Lux junior 2024

17. Internationales Forum für den lichttechnischen Nachwuchs
Ilmenau, 6. – 8. September 2024
DOI (proceedings): 10.22032/dbt.61636

DOI: 10.22032/dbt.xxxxx

Challenges in Modelling Photosynthesis as a Function of Spectral Photon Flux Density

Adrian Zwenger, Jens Balasus, Stefan Klir, Prof. Tran Quoc Khanh

Technische Universität Darmstadt, Laboratory of Adaptive Lighting Systems and Visual Processing, Hochschulstraße 4a, 64289 Darmstadt

adrian.zwenger@stud.tu-darmstadt.de

Abstract

To establish a sustainable and efficient crop cultivation, independent of season and climate, artificial lighting remains a critical factor. Artificial lighting can incur a significant environmental and financial cost. Thus, a key challenge is reducing power consumption for lighting without compromising crop quality and yield. In other terms, lighting needs to be optimized with regard to net photosynthetic activity, which is directly linked to crop quality and yield.

Determining an optimal lighting strategy is not trivial, as the photosynthetic process and its influencing factors have not been fully understood yet. Furthermore, interaction between the influencing factors is not only possible, but can occur non-linearly. This is the reason why using data-driven approaches and machine learning to model photosynthetic behaviour as an abstract black-box has become one of many focal points in photosynthesis research.

Currently, studying photosynthesis requires the examination of the net CO_2-assimilation-rate, which is labour- as well as resource-intensive. The aim of this work is to propose a framework for modelling the photosynthetic activity of C3- and C4-plants as a function of the spectral photon flux density, with the goal of eliminating the assimilation-rate measurement altogether. Challenges arising from the biochemical complexities are highlighted and possible workarounds determined. A special focus is placed on data pre-processing and augmentation steps, as well as proposed machine learning approaches.

Index Terms: Controlled-environment agriculture (CEA), photosynthesis, machine learning, UMAP, HDBSCAN

1 Introduction

Controlled-environment agriculture (CEA) is a form of agriculture that is optimised for year-round crop production regardless of location or season. It promises to be environmentally friendlier than traditional farming methods. This is achieved by restricting the farming process to the confines of the CEA, which limits the influence of environmentally detrimental farming side-effects [1]. Furthermore, production efficiency and product quality can be fine-tuned by adjusting relevant crop development factors within the CEA. These include, but are not limited to, temperature, humidity, nutrient supply, and artificial lighting or supplementation of natural light. It has been shown that this level of control can lead to larger yields and shorter production cycles in comparison to traditional field-based cultivation [2]. This, however, comes with an increased energy consumption that is significantly impacted by artificial lighting [3]. To improve the environmental and economical sustainability of CEA's, the landed costs and the energy consumption need to be decreased. This could be achieved by finding new lighting strategies that increase crop yield and quality while decreasing energy consumption. To address the above strategies, a simple verification method is required. The photosynthetic activity of a target crop can be used as a proxy to inspect a strategy's efficiency, as photosynthesis is directly linked to a plant's growth and nutrient development [4]. Choosing an optimal strategy thus is equivalent to choosing the lighting strategy that has the highest photosynthetic activity to energy consumption ratio.

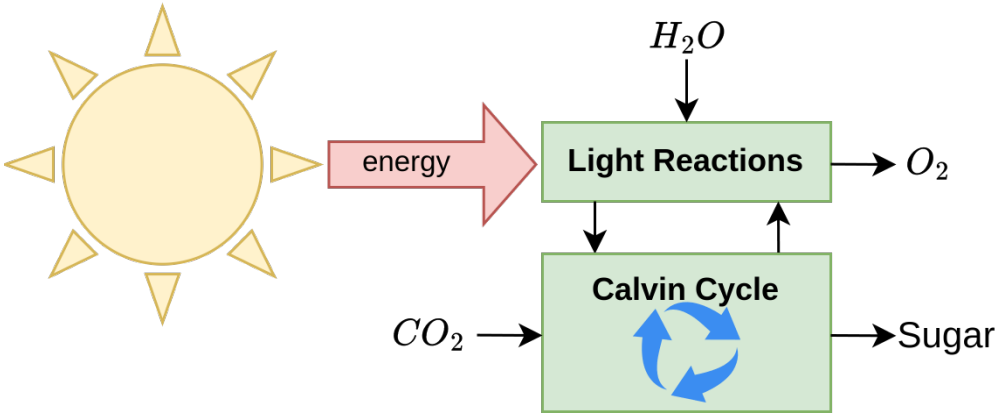

Figure 1: *This figure abstractly visualizes photosynthesis as a cross-defendant process of its light reactions and the Calvin cycle.*

Photosynthesis oversimplified is a biochemical process that fixates carbon dioxide (CO_2) to produce organic molecules, like starch or sugar, and releases oxygen (O_2) using energy absorbed from light, as is visualized in **Figure 1**. This process is comprised of the light reactions and the light independent Calvin Cycle. The light reactions and the Calvin Cycle depend on each other's products. If the plant is photosynthetically active, CO_2 is fixated within the plant and O_2 is released.

Conversely, if the plant is not photosynthetically active, the plant can no longer fixate CO_2 and O_2 can no longer be released.[5]

The rate at which CO_2 is fixated within the plant is called the CO_2-assimilation-rate and can be used as a direct physical proxy for a plant's photosynthetic activity. This is the main measurement principle of the gas exchange measurement method. While gas exchange measurements offer accurate estimates of photosynthetic activity, it is a technically complex, laborious, and expensive process which makes this method unsuitable for many use cases. Thus, finding alternative measurement principles to overcome the limitations of gas exchange measurement has become a focus of recent research. [6]

This work proposes a semi-automated, data-driven framework, with the aim to simplify photosynthetic activity measurements by replacing the CO_2-assimilation-rate measurement with machine learning (ML) based black-box models. To deal with the complexity in modelling photosynthesis inherent to the various expressions of the biochemical processes across different plant species, dimensionality reduction and manifold learning techniques are employed. The results are then leveraged to fine-tune and train a single interspecies model or a set of independent models for the different species or identified manifold sub-spaces.

2 State Of The Art In ML-Based Photosynthesis Modelling

There have been many different works on leveraging machine learning to model photosynthesis. For example, Gao et al. [7] modelled the leaf-level photosynthesis rate of cucumbers as a function of light intensity, CO_2-concentration and the ratio of red light to the total amount of photosynthetically active radiation (PAR). Support Vector Machines (SVM), Random Forests (RF), and a non-linear regression model based on the Radial Basis Function (RBF) were chosen. While Zhang et al. [8] used cucumbers as well, Artificial Neural Networks (ANN) were chosen as the ML model. Instead of cucumbers, Yang et al. [9] have chosen to model the photosynthetic rate for grapes. This work is of particular interest as the authors used a hyperspectral camera to attain spectral information in addition to chlorophyll fluorescence measurements. Bayesian Neural Network (BNN) and Partial Least Squares (PLS) regression models were created. While the previous works leveraged ML for leaf-level modelling, Wu et al. [10] used various vegetation indices in combination with multispectral imaging to model photosynthesis of rice on the canopy-level. The spectral images were taken by flying unmanned vehicles over rice fields. To train and verify their linear regression, SVM, Gradient Boosted Decision Tree (GBDT), RF and NN models, leaf-level gas exchange measurements were taken and then extrapolated to the complete rice fields. Of these models GBDT performed the best overall. A similar approach was chosen by Heckmann et al. [11], however, instead of leveraging spectral imaging to model

photosynthesis for a single plant species, the authors inspected multiple plant species. Using Principal Component Analysis (PCA), it was shown that the leaf reflectance spectra of the various species of interest lie in a low-dimensional space. Here it was shown that the reflectance spectra of C3 and C4 species are similar and that the spectra of CAM species were clearly distinct from the C3 and C4 species. Interspecies photosynthesis NN and PLS models were then trained and evaluated. Noteworthy, however, is that the authors results show that the intraspecies models performed worse than models trained for each inspected plant species independently. Similar to the previous approaches, Fu et al. [12] leveraged hyperspectral imaging to acquire the leaf reflectance spectrum. Using this spectrum, tobacco's maximum carboxylation rate of Rubisco and the maximum electron transport rate supporting Rubisco regeneration were estimated and used as a proxy for the photosynthesis rate. However, instead of training a single ML model a framework to create a mixed model based on combining an ensemble of ANN, SVM, LASSO, RF, PLS and Gaussian Process (GP) models with a stacked approach was created. The regression results of the ensemble where then used as input for the stacked regressor which in turn outputs the final prediction.

Whereas the previous works focused on leveraging ML methods without a priori knowledge about the biochemical expression of the photosynthetic process, Kaneko et al. [13] combined an Artificial Neural Network (ANN) model with mechanistic leaf-level photosynthesis models, such as the models defined by Farquhar et al. [14], to estimate canopy level photosynthesis of spinach. The mechanistic models are used to estimate the leaf-level photosynthetic rate. This and the Leaf Area Index (LAI) are used together as inputs for the ANN model. The authors claim that this combination of modelling approaches results in superior generalization capabilities and overcomes the shortcomings of regular ML models that usually show low predictability outside of the training data range. However, to be able to estimate the leaf-level photosynthesis, empirical constants specific to each plant species need to be determined. In the context of this work, the approach of Kaneko et al. [13] poses a feasible alternative to the framework proposed, however, this is only the case, if the constants are already known. Otherwise, extensive empirical studies of the plant species in question would be required, significantly increasing the complexity of photosynthesis measurements.

The current state of the art approach to model photosynthesis using machine learning as described in the literature can be summarized as follows: Often a single plant species [7], [8], [9], [10], [12], [13] is selected for which a photosynthesis model is to be created. Leafy green plants such as cucumbers [7], [8], grapes [9] or spinach [13] are frequently chosen. Then the photosynthetic rate is measured either at the leaf-, whole-plant-, or canopy-level. Whole-plant- and canopy-level measurements are performed by using large enough gas chambers [13] or by taking leaf-level gas-exchange measurements and extrapolating them [10]. The acquired data is split into training and test datasets. ML models are optimized on the training dataset and often evaluated by calculating the determination coefficient (R^2) as well as the root mean squared error (RMSE) as metrics for model selection [7], [8], [9], [11], [12], [13].

An ML approach combined with hyperspectral imaging [9], [10], [11], [12] and photosynthetic photon flux density (PPFD) measurements [13], can be effectively utilized for photosynthetic activity estimation. However, building a generalizable model is challenging due to interspecies differences in biochemical process expression. Additionally, measuring photosynthetic activity can be difficult, because measuring it directly via the assimilation rate is impractical and not always feasible. Therefore, a unified modelling framework, which aims at building models that simplify photosynthetic rate measurements, is needed. Furthermore, this framework needs to be able to accommodate the variations in data structure inherent to the biochemical differences between the plant species of interest.

3 Framework Proposal

To overcome the previously mentioned challenges, the proposed framework takes interspecies leaf reflectance spectra, ambient light spectra, chlorophyll fluorescence, and PPFD measurements as input and leverages dimensionality reduction and clustering to create a generalisable estimator. The chlorophyll fluorescence is of interest, as it can be used as a proxy to determine the electron transport rate which is closely linked to the photosynthetic activity [15].

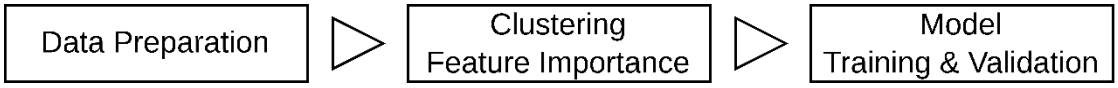

Figure 2: *Simplified Framework Structure*

The proposed framework roughly consists of three steps as is shown in **Figure 2**. First the measurements are aggregated into a single dataset and the spectral information is pre-processed. The dataset is then transformed into a low-dimensional space for data exploration using Uniform Manifold Approximation and Projection (UMAP) [16]. The transformed data is then clustered using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [17]. From this clustering the feature importance can be determined by fitting a classifier and determining the permutation feature importance. Having determined the feature importance, the dataset's dimensionality can be significantly reduced by dropping unimportant features before training the model. Finally, the regression model of choice can be trained and validated on either the interspecies dataset, datasets separated by species, or datasets separated by cluster label depending on the resulting model's performance. The procedure for identifying important features is depicted in **Figure 3**. The following subsections will elaborate on how the spectral data can be pre-processed and will introduce UMAP and HDBSCAN briefly.

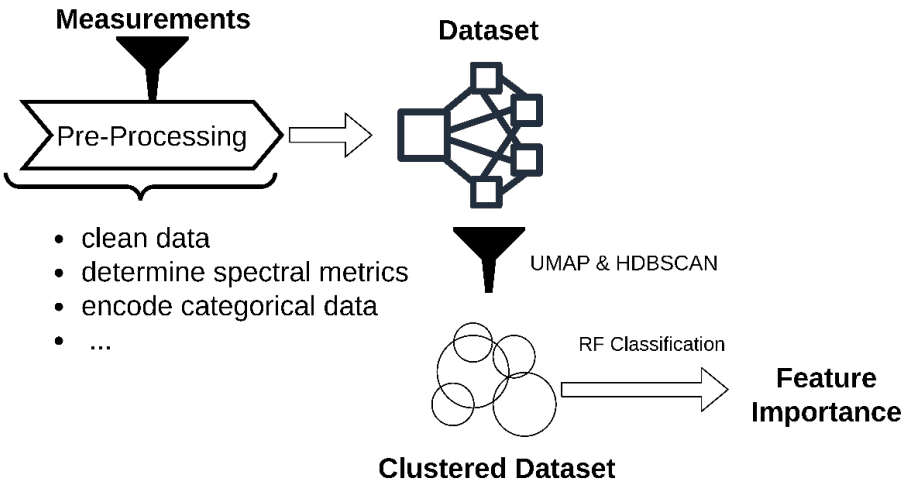

Figure 3: *This image visualises the framework process for identifying latent subspaces and determining a feature importance.*

Table 1: *Visualisation of dataset structure prior to pre-processing. Each row is a datapoint. The leaf reflectance and ambient light spectrum measurements are stored as row vectors in which columns store a metric at a certain wavelength. Here wavelengths between 380 nm and 780 nm are shown.*

			Leaf Reflectance Spectrum			Ambient Light Spectrum		
Species	PPFD	Chlorophyll Fluorescence	380 nm	...	780 nm	380 nm	...	780 nm
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...

3.1 Pre-processing spectra

The spectra obtained from reflectance or ambient light measurements are typically represented as a row vector per measurement. Each entry in these vectors corresponds to a specific metric at a particular wavelength such as the photon flux density. This leads to a dataset as is visualized in ***Table 1***. It includes all scalar valued measurements such as PPFD and the chlorophyll fluorescence, a column storing which plant species was measured for documentation purposes, as well as the leaf reflectance and ambient light spectrum measurements.

Using the spectrum vectors directly to train the machine learning model places the learning problem in a high dimensional space due to the high number of features in the

dataset. For example, the dataset in ***Table 1*** would consist of a total number of 163 features if a spectrometer with a resolution of 5 nm between 380 nm and 780 nm were to be used. The learning problem's high dimensionality can lead to a decrease in model efficiency and effectiveness. Furthermore, real world datasets tend to become sparser with increasing dimensionality. This exponentially increases the needed amount of data to discern meaningful pattern from noise. This phenomenon is known as the curse of dimensionality. To combat this, the spectra need to be represented in a low dimensional space. This can be achieved by applying varying weight functions to filter the spectrum and then calculating the ratio of the integrated weighted spectrum to the integrated spectrum. This way, a spectrum can be represented with k ratios denoted as R_k in the following. Let $S(\lambda)$ be the spectrum and $S_k(\lambda)$ be the weighted spectrum as functions of the wavelength λ then the ratio R_k can be written as:

$$R_k = \frac{\int_{-\infty}^{\infty} S_k(\lambda)\partial\lambda}{\int_{-\infty}^{\infty} S(\lambda)\partial\lambda}$$

In this framework normalised probability density functions of normal distributions are chosen as weighting functions to calculate S_k. Let $w_k(\lambda, \mu_k, \sigma_k)$ be the k-th weighting function with the centre wavelength μ_k and the standard deviation σ_k, which can thus be written as:

$$\mathrm{w_k}(\lambda, \mu_\mathrm{k}, \sigma_\mathrm{k}) = \frac{\mathrm{f}(\lambda, \mu_\mathrm{k}, \sigma_\mathrm{k})}{\mathrm{C}(\lambda)} = \frac{1}{\mathrm{C}(\lambda)\sigma_\mathrm{k}\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{\lambda-\mu_\mathrm{k}}{\sigma_\mathrm{k}}\right)^2}$$

The function $C(\lambda)$ normalises the values to ensure that $\sum R_k = 1$. This is achieved by choosing $C(\lambda)$ in such a way that the sum of all $f_k(\lambda, \mu_k, \sigma_k)$ evaluated at any λ equals to one:

$$\mathrm{C}(\lambda) = \sum_k f(\lambda, \mu_k, \sigma_k)$$

An example for a set of weighting functions is visualized in **Figure 4**. Here the functions were chosen to reduce the spectrum into ratios for blue, green and red light. For this, central wavelengths of 440 nm for blue, 570 nm for green and 710 nm for red was chosen. A standard deviation of 45 nm was chosen to visualize the effects of $C(\lambda)$. However, the spectra from measurements are not present as continuous functions, so R_k needs to be discretised. When measurements at N wavelengths λ_i are present, the ratio R_k is the scalar product of the weights vector $\overrightarrow{\mathrm{w_k}}$ and the spectrum vector $\vec{S}$ divided by the sum of elements in $\vec{S}$.

$$R_k = \frac{\sum_{i=1}^{N} w_k(\lambda_i, \mu_k, \sigma_k) S(\lambda_i)}{\sum_{i=1}^{N} S(\lambda_i)} = \frac{\overrightarrow{w_k} \cdot \vec{S}}{\sum_n \overrightarrow{S_n}}.$$

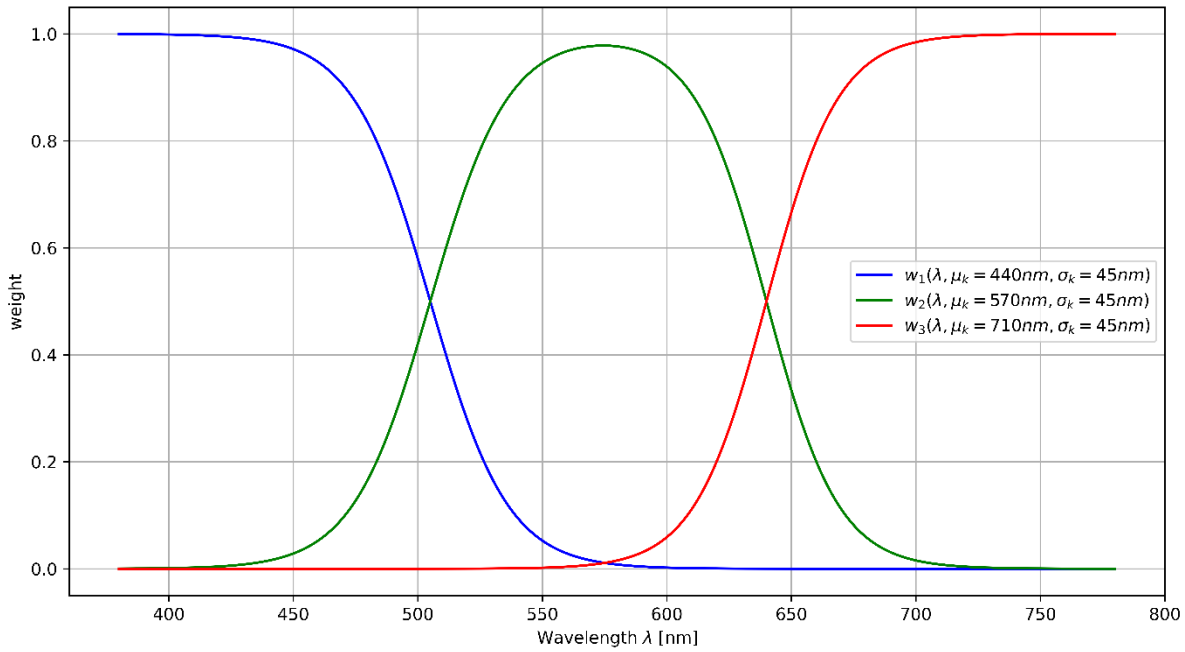

Figure 4: *Example weighting functions w_k for central wavelengths 440 nm, 570 nm and 710 nm with a chosen standard deviation of 45 nm.*

Now that the spectra have been pre-processed, the complete dataset as is visualised in ***Table 1*** can be explored to infer a feature importance, with which the curse of dimensionality can be further mitigated by dropping unimportant features.

3.2 Non-linear dimensionality reduction with UMAP

The proposed framework utilizes UMAP [16], a manifold learning technique for dimensionality reduction. UMAP preserves non-linear relationships in the data. This can lead to improved pattern identification and visualisation in biochemical applications [18], which is why UMAP was chosen over classical dimensionality reduction techniques like PCA. Another advantage of UMAP over alternative non-linear dimensionality reduction techniques, like T-distributed Stochastic Neighbour Embedding (t-SNE) [19], is that it is able to preserve more global data structure while offering superior runtime performance [16]. This is achieved by assuming that the data is uniformly distributed on some high-dimensional Riemannian manifold, that the Riemannian metric is locally approximately constant, and that the manifold is locally connected. In simpler words, the datapoints can be placed uniformly on some high dimensional surface, where the shape of said surface does not change much and no sharp edges, breaks or other distortions are present, and the distance between datapoints on said surface can be measured with a given distance metric. In this framework using the Euclidean distance metric for this step is proposed as the reflectance spectra lie in a low-dimensional space across plant species [11]. This

choice needs to be verified in future works as other distance metrics, like the Mahalanobis distance [20], might lead to better representations when compared to using the Euclidean distance. After placing the points on said surface, a nearest neighbour graph representation of the data is constructed. The topology of this graph is explored and a low-dimensional representation that best describes it is created and optimized. The result is often a two-dimensional or three-dimensional point cloud that represents the relationships present between the datapoints of the dataset.

3.3 Clustering with HDBSCAN

The result of applying UMAP to the dataset is a vector space in which similar datapoints are placed closer to each other than to dissimilar datapoints. This satisfies the core assumption of connectivity-based (hierarchical) clustering methods. Furthermore, if groups of similar datapoints are placed close to each other, then the density of points within a group is greater than its sparse surrounding. Thus, such groups could be identified using density-based clustering methods as well. As both approaches can be used in this case, Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [17], a density-based hierarchical clustering method, is chosen. This method applies the DBSCAN clustering method [21] over a range of values for the maximum distance between datapoints for them to be considered to be part of the same cluster. From these results the clusters are found that remain stable, which allows HDBSCAN to be more robust to parameter selection than DBSCAN and to find clusters with varying densities [22]. After segmenting the dataset into clusters, classification methods can be employed to identify the most important features that distinguish these clusters from one another.

3.4 Inferring feature importance with RF-classification

After reducing the dataset's dimensionality and grouping similar data points, it is crucial to identify what makes each group unique. Then the dataset can be further processed for model training with regards to this information. This can be achieved with descriptive statistics and classical methods like ANOVA [23]. However, this often requires manual interaction with the data and the results need to be manually interpreted as well. Therefore, an automated machine learning approach is preferable within the context of the proposed framework. A straightforward and effective method to achieve this is to label the datapoints with the cluster IDs and to fit and inspect a classification model. The classifiers permutation feature importance can then be used as the overall feature importance for the photosynthetic rate regression model. To determine the feature importance the classifier is fitted to the complete labelled dataset. As the goal is to identify what makes the clusters unique rather than building a generalizable classifier, overfitting is not an issue but rather something desirable.

After the RF has been fitted, values of a feature are shuffled and the resulting degradation in classification performance determined. The more the performance degrades the more important a feature is. Now knowing which features contribute to the differences of clusters within the dataset, it can be used for feature selection and dimensionality reduction as part of pre-processing the data before training and evaluating the black-box photosynthesis model.

4 Conclusion & Outlook

In conclusion, this paper illuminates the current state of machine learning approaches to modelling photosynthesis. Additionally, a framework for creating interspecies photosynthesis models as a function of PPFD has been proposed. This framework includes data pre-processing, dimensionality reduction, feature importance inference, and model training and selection steps. While this framework holds promise as a useful tool for developing photosynthesis models, its performance remains to be evaluated, necessitating data collection. Particularly, the side-effects of the chosen ML methods must be considered during framework performance evaluation. A known issue that might arise stems from the choice of ML methods used to determine the feature importance. Combining manifold learning for projecting the dataset into a lower-dimensional space with clustering might lead to segmentation of clusters within the dataset. This segmentation artificially increases the number of clusters and might bias the determined feature importance detrimentally. This could potentially be combatted by increasing the dimensionality of the lower-dimensional representation. As this, however, defeats the goal of reducing dataset dimensionality a trade-off will need to be found. Future works include data collection in cooperation with the biology faculty and framework evaluation.

5 References

[1] J. Luo, B. Li, and C. Leung, ‘A Survey of Computer Vision Technologies In Urban and Controlled-environment Agriculture’, Oct. 12, 2023, *arXiv*: arXiv:2210.11318. doi: 10.48550/arXiv.2210.11318.

[2] M. Gargaro, R. J. Murphy, and Z. M. Harris, ‘Let-Us Investigate; A Meta-Analysis of Influencing Factors on Lettuce Crop Yields within Controlled-Environment Agriculture Systems’, *Plants (Basel)*, vol. 12, no. 14, p. 2623, Jul. 2023, doi: 10.3390/plants12142623.

[3] C. F. Nicholson, M. Eaton, M. I. Gómez, and N. S. Mattson, ‘Economic and environmental performance of controlled-environment supply chains for leaf

lettuce', *European Review of Agricultural Economics*, vol. 50, no. 4, pp. 1547–1582, Sep. 2023, doi: 10.1093/erae/jbad016.

[4] A. Wingler and R. Henriques, 'Sugars and the speed of life—Metabolic signals that determine plant growth, development and death', *Physiol Plant*, vol. 174, no. 2, p. e13656, 2022, doi: 10.1111/ppl.13656.

[5] A. Stirbet, D. Lazár, Y. Guo, and G. Govindjee, 'Photosynthesis: basics, history and modelling', *Annals of Botany*, vol. 126, no. 4, pp. 511–537, Sep. 2020, doi: 10.1093/aob/mcz171.

[6] M. H. Siebers, N. Gomez-Casanovas, P. Fu, K. Meacham-Hensold, C. E. Moore, and C. J. Bernacchi, 'Emerging approaches to measure photosynthesis from the leaf to the ecosystem', *Emerging Topics in Life Sciences*, vol. 5, no. 2, pp. 261–274, 2021.

[7] P. Gao and J. Hu, 'A predictive model of photosynthesis for cucumber', in *International Conference on Computer Application and Information Security (ICCAIS 2021)*, SPIE, May 2022, pp. 286–291. doi: 10.1117/12.2637492.

[8] P. Zhang, Z. Zhang, B. Li, H. Zhang, J. Hu, and J. Zhao, 'Photosynthetic rate prediction model of newborn leaves verified by core fluorescence parameters', *Sci Rep*, vol. 10, no. 1, p. 3013, Feb. 2020, doi: 10.1038/s41598-020-59741-6.

[9] Z. Yang, J. Tian, Z. Wang, and K. Feng, 'Monitoring the photosynthetic performance of grape leaves using a hyperspectral-based machine learning model', *European Journal of Agronomy*, vol. 140, p. 126589, Oct. 2022, doi: 10.1016/j.eja.2022.126589.

[10] T. Wu *et al.*, 'Retrieving rice (Oryza sativa L.) net photosynthetic rate from UAV multispectral images based on machine learning methods', *Frontiers in Plant Science*, vol. 13, p. 1088499, 2023.

[11] D. Heckmann, U. Schlüter, and A. P. M. Weber, 'Machine Learning Techniques for Predicting Crop Photosynthetic Capacity from Leaf Reflectance Spectra', *Molecular Plant*, vol. 10, no. 6, pp. 878–890, Jun. 2017, doi: 10.1016/j.molp.2017.04.009.

[12] P. Fu, K. Meacham-Hensold, K. Guan, and C. J. Bernacchi, 'Hyperspectral leaf reflectance as proxy for photosynthetic capacities: An ensemble approach based on multiple machine learning algorithms', *Frontiers in Plant Science*, vol. 10, p. 454448, 2019.

[13] T. Kaneko *et al.*, 'A canopy photosynthesis model based on a highly generalizable artificial neural network incorporated with a mechanistic understanding of single-leaf photosynthesis', *Agricultural and Forest Meteorology*, vol. 323, p. 109036, 2022.

[14] G. D. Farquhar, S. von Caemmerer, and J. A. Berry, 'A biochemical model of photosynthetic CO2 assimilation in leaves of C3 species', *Planta*, vol. 149, no. 1, pp. 78–90, Jun. 1980, doi: 10.1007/BF00386231.

[15] N. R. Baker, 'Chlorophyll Fluorescence: A Probe of Photosynthesis In Vivo', *Annual Review of Plant Biology*, vol. 59, no. Volume 59, 2008, pp. 89–113, Jun. 2008, doi: 10.1146/annurev.arplant.59.032607.092759.

[16] L. McInnes, J. Healy, and J. Melville, 'UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction', Sep. 17, 2020, *arXiv*: arXiv:1802.03426. doi: 10.48550/arXiv.1802.03426.

[17] R. J. G. B. Campello, D. Moulavi, and J. Sander, 'Density-Based Clustering Based on Hierarchical Density Estimates', in *Advances in Knowledge Discovery and Data Mining*, J. Pei, V. S. Tseng, L. Cao, H. Motoda, and G. Xu, Eds., Berlin, Heidelberg: Springer, 2013, pp. 160–172. doi: 10.1007/978-3-642-37456-2_14.

[18] Y. Yang *et al.*, 'Dimensionality reduction by UMAP reinforces sample heterogeneity analysis in bulk transcriptomic data', *Cell Reports*, vol. 36, no. 4, Jul. 2021, doi: 10.1016/j.celrep.2021.109442.

[19] L. van der Maaten and G. Hinton, 'Visualizing Data using t-SNE', *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.

[20] 'Reprint of: Mahalanobis, P.C. (1936) "On the Generalised Distance in Statistics."', *Sankhya A*, vol. 80, no. 1, pp. 1–7, Dec. 2018, doi: 10.1007/s13171-019-00164-5.

[21] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, 'A density-based algorithm for discovering clusters in large spatial databases with noise', *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 226–231, Aug. 1996.

[22] L. McInnes, J. Healy, and S. Astels, 'hdbscan: Hierarchical density based clustering', *Journal of Open Source Software*, vol. 2, no. 11, p. 205, Mar. 2017, doi: 10.21105/joss.00205.

[23] S. Galbraith, J. A. Daniel, and B. Vissel, 'A Study of Clustered Data and Approaches to Its Analysis', *J. Neurosci.*, vol. 30, no. 32, pp. 10601–10608, Aug. 2010, doi: 10.1523/JNEUROSCI.0362-10.2010.