numbers of days of study before event days, to study the effect of the time period on the result, and also to determine the optimum number of days we need to look back in time to detect HABs. Increasing the number of time steps in the training sets adds to the complexity of the model and the number of training parameters. To have a fair comparison, the step (temporal resolution) was increased according to the time period so as to keep the length of the sequences below 7 days. These time variations and an example of the time sampling are shown in Figure 3.4.

Name	Time period
D3S1	Duration = 3, Step = 1
D5S1	Duration = 5, Step = 1
D7S2	Duration = 7, Step = 2
D9S2	Duration = 9, Step = 2
D11S2	Duration = 11, Step = 2
D13S2	Duration = 13, Step = 2
D15S4	Duration = 15, Step = 4
D17S4	Duration = 17, Step = 4
D19S4	Duration = 19, Step = 4
D21S4	Duration = 21, Step = 4
D23S4	Duration = 23, Step = 4
D25S4	Duration = 25, Step = 4

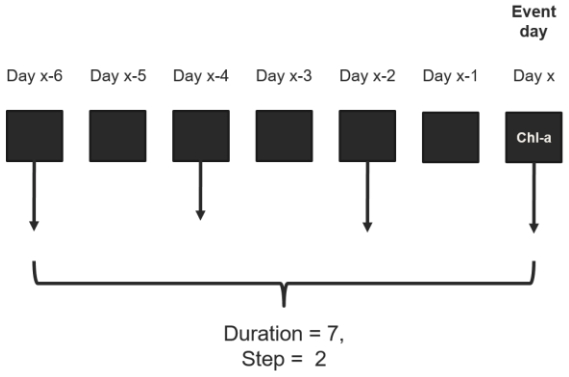

Figure 3.4 Variations in temporal window structure. These sequences vary the number of days in the past for both training and testing sets. An illustration of D7S2 is provided as an example.

3.3.6 Evaluation criteria and metrics

To evaluate the performance of each regression model, root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2) were used, and the equations of these metrics are shown in Equations (3.13)–(3.15), respectively. The performance of classification models were evaluated using the metrics accuracy and F1 score (Equations (3.16) and (3.17)). The final F1 score is the average for both classes, HAB and No HAB. In these equations, TP, TN, FP, and FN are the number of true positives, true negatives, false positives and false negatives.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad \textbf{(3.13)}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|(y_i - \hat{y}_i)\right| \quad \textbf{(3.14)}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \textbf{(3.15)}$$

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \quad \textbf{(3.16)}$$

$$F1 = \frac{2 \times Precision \times Recall}{\mathrm{Precision} + \mathrm{Recal}} \quad \textbf{(3.17)}$$

$$\mathrm{Precision} = \frac{TP}{TP + FP} \quad \textbf{(3.18)}$$

$$Recall = \frac{TP}{TP + FN} \quad \textbf{(3.19)}$$

3.4 Results

Two LSTM models, one for the estimation of chl-a concentration, and another for the classification of HAB or No HAB events, were trained on training sets with different time periods and steps shown in Figure 3.4. They were evaluated on both training and testing sets and the results are represented in Tables 3.3 to 3.6. Table 3.3 shows the evaluation metrics for the twelve time periods for the testing dataset. RMSE values on this table vary from 11.95 to 14.67 for different time periods, with D15S4 showing the lowest RMSE and D3S1 having the highest. As the period increases up to 15 days (D15S4), RMSE tends to decrease, before increasing afterward, reaching its lowest point at D15S4. This trend is also true for MAE, with D15S4 having the lowest MAE at 8.53 µg/L. A range of R^2 values was observed between 0.15 and 0.43, with the lowest at D3S1 and the highest at D15S4, following the same pattern as RMSE. Likewise, D15S4's classification accuracy and F1 scores were the highest, at 0.76 and 0.82, respectively (Table 3.5). Observed accuracy scores ranged from 0.66 to 0.76, and F1 scores ranged from 0.75 to 0.82.

Figures 3.5 (a) and 3.5 (b) illustrate the loss function plots of the classification and regression models, respectively. These models were trained in the D15S4 period. Figure 3.5 (b) shows that test loss was always lower than training loss, which was due to the fact that regularization (dropout) is only applied in training and not testing, and regularization loss is not included in the training loss. The model is neither underfitting nor overfitting, as both loss curves decrease gradually after about 10 epochs. To investigate this further, each model was also tested on its training set and the result are provided in Table 3.4, for comparison to Table 3.3. Mostly, RMSE, MAE and R2 on training data are either roughly the same or better the mentioned metrics on the testing set; it is reasonable for the metrics to be higher when evaluated on the training data

because they have been previously seen by the model, and the small gap between these metrics on each time period demonstrates that the models are not overfitted and properly trained.

The loss curves of the classification model also show that the model is not overfitted since the testing and training loss curves stay in the same range during training, and testing accuracy is slightly lower than training accuracy most of the time. Classification accuracy on both testing and training datasets can also be compared on Tables 3.5 and 3.6, respectively, and the accuracy and F1 scores on training are higher than the evaluation metrics on the testing data with a reasonable difference between them, showing that these models were also neither underfitted nor overfitted.

The performance of the machine learning models are shown in Table 3.7. For time-series inputs, D15S4 was selected as the best input since it had the best performance in the previous experiment. Among KNN, SVM and RF, the R^2 of RF was found to be the highest, values at 0.23, which was lower than the R^2 achieved using LSTM (0.43). RMSE and MAE were in the range of 13.43 to 14.42 and 9.34 to 10.69, respectively. Overall, R^2 of each model was higher when trained on the time-series input than single-input and the two other error metrics were lower. The classification accuracy of these models were comparable with the accuracy of the LSTM model (F1: 0.82, accuracy: 0.76), with a range of 0.61 to 0.81 for accuracy and 0.70-0.80 for accuracy and F1, respectively. Similar to the LSTM models, training results are shown in Table 3.7, to demonstrate how well the models are trained and whether the models are overfitted or underfitted to the data.

The scatter plots of predictions on training and testing from the D15S4 period are shown in Figure 3.6. Figure 3.7 depicts the scatter plots of true vs predicted chl-a values for each model and input type, single time and time-series.

Table 3.3 The performance of the LSTM model for chl-a predictions using twelve time window variations, on testing data.

	D3S1	D5S1	D7S2	D9S2	D11S2	D13S2	D15S4	D17S4	D19S4	D21S4	D23S4	D25S4
RMSE (µg/L)	14.67	13.15	12.65	12.99	12.08	12.45	**11.95**	12.45	12.77	12.71	13.43	12.64
MAE (µg/L)	10.76	9.42	9.17	9.78	9.26	9.11	**8.55**	9.00	9.06	9.31	9.48	8.86
R^2	0.15	0.25	0.35	0.27	0.33	0.41	**0.43**	0.35	0.28	0.37	0.30	0.34

Table 3.4 The performance of the LSTM model for chl-a predictions using twelve time window variations, on training data.

	D3S1	D5S1	D7S2	D9S2	D11S2	D13S2	D15S4	D17S4	D19S4	D21S4	D23S4	D25S4
RMSE (µg/L)	14.77	12.69	13.27	12.46	12.13	11.10	13.21	12.68	12.48	11.64	11.99	10.99
MAE (µg/L)	10.23	8.49	9.14	8.49	8.33	7.56	9.12	8.82	8.69	8.13	8.12	7.60
R^2	0.32	0.48	0.44	0.50	0.53	0.61	0.50	0.48	0.50	0.57	0.54	0.61

Table 3.5 The performance of the LSTM model for HAB/ No HAB classification using twelve time window variations, on testing data.

	D3S1	D5S1	D7S2	D9S2	D11S2	D13S2	D15S4	D17S4	D19S4	D21S4	D23S4	D25S4
Accuracy	0.66	0.68	0.71	0.73	0.73	0.73	**0.76**	0.73	0.67	0.66	0.67	0.70
F1 score	0.76	0.76	0.79	0.79	0.80	**0.82**	**0.82**	0.76	0.75	0.76	0.73	0.78

Table 3.6 The performance of the LSTM model for HAB/ No HAB classification using twelve time window variations, on training data.

	D3S1	D5S1	D7S2	D9S2	D11S2	D13S2	D15S4	D17S4	D19S4	D21S4	D23S4	D25S4
Accuracy	0.76	0.82	0.80	0.81	0.84	0.84	0.79	0.81	0.82	0.83	0.86	0.86
F1 score	0.85	0.88	0.86	0.87	0.89	0.89	0.86	0.87	0.88	0.89	0.90	0.90

Table 3.7 The performance of KNN, SVM, and RF on both chl-a estimations and classifications of HAB/No HAB, using single and time-series inputs on both training and testing data.

Testing / Training	RMSE (µg/L)	MAE (µg/L)	R^2	Accuracy	F1 score
KNN - single-time input	14.38	10.69	0.15	0.61	0.70
	15.00	10.30	0.27	0.82	0.87
KNN – time-series input	14.45	10.34	0.12	0.82	0.77
	14.78	10.08	0.29	0.81	0.86

SVM - single-time input	14.15	9.34	0.09	0.68	0.77
	14.81	9.47	0.17	0.76	0.84
SVM– time-series input	14.12	9.15	0.10	0.69	0.78
	14.86	9.41	0.17	0.77	0.85
RF - single-time input	14.42	10.15	0.14	0.68	0.77
	13.98	8.96	0.41	0.85	0.90
RF– time-series input	13.43	9.87	0.23	0.74	0.80
	12.36	8.42	0.48	0.88	0.91

(a) (b)

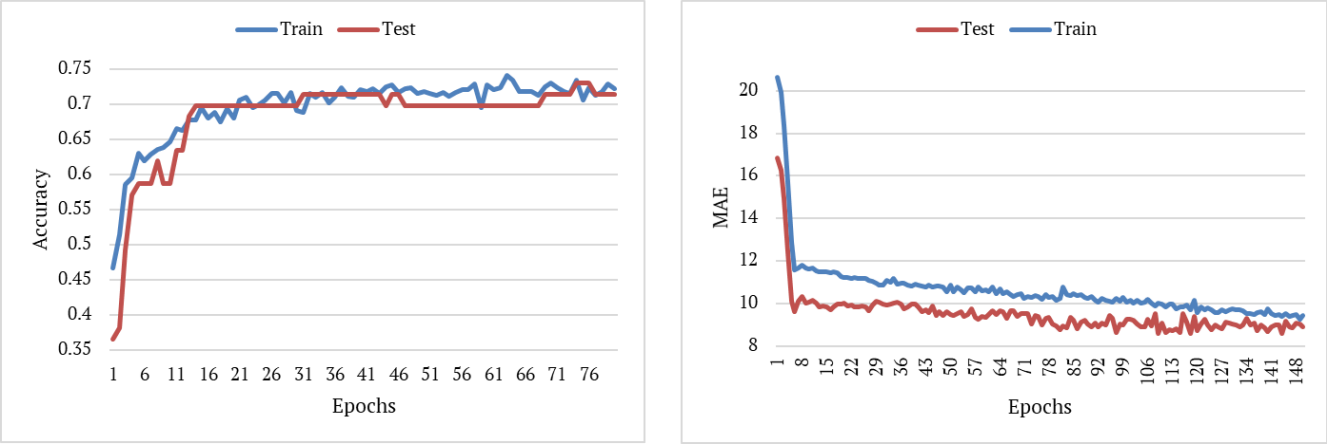

Figure 3.5 The loss curves of the a) classification LSTM model, and b) regression LSTM model. The metric during training for classification is accuracy and it is mean absolute error (MAE) for regression.

(a) (b)

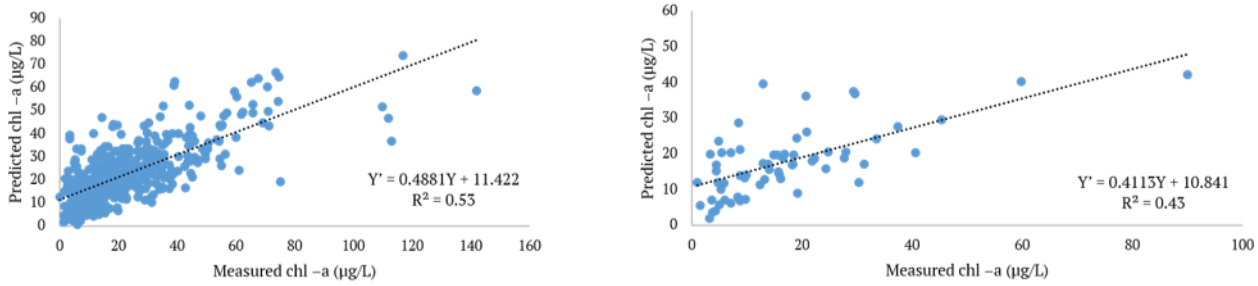

Figure 3.6 Scatter plots of measured vs predicted chl-a values using the LSTM model and time period D14S4 on a) train data, and b) test data.

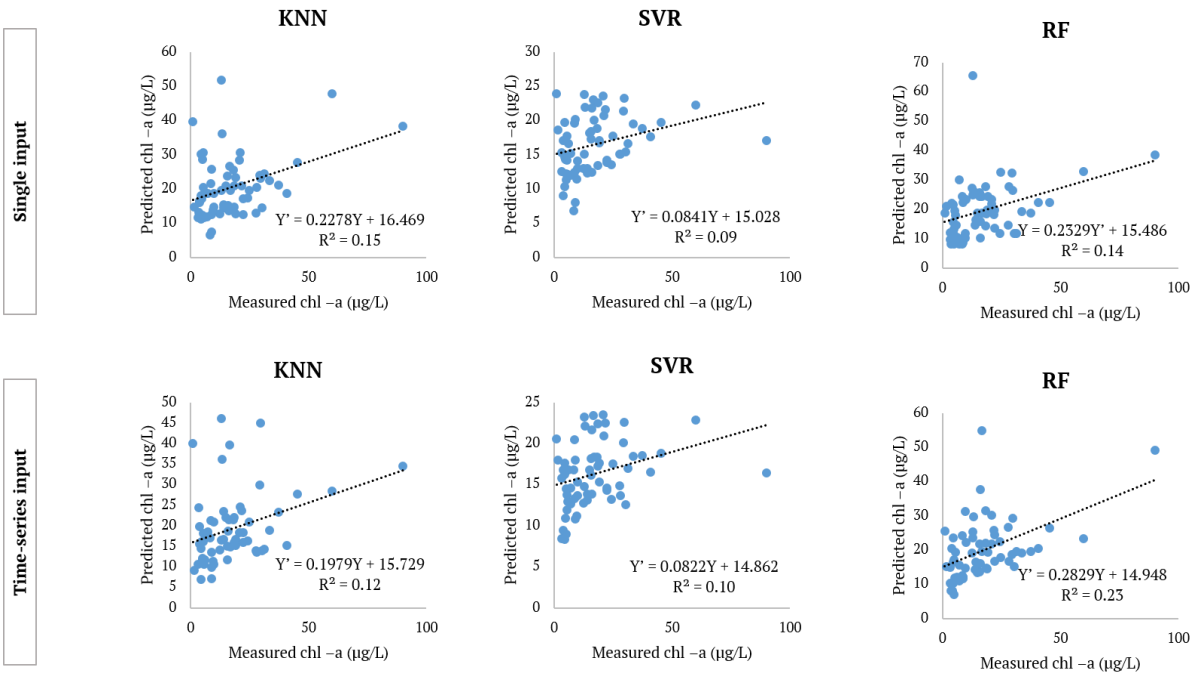

Figure 3.7 Scatter plots of measured vs predicted chl-a values using the KNN, SVR, and RF models with single and time-series inputs. The time-series input was a 1-D conversion of time period D15S4.

3.4.1 Feature importance analysis

Permutation importance was used to evaluate the importance of the studied features (Altmann et al., 2010). By assessing how much the model accuracy decreases when a data feature is absent, this algorithm provides insight into the importance of each data feature. Training models by eliminating features and assessing their performance is computationally intensive. So instead of performing the analysis during training, features were shuffled one at a time during testing, and their MAE is compared to the MAE of unshuffled test data. The model trained on the D15S4 period was selected for this analysis since it had the best performance compared to the other periods and it is more likely to provide insight into the most significant features. The result of this analysis is shown in Figure 3.8. The permutation importance score in this chart is the difference between the original MAE (8.55 µg/L based on Table 3.3) and the MAE obtained by testing the model on the test data with the shown feature randomly shuffled.

According to this chart, OCx, cloud cover, minimum temperature (TMIN), date, and B7 (2105 – 2155 nm) were the top five features, and B1 had the lowest importance score.

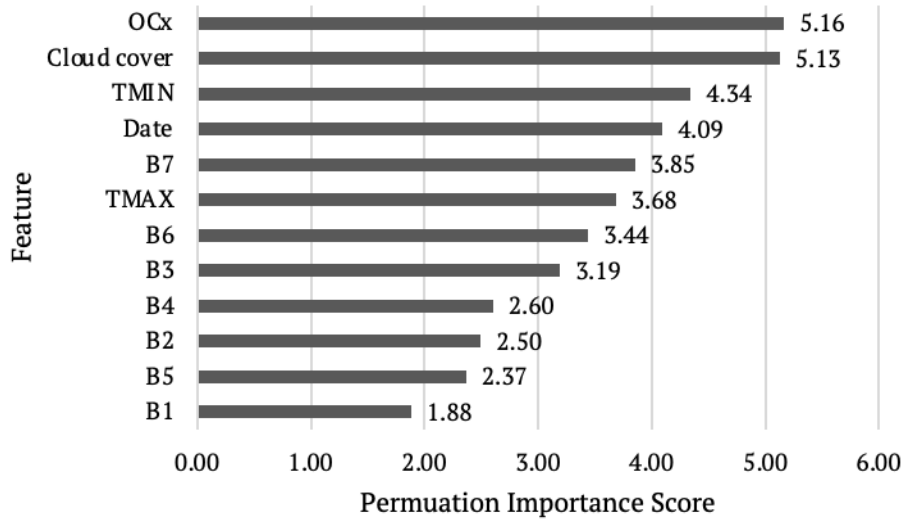

Figure 3.8 Permutation importance scores of the twelve features used as inputs in the training and testing datasets.

3.5 Discussion

All of the results for the twelve periods shown in Table 3.3, with the exception of D3S1, gave significantly better results than the machine learning models shown in Table 3.7. It is assumed that due to the capability of LSTM models in learning long-term dependencies and characterizing discriminating temporal features in data, they were able to make predictions much more effectively. We assume that in the first time periods, the interval was not long enough for there to be sufficient temporal variations for the LSTMs to effectively characterize the change in temporal features, and in the last time periods, the interval is so long that it did no carry relevant information to cause or not to cause a HAB. Therefore, a period of about two weeks (D13S2 and D15S4) seems to be the optimal interval to observe for making predictions.

Among the machine learning models, random forest had the best performance with an R^2 of 0.23 on the time-series input. This model's performance was improved by 70% by training it on the time-series input rather than a single-time input, signifying the importance of the temporal features. Looking at the scatter plots of RF predictions in Figure 3.7, there are several outliers in the predictions which reduced the R^2, but the RMSE and MAE from these predictions were in the same range as the RMSE and MAE from LSTM predictions. These outliers explain why classification accuracy from the ML and LSTM models are comparable but the regression results from LSTM are much better. The fact that LSTM's performance in classification is not significantly different from ML models was also shown in Hill et al. (2020)'s work, however, chl-a predictions (regression results) were not provided in this study.

SVM witnessed a small performance improvement after switching to the time-series input and KNN's performance was worsened. KNN's inferior performance on time-series input can be because of the known curse of dimensionality in KNN models (Cover & Hart, 1967). These models tend to face difficulty with high-dimensional data.

It is difficult to compare studies on chl-a estimations based on their results since different lakes have dissimilar behaviors. Factors such as water inflow/outflow, chl-a concentration levels, water depth, and surface area, can affect how a model performs. For instance, Kutser (2009) showed how complicated or in some cases impossible it is to monitor cyanobacterial blooms in shallow-water areas, and Vidot and Santer (2005) showed how atmospheric correction can become more difficult for larger water bodies.

Therefore, even though it's hard to compare directly, we can compare our work to the study by Chang et al. (2011) on chl-a predictions in Lake Okeechobee. Three models were developed for chl-a predictions in Lake Okeechobee using MODIS images from 2003 to 2004.

They compared Genetic Programming (GP), Artificial Neural Network (ANN), and Multiple Linear Regression (MLR) models and showed the best predictions were obtained by the GP model with an $R^2 = 0.57$. This R^2 is higher than the highest achieved on a validation dataset in this work (0.43). According to the histogram of chl-a values in this study in the period of 2003-2004, which is different from the period of our study (2011-2020), chl-a values were below 30 µg/L at all times and that lowers the probability of having a saturated model that performs worse on larger values. That also means having a balanced dataset that is equally trained on different ranges of data, but in our study, about 6% of chl-a values exceeded 50 µg/L, making it an unbalanced dataset and harder for the model to predict extreme values. In addition, the chl-a measurements in this paper were collected by the authors directly, which could have positively affected the quality of the data, whereas in this research the chl-a data sourced from a third party online database. Therefore the integrity of our data could not be as substantially and directly ensured.

3.6 Conclusions

In this study, chlorophyll-a concentrations in Lake Okeechobee were estimated using daily satellite imagery from MODIS, a satellite-based sensor that provides multispectral images in seven bands. The dataset was expanded by additional features; cloud cover, chl-a estimations using the OCx algorithm, temperature data, and the sine transform of timestamps. Long-short term memory, a recurrent neural network capable of learning temporal features, was trained on the dataset. Two important questions were answered, whether temporal features improve prediction accuracy, and if yes, how many days of data are required for the best predictions. To answer the first question, three machine learning models were trained on single time step and time-series inputs. The results showed that time-series data have invaluable information that

helps prediction accuracy, and it is noteworthy that LSTM outperformed the traditional machine learning models trained on both single inputs and time-series inputs, attesting to its ability to learn long-term dependencies in data. To answer the second question, LSTM models were trained on twelve different periods of time-series data, ranging from 3 to 25 days before chl-a measurements and their performances were compared; the results showed that fifteen days of data with a resolution of 4 days had the best performance. Additionally, a feature importance analysis was conducted to assess the value of each feature, and it was discovered that OCx, cloud cover, minimum temperature, date (sine transform of timestamp), and the seventh band of MODIS (2105 – 2155 nm) were the top five features. It was demonstrated that chl-a concentrations can be estimated using ML methods and satellite images and HABs can be detected with a reasonable accuracy even when a large dataset of chlorophyll measurements is not available, and that there is potential for the monitoring of other lakes with few field measurements. A possible future study would be leveraging the pre-trained model in this work to develop a transfer leaning method for the use of other lakes.

3.7 References

Altmann, A., Tolo¸si, L., Tolo¸si, T., Sander, O., & Lengauer, T. (2010). Data and text mining Permutation importance: A corrected feature importance measure. Bioinformatics, 26(10), 1340–1347.

Chang, N. B., Yang, Y. J., Daranpob, A., Jin, K. R., & James, T. (2011). Spatiotemporal pattern validation of chlorophyll-a concentrations in Lake Okeechobee, Florida, using a comparative MODIS image mining approach. 33(7), 2233–2260.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21–27.

Craig, S. E., Lohrenz, S. E., Lee, Z., Mahoney, K. L., Kirkpatrick, G. J., Schofield, O. M., & Steward, R. G. (2006). Use of hyperspectral remote sensing reflectance for detection and assessment of the harmful alga, Karenia brevis. Applied Optics, Vol. 45, Issue 21, Pp. 5414-5425, 45(21), 5414–5425.

DBHYDRO, South Florida Water Management District's corporate environmental database. (2021). Retrieved from https://www.sfwmd.gov/science-data/dbhydro

Gianella, F., Burrows, M. T., Swan, S. C., Turner, A. D., & Davidson, K. (2021). Temporal and spatial patterns of harmful algae affecting scottish shellfish aquaculture. Frontiers in Marine Science, 8.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment, 202, 18–27.

Hill, P. R., Kumar, A., Temimi, M., & Bull, D. R. (2020). HABNet: Machine learning, remote sensing-based detection of harmful algal blooms. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13, 3229–3239.

Hoagland, P., & Scatasta, S. (2006). The economic effects of harmful algal blooms. Ecology of Harmful Algae, 391–402.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.

Hu, C., Lee, Z., & Franz, B. (2012). Chlorophyll a algorithms for oligotrophic oceans: A novel approach based on three-band reflectance difference. Journal of Geophysical Research: Oceans, 117(1).

Khan, R. M., Salehi, B., Mahdianpari, M., Mohammadimanesh, F., Mountrakis, G., & Quackenbush, L. J. (2021). A meta-analysis on harmful algal bloom (Hab) detection and monitoring: A remote sensing perspective. Remote Sensing, 13(21).

Kutser, T. (2009). Passive optical remote sensing of cyanobacteria and other intense phytoplankton blooms in coastal and inland waters. International Journal of Remote Sensing, 30(17), 4401–4425.

Li, J., Gao, M., Feng, L., Zhao, H., Shen, Q., Zhang, F., Zhang, B. (2019). Estimation of chlorophyll-a concentrations in a highly turbid eutrophic lake using a classification-based MODIS land-band algorithm. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12(10), 3769–3783.

Liu, M., Ling, H., Wu, D., Su, X., & Cao, Z. (2021). Sentinel-2 and Landsat-8 observations for harmful algae blooms in a small eutrophic lake. Remote Sensing 2021, Vol. 13, Page 4479, 13(21), 4479.

McKinney, W. (2010). Data structures for statistical computing in python. Proceedings of the 9th Python in Science Conference, 56–61.

National Water Quality Council's Water Quality Portal. (2022). Retrieved from ww.waterqualitydata.us

NOAA National Centers for Environmental Information. (2022). Retrieved from www.ncdc.noaa.gov/

O'Reilly, C. M., Sharma, S., Gray, D. K., Hampton, S. E., Read, J. S., Rowley, R. J., Zhang, G. (2015). Rapid and highly variable warming of lake surface waters around the globe. Geophysical Research Letters, 42(24), 10,773-10,781.

South Florida Water Management District [SFWMD]. (2022).

Ventura, D. L. T., Martinez, J. M., de Attayde, J. L., Martins, E. S. P. R., Brandini, N., & Moreira, L. S. (2022). Long-term series of chlorophyll-a concentration in Brazilian Semiarid Lakes from Modis imagery. Water (Switzerland), 14(3).

Vidot, J., & Santer, R. (2005). Atmospheric correction for inland waters—Application to SeaWiFS. Retrieved November 22, 2022, from

Xu, D., Pu, Y., Zhu, M., Luan, Z., & Shi, K. (2021). Automatic detection of algal blooms using Sentinel-2 MSI and Landsat OLI images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14, 8497–8511.

Yussof, F. N., Maan, N., & Reba, M. N. M. (2021). LSTM networks to improve the prediction of harmful algal blooms in the west coast of Sabah. International Journal of Environmental Research and Public Health 2021, Vol. 18, Page 7650, 18(14), 7650.

Zhang, Y., Lin, S., Qian, X., Wang, Q., Qian, Y., Liu, J., & Ge, Y. (2011). Temporal and spatial variability of chlorophyll a concentration in Lake Taihu using MODIS time-series data. Hydrobiologia, 661(1), 235–250.

Conclusion

In this thesis, the feasibility of utilizing remote sensing and deep learning technologies for peanut phenotyping and water quality monitoring were assessed. Each of those problems necessitate the right choice of remote sensing platform and deep learning model; for rapid peanut phenotyping UAV-based hyperspectral images were used and the performance of an ensemble model was compared with the performance of an end-to-end deep learning model. In both cases, promising results were obtained; deep learning showed better results for the prediction of agronomic traits, and physiological traits were predicted more accurately with the ensemble model. It was also found that in case of a drought, eighteen days after the start of drought is the best day for hyperspectral data collection, as the model trained on the data from this day showed the most accurate predictions. It was shown and discussed that some drought tolerant genotypes such as Line-8 recover from drought after 29 days, and therefore 18 days after drought is when the highest effect of drought was seen on some plants, and that information helped the model achieve more accurate predictions for peanut traits. Additionally, a feature importance analysis was performed on both models, and the most important wavelengths were in the visible, red-edge and NIR regions. This study showed that it is possible to provide insight to breeders regarding yield, biomass and pod count 30-45 days before harvest, and it is also feasible to estimate photosynthesis and stomatal conductance for same-day measurement using hyperspectral images and machine learning, which makes getting estimates for peanut phenotypes a lot faster compared to traditional in-situ measurements. This research can be improved by acquiring more data in different locations and for multiple years. Another way it can be improved is to lower the temporal resolution by increasing the number of days the peanuts are imaged, and train a

recurrent neural network and evaluate weather temporal features would improve the prediction accuracy.

In a second study, satellite images from the MODIS dataset were used for monitoring Lake Okeechobee, estimating chl-a concentrations and determining whether there was an algal bloom or not. Satellite images are preferable for monitoring larger areas such as Lake Okeechobee, and also allow us to leverage pre-existing historic data. A research question being addressed in this chapter was discovering whether temporal features can help with the prediction of HABs. In other words, is studying same-day water reflectance enough for estimating chl-a concentrations and the predictions of HABs, or would it be beneficial to include the reflectance information, as well as the additional discussed features, from several days before. By training three machine learning models with both same-day and time-series inputs, it was demonstrated that temporal features improve prediction accuracy. However, using machine learning models to analyze time-series data is not the best approach, therefore, an LSTM model was trained on the time-series data and it was revealed that it outperforms ML models. In addition, twelve window frames were tested and it was shown that fifteen days of data preceding the event day is the ideal time frame. Since this time frame had a resolution of four days, satellite-based imaging platforms with a lower temporal resolution than MODIS, but higher spatial resolution can be employed using the same method for future studies, to evaluate the importance of spatial resolution in water quality monitoring and whether it affects the prediction accuracy. Moreover, a transfer learning method to use the pre-trained model developed in this study to the use of other lakes could be a valuable follow-up project.