Science of the Total Environment 932 (2024) 172741

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Discriminating bloom-forming cyanobacteria using lab-based hyperspectral imagery and machine learning: Validation with toxic species under environmental ranges

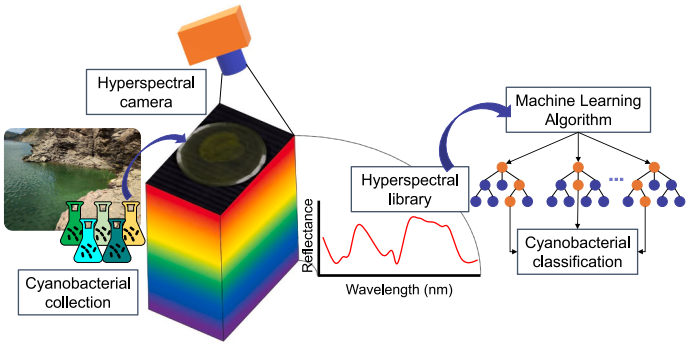
Claudia Fournier [a], Antonio Quesada [a,*], Samuel Cirés [a], Mohammadmehdi Saberioon [b]

[a] Departamento de Biología, Universidad Autónoma de Madrid, 28049 Madrid, Spain
[b] Section 1.4 Remote Sensing and Geoinformatics, German Research Centre for Geosciences (GFZ), Telegrafenberg, 14473 Potsdam, Germany

HIGHLIGHTS

- Cyanobacterial-derived spectral variability was induced during biomass growth.
- Five cyanobacterial genera were spectrally discriminated in 80-90% of cases.
- Spectral pre-processing assisted in classification and model robustness.
- Reflectance from wavelengths in both VIS and NIR ranges was essential.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Warish Ahmed

Keywords:
Algal blooms
Remote sensing
Near-infrared
Ensemble classifiers
Early warning

ABSTRACT

Cyanobacteria are major contributors to algal blooms in inland waters, threatening ecosystem function and water uses, especially when toxin-producing strains dominate. Here, we examine 140 hyperspectral (HS) images of five representatives of the widespread, potentially toxin-producing and bloom-forming genera *Microcystis*, *Planktothrix*, *Aphanizomenon*, *Chrysosporum* and *Dolichospermum*, to determine the potential of utilizing visible and near-infrared (VIS/NIR) reflectance for their discrimination. Cultures were grown under various light and nutrient conditions to induce a wide range of pigment and spectral variability, mimicking variations potentially found in natural environments. Importantly, we assumed a simplified scenario where all spectral variability was derived from cyanobacteria. Throughout the cyanobacterial life cycle, multiple HS images were acquired along with extractions of chlorophyll *a* and phycocyanin. Images were calibrated and average spectra from the region of interest were extracted using k-means algorithm. The spectral data were pre-processed with seven methods for subsequent integration into Random Forest models, whose performances were evaluated with different metrics on the training, validation and testing sets. Successful classification rates close to 90 % were achieved using either the first or second derivative along with spectral smoothing, identifying important wavelengths in both the VIS and NIR. *Microcystis* and *Chrysosporum* were the genera achieving the highest accuracy (>95 %), followed by *Planktothrix* (79 %), and finally *Dolichospermum* and *Aphanizomenon* (>50 %). The potential of HS imagery to

* Corresponding author.
E-mail address: antonio.quesada@uam.es (A. Quesada).

https://doi.org/10.1016/j.scitotenv.2024.172741
Received 7 September 2023; Received in revised form 28 March 2024; Accepted 22 April 2024
Available online 26 April 2024

discriminate among toxic cyanobacteria is discussed in the context of advanced monitoring, aiming to enhance remote sensing capabilities and risk predictions for water bodies affected by cyanobacterial harmful algal blooms.

1. Introduction

Cyanobacteria are photosynthetic prokaryotic organisms that synthesize chlorophyll-*a* (Chla) and other pigments, such as phycocyanin (PC), which is the pigment responsible for their known blue-green color (Whitton and Potts, 2012). They can be found in the water column of lakes and reservoirs, coexisting naturally with the phytoplankton community at relatively low abundances. However, under specific conditions (e.g., eutrophication, high water temperature, stability of the water column, etc.), they can undergo massive proliferation, leading to the formation of dense blooms (Huisman et al., 2018). When these blooms are dominated by toxin-producing cyanobacterial strains, they are referred to as harmful cyanobacterial algal blooms (CyanoHABs) (Huisman et al., 2018). Cyanotoxins have a wide range of adverse effects, with hepatotoxins such as microcystins (MCs) being the most common worldwide, followed by cytotoxins such as cylindrospermopsin (CYN), and neurotoxins, such as anatoxin-a (ATX) and paralytic shellfish poisoning toxins (PSPs), with *Microcystis*, *Dolichospermum* (formerly *Anabaena*), *Aphanizomenon*, *Planktothrix* and *Chrysosporum* being the main toxin-producing genera in freshwater ecosystems (Svirčev et al., 2019). In this manner, CyanoHABs have a profound impact on water quality, influencing multiple water uses such as recreation and water supply. As a result, they engender significant socio-economic consequences, with associated costs that reach up to millions of euros/dollars per year for some countries (Sanseverino et al., 2016). Furthermore, it is expected that eutrophication and climate change will increase the frequency and duration of cyanobacterial blooms (Chapra et al., 2017), as well as expand their distribution towards higher latitudes (Przytulska et al., 2017).

Due to the potential risks associated with CyanoHABs, a wide range of techniques has been developed to monitor cyanobacteria. These techniques vary from more traditional and time-consuming methods, such as microscopic enumeration or pigment extraction, to more automated approaches. Automation in recent decades in the field of cyanobacterial monitoring includes tools such as probes permanently installed in water bodies or satellite-based remote sensing, among others. Additionally, machine learning has emerged as an efficient way to process all the data coming from automation, boosting cyanobacterial monitoring and enabling broader data applications (Almuhtaram et al., 2021). Optical methods, including approaches based on the absorption, transmission, fluorescence, and reflectance of cyanobacteria when exposed to light, have been used in different manners to monitor algae and blooms at laboratory and field scales (Almuhtaram et al., 2021; Solovchenko, 2023). In this context, Rousso et al. (2020) showed how data-driven forecasting strategies, based on statistical relationships between input variables and bloom occurrence, had gained popularity in the research of CyanoHABs compared to process-based models, which require a detailed understanding of the underlying physical and biological processes. The shift towards data-driven approaches can be attributed to the evolution of more automated monitoring strategies and tools, as mentioned above, with in-situ fluorescence and remote sensing as the main contributors (Rousso et al., 2020). These methods are capable of generating large amounts of data, while machine learning techniques play a key role in enabling their processing and integration in predictive models.

Nowadays, key parameters associated with the occurrence of CyanoHABs, such as turbidity, Chla and PC, can be estimated using data from multispectral satellite imagery (Jiang et al., 2020; Pamula et al., 2023). Thus, remote sensing has become a well-recognized tool for the early warning of cyanobacterial blooms (Almuhtaram et al., 2021). However, while satellite multispectral imagery collects data in discrete bands from specific parts of the electromagnetic spectrum, HS technology provides reflectance data in narrower spectral bands, offering higher resolution. This higher resolution allows to expand remote sensing capabilities beyond early detection of cyanobacterial blooms, to early identification of potentially toxic genera dominating these proliferations (Kudela et al., 2015). In addition, HS remote sensing using satellites, drones, or local HS cameras can detect potential cyanobacterial risk not only in the early stages of bloom development, but also in low-risk areas associated with more strategic or exposed water bodies that could become a threat due to future displacement or proliferation of the cyanobacterial biomass. Another reason supporting the utility of HS imagery for cyanobacterial classification in the field is that each genus has specific traits. For instance, not all potentially toxic species exhibit the same ratio of toxin/biomass. *Microcystis* species, for example, are well-known for their high microcystin production (Dittmann et al., 2013; Li et al., 2023). As a result, varying levels of potential risk are expected across different genera. In this context, machine learning provides various techniques that are essential for properly handling, processing and analyzing reflectance data from HSl imagery. Despite its suggested potential for early identification of potentially toxic genera dominating developing blooms (Kudela et al., 2015), HS technology capabilities are still relatively unexplored in this field (Almuhtaram et al., 2021). Recent advances in HS technology have made HS cameras more affordable and commercially available, increasing their popularity, especially at industrial and laboratory scale (Liu et al., 2021). The range of applications is diverse, for instance, Salmi et al. (2021) utilized various vegetation indices, incorporating reflectance data from selected wavebands captured by a HS camera, to monitor the growth of different cyanobacteria and algae species, revealing strong correlations with fluorescence measurements. In another study, Adejimi et al. (2023) explored the suitability of HS transmittance for classifying cyanobacterial species in bioreactors, employing data from the visible and near-infrared (VIS/NIR) spectrum to train machine learning algorithms. However, while the existing studies have focused more on industrial perspectives, there is a lack of agreement regarding the applicability of HS imagery from an environmental standpoint (Almuhtaram et al., 2021).

This study aims to explore the potential of reflectance HS imagery combined with machine learning techniques for discriminating among five representative species of bloom-forming, potentially toxin-producing cyanobacterial genera when a wide spectral variability is considered, elucidating the importance of visible and near-infrared wavelengths on achieving accurate classifications. The ultimate goal is to demonstrate the capacity of HS technology to fine-tune existing early warning systems based on remote sensing tools by discerning specific cyanobacterial genera with potential toxicity. To the best of our knowledge, this study represents the first investigation to consider such technology along with the significant spectral and taxonomical variability, highlighting the novelty of both the methods employed and the findings obtained.

2. Materials and methods

The methodology used is illustrated in Scheme 1 and it can be summarized in five main steps. First, different cyanobacteria were grown under various conditions to induce pigment and spectral variability. Next, HS images were acquired to record the spectral information from the grown biomass, and some bio-optical characteristics were measured to confirm the induced variability. Finally, the resulting HS

images were pre-processed and integrated into machine learning models, whose performances were evaluated by different metrics.

2.1. Cyanobacterial collections and factorial experiment

Five cyanobacteria were selected from the collection of Universidad Autónoma de Madrid (UAM, Spain): (1) *Microcystis aeruginosa* (*M. aeruginosa*) (UAM284, MCs producing strain), (2) *Planktothrix agardii* (*P. agardii*) (UAM565, MCs producing strain), (3) *Aphanizomenom gracile* (*A. gracile*) (UAM529, PSPs producing strain), (4) *Chrysosporum ovalisporum* (*C. ovalisporum*) (UAM292, CYN producing strain), and (5) *Dolichospermum crassum* (*D. crassum*) (UAM502, non-toxic strain) (Table S1). Two collections were created for the study. The *mother collection* (Fig. S2) consisted of five cultures, each corresponding to one of the cyanobacteria mentioned above, maintained under their stock conditions at UAM culture collection (temperature 25 °C, light intensity of 8 μmol photons m^{-2} s^{-1}, and nutrients). *M. aeruginosa* and *P. agardhii* were cultivated in BG11 medium (Rippka, 1988), while the other species, which can fix atmospheric nitrogen via specialized cells (heterocyst), were cultured in $BG11_0$ medium (with the same composition as BG11 but lacking a source of nitrogen). The second collection was cultivated with biomass from the *mother collection* in a factorial experiment. Thus, the cyanobacterial biomass was cultivated in 300 mL of the corresponding medium in a 500 mL Erlenmeyer flask, starting with an initial biomass equivalent to an optical density at 750 nm (OD_{750}) of 0.1. The factorial experiment involved different light and nutrient conditions. Three levels of light intensity were defined for all cyanobacteria: 8, 30, and 120 μmol photons m^{-2} s^{-1}. Additionally, three nutrient levels were defined for non-N-fixing cyanobacteria: BG11 with nitrogen proportions of 1/20, 1/4, and undiluted (1.4, 7, and 28 mg of N/L, respectively). Thus, *M. aeruginosa* and *P. agardhii* were exposed to all factorial levels, experiencing changes in both nutrient and light conditions, while nitrogen-fixing cyanobacteria were subjected to variations in light conditions only. As mentioned above, the main goal of the factorial experiment is to induce pigment and spectral variability by manipulating nutrient and light availability, which directly affect the production of key photosynthetic pigments as well as other biological traits (e.g., cell size and morphology) (Wyman and Fay, 1986). Following other similar works (Legleiter et al., 2022), we assumed a simplified scenario in which the water column consisted only of pure water and the cyanobacterial biomass. Therefore, spectral variability resulting from other potential sources, such as colored dissolved organic matter (CDOM), suspended sediments, or other non-cyanobacterial algal constituents, was not considered in this experiment.

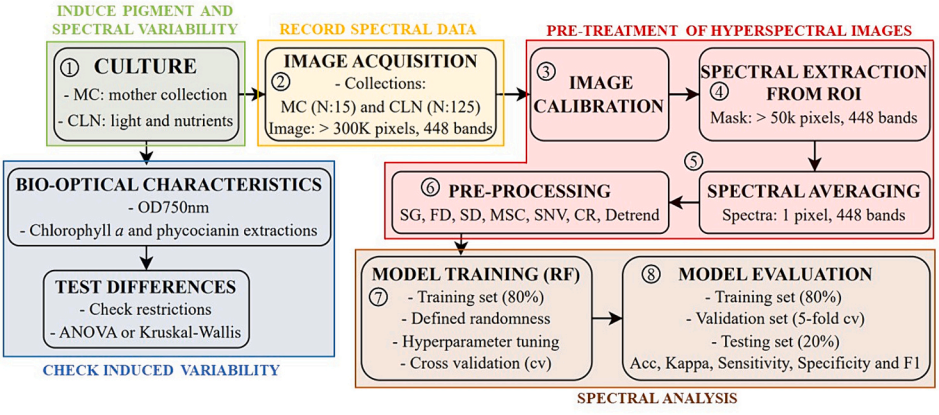
2.2. Bio-optical measurements

Turbidity, determined by measuring OD_{750}, served as a proxy for cyanobacterial growth. Additionally, for each culture used for image acquisition in the factorial experiment, Chla and PC were extracted, and their concentrations were measured. For Chla extraction, 10 mL of the culture was centrifuged at 4700 rpm for 15 min, and the resulting pellet was resuspended in 8 mL of 90 % (v/v) cold methanol at 4 °C (Cirés et al., 2011). The extraction took place overnight, in darkness and at 4 °C. After the extraction period, the extract was centrifuged again at 4700 rpm for 15 min, the OD_{665} (chl *a* absorbance peak) and OD_{750} (turbidity) were measured, and chl *a* concentration (μg L^{-1}) was calculated using equations derived from Marker (1980) (Supplementary material – Eq. S4-a). Similarly, for PC extraction, 10 mL of the culture were centrifuged at 4700 rpm for 15 min, followed by resuspension in 5 mL of phosphate buffer (0.1 M, pH = 6). PC was extracted according to Lawrenz et al. (2011) by sonication (5 s duration, 8-W pulses for 30 s), and then in darkness, at 4 °C for 48 h. After extraction, 2 mL of the extract were centrifuged at 10,870*g* (7,130 rpm) for 5 min. The OD_{620} (PC absorbance peak) and OD_{750} (turbidity) were measured, and PC concentration (μg L^{-1}) was calculated using equations detailed in Lawrenz et al. (2011) (Supplementary material – Eq. S4-b). All absorbance measurements were taken using a Hitachi U-2000 Dual-Beam UV–Vis Spectrophotometer.

Statistical tests were employed to examine differences in pigment ratios among groups of environmental conditions for each cyanobacterial taxa. If the assumptions of normality and variance homogeneity were met, assessed using the Shapiro-Wilk and Levene tests respectively (Levene, 1960; Shapiro and Wilk, 1965), an ANOVA test was applied (Fisher, 1925), otherwise Kruskal-Wallis test was used (Kruskal, 1952).

2.3. Image acquisition

A line-scanning HS imaging system with push-broom configuration was used to acquire the HS images. It was located in a dark room and covered with a blackout cloth, minimizing light interference from the room (Fig. S3). The system comprises a HS camera (Specim FX10), and a scanner (Specim LabScanner 40 × 20). The camera operates in the visible and near-infrared (VIS/NIR) region, recording reflectance data for 448 bands from 397 to 1004 nm, with a spectral resolution FWHM (Full Width at Half Maximum) of 5.5 nm (mean). The camera was mounted in the scanner, that also includes a moving table where the sample was placed, and a light source (170 W halogen dual illumination) positioned at a 45-degree angle to the sample location to reduce shadowing effects. Each sample consisted of 10 mL of the cyanobacterial culture contained within a 50 mm diameter petri dish. The samples were placed one by one in a fixed position on the moving table. Some parameters were adjusted to convey spectral and spatial resolution, such as the frame rate (24 Hz), exposure time (12 ms), positioning speed (12 mm/s), and scanning speed (5 mm/s). The procedure was controlled and implemented by an image acquisition software (Specim Lumo software,



Scheme 1. Workflow followed in this study. For details read material and methods.

Spectral Imaging Ltd, Oulu, Finland), and the results were checked in real time through fast visualization using a HS analysis software (Specim Insight software, Spectral Imaging Ltd., Oulu, Finland). The output consisted of the raw image, created from multiple congruent and overlapping images, the dark reference (0 % reflectance) generated by closing the shutter of the camera, and white reference (99 % reflectance) obtained from a Spectralon panel. Due to the low signal-to-noise ratio at the two ends of the spectral ranges, only wavelengths from 400 to 1000 nm (442 bands) were used to train the models.

Two datasets were created. The primary, referred to as CLN (from Cyanobacteria, Light, and Nutrients), consisted of 125 HS images captured throughout the entire life cycle of the cultures from the factorial experiment (light and nutrient conditions explained in Section 2.1). The second dataset, referred to as MC (from Mother Collection), was composed of 15 HS images (three of each culture) captured when great biomass was reached, and cultures were between exponential and stationary phases. The main goal of generating the MC dataset was to provide images of cultures grown under standard and optimal conditions. Although the biomass of each genus came from the same stock and can be seen as triplicates, their average reflectance showed some differences.

2.4. Treatment of hyperspectral images

Spectral calibration of the acquired HS images was computed applying Eq. (1).

$$I_i = \frac{R_i - D_i}{W_i - D_i} \tag{1}$$

where I is the corrected HS image, R is the raw HS image, W and D are the white and dark references respectively, and i corresponds to the pixel index. This calibration was performed using a Python command-line software developed in-house, designed to automatically compute Eq. (1) across all images.

Spectral extraction consisted of the automatic selection of the pixels from the Region of Interest (ROI) containing cyanobacterial biomass, followed by the extraction of reflectance data from each pixel. The extraction was achieved by creating a mask over the ROI using the k-means algorithm (Lloyd, 1982), an unsupervised clustering method that separates n observations into k clusters based on their similarities. Each observation is assigned to the cluster with the nearest mean by minimizing the within-cluster sum of squared distances (Eq. (2)). Thus, pixels in each cluster are spectrally similar to the pixels in their own group, and spectrally different from pixels in other groups.

$$\sum_{i=0}^{n} \min_{\mu_j \in C} \left(\left| x_i - \mu_j \right|^2 \right) \tag{2}$$

where n is the number of clusters, C is each cluster, μ its centroid, and x is every data point.

The spectra from all the pixels included in each mask were averaged to obtain a single representative spectrum for each sample, that was then stored in a common dataset.

Spectral exploration was performed on the resulting dataset to identify distinct and well-defined groups, applying techniques such as the Principal Component Analysis (PCA) to visualize the data in a two-dimensional space that captures maximum variability, enabling a better understanding of the potential grouping patterns (Jolliffe, 1986). The averaged spectra and standard deviation of each cyanobacterial taxa were displayed to identify differences and similarities.

2.5. Spectral analysis: model training and evaluation

Spectral pre-processing techniques are known to reduce the noise in the data (Gholizadeh et al., 2015), while enhancing important features and spectral data quality (Saberioon et al., 2019), significantly improving the accuracy and reliability of machine learning predictions. Importantly, we refer to noise as any variations within the data that do not correspond to specific cellular features but come from other sources, such as vibrations or reflections during image acquisition.

Following previous successful experiences with HS imagery (Saberioon et al., 2019), seven different pre-processing methods were applied and combined to achieve optimal results in the classification model. The main goal of exploring these different pre-processing methods was to identify and select the one that effectively enhanced critical features for classification, minimizes noise, and strengthens the robustness of the model. Specifically, the Savitzky-Golay (SG) smoothing technique was applied using a second-order polynomial fit and 11 smoothing points. SG was then combined with six other techniques, including first (FD) and second (SD) derivatives, Multiplicative Scatter Correction (MSC), Standard Normal Variate (SNV), Continuum Removal (CR), and Detrend. Further information about these pre-processing techniques can be found in Table S5.

The Random Forest algorithm was selected as classification model, due to its outstanding performance in previous studies after spectral pre-processing (Belgiu and Drăgu, 2016), demonstrating its capacity to handle high-dimensional and correlated data, identify complex non-linear relationships, and maintain high accuracy even with noisy or imbalanced datasets. Random Forest is an ensemble classifier that builds multiple decision trees using for each one a random selection of samples and variables (Breiman, 2001). Each decision tree in the forest makes a prediction, and the final result is determined by the mode of all the individual tree predictions. The different combinations of the datasets mentioned in Section 2.2 (CLN, CLN + MC, and CLN + MC filtered) were randomly split 80 %/20 %, maintaining a proportional representation of each class. As the three combinations had a shared set of samples, a common core of randomly selected data pairs was maintained in the splitting to avoid bias potentially introduced by different data partitions. Beyond this, instead of a single split, a multiple train-test split approach was followed. Each model was trained on multiple training sets, and the performance was evaluated on the corresponding test sets. The results were then averaged for each model. Randomness was fixed during model training to enable reproducibility. This iterative process aimed to reduce the variance that may occur in the data splitting process, especially when dealing with limited data, and thus to provide a more robust estimate of the performance of the models. Moreover, hyperparameters, such as the number of trees in the forest and the maximum depth of these trees, were optimized using a random search approach with five-fold cross-validation. Thus, each validation set consisted of approximately 16 % of the samples from the original dataset (i.e., 20 % of the 80 % of the training set). The search was conducted over multiple iterations with a standardized randomness, to maintain reproducibility and enable comparison.

The evaluation of the classification models was performed on the training, validation, and testing sets. The training set, containing 80 % of the data, was used to train the models, therefore optimal performance of the classifiers is expected on this set. The validation set, obtained through 5-fold cross-validation, was used for hyperparameter tuning and model selection. As a result, the performance of the classifiers is expected to be slightly lower on this set, which may be influenced by factors such as data limitations. Finally, the testing set, consisting of the remaining 20 % of the data, was used to estimate the capacity of the model to generalize to unseen data. Thus, the Accuracy (Acc) was analyzed in the training and validation sets, while Cohen's Kappa coefficient (Kappa) was computed for all sets. These metrics were calculated with Eqs. (3) and (4) respectively.

$$Acc = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{3}$$

where TP and TN are true positives and true negatives respectively; and FP and FN are false positives and false negatives respectively.

$$kapa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \quad (4)$$

where Pr(a) is the probability of observed agreement and Pr(e) is probability of random agreement.

Since a single metric may not capture all aspects of the performance of a model, it is often necessary to use multiple metrics simultaneously during the evaluation process (Sokolova and Lapalme, 2009). In this case, specificity and F1 score (a combination of precision and sensitivity metrics) were also calculated in the testing set. These metrics were calculated with Eqs. (5), (6), (7) and (8) respectively.

$$Specificity = \frac{(TN)}{(\mathrm{TN} + FP)} \quad (5)$$

$$F1 = 2 \times \frac{(precision \times sensitivity)}{(precision + sensitivity)} \quad (6)$$

$$Precision = \frac{(TP)}{(TP + FP)} \quad (7)$$

$$Sensitivity = \frac{(TP)}{(TP + FN)} \quad (8)$$

where again TP and TN are true positives and true negatives respectively; and FP and FN are false positives and false negatives respectively.

2.6. Computational tools and libraries

All analyses were carried out with in-house Python (version 3.9.2) scripts developed and executed in the Jupyter Notebook environment, an open-source web-based platform for interactive computing (Perkel, 2018). Statistical analyses were carried out with functions from the stats module of SciPy library (version 1.10.1), spectral extraction and averaging with Spectral Python (SPy) module (version 0.21), pre-processing with functions from PySpectra package (version 0.0.1.2) and SciPy library (version 1.10.1), and machine learning models with functions from the Scikit-learn library (version 1.2.2) (Pedregosa et al., 2011).

3. Results and discussion

In the following subsections, all results are presented with specific references to related studies, allowing for comparison and discussion of specific results in their relevant context. At the end of the section, overall novelties, limitations, and research directions are addressed.

3.1. Visual and pigment variability during cyanobacterial growth

Significant visual variations were observed during cyanobacterial growth, both within and between groups. Fig. S6 shows the evolution of OD_{750} for each cyanobacterial genus. Notably, *C. ovalisporum* color ranged from greenish brown under low light conditions to dark brown under high light intensities. Similarly, *M. aeruginosa* color varied from dark green under high nutrient and low light conditions, to yellowish tones under low nutrient and high light conditions (Fig. S7). In certain cultures, the development of yellowish colors indicated a stress response, particularly in cases where low nutrient levels and high light intensities coincided. Xi et al. (2015) discussed the applicability of HS absorbance measurements in discriminating among phytoplanktonic groups. In their experiments, they subjected the cultures to similar light intensities as those used in this study, obtaining an effective induced variability, however they did not take into account nutrient variations. These variations were also observed in the pigment concentrations. Table 1 presents the average and standard deviations of PC and Chla concentrations, and their ratios across the different conditions. Overall, a consistent trend emerged among all cyanobacteria, where increasing light intensity led to decreased PC and Chla concentrations. Moreover, in non-N-fixing cyanobacteria, higher nutrient availability resulted in the opposite outcome, with higher concentrations of both pigments. However, exceptions were observed in *P. agardhii* and *D. crassum*, which could be due to insufficient light levels required for their maximum growth rates, causing pigment synthesis to decrease as growth becomes light limited (Wyman and Fay, 1986). On the other hand, the PC/Chla ratio exhibited variations across all cyanobacteria and conditions, with statistically significant differences observed in *M. aeruginosa* under different nutrient conditions, and *P. agardhii* under different light

Table 1
Mean and standard deviation values of pigment concentrations and pigment ratios grouped by conditions from CLN dataset.

			[PC](mg L^{-1})		[Chla] (mg L^{-1})		PC/Chla	
			Mean	SD	Mean	SD	Mean	SD
Microcystis	Light	Low	11.0	12.8	3.7	3.4	2.7	0.9
		Medium	6.2	6.3	2.3	2.0	2.4	0.9
		High	3.7	3.4	1.3	0.9	2.7	1.1
	Nutrients***	Low	1.9	1.0	1.0	0.4	2.0	1.0
		Medium	4.9	3.2	2.2	1.4	2.4	0.9
		High	13.8	12.3	4.1	3.5	3.3	0.5
Planktothrix	Light***	Low	4.8	2.8	2.6	0.5	1.8	0.8
		Medium	2.2	0.8	1.0	0.3	2.1	0.6
		High	1.9	0.5	0.6	0.2	3.2	1.0
	Nutrients	Low	1.8	0.5	1.4	0.9	1.8	0.9
		Medium	4.2	2.6	1.7	0.9	2.7	1.2
		High	3.8	2.5	1.7	1.0	2.3	0.6
Chrysosporum	Light	Low	9.6	4.7	4.1	2.3	2.4	0.4
		Medium	7.7	2.7	3.6	1.4	2.2	0.3
		High	4.7	1.3	2.5	0.7	2.0	0.5
Aphanizomenon	Light	Low	14.4	10.3	4.3	3.1	3.3	0.4
		Medium	3.6	1.8	1.3	0.6	2.7	0.7
		High	1.7	0.2	0.7	0.1	2.4	0.5
Dolichospermum	Light	Low	12.1	6.5	4.4	2.6	2.8	0.5
		Medium	2.7	1.0	1.3	0.3	2.1	0.7
		High	4.2	2.0	2.4	1.1	2.4	1.5

A symbol is included when intragroup differences in the PC/Chla ratio were statistically significative regarding culture conditions. SD: standard deviation.
*** p-Value < 0.01.

conditions. In the case of *Microcystis*, the ratio increased with higher nutrient availability, primarily driven by a substantial increase in the average PC concentration (from 1.9 to 13.8 mg L^{-1}, more than a seven-fold increase), which was greater than the increase in the average Chla (from 1 to 4.1 mg L^{-1}, approximately a four-fold increase). Conversely, in *Planktothrix* the ratio increased with higher light intensity, primarily due to a decrease in the average Chla (from 2.6 to 0.6 mg L^{-1}, a four-fold decrease), which was greater than the decrease in the average PC concentration (from 4.8 to 1.9 mg L^{-1}, a two to three-fold decrease). Under the light of these results, it can be assumed that the wide range of physicochemical conditions applied led to variations up to two orders of magnitude in pigment concentrations (from minimum to maximum values). These variations were consistent with the changes in culture colors observed during the experiment, confirming that great variability had been effectively induced.

3.2. Spectral pre-treatment

After image acquisition, final images (Fig. 1a) consisted of >320,000 pixels (313 lines, and 1024 pixels per line), each one integrating reflectance data from 448 bands. Applying the k-means algorithm generated a mask for each image, where the pixels with cyanobacterial biomass were selected as shown in Fig. 1b. With this masking process, the quantity of pixels effectively decreased (e.g., to 50,000 pixels in the case of the image displayed in Fig. 1b). Finally, the spectra from every pixel in the ROI was averaged into one mean spectrum (Fig. 1c), which conserved the reflectance from the 448 bands, and it was assumed to be representative of the sample. Importantly, the standard deviation was constant for all wavelengths except for those between 400 and 450 nm, where the standard deviation increased, which could be due to some interferences during the hyperspectral image acquisition (e.g., light reflection from the metal platform or Petri dish). The spectral extraction process could be automated if the number of clusters (k) required for each image was the same, as it occurred when biomass was concentrated enough. However, most images with lower biomass concentration had a different optimal number of clusters, which should be taken into consideration for future experiments as it implies an increased amount of work.

In any case, this method for spectral extraction proved to be practical and innovative compared with other studies with phytoplanktonic biomass where the ROI was manually selected with a default square of pixels or commercial software, which apart from being more time-consuming, is also prone to losses of spectral information (Salmi et al., 2022). It is important to note that during the image acquisition process, in cases where the biomass concentration was too low, the k-means algorithm was unable to effectively discriminate the cyanobacterial biomass from the background. Consequently, not all pixels containing biomass were included in the mask, while some pixels without biomass were mistakenly included. It was observed that, on average, a minimum OD_{750} of 0.3 was required to generate an accurate mask when placing 10 mL of cyanobacterial culture within a 50 mm diameter petri dish for HS image acquisition. This aspect has been observed in other studies with HS imagery of cyanobacterial and algal biomass, where similar minimum concentrations for detection were determined (Salmi et al., 2021), or even ratios among absorbance at specific wavelengths (751/676 nm) were used as indicators for minimum concentrations (Salmi et al., 2022). Consequently, during spectral extraction some noise could be introduced. Thus, three combinations of the obtained datasets were selected for the following analysis: (1) CLN images (125 images), (2) CLN and MC images (140 images), and (3) CLN and MC images after applying a quality filter (96 images). The latter (data set 3) was used to illustrate the patterns observed during the exploration, although all data set combinations were consistent with the results. The quality filter rejected images that met either of the following conditions: (1) low biomass concentration that could not be properly distinguished from the background when the mask for spectral extraction was applied (the main reason of image rejection), and (2) samples consisting primarily of white, dead biomass with abnormally high reflectance and no distinct spectral shape.

3.3. Data exploration – comparative of five genera and grouping patterns

The averaged spectra from the five cyanobacteria appeared to follow a similar pattern, presenting characteristic hills and valleys in the range from 400 to 700 nm (Fig. 2a), being consistent with other HS studies where cyanobacteria spectra have been explored, not just in terms of reflectance (Legleiter et al., 2022; Salmi et al., 2021), but also in terms of absorbance (Malhotra and Örmeci, 2022; Salmi et al., 2022), and transmittance (Adejimi et al., 2023). One notable valley between 620 and 640 nm coincides with the maximum absorbance of PC, while another clear valley at 680 nm aligns with the maximum absorbance of Chla in vivo. Remote sensing commonly utilizes this property and the relation between these two pigments to differentiate cyanobacteria from other algal groups (Dev et al., 2022), which will be further discussed in subsequent sections. It is important to note that the reflectance pattern observed in the blue bands (between 400 and ~450 nm) looks different

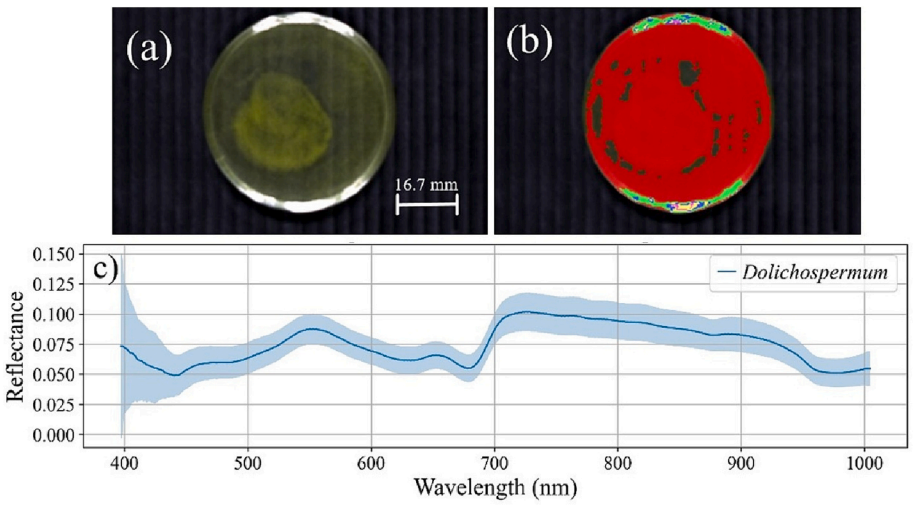


Fig. 1. General visualization of (a) calibration, (b) spectral extraction (red mask corresponds to the pixels containing cyanobacterial biomass), and (c) mean VIS/NIR spectral reflectance from all pixels in the ROI. Image from *Dolichospermum crassum* is used as example.

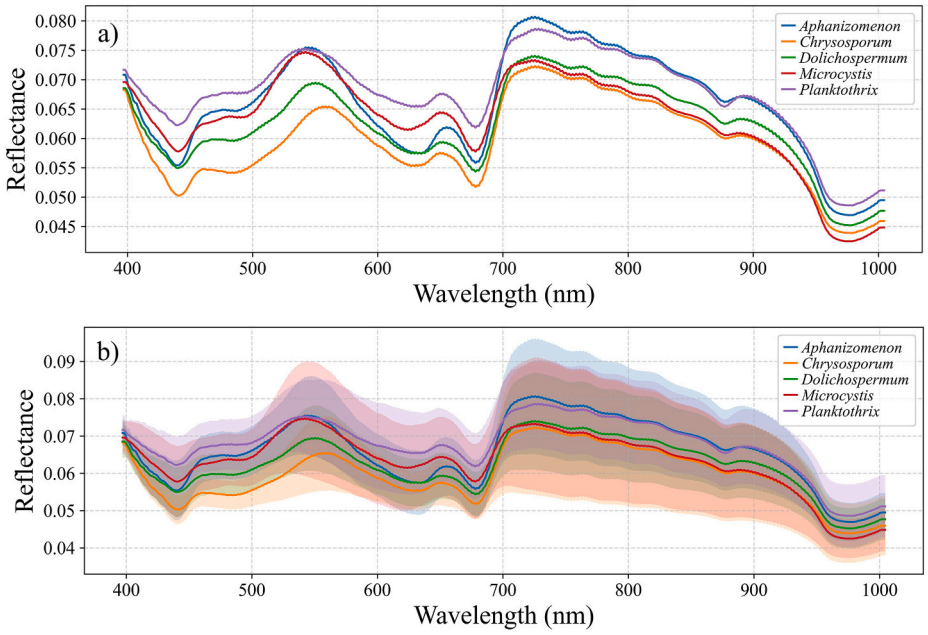


Fig. 2. Mean of VIS/NIR spectral reflectance of the five cyanobacteria. (a) Averaged spectra only. (b) Averaged spectra including standard deviations. Spectra belong to data set 3, with CLN and MC images after applying a quality filter N: 96 images.

from those in the previously mentioned publications, which is probably related to the great standard deviation observed when the mask was applied (Fig. 1-c). Finally, in the near-infrared bands, all the spectra appear to be similar, but slight differences still distinguish them from one another, presenting some maximums and minimums between 750 and 800 nm and around 870 nm. These averaged spectra show some overlapping areas, but there is never complete overlap between any two cyanobacterial taxa, which could assist in classification. However, when the standard deviations are included (Fig. 2-b), the scenario changes significantly, resulting in spectral overlapping among all groups, which could pose a challenge in classifying them based on their HS data.

Principal Component Analysis (PCA) resulted in two principal components that accounted for >93 % of the variability (Fig. 3). All samples were distributed along the first and second components, and no clear groups could be visually detected. However, some patterns might be present. For example, *C. ovalisporum* samples (red dots) were distributed in the higher area of PC2, and most of the *M. aeruginosa* samples (purple dots) were in the lower part of PC2. This lack of clear aggregation brings us to the same point as Fig. 2b, representing a potential challenge for the classification task.

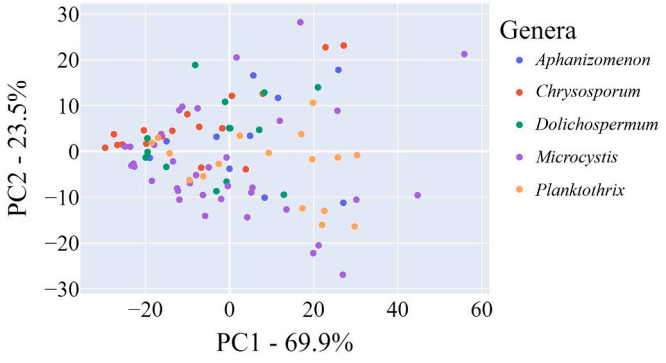


Fig. 3. Principal Component Analysis of the VIS/NIR spectral reflectance of the five cyanobacteria. Spectra belong to data set 3, with CLN and MC images after applying a quality filter N: 96 images.

Two other interesting findings emerged. Firstly, the most relevant wavelengths in both the first and second components of PCA were within the 400–700 nm (VIS) range. This aligns with the specific pigment activity of cyanobacteria, which focuses on wavelengths within this range as previously mentioned. Secondly, when exploring the VIS/NIR spectral reflectance of each genus separately, an intragroup pattern based on the life stage was observed (Fig. S8). Thus, three types of spectra were identified within each group: (1) an initial stage with low reflectance but a discernible shiny shape, (2) a transitional stage between exponential and stationary growth stages with well-defined shape and slightly higher reflectance, and (3) a later stage with higher reflectance but decreasing shape definition. These findings imply that the spectral response of cyanobacterial biomass could potentially provide insights into its life stage, which could be potentially useful to discern between healthy cyanobacterial populations and those in decay, but further research is necessary to validate this observation.

3.4. Classification model

To ensure the creation of a reliable classification model and to explore the impact of spectral pre-processing on classification accuracy, seven different pre-processing methods were used for the training of the classifier. Fig. 4 shows the pre-processing combinations used to train the models for all samples. The variability included in each data split across the different fixed randomizations is shown in Fig. S9.

The trained models and evaluation metrics from the three combinations of datasets and pre-processing methods are presented in Table 2, resulting in a total of 24 models. Accuracy was considered a reliable indicator of the performance, consistent with other metrics such as specificity and F1 in all cases. Kappa coefficient exhibited high sensitivity and produced the most stringent results, while specificity tended to overestimate the performance of the models. Overall, significant differences were observed between the performance of the classifier on the training set and that on the validation and testing sets. The metrics reached near-maximum values on the training set, suggesting potential overfitting to the training data. Nevertheless, performance on the validation and test sets yielded more moderate results, with softer differences between them. The consistency of performance between the validation and testing sets was systematically improved by various

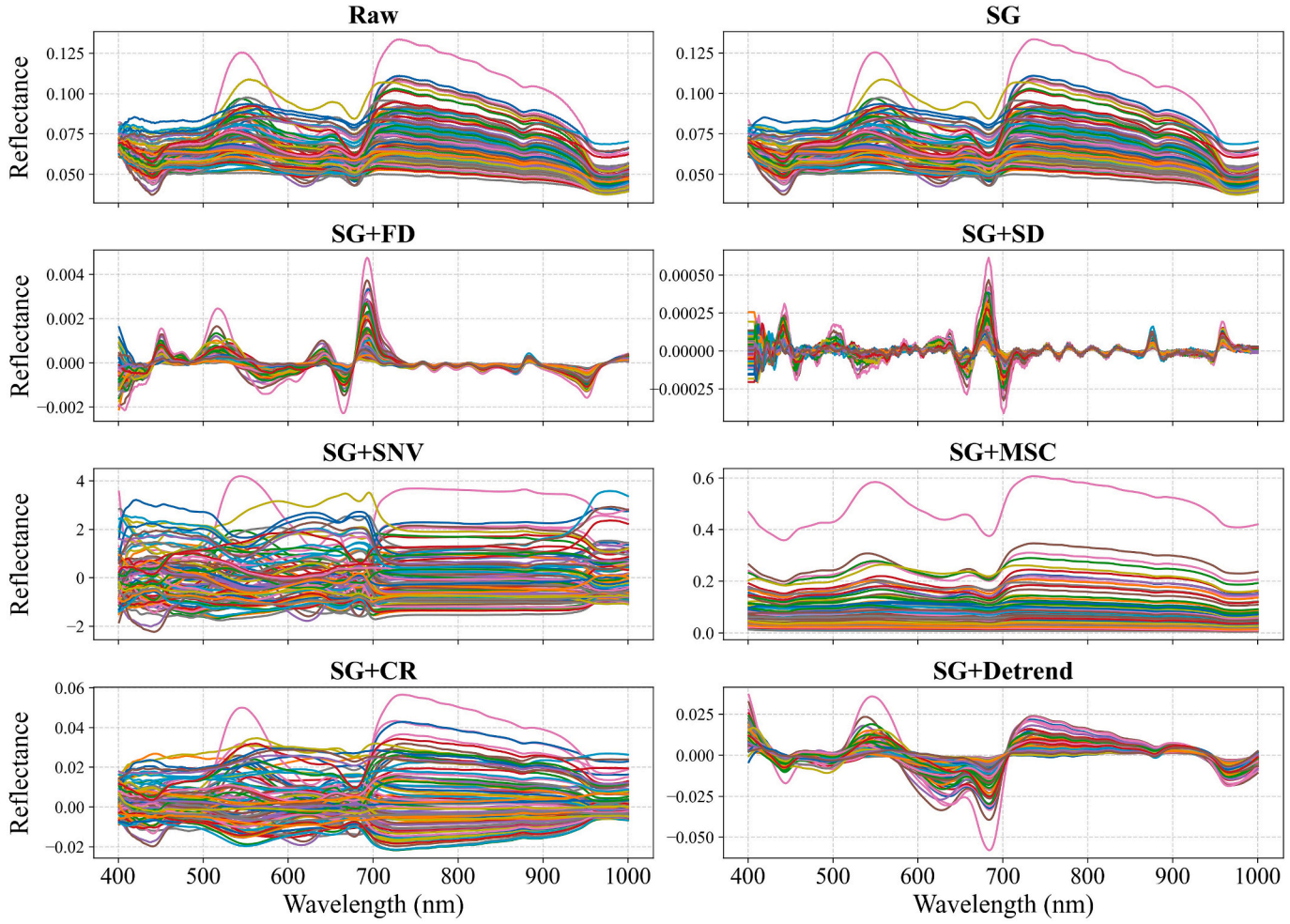


Fig. 4. Representative VIS/NIR average spectra of the different combinations of pre-processing methods used for the training. Spectra belong to data set 3, with CLN and MC images after applying a quality filter N: 96 images. SG: Savitzky-golay; FD: First Derivative; SD: Second Derivative; SNV: Standard Normal Variate; MSC: Multiplicative Scatter Correction; CR: Continuum Removal; Detrend: Detrend.

preprocessing methods in all data sets. For example, when SG was coupled with FD or SD, the differences between the training, validation, and testing sets were much less pronounced, resulting in improved consistency. These results highlight the importance of using effective preprocessing methods when working with HS data for cyanobacterial classification, and suggest that FD and SD are good options in this regard. Beyond the global classification accuracy, specific classification accuracy across classes is displayed in Table S10. Classification accuracies were highest for the genera *Microcystis* and *Chrysosporum*, exceeding 90 % when models were trained with both first and second derivatives. *Planktothrix* followed closely, achieving an accuracy of around 80 %. Models had greater difficulty in classifying the genera *Aphanizomenon* and *Dolichospermum*, with accuracies still >50 %. Since *Microcystis* and *Planktothrix* were sampled across more environmental conditions (nutrients and light) than the other genera, a greater number of images were available. This abundance of data may have contributed to their higher classification accuracies. Although *Chrysosporum* was less represented compared to these two genera, the models showed a strong ability to discriminate its images, likely due to the distinct brownish tone characteristic of *Chrysosporum* cultures during cyanobacterial growth. Conversely, *Aphanizomenon* and *Dolichospermum*, which were primarily subjected to light changes, yielded fewer images, and cultures of these genera did not show a distinct tone, likely posing a greater challenge for the classification.

In this case, the combination of SG and FD techniques consistently yielded the best results. Following this, SG coupled with the SD also performed well, outperforming SG-FD when the models were trained with the third dataset (CLN + MC filtered). In general, the performance of the classification models was either maintained or improved after applying the pre-processing methods, with the exception of the MSC, which showed the worst results across all metrics and datasets. Regarding the other results, accuracy ranged from 0.43 to 0.92. The highest values were achieved when using filtered data pre-processed with SG and FD or SD techniques. Overall, the RF classifier with data preprocessed using SG and FD performed the best when considering all datasets (Accuracy: 0.81–0.88, Kappa: 0.75–0.84). These results indicate that the combination of SG and FD techniques can successfully reduce noise and baseline effects, and enhance features related to different cyanobacteria, as is supported by other studies with HS measurements of cyanobacteria. For example, Malhotra and Örmeci (2022) effectively applied Savitzky-Golay coupled with the first derivative of absorbances to enhance the features of the spectra and decrease detection limits, or Adejimi et al. (2023) that found the first derivative as an effective technique to pre-process HS transmittance data for training Support Vector Machines for cyanobacterial inter-genera classification. In general, the results were similar across all datasets, with no significant differences observed. The high performance demonstrated by the models trained on the CLN dataset, which included images that did not pass any quality filters, suggest that the combination of SG and FD with RF can yield robust results that are not likely to be affected by the quality of the input. However, two aspects should be considered. Firstly, the accuracies of the CLN + MC dataset were slightly lower, despite

Table 2
Performance for classification models on the training, validation and testing sets. Eight combinations of spectral pre-processing and three datasets considered. Corresponding classification accuracies across classes is available in Table S10.

Dataset	Pre-processing-model	Training		Validation	Testing			
		Accuracy	Kappa	Kappa	Accuracy	Kappa	Specificity	F1
CLN (N: 125)	Raw-RF	0.98	0.98	0.34	0.63	0.46	0.90	0.63
	SG-RF	0.96	0.94	0.37	0.62	0.43	0.90	0.62
	SG + FD-RF	**0.99**	**0.99**	**0.75**	**0.87**	**0.82**	**0.97**	**0.87**
	SG + SD-RF	**0.99**	**0.99**	**0.61**	**0.78**	**0.68**	**0.94**	**0.76**
	SG + SNV-RF	0.98	0.97	0.38	0.67	0.52	0.91	0.67
	SG + MSC-RF	0.90	0.87	0.11	0.40	0.12	0.85	0.40
	SG + CR-RF	1.00	1.00	0.27	0.60	0.43	0.90	0.60
	SG + Detrend-RF	0.99	0.99	0.50	0.70	0.56	0.92	0.70
CLN + MC (N: 140)	Raw-RF	0.99	0.99	0.37	0.53	0.36	0.88	0.53
	SG-RF	0.98	0.98	0.36	0.56	0.34	0.88	0.52
	SG + FD-RF	**1.00**	**1.00**	**0.73**	**0.81**	**0.75**	**0.95**	**0.81**
	SG + SD-RF	**1.00**	**1.00**	**0.70**	**0.75**	**0.66**	**0.94**	**0.75**
	SG + SNV-RF	0.98	0.97	0.36	0.57	0.40	0.89	0.57
	SG + MSC-RF	0.87	0.82	0.14	0.34	0.04	0.84	0.35
	SG + CR-RF	1.00	1.00	0.23	0.43	0.21	0.85	0.43
	SG + Detrend-RF	1.00	1.00	0.55	0.65	0.52	0.91	0.65
CLN + MC filtered (N: 96)	Raw-RF	0.99	0.98	0.44	0.62	0.46	0.91	0.62
	SG-RF	0.99	0.99	0.42	0.60	0.44	0.90	0.60
	SG + FD-RF	**0.99**	**0.99**	**0.68**	**0.88**	**0.84**	**0.97**	**0.88**
	SG + SD-RF	**0.99**	**0.99**	**0.75**	**0.92**	**0.88**	**0.98**	**0.92**
	SG + SNV-RF	0.99	0.99	0.43	0.60	0.44	0.90	0.60
	SG + MSC-RF	0.98	0.97	0.12	0.30	−0.02	0.83	0.30
	SG + CR-RF	1.00	1.00	0.41	0.50	0.33	0.88	0.50
	SG + Detrend-RF	0.99	0.99	0.57	0.73	0.62	0.94	0.73

CLN: imagery collection from the cultures subjected to the factorial experiment. MC: imagery collection from cultures grown under same light and nutrient conditions. Raw: not preprocessed. RF: Random Forest. SG: Savitzky-Golay. FD: First Derivative. SD: Second Derivative. SNV: Standard Normal Variate. MSC: Multiplicative Scatter Correction. CR: Continuum Removal. Detrend: detrend. In bold FD and SD results for clear comparison of top-performing models.

containing additional images intended to improve the classification capacity. Secondly, when a quality filter was applied, the number of images was significantly reduced, however best performances were still achieved, but the discrepancy between the results of the validation and testing sets increased. This suggests that while high quality data can compensate for lower data availability, it may also affect hyperparameter tuning and model optimization. Regarding the performance of the other pre-processing methods, the combination of SG and SD yielded the best results (Accuracy: 0.75–0.92, Kappa: 0.66–0.88) followed by SG and Detrend (Accuracy: 0.65–0.73, Kappa: 0.52–0.62). On the other hand, the other pre-processing techniques, including SG, SG with SNV, and SG with CR, did not show any significant improvement in model performance compared to the raw data. Based on these results, models were retrained with data preprocessed solely using the first derivative to assess the added value of SG (Table S11). Notably, although results between Raw-RF and SG-RF showed no clear impact on performance, preprocessing with only the FD led to a decrease in consistency among metrics across the training, validation, and testing sets, as well as in the overall performance.

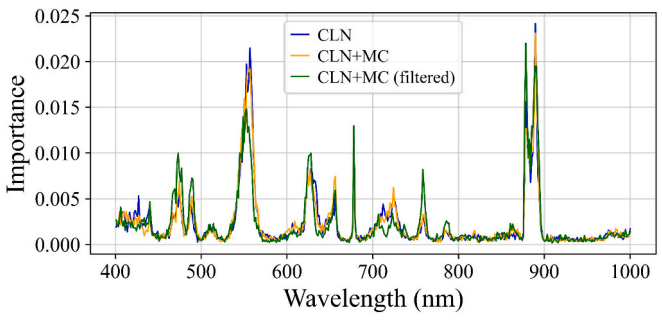


Fig. 5. Importance of wavelengths in the models trained with the SG and FD pre-processed data from CLN, CLN + MC, and CLN + MC filtered. CLN: imagery from the cultures subjected to the factorial experiment. MC: imagery from cultures grown under the same light and nutrient conditions.

Finally, the importance of each wavelength in the best models (i.e., the ones pre-processed with SG and FD) are displayed in Fig. 5. The importance of all wavelengths sum up to 1 according to the impurity decrease within each tree, it can be understood as the relative importance of each wavelength for a particular model. As anticipated based on the results from the PCA, a vast majority of the significant wavelengths were within the VIS range, with some of them coinciding with the maximums of absorption of photosynthetic pigments. These findings align with those from similar studies where the most important features are mostly between 550 and 570 nm, and 640 and 690 nm (Adejimi et al., 2023). Nevertheless, important features were also in spectral regions within the NIR range. The first and most prominent of these regions was located between 870 and 900 nm, where numerous wavelengths demonstrated substantial importance in every model. Additionally, several other regions between 700 and 800 nm exhibited moderate importance, especially when the models were trained on data filtered based on image quality suggesting that, while a larger dataset may result in higher classification accuracies, a minimum quality of the images is critical to capture small details from the NIR range. The reasons for this relevance are still to be explored. One possible explanation is that the relevance of chlorophyll extends into this range. The "red edge" consists of the abrupt change in spectral behavior of chlorophyll from high absorbance in the VIS to high reflectance in the NIR, a property commonly used in vegetation analysis (where NIR reflectance is used) (Haboudane et al., 2004). Another hypothesis is that the specific structure of cells, or derived from cell aggregation, may result in different light absorption and scattering behavior from each taxa, producing unique reflectance properties in the NIR. Finally, and regarding the results in the blue bands mentioned in the previous sections (in relation to Figs. 1-c and 2), it is reassuring that the models that performed best in classifying the images did not identify any particularly relevant wavelength patterns in this region. However, the discrepancy between our findings and those of other publications concerning this spectrum range (~400 nm) should be considered for future uses of the

hyperspectral library created.

Overall, the models trained on data pre-processed with SG and FD achieved accuracies up to 0.81 to 0.88, meaning that these classifiers were able to accurately determine the cyanobacterial taxa of the new provided images in 81 to 88 % of the cases, with VIS and NIR wavelengths playing a critical role, and with impressive performance in identifying *Microcystis* and *Chrysosporum* genera. These results are particularly surprising and showcase the impressive capabilities of our models in classifying unfamiliar images. Moreover, it is important to consider that the data used to train these models came primarily from the CLN dataset, which consisted of cultures that were grown under varying light and nutrient conditions. As anticipated, these varying conditions induced a high degree of variability in appearance, pigment concentrations and ratios, and spectral behavior, which resulted in spectral ranges from the five cyanobacteria to overlap, potentially threatening the performance of the classifiers. Nevertheless, the results indicate that, even with the induced high variability, the Random Forest algorithm combined with SG and FD pre-processing methods can successfully differentiate among these cyanobacteria with high accuracy levels.

Our study offers a promising tool validated at laboratory scale in relevant bloom-forming potentially toxic taxa. This could pose a baseline work towards the ultimate goal of effectively applying HS remote sensing in the field, once some challenges are met via future field studies. Firstly, although great variability was induced, in natural ecosystems this variability is likely to be greater. For instance, in the field, cyanobacteria exhibit various morphologies such as unicellular, filamentous, or colonial forms, that can be lost when cultured in the lab (Salmaso et al., 2015). For example, in aquatic ecosystems, *Microcystis* forms dense floating aggregations of biomass by growing as colonies, structure that tends to be lost in lab cultures, where it grows spreading homogenously throughout the medium. Another example of natural variability in cyanobacteria is the presence of gas vesicles in some genera. For instance, *Planktothrix* genus has been observed to have different amounts of gas vesicles, which can affect their position in the water column modifying their interaction with the environment and basic resources such as light and nutrients, likely resulting in a greater range of spectral responses (D'Alelio et al., 2011). On the other hand, the application of machine learning algorithms is becoming increasingly popular in the field of cyanobacterial blooms. However, while various datasets are emerging due to more automated sampling methods, usually only one machine learning technique is applied to analyze each of them making comparison challenging (Rousso et al., 2020). Here we demonstrate the potential of Random Forest to handle HS data from cyanobacterial cultures, providing accurate classification models; in future studies, different machine learning algorithms should be trained on this data, not only to improve the accuracy of the results, but also to facilitate comparison. A final consideration to bear in mind is that the experiment was carried out with one representative species for each genus, therefore it is not possible to determine whether these results are genus, species or strain specific. As suggested by other authors (Legleiter et al., 2022), a possible solution to these limitations could be the creation or expansion of a cyanobacterial hyperspectral library, such as the one created in this study, to include images from more genera and species, thus becoming more complete and representative. Ideally, not only images from cultures in the lab, but also fresh cyanobacterial biomass or field images should be incorporated into this library. In fact, this could facilitate the transition from lab-based analyses to practical field applications of HS imagery, as it has been pointed by Legleiter et al. (2022) that showed the potential from cyanobacterial HS reflectance data acquired in the lab, to be extrapolated into the field. This fact, combined with the increasing sources of hyperspectral data from water bodies, such as the provided by unmanned autonomous vehicles (Kislik et al., 2018), or ground-based multispectral cameras (Zhao et al., 2021), along with the increased quality of these data, supported by HS technology aimed to improve remote sensing estimations from satellite imagery (Goyens et al., 2021), highlights the capabilities of these techniques in the early management of cyanobacterial blooms.

4. Conclusions

Advances in cyanobacterial early warning are critical given the expected worsening of CyanoHABs under future human and climate change scenarios. Based on the analysis of a HS imagery collection created with images from five bloom-forming cyanobacterial taxa with toxicological relevance in the context of CyanoHABs, we conclude that:

- Hyperspectral imagery of cyanobacteria, combined with machine learning techniques, has the potential to effectively discriminate among the five taxa.
- The classification remains accurate even when a wide pigment and spectral variability is induced by growing the cultures under different light and nutrient conditions, and when images from different stages of the cyanobacterial life cycle are considered.
- The Random Forest algorithm is shown to be effective for cyanobacterial classification, particularly when trained on spectral data pre-processed with smoothing techniques and first derivatives.
- Reflectance from wavelengths in both the visible and near-infrared regions play a critical role in successful classification.

Our findings highlight the importance of considering hyperspectral technology to enable the early identification of cyanobacterial genera with potential toxicity, which may facilitate timely and effective prevention of the associated risks.

CRediT authorship contribution statement

Claudia Fournier: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Antonio Quesada:** Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing – original draft, Writing – review & editing. **Samuel Cirés:** Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing – review & editing, Writing – original draft. **Mohammadmehdi Saberioon:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors would like to thank the EU, the Spanish Ministry of Science and Innovation and the German Federal Ministry of Education and Research (BMBF) for funding, in the frame of the collaborative international consortium AIHABs financed under the ERA-NET AquaticPollutants Joint Transnational Call (GA N° 869178). Agencia Estatal de Investigación (Spain) funded the project through the grant PCI2021-121915. Authors also acknowledge funding from Helmholtz Information and Data Science Academy (HIDA) (Foundation grant No 15052). The authors thank the reviewers for their valuable comments and feedback.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2024.172741.

References

Adejimi, O.E., Sadhasivam, G., Schmilovitch, Z., Shapiro, O.H., Herrmann, I., 2023. Applying hyperspectral transmittance for inter-genera classification of cyanobacterial and algal cultures. Algal Res. 71, 103067 https://doi.org/10.1016/j.algal.2023.103067.

Almuhtaram, H., Kibuye, F.A., Ajjampur, S., Glover, C.M., Hofmann, R., Gaget, V., Owen, C., Wert, E.C., 2021. State of knowledge on early warning tools for cyanobacteria detection. Ecol. Indic. 133 https://doi.org/10.1016/j.ecolind.2021.108442.

Belgiu, M., Drăgu, L., 2016. Random forest in remote sensing: a review of applications and future directions. ISPRS J. Photogramm. Remote Sens. 114, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Chapra, S.C., Boehlert, B., Fant, C., Bierman, V.J., Henderson, J., Mills, D., Mas, D.M.L., Rennels, L., Jantarasami, L., Martinich, J., Strzepek, K.M., Paerl, H.W., 2017. Climate change impacts on harmful algal blooms in U.S. freshwaters: a screening-level assessment. Environ. Sci. Technol. 51, 8933–8943. https://doi.org/10.1021/acs.est.7b01498.

Cirés, S., Wörmer, L., Timón, J., Wiedner, C., Quesada, A., 2011. Cylindrospermopsin production and release by the potentially invasive cyanobacterium *Aphanizomenon ovalisporum* under temperature and light gradients. Harmful Algae 10, 668–675. https://doi.org/10.1016/j.hal.2011.05.002.

D'Alelio, D., Gandolfi, A., Boscaini, A., Flaim, G., Tolotti, M., Salmaso, N., 2011. *Planktothrix* populations in subalpine lakes: selection for strains with strong gas vesicles as a function of lake depth, morphometry and circulation. Freshw. Biol. 56, 1481–1493. https://doi.org/10.1111/j.1365-2427.2011.02584.x.

Dev, P.J., Sukenik, A., Mishra, D.R., Ostrovsky, I., 2022. Cyanobacterial pigment concentrations in inland waters: novel semi-analytical algorithms for multi- and hyperspectral remote sensing data. Sci. Total Environ. 805 https://doi.org/10.1016/j.scitotenv.2021.150423.

Dittmann, E., Fewer, D.P., Neilan, B.A., 2013. Cyanobacterial toxins: biosynthetic routes and evolutionary roots. FEMS Microbiol. Rev. 37, 23–43. https://doi.org/10.1111/j.1574-6976.2012.12000.x.

Fisher, R.A., 1925. Intraclass correlations and the analysis of variance. In: Statistical Methods for Research Workers, pp. 187–210.

Gholizadeh, A., Boruvka, L., Saberioon, M.M., Kozák, J., Vašát, R., Nemecek, K., 2015. Comparing different data preprocessing methods for monitoring soil heavy metals based on soil spectral features. Soil Water Res. 10, 218–227. https://doi.org/10.17221/113/2015-SWR.

Goyens, C., de Vis, P., Hunt, S.E., 2021. Automated generation of hyperspectral fiducial reference measurements of water and land surface reflectance for the hypernets networks. In: International Geoscience and Remote Sensing Symposium (IGARSS) 2021-July, pp. 7920–7923. https://doi.org/10.1109/IGARSS47720.2021.9553738.

Haboudane, D., Miller, J.R., Pattey, E., Zarco-Tejada, P.J., Strachan, I.B., 2004. Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies: modeling and validation in the context of precision agriculture. Remote Sens. Environ. 90, 337–352. https://doi.org/10.1016/j.rse.2003.12.013.

Huisman, J., Codd, G.A., Paerl, H.W., Ibelings, B.W., Verspagen, J.M.H., Visser, P.M., 2018. Cyanobacterial blooms. Nat. Rev. Microbiol. 16, 471–483. https://doi.org/10.1038/s41579-018-0040-1.

Jiang, G., Loiselle, S.A., Yang, D., Ma, R., Su, W., Gao, C., 2020. Remote estimation of chlorophyll a concentrations over a wide range of optical conditions based on water classification from VIIRS observations. Remote Sens. Environ. ISSN: 0034-4257 241, 111735 https://doi.org/10.1016/j.rse.2020.111735.

Jolliffe, I.T., 1986. Graphical representation of data using principal components. In: Principal Component Analysis. Springer, New York, NY, pp. 64–91. https://doi.org/10.1007/978-1-4757-1904-8_5.

Kislik, C., Dronova, I., Kelly, M., 2018. UAVs in support of algal bloom research: a review of current applications and future opportunities. Drones 2, 35. https://doi.org/10.3390/drones2040035.

Kruskal, W.H., 1952. A nonparametric test for the several sample problem. Ann. Math. Stat. 23, 525–540. https://doi.org/10.1214/aoms/1177729332.

Kudela, R.M., Palacios, S.L., Austerberry, D.C., Accorsi, E.K., Guild, L.S., Torres-Perez, J., 2015. Application of hyperspectral remote sensing to cyanobacterial blooms in inland waters. Remote Sens. Environ. 167, 196–205. https://doi.org/10.1016/j.rse.2015.01.025.

Lawrenz, E., Fedewa, E.J., Richardson, T.L., 2011. Extraction protocols for the quantification of phycobilins in aqueous phytoplankton extracts. J. Appl. Phycol. 23, 865–871. https://doi.org/10.1007/s 10811-010-9600-0.

Legleiter, C.J., King, T.V., Carpenter, K.D., Hall, N.C., Mumford, A.C., Slonecker, T., Graham, J.L., Stengel, V.G., Simon, N., Rosen, B.H., 2022. Spectral mixture analysis for surveillance of harmful algal blooms (SMASH): a field-, laboratory-, and satellite-based approach to identifying cyanobacteria genera from remotely sensed data. Remote Sens. Environ. 279, 113089 https://doi.org/10.1016/j.rse.2022.113089.

Levene, H., 1960. Robust tests for equality of variances. In: Contributions to Probability and Statistics, pp. 278–292.

Li, Z., Zhu, X., Wu, Z., Sun, T., Tong, Y., 2023. Recent advances in cyanotoxin synthesis and applications: a comprehensive review. Microorganisms. https://doi.org/10.3390/microorganisms11112636.

Liu, J.Y., Zeng, L.H., Ren, Z.H., 2021. The application of spectroscopy technology in the monitoring of microalgae cells concentration. Appl. Spectrosc. Rev. 56, 171–192. https://doi.org/10.1080/05704928.2020.1763380.

Lloyd, S.P., 1982. Least squares quantization in PCM. IEEE Trans. Inf. Theory 28, 129–137. https://doi.org/10.1109/TIT.1982.1056489.

Malhotra, A., Örmeci, B., 2022. Monitoring of cyanobacteria using derivative spectrophotometry and improvement of the method detection limit by changing pathlength. Water Supply 22, 2914–2928. https://doi.org/10.2166/ws.2021.427.

Marker, A., 1980. The Measurement of Photosynthetic Pigments in Freshwaters and Standardization of Methods: Conclusions and Recommendations.

Pamula, A.S.P., Gholizadeh, H., Krzmarzick, M.J., Mausbach, W.E., Lampert, D.J., 2023. A remote sensing tool for near real-time monitoring of harmful algal blooms and turbidity in reservoirs. J. Am. Water Resour. Assoc. 59 (5), 929–949. https://doi.org/10.1111/1752-1688.13121.

Pedregosa, F., Michel, V., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., Perrot, É., Duchesnay, É., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Perkel, J.M., 2018. Why Jupyter is data scientists' computational notebook of choice. Nature 563, 145–146. https://doi.org/10.1038/d41586-018-07196-1.

Przytulska, A., Bartosiewicz, M., Vincent, W.F., 2017. Increased risk of cyanobacterial blooms in northern high-latitude lakes through climate warming and phosphorus enrichment. Freshw. Biol. 62, 1986–1996. https://doi.org/10.1111/fwb.13043.

Rippka, R., 1988. [1] Isolation and purification of cyanobacteria. Methods Enzymol. 167, 3–27. https://doi.org/10.1016/0076-6879(88)67004-2.

Rousso, B.Z., Bertone, E., Stewart, R., Hamilton, D.P., 2020. A systematic literature review of forecasting and predictive models for cyanobacteria blooms in freshwater lakes. Water Res. 182, 115959 https://doi.org/10.1016/j.watres.2020.115959.

Saberioon, M., Císař, P., Labbé, L., Souček, P., Pelissier, P., 2019. Spectral imaging application to discriminate different diets of live rainbow trout (*Oncorhynchus mykiss*). Comput. Electron. Agric. 165, 104949 https://doi.org/10.1016/j.compag.2019.104949.

Salmaso, N., Naselli-Flores, L., Padisák, J., 2015. Functional classifications and their application in phytoplankton ecology. Freshw. Biol. 60, 603–619. https://doi.org/10.1111/fwb.12520.

Salmi, P., Eskelinen, M.A., Leppänen, M.T., Pölönen, I., 2021. Rapid quantification of microalgae growth with hyperspectral camera and vegetation indices. Plants 10, 341. https://doi.org/10.3390/plants10020341.

Salmi, P., Calderini, M., Pääkkönen, S., Taipale, S., Pölönen, I., 2022. Assessment of microalgae species, biomass, and distribution from spectral images using a convolution neural network. J. Appl. Phycol. 34, 1565–1575. https://doi.org/10.1007/s10811-022-02735-w.

Sanseverino, I., Conduto, D., Pozzoli, L., Dobricic, S., Lettieri, T., 2016. Algal Bloom and Its Economic Impact. European Commission, Joint Research Centre Institute for Environment and Sustainability.

Shapiro, S.S., Wilk, M.B., 1965. An analysis of variance test for normality (complete samples)†. Biometrika 52, 591–611. https://doi.org/10.1093/biomet/52.3-4.591.

Sokolova, M., Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Inf. Process. Manag. 45, 427–437. https://doi.org/10.1016/j.ipm.2009.03.002.

Solovchenko, A., 2023. Seeing good and bad: optical sensing of microalgal culture condition. Algal Res. 71, 103071 https://doi.org/10.1016/j.algal.2023.103071.

Svirčev, Z., Lalić, D., Bojadžija Savić, G., Tokodi, N., Drobac Backović, D., Chen, L., Meriluoto, J., Codd, G.A., 2019. Global geographical and historical overview of cyanotoxin distribution and cyanobacterial poisonings. Arch. Toxicol. 93, 2429–2481. https://doi.org/10.1007/s00204-019-02524-4.

Whitton, B.A., Potts, M., 2012. Introduction to the cyanobacteria. In: Ecology of Cyanobacteria II: Their Diversity in Space and Time. https://doi.org/10.1007/978-94-007-3855-3_1.

Wyman, M., Fay, P., 1986. Underwater light climate and the growth and pigmentation of planktonic blue-green algae (Cyanobacteria) I. The influence of light quantity. Proc. R. Soc. Lond. B Biol. Sci. 227, 367–380. https://doi.org/10.1098/rspb.1986.0027.

Xi, H., Hieronymi, M., Röttgers, R., Krasemann, H., Qiu, Z., 2015. Hyperspectral Differentiation of Phytoplankton Taxonomic Groups: A Comparison Between Using Remote Sensing Reflectance and Absorption Spectra, vol. 7, pp. 14781–14805. https://doi.org/10.3390/rs71114781.

Zhao, H., Li, J., Yan, X., Fang, S., Du, Y., Xue, B., Yu, K., Wang, C., 2021. Monitoring cyanobacteria bloom in Dianchi Lake based on ground-based multispectral remote-sensing imaging: preliminary results. Remote Sens. 13, 3970. https://doi.org/10.3390/RS13193970.