Estimating Chlorophyll Content from Hyperspectral Data Using Gradient Features*

Bogdan Ruszczak[1,4][0000−0003−1089−1778], Agata M. Wijata[2,4][0000−0001−6180−9979], and Jakub Nalepa[3,4][0000−0002−4026−1569]

[1] Faculty of Electrical Engineering, Automatic Control and Informatics, Department of Informatics, Opole University of Technology, Opole, Poland
[2] Faculty of Biomedical Engineering, Silesian University of Technology, Zabrze, Poland
[3] Faculty of Automatic Control, Electronics and Computer Science, Department of Algorithmics and Software, Silesian University of Technology, Gliwice, Poland
[4] KP Labs, Gliwice, Poland
b.ruszczak@po.edu.pl, {awijata, jnalepa}@ieee.org

Abstract. Non-invasive estimation of chlorophyll content in plants plays an important role in precision agriculture. This task may be tackled using hyperspectral imaging that acquires numerous narrow bands of the electromagnetic spectrum, which may reflect subtle features of the plant, and inherently offers spatial scalability. Such imagery is, however, high-dimensional, therefore it is challenging to transfer from the imaging device, store and investigate. We propose a machine learning pipeline for estimating chlorophyll content from hyperspectral data. It benefits from the Savitzky-Golay filtering to smooth the (potentially noisy) spectral curves, and from gradient-based features extracted from such a smoothed signal. The experiments revealed that our approach significantly outperforms the state of the art according to the widely-established estimation quality metrics obtained for four chlorophyll-related parameters.

Keywords: machine learning · chlorophyll content · feature engineering · hyperspectral image · regression.

1 Introduction

The agricultural sector has evolved over the years, in response to a growing demand for food, fiber and fuel [13]. The limited availability of land requires targeted management of resource production and leads to the increasing adoption of precision agriculture [17]. In this context, remote sensing can easily become a tool for identifying soil and crop parameters, due to its intrinsic scalability [17,

* This work was partially supported by The National Centre for Research and Development of Poland (POIR.04.01.04-00-0009/19). AMW was supported by the Silesian University of Technology, Faculty of Biomedical Engineering grant (07/010/BK_23/1023). JN was supported by the Silesian University of Technology Rector's grant (02/080/RGJ22/0026).

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_18

9]. In the case of agriculture, methods using multi- and hyperspectral remote sensing, capturing multispectral and hyperspectral images (MSI and HSI) are used, and non-invasive extraction of the chlorophyll content plays an increasingly important role, as it can ultimately lead to improving agricultural practices [7].

The majority of vegetation indices (VIs) were designed for multispectral sensors [14]. However, the wide bands in such imagery result in limited accuracy in the early detection of negative plant symptoms [1]. The use of HSIs, which are characterized by high spectral resolution, allows for the extraction of more details in the spectral response of an object [10]. Here, estimating chlorophyll content from hyperspectral data is commonly carried out by calculating the value of narrow-band VIs [19]. They include the Normalized Difference Vegetation Index (NDVI), Optimal Soil Adjusted Vegetation Index, Ratio Index and Difference Index [19]. Another parameter is the maximum quantum yield of photochemistry (Fv/Fm) [15]. Also, the Soil and Plant Analyzer Development (SPAD) tool is exploited, which measures the relative level of chlorophyll content in a crop taking into account the level of chlorophyll in the canopy. Finally, the performance index (PI) makes it possible to estimate the level of chlorophyll too [10].

Some approaches operate directly on the HSI data to estimate the chlorophyll content, hence they omit the stage of determining VIs. In such techniques, selected hyperspectral bands are analyzed—the encompass the Red and Near-Infrared (NIR) combination [6], the Blue, Green, Red, and NIR combined range [16], the Blue, Green and Red channels [17], or just the NIR band [4]. The bands undergo feature extraction (often followed by feature selection) in classic machine learning algorithms, whereas deep learning models benefit from automated representation learning over such data. The former group of techniques span across a variety of feature extractors and regression models, including continuous wavelet transforms (CWTs) [17], partial least square regression (PLSR) [18], kernel ridge regression [12] or regression using random forests. On the other hand, convolutional neural networks (CNNs) [14] and generative adversarial nets (GAN) [16] have been utilized for chlorophyll estimation as well.

Unfortunately, the data-driven algorithms are commonly validated over the in-house (private) data, following different validation procedures. This ultimately leads to the reproducibility crisis, and to inability to confront the existing approaches in a fair way [8]. In our recent work, we addressed this research gap and introduced a benchmark dataset, together with the validation procedure and a set of suggested metrics which should be used to quantify the generalization capabilities of machine learning models for chlorophyll estimation [11]. Here, we exploit this dataset and validation procedure to understand the abilities of the proposed processing chain, and to compare its estimation performance with 15 baseline models (for clarity, we focus on the best algorithms from [11]).

We tackle the problem of automated analysis of hyperspectral data using machine learning algorithms in the context of estimating the chlorophyll content (Sect. 2). We show that appropriately designed feature extractors fed into well-established supervised regression models can dramatically enhance their operational capabilities. Our experiments (Sect. 3), performed over the recent

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_18

dataset and following the validation suggested in [11], indicated that smoothing the spectral curves using the Savitzky–Golay filtering and extracting gradient-based features lead to significant improvements in the estimation quality when compared to the current state of the art. Also, we executed extensive computational experiments to understand the impact of the hyperparameter selection on the regression engine, and to optimize the hyperparameters of the system.

2 Materials and Methods

In this section, we summarize the dataset and the quantitative metrics used to assess the investigated algorithms (Section 2.1), together with our machine learning pipeline for estimating chlorophyll content from HSI (Section 2.2).

2.1 Dataset

We exploit the CHESS (**CH**lorophyll **ES**timation Data**S**et) dataset introduced in our recent study [11]—it was collected in the Plant Breeding and Acclimatization Institute —National Research Institute (IHAR-PIB) facility located in Central Poland (Jadwisin, Masovian Voivodeship) during the 2020 campaign (June—July, with three rounds of data acquisition, 4 weeks apart from each other). There were three flights over two sets of 12 plots resulting in 72 HSIs (150 bands, 460–902 nm, with the 2.2 cm ground sampling distance). In the plots, there were two potato varieties planted: *Lady Claire* (12 plots) and *Markies* (12 plots). The image data is accompanied with the in-situ measurements for each plot: (*i*) the SPAD, (*ii*) the maximum quantum yield of the PSII photochemistry (Fv/Fm), (*iii*) the performance of the electron flux to the final PSI electron acceptors, and (*iv*) relative water content (RWC), reflecting the degree of hydration of the leaf's tissue. In [11], we introduced the training-test split in which both subsets are equinumerous, and they are stratified following the distribution of each ground-truth parameter independently, so that both training and test subsets (each containing 36 plots) maintain a similar parameter's distribution.

To quantify the regression performance, we use the metrics, as suggested in [11]: the coefficient of determination R^2 which should be maximized ($\uparrow$; R^2 with one being the perfect score; its negative values indicate a worse fit than the average), mean absolute percentage error (MAPE), mean squared error (MSE) and mean absolute error (MAE)—all those measures should be minimized ($\downarrow$).

2.2 Estimating Chlorophyll Content Using Machine Learning

We exploit a processing chain, in which the input HSI undergoes feature extraction, and the features are fed to a regression model to predict the value of each parameter (we train four independent models). The algorithms at each step of the pipeline can be conveniently replaced by other techniques. We build upon several insights concerning the shape of the median spectral curves extracted for the separate fields of interest (the spectral curves for all pixels are aggregated to

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_18

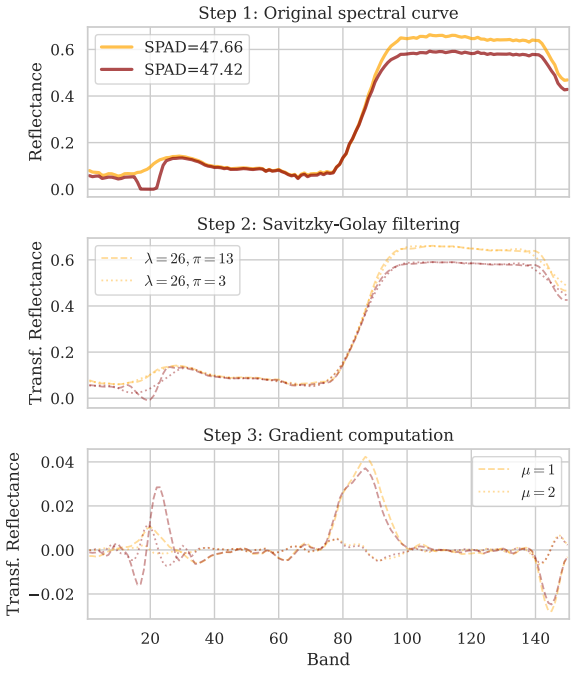

Fig. 1. The flowchart presenting the proposed feature engineering process.

generate a single median curve for the field). It is of note that some of the curves extracted for CHESS that should represent similar measurements (e.g., similar ground truth) do not look alike, and there exist noisy curves—this could be related to the difficulties in capturing enough light for selected spectral bands, which could easily lead to a narrower tonal range of the photosensitive camera.

To tackle the issue of the noisiness of spectral curves (and to increase its signal-to-noise ratio through removing high-frequency noise from the signal), the feature extraction stage is preceded by the filtering of spectral data using the Savitzky-Golay filter which may be considered as a generalized moving average filter. We aim at eliminating the influence of random noise and at reducing the drift phenomenon on the spectral reflection coefficient [5]. The Savitzky-Golay filter is given as the discrete convolution (h denotes the signal):

$$y[\pi] = \sum_{m=-\lambda}^{\lambda} h[m] * [\pi - m] = \sum_{m=\pi-\lambda}^{\pi+\lambda} h[\pi - m] * [m], \tag{1}$$

where $2\lambda+1$ is the length of the approximation interval, and π is the polynomial order, both being the tunable hyperparameters. In Fig. 1, we present the example of the original spectral curves for two fields with similar SPAD values amounting to 47.66 and 47.42 (first row) which underwent Savitzky-Golay filtering for two

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_18

hyperparameter sets ($\lambda = 26$, $\pi = 13$ and $\lambda = 26$, $\pi = 3$), resulting in slightly different shapes of the smoothed signal (second row). In both cases, the small "noisy" variations of the original curve were removed across the entire spectrum.

Once the original spectral curve is filtered, we extract the gradients which constitute the feature vectors for each field. We build upon the observation that the detection of anomalies in the signal data can be supported by the use of a gradient elaborated over the spectral curve, which allows us to capture the subtle characteristics in such data [2]. The feature extraction may be performed μ times, with $\mu = 0$ denoting the original curve (i.e., the feature vector includes the original reflectance values). Such feature vectors of size $\mathcal{B}$, where $\mathcal{B}$ is the number of hyperspectral bands (here, $\mathcal{B} = 150$), are fed to the regression model.

3 Experimental Validation

We investigate the linear machine learning models: (*i*) linear regressors with L_1, and (*ii*) L_2 regularization, (*iii*) support vector machines with linear kernel, and (*iv*) the elastic net with regularization, all implemented in `Scikit-learn`. Savitzky-Golay filtering was implemented in `SciPy`, and feature extraction in `NumPy`. We focus on the linear regression models to avoid heavily parameterized techniques—due to this assumption, we were able to extensively evaluate thousands of models in a reasonable time (27 840 model's configurations). At the same time, we maintained the high interpretability of the study.

For each model, we optimized its hyperparameters (this fine-tuning was performed following the 5-fold cross-validation strategy over the training set):

- Linear regression with L1 regularization, for: $\alpha \in \{10^{-15}, 10^{-14}, \ldots, 10^{15}\}$,
- Linear regression with L2 regularization, for: $\alpha \in \{10^{-15}, 10^{-14}, \ldots, 10^{15}\}$,
- Support vector machine with linear kernel, for: $C \in \{2^{-5}, 2^{-4}, \ldots, 2^{8}\}$, and the maximum number of iterations $\mathcal{I} \in \{500, 1000, 2500\}$,
- Elastic net with L_1 regularization, for $\alpha \in \{10^{-15}, 10^{-14}, \ldots, 10^{15}\}$, and $L_{1ratio} \in \{0.05, 0.1, \ldots, 0.9\}$.

Similarly, the feature extraction is a parameterized step, as it may be performed multiple times. We denoted the number of gradient runs as μ, and by $\mu = 0$ we report the results obtained for the regression models operating over the original curves. Therefore, the full configuration for the performed experimental search was as follows: $\mu \in \{0, 1, 2, 3\}, \lambda \in \{1, 2, \ldots 30\}, \pi \in \{1, 2, \ldots, 29\}$.

In Fig. 2, we depict the impact of π, λ, and μ on the R^2 coefficient for the elastic net models predicting the SPAD parameter. Here, we focused on a single machine learning model (with default parameterization) to verify the importance of signal filtering and feature extraction on the regression capabilities of the algorithm. Albeit the insights learned from this experiment may not be generalizable to other models, we anticipate that a similar trend would be observed, as feature engineering constitutes one of the most important aspects of building machine learning pipelines [3]. We can observe that the exhaustive traversal of the search spaces allows to indicate their most promising regions for the (λ, π)

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_18

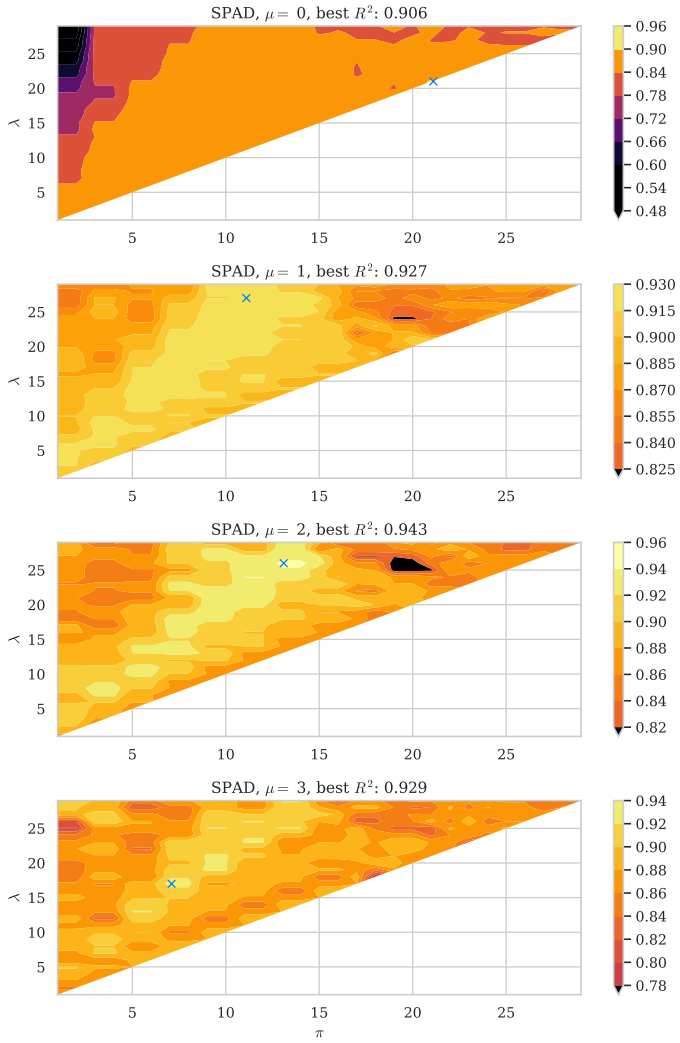

Fig. 2. The R^2 metric obtained for various filtering and feature extraction settings. The light blue $\times$ marker indicates the best configuration for each gradient level (μ).

configurations, which remain consistent for the gradient-extraction levels (μ). However, the exact position of the best parameterization differs across μ's.

The optimized models outperform *Baseline* (the best-known R^2 values from the literature [11])—in Table 1, we report the optimized hyperparameters for the models offering the best regression. The R^2 measure notably increased for all parameters, whereas the regression errors, e.g., MAE, decreased by 53.5%, 23.3%, 30.9%, and 25.0% for SPAD, FvFm, PI, and RWC. The experiments

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_18

indicated that appropriate feature engineering, which involves Savitzky-Golay filtering of the median spectral curves followed by feature extraction, allows the elaboration of high-quality models for estimating chlorophyll-related parameters.

Table 1. The best results for all quality metrics, elaborated for the best parameterization (Savitzky-Golay filtering, feature extraction and regression models).

Param.	Model configuration	μ	π	λ	MAPE ↓	R^2 ↑	MAE ↓	MSE ↓
SPAD	Elastic net ($\alpha = 10^{-1}$, $L_{1ratio} = 0.5$)	2	26	12	0.035	0.943	0.756	3.012
	Linear regr. with L_2 ($\alpha = 2.5 \times 10^{-5}$)		—		0.072	0.827	1.625	9.095
FvFm	Linear regr. with L_2 ($\alpha = 10^{-3}$)	0	22	17	0.030	0.764	0.016	0.001
	Linear regr. with L_2 ($\alpha = 5 \times 10^{-5}$)		—		0.036	0.727	0.021	0.001
PI	SVM with linear kernel ($C = 2$, $\mathcal{I} = 10^3$)	1	29	20	0.401	0.837	0.194	0.083
	Linear regr. with L_2 ($\alpha = 10^{-11}$)		—		0.532	0.677	0.280	0.169
RWC	Linear regr. with L_1 ($\alpha = 10^{-1}$)	2	22	11	0.010	0.911	0.706	1.089
	Linear regr. with L_2 ($\alpha = 10^{-3}$)		—		0.013	0.859	0.941	1.731

4 Conclusions

We exploited HSIs for the non-invasive estimation of chlorophyll-related parameters in plants and proposed a machine learning technique for this task. To deal with subtle signal noise, we utilized the Savitzky-Golay smoothing filter that is followed by the gradient-based feature extractor. The experiments revealed that our techniques outperform the state of the art, as quantified using four chlorophyll-related parameters. The coefficient of determination (R^2) achieved by our techniques reached 0.943 (compared to 0.827 reported for the best model in [11], therefore we obtained the improvement of 14%), 0.764 (0.727, improvement of 5%), 0.837 (0.667, improvement of 25%), 0.911 (0.859, improvement of 6%) for SPAD, FvFm, PI, and RWC. Also, we showed that employing the Savitzky-Golay smoothing brings improvements in the generalization of a model trained over the gradient-based features extracted from such filtered signal.

References

1. Adão, T., Hruška, J., Pádua, L., Bessa, J., Peres, E., Morais, R., Sousa, J.J.: Hyperspectral imaging: A review on UAV-based sensors, data processing and applications for agriculture and forestry. Remote Sensing **9**(11) (2017)
2. Chen, J., Lu, M., Chen, X., Chen, J., Chen, L.: A spectral gradient difference based approach for land cover change detection. ISPRS Journal of Photogrammetry and Remote Sensing **85**, 1–12 (2013)
3. Chicco, D., Oneto, L., Tavazzi, E.: Eleven quick tips for data cleaning and feature engineering. PLOS Computational Biology **18**(12), 1–21 (12 2022)
4. Gorretta, N., Nouri, M., Herrero, A., Gowen, A., Roger, J.M.: Early detection of the fungal disease "apple scab" using SWIR hyperspectral imaging. In: 2019 10th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS). pp. 1–4 (2019)

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_18

5. Guo, C., Liu, L., Sun, H., Wang, N., Zhang, K., Zhang, Y., Zhu, J., Li, A., Bai, Z., Liu, X., Dong, H., Li, C.: Predicting Fv/Fm and evaluating cotton drought tolerance using hyperspectral and 1D-CNN. Frontiers in Plant Science **13**, 3700 (oct 2022)
6. Huynh, N.H., Böer, G., Schramm, H.: Self-attention and generative adversarial networks for algae monitoring. European Journal of Remote Sensing **55**(1), 10–22 (2022)
7. Jin, X., Li, Z., Feng, H., Ren, Z., Li, S.: Deep neural network algorithm for estimating maize biomass based on simulated sentinel 2A vegetation indices and leaf area index. The Crop Journal **8**(1), 87–97 (2020)
8. Nalepa, J., Myller, M., Kawulok, M.: Validating hyperspectral image segmentation. IEEE Geoscience and Remote Sensing Letters **16**(8), 1264–1268 (2019)
9. Ponnusamy, V., Natarajan, S.: Precision agriculture using advanced technology of iot, unmanned aerial vehicle, augmented reality, and machine learning. In: Gupta, D., Hugo C. de Albuquerque, V., Khanna, A., Mehta, P.L. (eds.) Smart Sensors for Industrial Internet of Things: Challenges, Solutions and Applications. pp. 207–229. Springer International Publishing, Cham (2021)
10. Ruszczak, B., Boguszewska-Mańkowska, D.: Deep potato – the hyperspectral imagery of potato cultivation with reference agronomic measurements dataset: Towards potato physiological features modeling. Data in Brief **42**, 108087 (2022)
11. Ruszczak, B., Wijata, A.M., Nalepa, J.: Unbiasing the estimation of chlorophyll from hyperspectral images: A benchmark dataset, validation procedure and baseline results. Remote Sensing **14**(21) (2022)
12. Singhal, G., Bansod, B., Mathew, L., Goswami, J., Choudhury, B., Raju, P.: Estimation of leaf chlorophyll concentration in turmeric (curcuma longa) using high-resolution unmanned aerial vehicle imagery based on kernel ridge regression. Journal of the Indian Society of Remote Sensing **47**, 1–12 (03 2019)
13. Sishodia, R.P., Ray, R.L., Singh, S.K.: Applications of remote sensing in precision agriculture: A review. Remote Sensing **12**(19) (2020)
14. Wang, J., Xiao, X., Bajgain, R., Starks, P., Steiner, J., Doughty, R.B., Chang, Q.: Estimating leaf area index and aboveground biomass of grazing pastures using Sentinel-1, Sentinel-2 and Landsat images. ISPRS Journal of Photogrammetry and Remote Sensing **154**, 189–201 (2019)
15. Wen, S., Shi, N., Lu, J., Gao, Q., Yang, H., Gao, Z.: Estimating Chlorophyll Fluorescence Parameters of Rice (Oryza sativa L.) Based on Spectrum Transformation and a Joint Feature Extraction Algorithm. Agronomy **13**(2) (2023)
16. Yan, T., Xu, W., Lin, J., Duan, L., Gao, P., Zhang, C., Lv, X.: Combining Multi-Dimensional Convolutional Neural Network (CNN) With Visualization Method for Detection of Aphis gossypii Glover Infection in Cotton Leaves Using Hyperspectral Imaging. Frontiers in Plant Science **12** (2021)
17. Yue, J., Zhou, C., Guo, W., Feng, H., Xu, K.: Estimation of winter-wheat aboveground biomass using the wavelet analysis of unmanned aerial vehicle-based digital images and hyperspectral crop canopy images. International Journal of Remote Sensing **42**(5), 1602–1622 (2021)
18. Zhang, J., Sun, H., Gao, D., Qiao, L., Liu, N., Li, M., Zhang, Y.: Detection of canopy chlorophyll content of corn based on continuous wavelet transform analysis. Remote Sensing **12**(17) (2020)
19. Zhang, Y., Xia, C., Zhang, X., Cheng, X., Feng, G., Wang, Y., Gao, Q.: Estimating the maize biomass by crop height and narrowband vegetation indices derived from UAV-based hyperspectral images. Ecological Indicators **129**, 107985 (2021)

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_18