The International Society of Precision Agriculture presents the
16th International Conference on
Precision Agriculture
21–24 July 2024 | Manhattan, Kansas USA

Airborne Spectral Detection of Leaf Chlorophyll Concentration in Wild Blueberries

Christopher Ewanik[1], Kallol Barai[2], Yong-Jiang Zhang[2,5], Umesh R. Hodeghatta[3], Vikas Dhiman[4]

[1]Data Science and Engineering, University of Maine, USA
[2]School of Biology and Ecology, University of Maine, USA
[3]CPS-Applied Machine Intelligence, Northeastern University, USA
[4]Electrical and Computer Engineering, University of Maine, USA
[5]Climate Change Institute, University of Maine, USA

A paper from the Proceedings of the
16th International Conference on Precision Agriculture
21-24 July 2024
Manhattan, Kansas, United States

ABSTRACT

Leaf chlorophyll concentration (LCC) detection is crucial for monitoring crop physiological status, assessing the overall health of crops, and estimating their photosynthetic potential. Fast, non-destructive, and spatially extensive monitoring of LCC in crops is critical for accurately diagnosing and assessing crop health in large commercial fields. Advancements in hyperspectral remote sensing offer non-destructive and spatially extensive alternatives for monitoring plant parameters such as LCC. However, the LCC prediction model may vary from one crop to another due to differences in structural and physiological properties. The wild blueberry crop has diverse genotypes grown in semi-natural systems, making precision management difficult. Here, we aimed to test the performance of the remote LCC detection models in wild blueberries using machine learning (ML). Hyperspectral data ranging from 400nm-1000nm were collected using an unmanned aerial vehicle (UAV) from two adjacent irrigated and non-irrigated commercial fields covering different growth stages. LCC indicated by SPAD values were collected using the SPAD-502 Chlorophyll meter at the field site and then converted to LCC values using SPAD to LCC conversion models. In the preliminary data analysis of the UAV-based hyperspectral data for LCC prediction, different ML techniques were used. While previous research has used ML and ensembles for similar tasks, this research focused on using various preprocessing techniques to attempt to create more learnable features from the data that could be used in ensemble structures. The dimensionality of the dataset was reduced using Non-negative Matrix Factorization (NMF) and Gaussian Mixture Model (GMM) methods. Thirty-four different chlorophyll vegetation indices were feature-engineered to create an additional dataset for the ensemble structure. Various ML models were implemented, splitting the data into 80/20 for training and testing. The best single learning model was an ElasticNet regression trained on a GMM dataset with a coefficient of determination (R2) of 0.79 and a normalized root mean square error (nRMSE) of 3.48 %. PyTorch was used to combine six base models and differently preprocessed datasets into an optimal weights meta-learner architecture that achieved a better performance of an R2 of 0.89 and a nRMSE of 2.53%. Work

The authors are solely responsible for the content of this paper, which is not a refereed publication. Citation of this work should state that it is from the Proceedings of the 16th International Conference on Precision Agriculture. EXAMPLE: Last Name, A. B. & Coauthor, C. D. (2024). Title of paper. In Proceedings of the 16th International Conference on Precision Agriculture (unpaginated, online). Monticello, IL: International Society of Precision Agriculture.

is ongoing on developing a neural vegetation index (NVI) that searches for wavelengths in the space ratio of linear functions of reflectance to automate the process of index development. During training, a neural network searches the space of functions that minimizes a given loss function. The general framework of NVI could also work across species and different nutrients.

Keywords.
Chlorophyll prediction, Vegetation index, hyperspectral machine learning

INTRODUCTION

The quantification of leaf chlorophyll concentration (LCC) holds significant importance for agricultural practices, as it enables the continuous assessment of crop physiological status, determination of overall crop health, and estimation of photosynthetic potential (Gitelson et al., 2003). This information is particularly valuable for monitoring large commercial fields, where rapid, non-destructive, and spatially extensive techniques are essential for accurate diagnosis and assessment.

Although traditional methods of wet extraction analysis through field sampling provide accurate estimations of Leaf Chlorophyll Content (LCC), these methods are not always practical for estimating LCC over large areas of vegetation. However, non-destructive measurement of leaf spectral reflectance offers an instantaneous and alternative method for assessing the LCC of plants over a large spatial scale (Lu et al., 2018). This method involves measuring the light reflected by leaves, which varies according to the chlorophyll content (Lu et al., 2018; Zhang et al., 2014). By analyzing the spectral reflectance of the leaves, we can determine the LCC of the plants. This non-destructive approach can be particularly useful for quick, accurate, large-scale vegetation health evaluations.

UAV (Unmanned aerial vehicle)- based hyperspectral remote sensing has emerged as a promising avenue for monitoring various plant parameters, including LCC, due to its non-destructive nature and ability to cover large spatial extents (Hu et al., 2023). However, the challenge lies in developing LCC prediction models that can accommodate the inherent variability in structural and physiological properties across different crops.

The wild blueberry crop has diverse genotypes grown in semi-natural systems (Barai et al., 2022), so implementing precision management becomes challenging. To address this challenge, we have investigated the application of machine learning (ML) techniques for UAV-based remote LCC detection in wild blueberries. The objective is to evaluate the performance of ML models in accurately predicting LCC, considering the unique characteristics and variability inherent in wild blueberry crops. This research endeavors to contribute to advancing remote sensing applications in agriculture, specifically in the context of wild blueberries, by enhancing the robustness and adaptability of LCC prediction models by utilizing machine learning methodologies.

METHODS

Study Site

The study was conducted on commercial blueberry fields in Deblois, Maine (Longitude: -68.0001° N, Latitude: 44.7350° W). These commercial crop fields contain many different genotypes of wild blueberry plants growing within a particular field.

UAV-based Hyperspectral Data Collection and Ground Sampling

The hyperspectral image acquisitions and ground measurements were conducted one time in the summer of 2019 and three times in 2022 and 2023, covering different developmental stages within large commercial blueberry fields. Image acquisition and field ground data collection dates were carried out on sunny days. A total of 30 genotypes (15 in each field) in 2019 (crop year), and 40 genotypes (20 in each field) were systematically selected in 2022 (vegetative growth year) and 2023 (crop year) to cover the entire field and a wide range of genotypes based on morphological differences. Six wild blueberry stems were randomly selected from each genotype area to measure chlorophyll content. LCC indicated by SPAD values were collected using the SPAD-502 chlorophyll meter (Konica Minota Inc., Japan) at the field site and then converted to LCC (μg/cm2) values using the SPAD to LCC conversion model following Zhu et al. (2012).

We used a Headwall Photonics Micro A-Series Sensor (Bolton, MA, USA) hyperspectral imaging spectrometer attached to a DJI Matrice 600 Pro UAV for data collection. Data collection was conducted between 12:00 PM ± 2 hours local time. The sensor captured 324 spectral bands uniformly distributed between 400 to 1000 nm in the visible and near-infrared electromagnetic spectrum. After processing the imagery with the Headwall Spectral View application, we used ENVI software (version 5.5 64-bit) to identify and extract pixels of ground-sampled genotypes. These delineations were used as samples in both training and validation data sets. All downstream analyses were performed in Python.

Preliminary data analysis and preprocessing

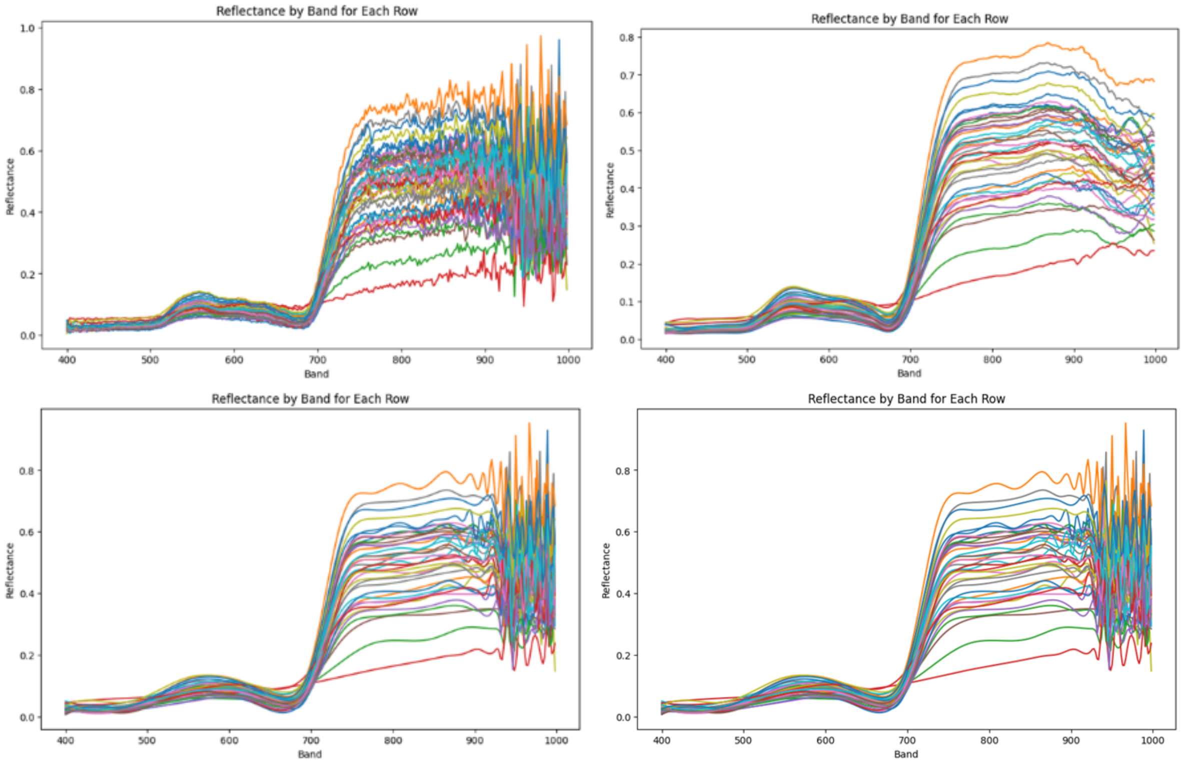


Figure 1: Left to right top first: (a) raw reflectances (b) Savitzky-Golay smoothed reflectances (c) univariate spline reflectances (d) Gaussian filter reflectances.

To understand the nature of data collected we first visualized the reflectance data from the hyperspectral imaging spectrometer. Recall that we get 324 spectral bands for each genotype (30 for 2019) data. We plotted the raw reflectance with wavelength on the x-axis and reflectance on the y-axis (Figure 1(a)). We observed that raw reflectance is noisy as neighboring wavelengths show fluctuations in the reflectance value. To remove noise, we tried 3 different methods of smoothing as a pre-processing step, (1) Savitzky-Golay smoothing (2) Univariate Spline smoothing and (3) Gaussian filter smoothing. The corresponding results after smoothing are shown in Figure 1(b-d) respectively.

We further observed that the visible wavelengths (400-700 nm) had much lower reflectance compared to the near-infrared wavelengths (700-1000 nm) (Figure 2). Within the visible spectrum, the green wavelength (~540 nm) had high reflectance, while the blue (~450 nm) and red (~670 nm) regions had comparatively low reflectance. On the other hand, the near-infrared region (720-1000 nm) had continuous high reflectance properties.

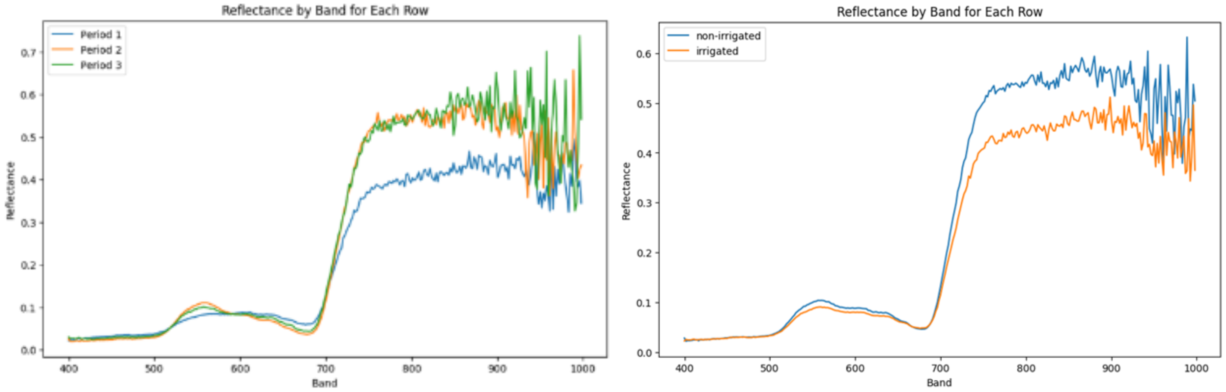


Figure 2: Left: Mean reflectances by period. Right: reflectances by irrigation status.

Model Development

In the preliminary data analysis of the UAV-borne hyperspectral imagery for LCC prediction, different ML techniques were used, using Scikit-Learn in Python. Splitting the data into 80/20 for training and testing, we utilized various machine-learning methods such as kernel ridge regression (KRR), ElasticNet, and XGBoost. We additionally used partial least square regression (PLSR). PLSR is a form of regularized linear regression where the number of components controls the strength of the regularization, which is suitable for predictors with high collinearity. Kernel ridge regression (KRR) combines ridge regression (linear least squares with 12-norm regularization) by learning a linear/ non-linear function in the space induced by the respective kernel and the data. Elastic net regression aims to select the predictor variables most important for predicting the target variable while using regularization to avoid overfitting the model. XGBoost is a decision-tree-based ensemble algorithm that uses gradient boosting and hardware optimizations to produce quick, accurate results for regression, classification, ranking, and time series tasks.

While previous research has used ML and ensembles for similar tasks, this research focused on using various preprocessing techniques to attempt to create more learnable features from the data that could be used in ensemble structures. A general challenge of working with hyperspectral data is the collinearity of bands. The dimensionality of the dataset was reduced using Non-negative Matrix Factorization (NMF) and Gaussian Mixture Model (GMM) methods. Thirty-four different chlorophyll vegetation indices were calculated to create an additional dataset for the ensemble structure. To tune utilized hyperparameters, Optuna, an automatic hyperparameter optimization software framework, was used to construct the search spaces for the hyperparameters dynamically. Finally, this study incorporated PyTorch to calculate the optimal combination of base learners. We first developed a rough estimate for a base learner coefficient and then used PyTorch and gradient descent to tune the optimal coefficients. We also explored the optimal wavelength bands for simple differences (R_{Y1}-R_{Y2}), simple ratios (R_{Y1}/R_{Y2}), and normalized differences (|R_{Y1}-R_{Y2}|/ (R_{Y1}-R_{Y2})).

The MDATT Index is the ratio of reflectance differences, defined as

$$MDATTIndex = \frac{R_{\gamma 1} - R_{\gamma 2}}{R_{\gamma 1} - R_{\gamma 3}}$$

calculate a few other simple two-band equations (see Appendix) that are searched to find the combination that produces the highest $R2$ (Lu, et al., 2018).

We observed a similarity between MDATT Index and Neural networks. MDATT Index searches for wavelengths that are ratios of linear functions (specifically differences) of reflectance. Similarly, Neural networks are linear functions chained together interspersed with non-linearities. During training, a Neural network searches the space of functions that minimizes a given loss function.

We designed a Neural Network that can be expressed as a ratio of two linear functions (technically, affine function) of reflectance to predict its target nutrient,

$$NVI(\boldsymbol{R}, \boldsymbol{w}, \boldsymbol{u}) = \frac{w_1 R_{\lambda 1} + w_2 R_{\lambda 2} + \cdots + w_n R_{\lambda n} + w_0}{u_1 R_{\lambda 1} + u_2 R_{\lambda 2} + \cdots + u_n R_{\lambda n} + u_0}$$

Here $\boldsymbol{w} = [w_1, w_2, \dots, w_n]$ and $\boldsymbol{u} = [u_1, u_2, \dots, u_n]$ are weights that will be optimized during training process. Note that the weights can take negative values as well, so they form a larger search space than MDATT Index. We use Stochastic Gradient Descent for optimization of the following loss function,

$$L(\boldsymbol{w}, \boldsymbol{u}, \theta) = \sum_{c_i, \boldsymbol{R_i}} \left|\left| c_i - MLP_\theta\left(NVI(\boldsymbol{R_i}, \boldsymbol{w}, \boldsymbol{u})\right) \right|\right|_2 + \lambda |\boldsymbol{w}|_1 + \lambda |\boldsymbol{u}|_1,$$

where c_i is the observed chlorophyll content corresponding to the reflectance $\boldsymbol{R_i}$ in the training dataset $\mathcal{D}$. Also, $MLP_\theta(.)$ denotes a multi-layer perceptron for regression. Note that we can use any other differentiable regressor than a $MLP_\theta(.)$. The use of a differentiable regressor allows us to train the Neural vegetation index simultaneously with the $MLP_\theta(.)$. In practice, we can use multi-dimensional NVI() that can feed as a vector into MLP(). In our experiments, we found that 4-dimensional NVI() output provides a good trade-off between computation and accuracy.

RESULTS:

When investigating the correlation coefficients of the wavelengths to the LCC, it was found that there were various sensitive regions (Figure 2). The green wavelength region had high positive correlations, while the blue and red regions of the visible spectrum had low or negative correlations. Typically, green vegetation exhibits high reflectance in the visible green and near-infrared regions (400-1000 nm) of the electromagnetic spectrum . In contrast, the visible blue and red regions have high absorbance (Figure 2).

Various ML models were implemented, such as PLSR, ElasticNet, XGBoost, and KRR on the original dataset, vegetation indices (ensemble structure), and two-dimensionality reduction methods (GMM and NMF). The overall performance of the PLSR models was found unsatisfactory in all the datasets. The model performance was poor when utilizing the original unprocessed and ensemble indices datasets. The performance of the ML models on the GMM and NMF processed dataset (dimensionality reduced) was improved. The best single learning model was an ElasticNet regression with a coefficient of determination (R^2) of 0.79 and a normalized root mean square error (nRMSE) of 3.48%, which was trained on the dataset reduced with the GMM dimensionality reduction method (Table 1).

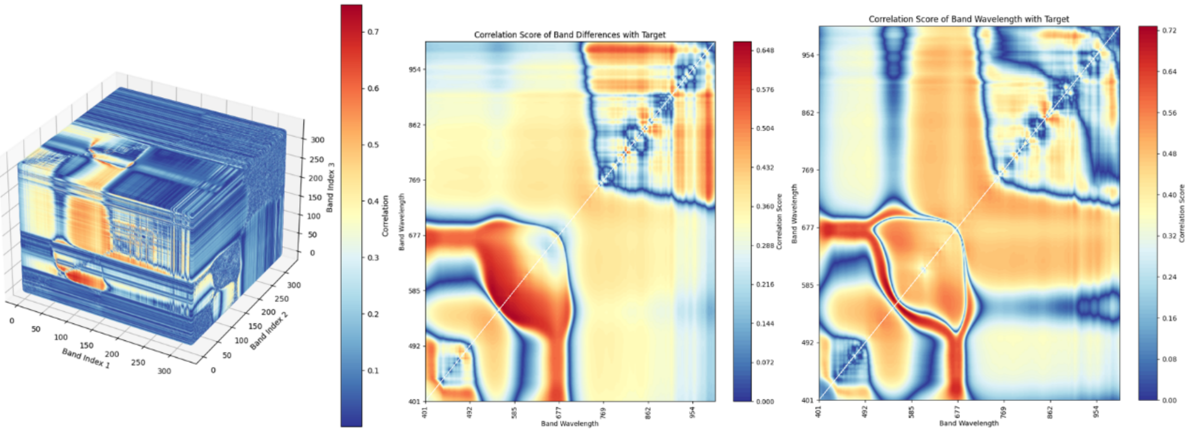


Figure 3: Left to right: (a) MDATT Index correlation heatmap (b) ND Index correlation heatmap (c) SD Index correlation heatmap

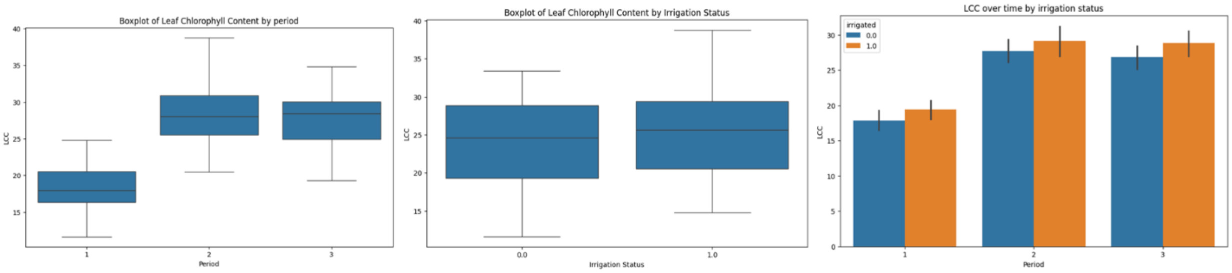
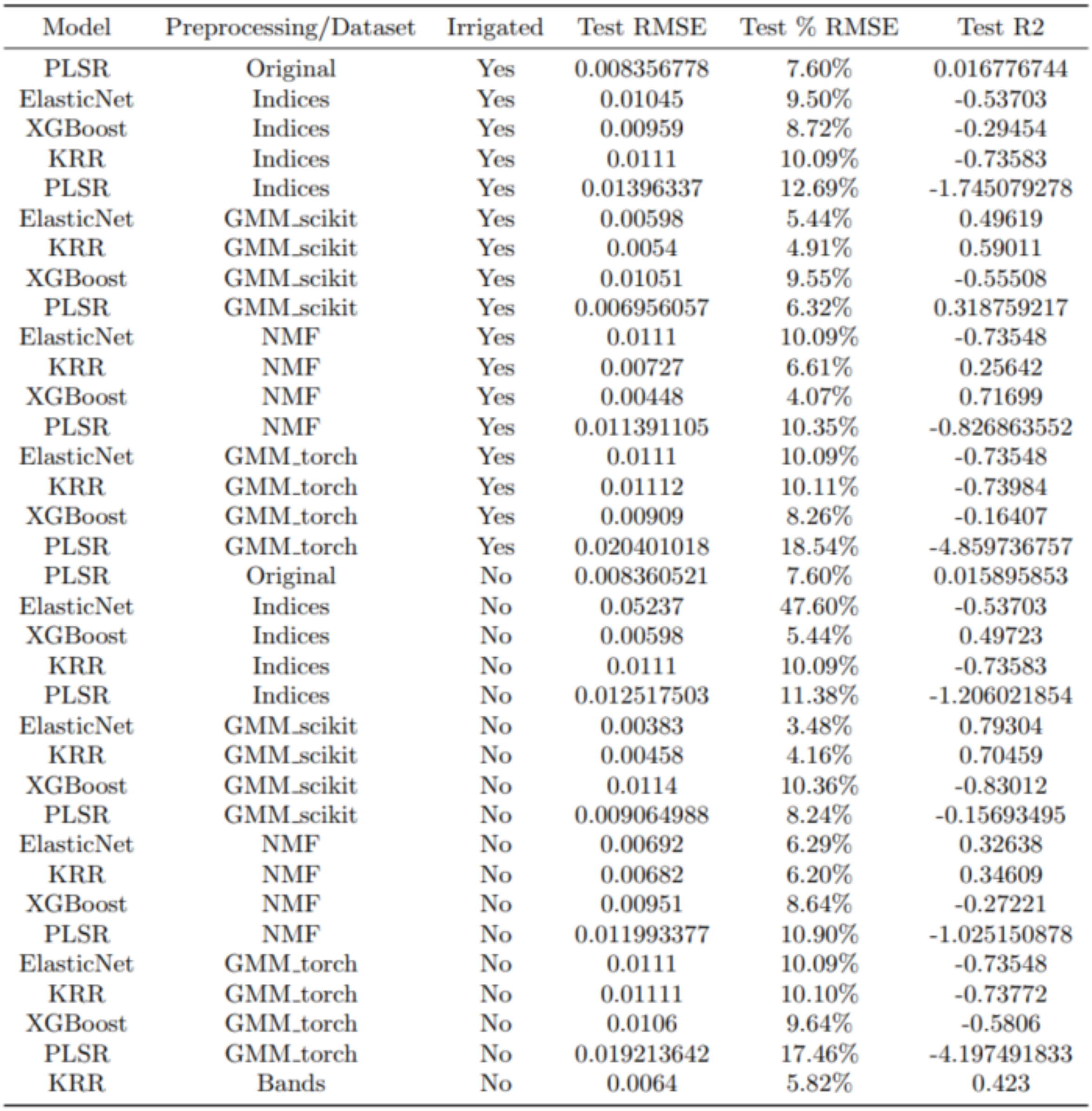


Figure 4: Leaf chlorophyll concentration of wild blueberry plants (a) at different growth periods (b), under different irrigation status (irrigated or not), and (c) combined.

Table 1: Results of Machine Learning Model Performance



Model	Preprocessing/Dataset	Irrigated	Test RMSE	Test % RMSE	Test R2
PLSR	Original	Yes	0.008356778	7.60%	0.016776744
ElasticNet	Indices	Yes	0.01045	9.50%	-0.53703
XGBoost	Indices	Yes	0.00959	8.72%	-0.29454
KRR	Indices	Yes	0.0111	10.09%	-0.73583
PLSR	Indices	Yes	0.01396337	12.69%	-1.745079278
ElasticNet	GMM_scikit	Yes	0.00598	5.44%	0.49619
KRR	GMM_scikit	Yes	0.0054	4.91%	0.59011
XGBoost	GMM_scikit	Yes	0.01051	9.55%	-0.55508
PLSR	GMM_scikit	Yes	0.006956057	6.32%	0.318759217
ElasticNet	NMF	Yes	0.0111	10.09%	-0.73548
KRR	NMF	Yes	0.00727	6.61%	0.25642
XGBoost	NMF	Yes	0.00448	4.07%	0.71699
PLSR	NMF	Yes	0.011391105	10.35%	-0.826863552
ElasticNet	GMM_torch	Yes	0.0111	10.09%	-0.73548
KRR	GMM_torch	Yes	0.01112	10.11%	-0.73984
XGBoost	GMM_torch	Yes	0.00909	8.26%	-0.16407
PLSR	GMM_torch	Yes	0.020401018	18.54%	-4.859736757
PLSR	Original	No	0.008360521	7.60%	0.015895853
ElasticNet	Indices	No	0.05237	47.60%	-0.53703
XGBoost	Indices	No	0.00598	5.44%	0.49723
KRR	Indices	No	0.0111	10.09%	-0.73583
PLSR	Indices	No	0.012517503	11.38%	-1.206021854
ElasticNet	GMM_scikit	No	0.00383	3.48%	0.79304
KRR	GMM_scikit	No	0.00458	4.16%	0.70459
XGBoost	GMM_scikit	No	0.0114	10.36%	-0.83012
PLSR	GMM_scikit	No	0.009064988	8.24%	-0.15693495
ElasticNet	NMF	No	0.00692	6.29%	0.32638
KRR	NMF	No	0.00682	6.20%	0.34609
XGBoost	NMF	No	0.00951	8.64%	-0.27221
PLSR	NMF	No	0.011993377	10.90%	-1.025150878
ElasticNet	GMM_torch	No	0.0111	10.09%	-0.73548
KRR	GMM_torch	No	0.01111	10.10%	-0.73772
XGBoost	GMM_torch	No	0.0106	9.64%	-0.5806
PLSR	GMM_torch	No	0.019213642	17.46%	-4.197491833
KRR	Bands	No	0.0064	5.82%	0.423

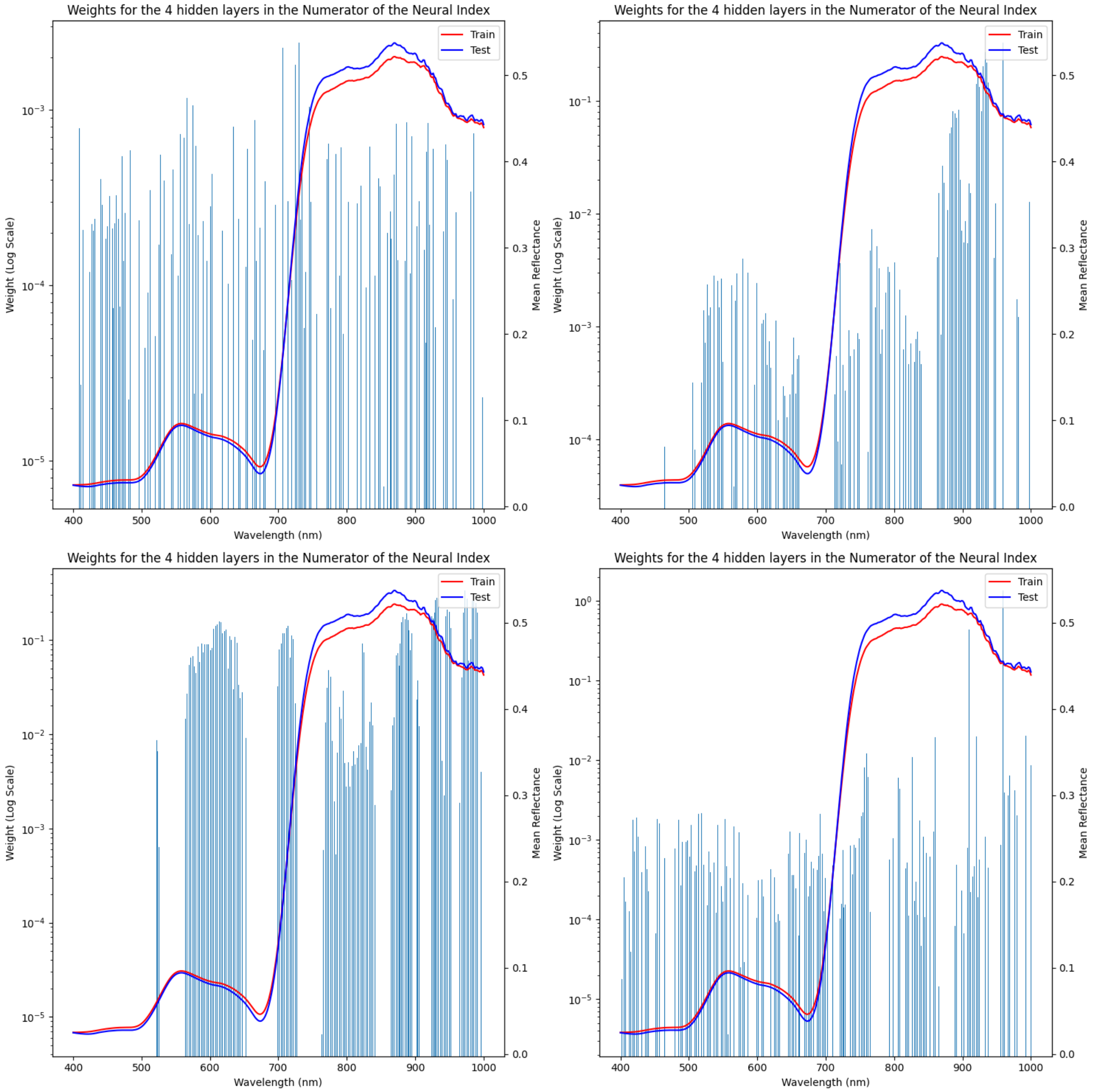


Figure 5: Neural vegetation index weights w, obtained after optimization.

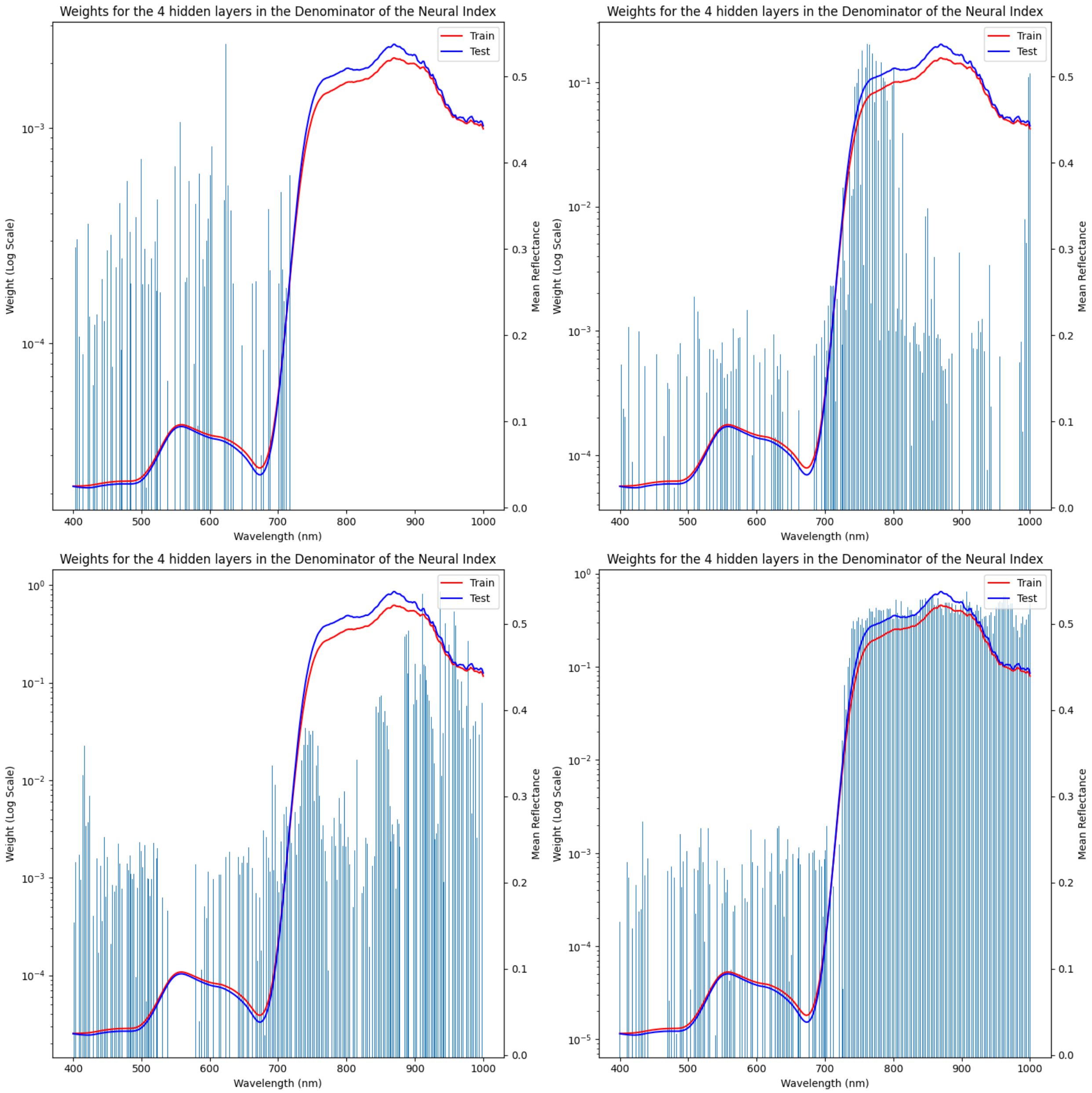


Figure 6: Neural vegetation index denominator weights, u, obtained after optimization.

DISCUSSION

In this study, advanced technologies such as machine learning and remote imaging spectroscopy were utilized to accurately predict the leaf chlorophyll concentration of genotypes in wild blueberry fields. We tested three different approaches: using all available spectral bands, selected vegetation indices, and dimensionality-reduced datasets as predictor variables. The findings showed that machine learning approaches outperformed PLSR when all spectral bands were used as input predictors. Moreover, using dimensionality reduction preprocessed data as input predictors further enhanced the ML model performance. Ultimately, the meta-learning model that combined the best-performing models achieved the highest level of predictive performance.

Upon analysis of the mean reflectance properties across various wavelengths, it was observed that the visible green and near-infrared regions of the electromagnetic spectrum tend to exhibit high reflectance. In contrast, the visible blue and red regions demonstrate high absorbance, which can be attributed to the absorption of light by chlorophyll in the visible range (Huete, 2004). Additionally, cellular structures such as cell walls and internal components are responsible for the strong reflectance in the NIR region (Huete, 2004). This also explains the high correlation pattern between LCC and wavelengths at different regions.

Partial Least Squares Regression (PLSR) has gained popularity as a statistical technique to establish relationships between hyperspectral reflectance and various biochemical factors in plants. However, it is important to note that its performance may significantly vary across different plant species, regions, and growth environments. In some cases, PLSR may yield highly accurate predictions, while in others, its performance may be suboptimal due to factors such as spectral interference, signal noise, and variations in plant physiology (Fu et al., 2019). This might be the reason for the low performance of PLSR that we see in our study.

The results of the machine learning models suggest the importance of dimensionality reduction to improve the high collinear hyperspectral data. Studies have found that selecting important bands for modeling through dimensionality reduction algorithms can lead to better model performance compared to using full-spectrum models (Wang et al., 2022). To enhance the performance of our model, we employed advanced techniques known as ensemble or meta-learning. This involved combining multiple weaker learners to create a stronger one. By doing so, we were able to leverage the strengths of each individual learner and compensate for their weaknesses, resulting in a more accurate and reliable model. This process allowed us to achieve superior results compared to using a single strong learner, as the ensemble technique reduces the risk of overfitting and increases the stability of the model. Overall, the use of ensemble/meta-learning proved to be a valuable strategy in improving our model's performance, which is parallel to some recent studies (Fu et al., 2019; Sterling & Di Rienzo, 2022).

Though we found a satisfactory performance from our meta-learner model, its real-life application to monitor the field-level spatial heterogeneity of LCC for precision agriculture could be complex due to high computational need.

Our proposed method of neural vegetation index achieves the highest validation R2 score of 0.76. During training, a neural network searches the space of functions that minimizes a given loss function. The general framework of NVI will also be tested across species and different nutrients. The optimal weights **w** and **u** are shown in Figure 5 and Figure 6. Note that weights focus on different parts of the spectrum but the weights for the infra-red spectrum are in general larger.

Photosynthetically active radiation (400 -770nm) absorption by leaves depends on photosynthetic pigment concentrations such as chlorophyll, which can impact the efficiency of CO^2 assimilation and primary production (Richardson et al., 2002). Accurate estimation of leaf chlorophyll content is required for monitoring vegetation stress, physiological conditions, and response to environmental factors (Croft et al., 2015). Integration of UAV-based remote sensing of hyperspectral data with machine learning shows promising results for estimating the spatial variability of leaf chlorophyll status.

ACKNOWLEDGMENT

We would like to thank Wyman’s for providing field access. We would also like to acknowledge Peter Nelson and Catherine Chan for helping with the data collection. This project was supported by the USDA National Institute of Food and Agriculture, Hatch Project Number ME0-22021, through the Maine Agricultural and Forest Experiment Station. This research was also supported by the Wild Blueberry Commission of Maine, the Maine Department of Agriculture, Conservation and Forestry (SCBGP; 20200918*0996), the NASA Maine Space Grant Consortium grant (SG-24-06), and the UMaine Faculty Summer Research Award.

REFERENCES

Barai, K., Calderwood, L., Wallhead, M., Vanhanen, H., Hall, B., Drummond, F., & Zhang, Y.-J. (2022). High Variation in Yield among Wild Blueberry Genotypes: Can Yield Be Predicted by Leaf and Stem Functional Traits? *Agronomy*, *12*(3), 617.

Croft, H., Chen, J. M., Zhang, Y., Simic, A., Noland, T. L., Nesbitt, N., & Arabian, J. (2015). Evaluating leaf chlorophyll content prediction from multispectral remote sensing data within a physically-based modelling framework. *ISPRS Journal of Photogrammetry and Remote Sensing*, *102*, 85–95. https://doi.org/10.1016/j.isprsjprs.2015.01.008

Fu, P., Meacham-Hensold, K., Guan, K., & Bernacchi, C. J. (2019). Hyperspectral Leaf Reflectance as Proxy for Photosynthetic Capacities: An Ensemble Approach Based on Multiple Machine Learning Algorithms. *Frontiers in Plant Science*, *10*. https://doi.org/10.3389/fpls.2019.00730

Gitelson, A. A., Gritz, Y., & Merzlyak, M. N. (2003). Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. *Journal of Plant Physiology*, *160*(3), 271–282.

Hu, J., Yue, J., Xu, X., Han, S., Sun, T., Liu, Y., Feng, H., & Qiao, H. (2023). UAV-Based Remote Sensing for Soybean FVC, LCC, and Maturity Monitoring. *Agriculture*, *13*(3), Article 3. https://doi.org/10.3390/agriculture13030692
Huete, A. R. (2004). 11—REMOTE SENSING FOR ENVIRONMENTAL MONITORING. In J. F. Artiola, I. L. Pepper, & M. L. Brusseau (Eds.), *Environmental Monitoring and Characterization* (pp. 183–206). Academic Press. https://doi.org/10.1016/B978-012064477-3/50013-8

Lu, S., Lu, F., You, W., Wang, Z., Liu, Y., & Omasa, K. (2018). A robust vegetation index for remotely assessing chlorophyll content of dorsiventral leaves across several species in different seasons. *Plant Methods*, *14*(1), Article 1.

Richardson, A. D., Duigan, S. P., & Berlyn, G. P. (2002). An evaluation of noninvasive methods to estimate foliar chlorophyll content. *New Phytologist*, *153*(1), 185–194. https://doi.org/10.1046/j.0028-646X.2001.00289.x

Sterling, A., & Di Rienzo, J. A. (2022). Prediction of South American Leaf Blight and Disease-Induced Photosynthetic Changes in Rubber Tree, Using Machine Learning Techniques on Leaf Hyperspectral Reflectance. *Plants*, *11*(3), Article 3. https://doi.org/10.3390/plants11030329

Wang, T., Gao, M., Cao, C., You, J., Zhang, X., & Shen, L. (2022). Winter wheat chlorophyll content retrieval based on machine learning using in situ hyperspectral data. *Computers and Electronics in Agriculture*, *193*, 106728. https://doi.org/10.1016/j.compag.2022.106728

Zhang, J., Huang, W., & Zhou, Q. (2014). Reflectance variation within the in-chlorophyll centre waveband for robust retrieval of leaf chlorophyll content. *PloS One*, *9*(11), e110812.

Zhu, J., Tremblay, N., & Liang, Y. (2012). Comparing SPAD and atLEAF values for chlorophyll assessment in crop species. *Canadian Journal of Soil Science*, *92*, 645–648. https://doi.org/10.4141/cjss2011-100