Bioresource Technology 435 (2025) 132943

Contents lists available at ScienceDirect

Bioresource Technology

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

Case Study

Artificial intelligence based hyperspectral biomass estimator for cyanobacteria cultivation☆

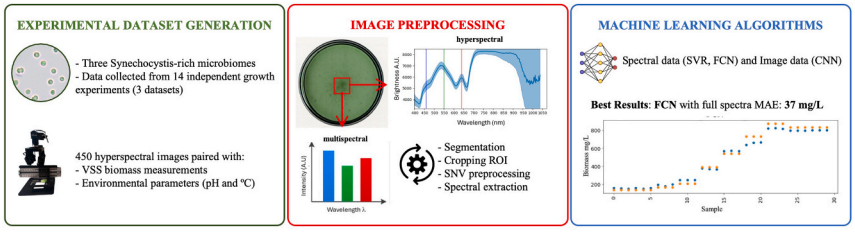
J.M. Fernández Montenegro [a], I. Pérez Couñago [a], S. Iglesias Moreira [a], F. Rodríguez Lorenzo [a], P. Villar Sola [b], E. Pancorbo González [b], B. Altamira-Algarra [c], J. Garcia [d], S. Muíños Lanín [a], E. Gonzalez-Flo [c,*]

[a] Smart Systems and Smart Manufacturing (S3M) Department, AIMEN technology centre, Porriño, Spain
[b] Environmental Technologies (Entech) Department, AIMEN technology centre, Porriño, Spain
[c] GEMMA - Group of Environmental Engineering and Microbiology, Department of Civil and Environmental Engineering, Escola d'Enginyeria de Barcelona Est (EEBE), Universitat Politècnica de Catalunya·BarcelonaTech, Av. Eduard Maristany 16, Building C5.1, E-08019 Barcelona, Spain
[d] GEMMA - Environmental Engineering and Microbiology Research Group, Department of Civil and Environmental Engineering, Universitat Politècnica de Catalunya·BarcelonaTech, c/ Jordi Girona 1-3, Building D1, E-08034 Barcelona, Spain

HIGHLIGHTS

- Dataset of 205 biomass measurements and 450 hyperspectral images was generated.
- Three machine learning models trained to predict biomass from spectral and image data.
- FCN model achieved the best accuracy with a mean absolute error of 37.33 mg/L.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Non-invasive monitoring
Machine learning
Photobioreactors operation
Cyanobacteria monitoring

ABSTRACT

Hyperspectral imaging combined with machine learning offers an innovative approach for biomass monitoring at laboratory and industrial scale, but a proof-of-concept linking hyperspectral data to biomass prediction remains limited. This study fills that gap by creating a dataset with 205 biomass measurements and 450 hyperspectral images from three cyanobacteria-rich microbiomes. Data were acquired using a compact push-broom camera, followed by image preprocessing to extract spectral information for training three machine learning algorithms based on either spectral or image data. The Fully Connected Neural Network model achieved the highest accuracy, predicting biomass levels with a mean absolute error of 37 mg/L (below 4 %). Notably, a simplified multispectral model using only three wavelengths reached comparable accuracy, highlighting the potential of low-cost multispectral systems. This study demonstrates the feasibility of hyperspectral and multispectral imaging for biomass estimation in cyanobacterial cultures and supports the development of real-time, non-invasive monitoring tools for photosynthetic bioprocesses.

☆ This article is part of a special issue entitled: 'PHOTOREFINERIES 2024' published in Bioresource Technology.

* Corresponding author.
E-mail address: eva.gonzalez.flo@upc.edu (E. Gonzalez-Flo).

https://doi.org/10.1016/j.biortech.2025.132943
Received 28 January 2025; Received in revised form 23 June 2025; Accepted 4 July 2025
Available online 5 July 2025

1. Introduction

In situ and continuous monitoring techniques are essential to ensure efficient operation of photosynthetic photobioreactors (PBRs) for bioproducts production. These bioproducts can range from biomass itself to specific compounds, such as pigments, bioplastics, or lipids, among others (Altamira-Algarra et al., 2024b; Senatore et al., 2023). Accurate biomass quantification is particularly important for optimizing productivity and process control. However, current biomass measurement techniques are often time-consuming, costly, or require destructive sampling. Direct biomass measurements such as dry weight determination and cell counting offer high precision but are labor-intensive and invasive. In contrast, indirect methods like optical density at 750 nm (OD_{750}), turbidity, or chlorophyll-based fluorescence are rapid and non-invasive but can lack robustness and reproducibility due to physiological changes in the culture or environmental variability (Kudela et al., 2015; Ogashawara et al., 2013).

Hyperspectral and multispectral imaging systems provide a promising alternative. These technologies enable the acquisition of both spatial and spectral information, allowing detailed characterization of cultures in a non-invasive and continuous manner.

First, unlike single-point optical methods, hyperspectral imaging captures reflectance across a wide spectral range, enabling detection of multiple biological signals including pigments, biomass concentration, and physiological state (Lorenzo et al., 2022; Pääkkönen et al., 2024). By correlating specific spectral bands with known biomarkers, such as chlorophyll levels or other pigments, it becomes possible to estimate culture's condition and assess its overall productivity. This level of detail is often unattainable with conventional methods that rely on single measurements of chlorophyll, optical density or turbidity. For example, turbidity is often used to extrapolate volatile suspended solids (VSS) in experimental setups (Altamira-Algarra et al., 2024a), but this estimation introduces high variability and lacks precision (Altamira-Algarra et al., 2024a). In contrast, hyperspectral imaging provides a more consistent and reliable indicative estimation of biomass.

Secondly, hyperspectral imaging allows for real-time monitoring of biomass without the need for destructive sampling, a common limitation of current techniques such cell counting or dry weight determination. This non-invasive nature not only preserves the culture but also reduces the need of human manipulation, minimizing the risks of contamination. Lastly, hyperspectral imaging has the potential to revolutionize biomass monitoring. By integrating this technology into sensors, it enables continuous tracking of culture parameters providing real-time data. Furthermore, when combined with machine learning algorithms, hyperspectral imaging can play a pivotal role in developing predictive models for biomass estimation. These models can streamline the monitoring process, enhance decision-making, and support more informed, data-driven strategies in industrial applications, ensuring greater efficiency and sustainability in production systems.

Despite its advantages, their application for biomass measurement has been explored in only a few studies, primarily focusing on microalgae species such as *Chlorella vulgaris*, *Desmodesmus maximus*, and *Tetradesmus obliquus* (Murphy et al., 2014; Pääkkönen et al., 2024; Salmi et al., 2022; Xu et al., 2020). Hyperspectral imaging techniques to facilitate real-time estimation of culture metrics in cyanobacteria cultures has been only used in one study (Lorenzo et al., 2022). In this publication, those techniques were applied at both lab and pilot scales, where the authors correlated biomass (measured as cell dry weight) with spectral data obtained from hyperspectral cameras. Specifically, authors identified unique spectral signatures based on the absorption of light by different pigments and linked these spectral signatures to biomass levels. The results of this study highlighted the potential of hyperspectral imaging for tracking biomass growth in microalgal cultures and its versatility across different scales. However, it was limited by a small sample size, focusing solely on single strain cultures of the cyanobacterium *Synechocystis* sp. with a restricted number of samples.

Despite the potential of hyperspectral imaging, further research is needed to refine these methods for biomass estimation. This study aims to address this gap by developing an affordable biomass estimator based on coupling spectral imaging and Artificial Intelligence (AI) for cyanobacteria cultivation. To do so, it is imperative to have a broad dataset of hyperspectral images at different biomass levels and the corresponding biomass measurements. To the authors' knowledge, there is no dataset that contains such information. To address this challenge, we cultured three cyanobacteria-rich microbiomes dominated by the cyanobacterium *Synechocystis* sp. to generate comprehensive datasets that cover biomass levels across cyanobacterial growth phases. Each dataset is linked to one microbiome and includes data from multiple growth curves experiments, with associated environmental parameters such as temperature and pH. As a proof of concept, machine learning models were trained on both spectral and image-based data to predict biomass concentration, demonstrating a promising approach for continuous, non-invasive monitoring in photobioreactors (PBRs).

2. Materials and methods

2.1. Experimental data

Three microbiomes rich in *Synechocystis* sp. collected in (Altamira-Algarra et al., 2023), named R1, *R*3 and CW2 were used as inoculum for 3 L glass vertical cylindrical photobioreactors (PBRs) of 2.5 L working volume (Fig. 1a). The cultures started by inoculating biomass to reach a concentration of 100 mg/L volatile suspended solids (VSS). BG-11 medium was used with modified concentrations of bicarbonate (as a source of inorganic carbon), nitrogen, and phosphorus (100 mg IC/L, 50 mg N/L, and 0.1 mg P/L), aimed at favoring the growth of cyanobacteria over competing phototrophic organisms (Altamira-Algarra et al., 2024b). Illumination of the reactors was kept at 30 klx (approx. 420 μmol photons/$m^2 \cdot s^1$) by 200 W LED floodlight (placed at 15 cm from the reactors). This illumination followed a 15:9-hour light-to-dark cycle during growth. pH levels were continuously monitored using a pH probe (model HI1001, HANNA instruments, Italy) placed inside the reactors. pH of the cultures was controlled within a range of 7–9 by pulse CO_2 injection into the reactors. Reactors were continuously agitated by a magnetic stirrer ensuring complete mixing. Culture temperature was kept at around 30 °C. Additional information on the reactors and their operation can be found in (Altamira-Algarra et al., 2024a,b).

Three distinct datasets corresponding to the different microbiomes were collected. Each dataset includes multiple growth curves, with each curve representing a separate experiment. For each dataset, several growth curves were analyzed (experiments), as summarized in Table 1. Dataset 1 included five different growth curves for microbiome *R*3. Samples were collected three times per day, and for each sample, biomass as VSS, pH, temperature, and hyperspectral imaging data were analyzed in both transmittance and reflectance modes. Dataset 2 and 3 consisted of microbiome R1 and microbiome CW2, respectively; in both datasets, one sample was taken per day and analyse for VSS, pH, temperature, and hyperspectral imaging in transmittance mode. Each experiment involved a single inoculation of cyanobacteria, and subsequent data points were collected as the cells grew under the same initial conditions.

2.2. Sample collection and analysis

2.2.1. Biomass analysis, pH and temperature

Biomass was quantified as VSS via a drying process of a 10 mL sample extracted from the photobioreactor and determined according to procedures in (Amerian Public Health Association, 2012). pH levels were continuously monitored using a pH probe (model HI1001, HANNA instruments, Italy) placed inside the reactors. The pH data were recorded at 5 min intervals using software PC400 (Campbell Scientific). Temperature was monitored using a TESTO 174H data logger, equipped with

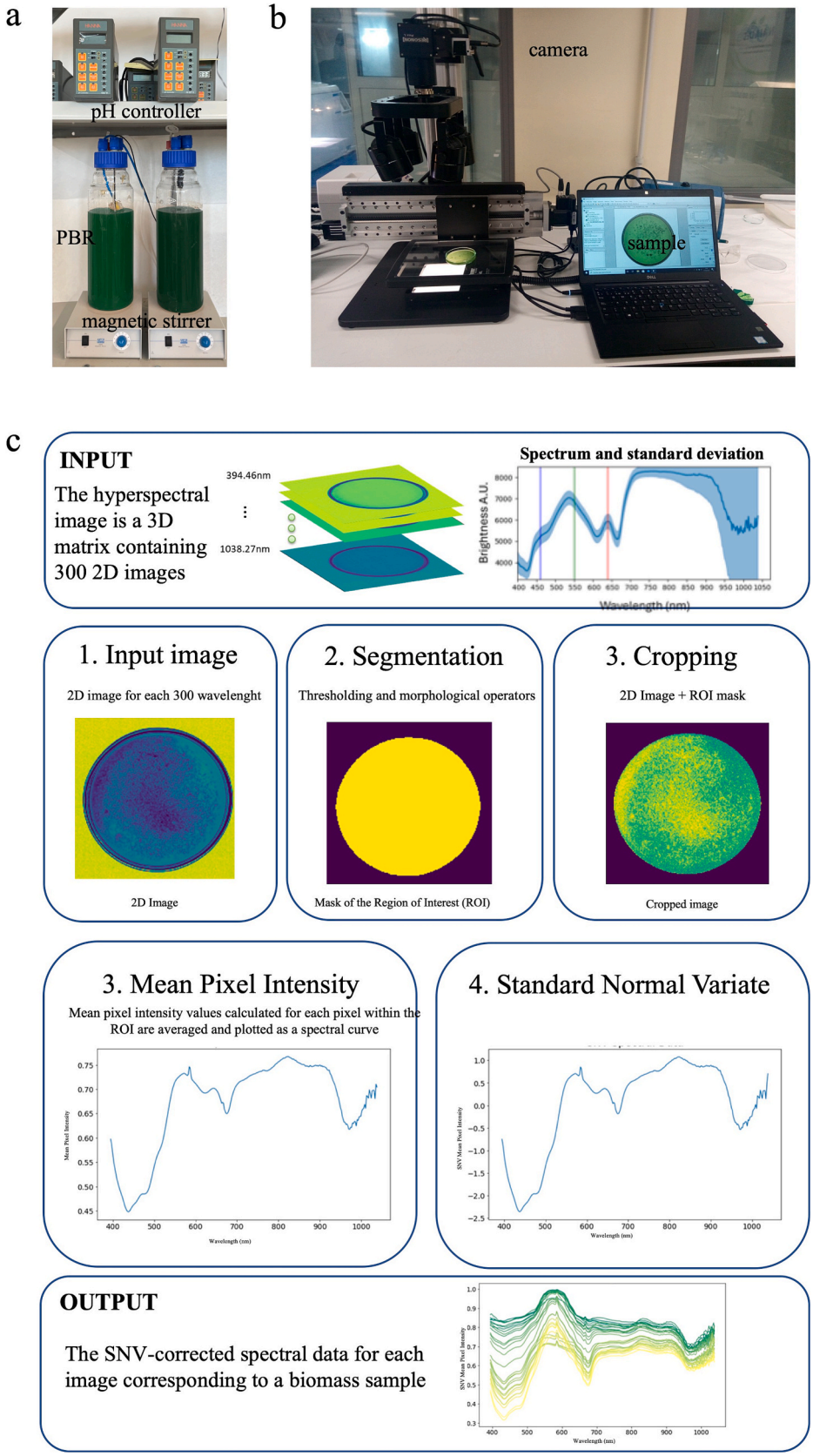


Fig. 1. Hyperspectral imaging setup. a) Photobioreactors setup for cyanobacterial microbiomes growth experiments. b) Hyperspectral camera and setup for image acquisition. c) Hyperspectral imaging processing steps: 1. Input image consisting on a 3D matrix containing the 300 2D image, together with the mean spectrum (represented in dark blue) and the standard deviation (depicted in a lighter blue shade). 2. Segmentation; 3. Cropping; 3 Mean pixel intensity spectra for each pixel in the region of interest. Standard Normal Variate of the spectra; Output spectra.

Table 1
Overview of datasets used for image acquisition showing microbiome type, experiment (growth curve) total growth days, number of biomass samples, and the number of hyperspectral images. For all samples, images were taken in triplicates.

	Microbiome	Experiment	Growth days	Samples	N° of hypercubes
Set 1	R3	1	19	29	59
		2	19	29	71
		3	50	53	112
		4	44	55	138
		5	28	19	30
Set 2	CW2	6	7	3	5
		7	7	2	2
		8	7	2	3
		9	7	3	5
		10	7	3	5
Set 3	R1	11	7	3	6
		12	7	3	4
		13	7	3	4
		14	7	3	6

a TC temperature sensor and an internal capacitive humidity sensor. The device operates within a measurement range of –20 to + 70 °C, with a temperature accuracy of ± 0.5 °C and a relative humidity range of 0–100 % RH (non-condensing atmosphere). Temperature sensor was placed adjacent to the photobioreactor (PBR). pH and temperature readings were recorded each time a sample was extracted for biomass analysis and imaging.

2.2.2. Hyperspectral image acquisition

Samples (5—10 mL) for image acquisition were taken from each PBR periodically as depicted in Table 1. The hyperspectral measurements (see supplementary materials) were acquired in 300 specific wavelengths covering the visible and part of the near infrared (IR) spectra (394.46 nm to 1038.27 nm) in two modes: transmittance and reflectance. These 300 images at specific wavelengths constitute what it is called a hypercube. For all samples, hyperspectral hypercubes were captured in duplicate or triplicate by re-imaging the same sample after shaking. In these cases, two or three images correspond to a single biomass measurement. For Dataset 1, hypercubes were acquired in both transmittance and reflectance modes, whereas for Datasets 2 and 3, imaging was performed only in transmittance mode. The acquisition of 300 spectral bands, ranging from 400 to 1000 nm, provides a detailed spectral signature essential for distinguishing subtle variations in the optical properties of the samples (see supplementary material). This high spectral resolution enables the identification of specific absorption and reflectance features that correlate with the chemical and physical characteristics relevant to our study. Moreover, it facilitates advanced feature selection and dimensionality reduction techniques, enhancing the performance and generalizability of Machine Learning models. Therefore, the 300-band configuration, combined with an extensive dataset of 450 hyperspectral cubes and 205 correlated VSS measurements (Table 1), was instrumental in building a robust predictive model.".

2.2.2.1. Hyperspectral technology. Images were acquired with the Pika L camera system from Resonon based on push broom technology. A detailed scheme of the technology is shown in Fig. 1b. It consists of a small and compact hyperspectral sensor adapted to field measurements. The sensor covers a spectral range from 400 to 1,000 nm, with 281 spectral channels. It has a spectral resolution of 2.7 nm FWHM, with a spectral bandwidth of approximately 2.1 nm. The camera features spatial pixels of 900 and a max frame rate of 249 fps. The sensor has a quantum efficiency higher than 55 % (55 % @ 700 nm and reaching more than 70 % between 500 and 600 nm). Global system quantum efficiency (Q.E.) is higher than 55 % in the range from 500 to 600 nm. A 20 mm focal lens objective allowed working directly on samples collected directly from the PBRs. For analysis and acquisition, Spectronon software from Resonon was used, as it has a complete library with image processing tools and allows integration of Python customized plugins.

2.2.2.2. Transmittance and reflectance acquisition modes. The Benchtop System from Resonon (Fig. 1b) assembled with a Pika-L hyperspectral camera (https://resonon.com/Pika-L) can acquire images either in transmittance (transmitted light) as well as reflectance (reflected light) modes. Samples were located in round borosilicate glass Petri dishes with a diameter of 55 mm and a height of 15 mm that sit onto a clear stage and can be illuminated from either above (reflectance) or below (transmittance). During data acquisition the stage moves, translating the sample beneath the imager. The imager and stage are controlled using Spectronon software. Altogether, this system combined with the hyperspectral camera is designed to provide high-quality imaging, with features such as low stray light, low optical distortions, and excellent imaging quality.

2.2.2.3. Calibration and image parameter adjustment. The calibration process for hyperspectral system consists in the correction of the flat field illumination and the correction of image consistency (the parameters of the stepper motor that control the table in which the sample is placed are defined in order to ensure good image reconstruction and avoid distortion). For the correction of flat field illumination, it is necessary to acquire one dark frame (dark reference), recorded with no light incident on the camera sensor (covering the lens of the hyperspectral camera), and to record a white frame with a material of known reflectance, in order to reach the maximum illumination. This process is done once before each experimental campaign. For white reference acquisition, in reflection configuration, we use a white Teflon plate, which has high reflectivity, placed in the sample location; for transmission configuration, we directly capture light coming from light source (light is guided with optical fiber from halogen source to the homogenization panel). Data are corrected automatically by the hyperspectral system following this equation:

$$C = r_{ref} \cdot \frac{R - D}{W - D}$$

Where C is the corrected frame, rref (scalar) is the reflectance of the reference material (often assumed to be 1), R is the raw camera frame to be corrected, W is the white reference frame, and D is the dark reference frame.

Even if the white background may contribute to minimal reflectance interference, the calibration and normalisation process mitigate its effects. Additionally, to reduce potential other possible source of errors, all the images were acquired under the same controlled illumination conditions to ensure repeatability in the measurements.

2.3. Data-sets preparation for ML models

2.3.1. Hyperspectral image pre-processing

Image pre-processing was applied to obtain the spectral information contained within the hyperspectral images. Each captured image consists of a 3D matrix (cube) containing three hundred 2D images, one per wavelength (Fig. 1c). The standard deviation on the spectra approximates ± 5 % of the central value up to 800 nm. Beyond 800 nm, the error magnitude increases significantly. (Data estracted from Resonon Software).

The representative spectral information of each cube is measured as the mean pixel intensity for each wavelength. As a result, three hundred mean pixel intensity values constitute the spectral data (Fig. 1c).

A slight increase in spectral variability was observed beyond 800 nm. This is likely due to several combined factors, including reduced quantum efficiency of the CMOS sensor used (Pika L, Resonon Inc.), which

declines significantly in the NIR range (see supplementary material), as well as decreased optical transmission from the grating and lens components (see supplementary material)(Holst & Lomheim, 2018). Additionally, water absorption bands in the NIR, particularly around 840 nm and 970 nm, can reduce reflectance signal and increase noise (Gao et al., 2009; Raj et al., 2021). These factors contribute to a lower signal-to-noise ratio in this spectral region. Further details and discussion of these potential sources of variability are provided in Supplementary Note S1.

2.3.1.1. Segmentation. The segmentation process involves identifying the pixels corresponding to the region of interest. Different approaches were used based on the image characteristics using either thresholding or a circle detection algorithm.

For the thresholding-based segmentation, the process was applied to the image corresponding to the 601.39 nm wavelength from the hyperspectral hypercubes. This wavelength was selected because it provided a clear contrast between the sample (Petri dish content) and the background, facilitating accurate segmentation (Fig. 1c). Other wavelengths could also serve this purpose, as several displayed sufficient separation between sample and background. Before thresholding, the selected wavelength image was normalized to a grayscale scale from 0 to 255. A binary threshold was then applied: pixels with intensity values below a set threshold (e.g., 43 for transmittance mode) were assigned a value of 0 (black), while those above were set to 255 (white), resulting in a binary mask isolating the ROI. To further refine segmentation, morphological operations with circular shapes were used to close small gaps in the ROI and eliminate stray background pixels. Following this, an erosion operation was applied to remove remaining edge pixels, particularly to eliminate the border of the Petri dish. Erosion is a standard morphological technique, readily available in libraries such as OpenCV, and is computationally efficient for this type of binary image processing. This method was first applied to one 2D image and then extended to all 300 2D images. Thanks to its simplicity and low computational cost, the method is suitable for automated processing, provided the images are acquired under consistent and controlled conditions (e.g., stable illumination, absence of shadows). If acquisition conditions change significantly, adjustments to the selected wavelength or threshold values may be necessary.

When thresholding method was not able to recognize the region of interest, Circular Hough Transform was used (Su et al., 2012). The process involved three steps: first, applying a Gaussian blur to smooth the image, then using the Canny edge detection algorithm to identify edges, and finally applying the Hough Circles transform to detect circles based on adjustable parameters. An iterative process refined these parameters until six circles were detected. Although the detected radius was not always exact, a fixed smaller radius was used to capture the relevant region.

2.3.1.2. Cropping. Following segmentation, the background pixel values were set to zero. A fixed square of 754 x 754 pixels is defined, sufficient to enclose the entire region of interest, while reducing the overall image size, thereby reducing computational demands.

2.3.1.3. Mean pixel intensity. After cropping, a reduced version of the hyperspectral image was obtained. The mean pixel intensity for all non-zero pixels was calculated for each 2D image and the spectral data was obtained. From each hyperspectral image the spectra consist on the mean pixel intensity of each of the 300 wavelengths. The resulting data was referred to as RAW data.

2.3.1.4. Standard Normal Variate. Standard Normal Variate (SNV), spectral pre-processing method, was applied to mitigate the effects of illumination variability (Barnes et al., 1989; Shahrimie et al., 2016). Standard Normal Variate (SNV) method because it effectively corrects for multiplicative scattering effects and offset variations in spectral data, which are common issues in hyperspectral imaging of biological samples. SNV was chosen for its simplicity, wide use in the literature, and suitability for datasets where light scattering and baseline shifts are present but chemical features are subtle. The resulting spectra after this correction are referred to as SNV-preprocessed data.

2.3.1.5. Data normalization. Data were normalized to ensure that values were in a certain range in order to later accelerate machine learning algorithms convergence (Cabello-Solorzano et al., 2023). During most part of the study, all the data comprising spectra (300 wavelengths), pH, temperature and biomass, was normalized using a standard scaler, which removes the mean and divides by the standard deviation. All data were normalized to have zero mean and a standard deviation of one. Since the normalization technique could affect the model training (Huang et al., 2023), the min–max scaler was applied during the model refinement to check if the results could be improved. The min–max scaler scales the values between zero and one, by subtracting the minimum and dividing by the difference between the maximum and the minimum.

2.3.1.6. Correlation analysis. To assess the relationship between spectral data and biomass measured with VSS, Pearson's correlation analysis was applied.

2.4. Machine learning models

Three different Machine Learning models (Support Vector Regression (SVR), Fully Connected Network (FCN) and Convolutional Neural Network (CNN)) were employed to estimate biomass from hyperspectral images. These models differ both in the algorithm used and the type of input data. Specifically, the SVR and FCN models utilized spectral information, while the CNN model processed multispectral images.

The selection of these models was based on the dataset characteristics and the objective of evaluating different approaches for biomass estimation. The SVR method was selected for its ability to deliver accurate predictions with a relatively small datasets, typically consisting of hundreds of samples (Rivas-Perea et al., 2012). Given the increasing success of deep learning models for complex, high-dimensional data, a Fully Connected Network (FCN) was also tested using spectral information to evaluate its potential for improved feature extraction, despite the limited dataset size (Alzubaidi et al., 2021). Additionally, an image-based CNN model was developed to explore whether spatial information from hyperspectral images could provide added predictive value beyond the averaged spectral features.

Both the FCN and SVR models were trained using two types of input data: the full hyperspectral information (300 wavelengths) and a reduced multispectral version, where the spectral data was condensed to only 3 selected wavelengths of interest (439, 620 and 650 nm) (Lorenzo et al., 2022) (see supplementary materials).

For the CNN algorithm the hyperspectral images were reduced to a multispectral version focused on the 3 wavelengths of interest (439, 620 and 650 nm) to reduce the training computational cost. In this case the images were cropped to a dimension of 58x58 (Lorenzo et al., 2022). Table S1 shows the values of parameters used for the SVR model and the values of the layers and output shapes for the FCN and CNN models.

Regarding the data splitting strategy, the dataset was divided by entire experiments rather than by randomly selecting individual data points. Each experiment represented a distinct growth batch, with data collected over a 1–2 year period. This splitting strategy was chosen to mimic real-world application scenarios, where predictive models must generalize to new, unseen growth batches. While splitting by experiments could introduce bias if the training set lacks variability, this approach also reveals such biases, as poor generalization becomes evident in the test performance. To mitigate this risk, multiple

experiments were included in the training set, ensuring diversity and robustness.

For each round of model development, the data allocated for training was further divided into training and internal validation subsets (80 %–20 %) to monitor model performance and prevent overfitting. The final evaluation was performed on separate experiments that were not used during training, serving as an independent test set.

The metrics used as model performance indicators were the Mean Absolute Error (MAE) and the Mean Squared Error (MSE), R^2 and percentage of prediction error.

Although the dataset includes 205 biomass-labeled samples and 450 hypercubes, it spans a range of environmental conditions and microbiome types, enhancing its representativeness. While datasets with fewer than 1,000 samples are generally considered limited for ML training (Li et al., 2022), the structured design of this dataset and the use of model evaluation strategies such as independent test sets and standard error metrics (MAE, MSE) support a meaningful assessment of model performance and generalization capability.

2.5. Machine learning setup

The global dataset included 205 biomass measurements and 450 hyperspectral images. The data was processed at three Learning Rounds during data acquisition (training & testing with ground truth). Learning Round 1 processed data from experiments 1 to 4. Learning Round 2 processed data from experiments 1 to 4 and experiments 6 to 14. Learning Round 3 processed all the data. At each Learning Round all the models were replicated to compare their performance using RAW data and SNV data. The final biomass estimation model was derived from the 3rd Learning Round moment. Table 2 provides detail on the data used for each machine learning moment (including both testing and training).

The models were developed through successive learning rounds of training, progressively incorporating additional data to improve prediction accuracy (Table 2). At each learning round, model performance was monitored to assess improvements and identify tendencies in biomass estimation. This iterative approach was necessary, as the initial dataset was considered insufficient to develop a robust model.

Each round produced two biomass models: one based on RAW data and one SNV data. The final biomass estimation model was derived from the third round, with both RAW and SNV data used for model development. Table 2 provides detail on the data used for each machine learning round (including both testing and training).

Table 2
Overview of data used in each machine learning round, including the number of images for training and testing.

Learning round	Training /Testing	Input data	Number of images
1	Training	Experiment 1, Experiment 2, Experiment 4	268
	Testing	Experiment 3	112
2	Training	Training from Round 1	268
	Testing	Experiment 6, Experiment 7, Experiment 8, Experiment 9, Experiment 10, Experiment 11, Experiment 12, Experiment 13, Experiment 14	40
3	Training	Experiment 1, Experiment 2, Experiment 3, Experiment 4, Experiment 6, Experiment 7, Experiment 8, Experiment 9, Experiment 10, Experiment 11, Experiment 12, Experiment 13, Experiment 14	420
	Testing	Experiment 5	30

3. Results and discussion

3.1. Experimental data description

Three cyanobacteria-rich microbiomes from (Altamira-Algarra et al., 2023) were subjected to different growth experiments in 3L PBRs. As detailed in Fig. 2a, all the three microbiomes were dominated by the cyanobacterium *Synechocystis* sp.

Dataset 1 (Fig. 2b, Table 1) consisted of 5 growth experiments of microbiome *R*3 starting from 100 mg/L of and reaching up to 1000–1200 mg/L. Experiments 1–5 had a time span between 20- and 50-days. Additional details on microbiome *R*3 can be found in (Altamira-Algarra et al., 2024b). Dataset 2 (microbiome R1) and Dataset 3 (microbiome CW2) each comprised 5 and 4 one-week growth experiments, respectively. Dataset 2 reached biomass levels between 500 and 1200 mg/L, while Dataset 3 reached up to 700 mg/L. Notably, Dataset 2 exhibited greater variability in the different experiments.

A total of 205 biomass concentrations were obtained from the three datasets (Table 1). The distribution of biomass values across the three datasets is shown in Fig. 2c. The histogram shows that most biomass values fall within the lower range of 0 to 200 mg/L, with a more balanced distribution observed between 200 and 800 mg/L. Fewer samples are available for biomass values exceeding 800 mg/L, and very few samples (8) was recorded having values greater than 1000 mg/L. No data was available for biomass values above 1200 mg/L.

3.2. Spectral analysis and biomass correlation

A linear correlation analysis was performed to evaluate the relationship between the spectra obtained from hyperspectral images and the experimentally measured biomass values. For experiments 1–5 (Dataset 1), spectra were generated using both transmittance and reflectance acquisition modes. Additionally, spectra were analyzed in both RAW and SNV (Standard Normal Variate) formats, so a total of 4 spectra combinations were obtained per experiment. For illustrative purposes, Fig. 3a-d showcases the spectra alongside their corresponding Pearson correlation coefficients with the biomass measurements for Experiment 4 in Dataset 1. Data shows an inverse correlation in all cases between the biomass values and the spectral wavelengths below 700 nm.

The spectral responses of the phycocyanin and chlorophyll-a pigments are observed as valleys at 620 nm and 670 nm, respectively (Fig. 3a-d). In these plots, as the biomass concentration increases, the absorption of both phycocyanin at 620 nm and chlorophyll-a at 670 nm also increases.

By extracting direct reflectance values from the spectra at 620 nm, which corresponds to the absorption peak of phycocyanin, we can track variations in phycocyanin reflectance across consecutive days of growth. Fig. S1 (see supplementary materials) demonstrates a negative correlation between biomass and reflectance at this wavelength. This relationship was initially identified in (Lorenzo et al., 2022), though the analysis was based on a limited dataset, which only provided preliminary evidence of the potential negative correlation between biomass and spectral reflectance. Our data-set further supports a strong negative correlation between 620 nm and 670 nm peaks and biomass. On the other hand, a peak observed at 540–550 nm, characteristic of blue-green light, has been previously linked to the concentration of phycocyanin and can be used to monitor cyanobacterial growth (Lorenzo et al., 2022). However, this peak does not exhibit a strong correlation across all spectra. Therefore, relying solely on this wavelength for monitoring cyanobacterial growth may not provide reliable results in all scenarios.

The selection between transmission and reflectance modes was based, firstly, on the more accurate segmentation of the images (higher contrast image and reflection effect removing in transmittance configuration), and secondly, on the fact that the SNV Reflectance data shows a decrease in correlation in wavelengths between 500–600 nm (Fig. 3d).

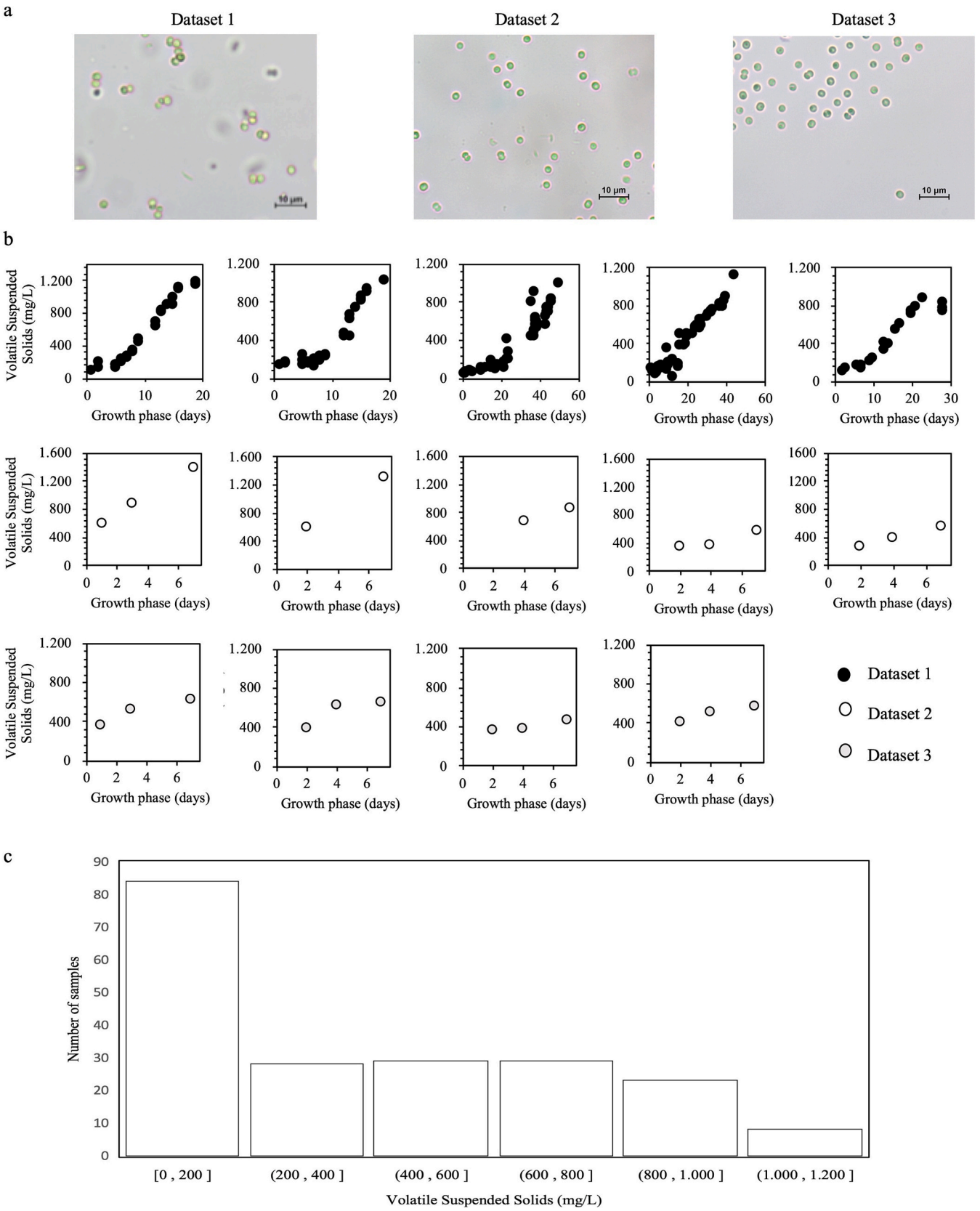
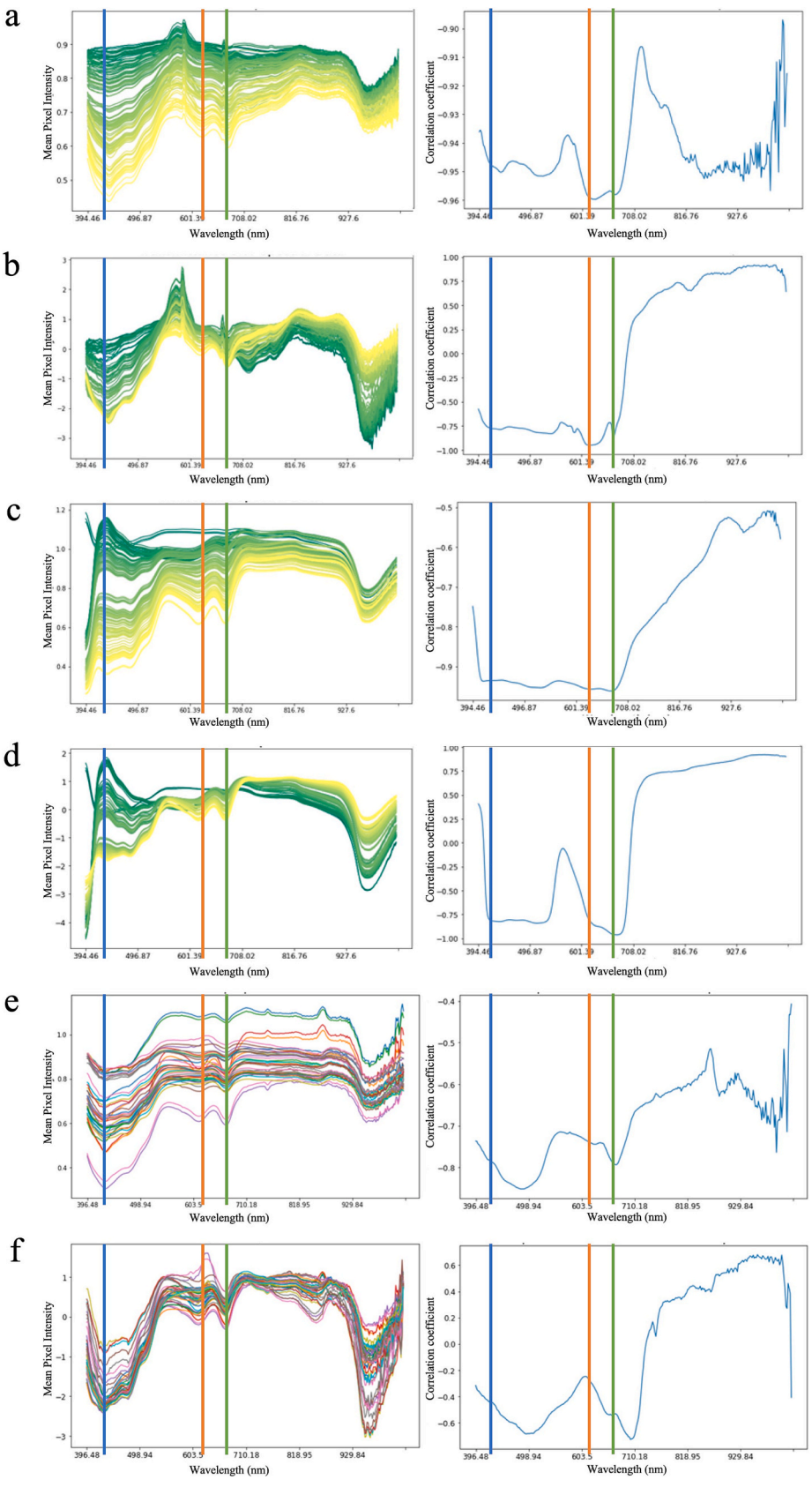


Fig. 2. Experimental datasets. a) Microscopic images of microbiomes R3, R1 and CW. b) Growth curves of the different experiments. Dataset 1 comprises 5 different experiments (plots 1-5), dataset 2 and dataset 3 consists on the repetition of 4 growth -phases of 7 days (plots 6 and 7 respectively). c) Histogram showing the distribution of biomass values across the three datasets.



(caption on next page)

Fig. 3. Spectral analysis and biomass correlation. Left plots: a) RAW Transmittance, b) SNV Transmittance, c) RAW Reflectance and d) SNV Reflectance of dataset 1 (experiment) 4. Note that each line is the result from an image taken at a certain moment in the growth curves. Right plots: corresponding Pearson correlation values between spectra and biomass measurements. e) Left plot: transmittance SNV data for dataset 2. Right plot: corresponding Pearson correlation values. f) Left plot: transmittance SNV data for dataset 3. Right plot: corresponding Pearson correlation values. Color key: dark green corresponding to biomass of less than 200 mg/L, while bright yellow corresponds to approximately 1000 mg/L of biomass. Blue vertical line at 439 nm. Orange vertical lines at 620 nm. Green vertical lines at 650 nm. Selected wavelengths were extracted based on the relevant spectral ranges identified in Lorenzo et al. (2022).

For those reasons, all subsequent experiments were conducted in Transmittance mode.

Fig. 3e-f shows the spectra and correlations for dataset 2 and dataset 3. The peak at 540–550 nm and absorption at 620 and 670 nm are also detected, although the correlation coefficients at these specific wavelengths are weaker than in dataset 1. The correlation data represent two distinct microbiome sets, R1 and CW2, which share similar microscopic microorganisms' composition but differ in their origins (Altamira-Algarra et al., 2023).

Correlation analysis has been performed across all datasets (for both RAW and SNV spectral data) to corroborate the results from individual datasets. While strong correlations observed within individual datasets may be influenced by the limited sample sizes, the overall correlation across the entire dataset remains consistent. Fig. S1 (see supplementary materials) shows that both the raw and SNV-transformed spectral data maintain strong correlations with biomass, with most p-values below 0.001. The only exceptions were observed in the SNV-transformed data at 690.82 nm and 692.97 nm, where the correlation coefficients were −0.021 and 0.102, with p-values of 0.642 and 0.029, respectively. Despite these exceptions, the vast majority of wavelengths show statistically significant correlations, supporting the relevance of the spectral data for biomass prediction.

3.3. Machine learning models

The first learning round was done with dataset 1, which included data from a single microbiome (*R*3) as both training and testing sets. Testing provide good estimation statistics, with errors below 0.3 (prediction error (p.e.) below 10 %) in Transmittance mode. Then, the same model trained with dataset 1 was used in a second learning round, but in this case the testing was done with dataset 2 and 3, by consisting of microbiomes R1 and CW2 respectively. In this scenario estimation statistics performed poorly, with values of MAE above 0.3 (p.e. 10 %) (see supplementary materials), indicating the difficulties of the trained model to predict new data from a different microbiome with a slightly different microorganism composition. In response to these challenges, a 3rd learning round was performed. In this iteration, the training phase incorporated experiments from all three datasets, and the testing was done on dataset 1 (experiment 5). To evaluate the impact of SNV, the models were trained and tested using raw (non-pre-processed) spectral data and compared to the models trained with SNV-pre-processed spectra. The performance metrics for both cases are presented in Table S2. In this learning round, from the two data types used for modelling, the SNV data performed better, indeed SNV metrics were below 0.2 (p.e. 7 %) (see supplementary materials), suggesting improved predictive capabilities compared to the initial attempt and demonstrating that SNV pre-processing provided a consistent improvement in prediction accuracy.

The best model from the 3rd learning round underwent several refinements to improve prediction accuracy. The first refinement involved replacing data standardization with normalization, scaling the values between 0 and 1. Table S2 (see supplementary materials) compares the results from the 3rd learning round (using standardized data) with those from the first data refinement (using normalized data). For clarity, metrics are presented in biomass units (mg/L) to facilitate a more comprehensive understanding of the error predictions. For both models, with and without data normalisation, the MAE values indicate a consistent prediction error of 59 mg/L of biomass (p.e. below 5 %). Fig. 3a illustrates the real vs. predicted output values from this model.

The VSS technique for biomass analysis is subject to inherent errors and limitations, typically around 10 %, which can affect the accuracy of biomass estimations (APHA, 2005; Sagrado et al., 2005). Given the 59 mg/L (p.e. below 5 %) prediction error, this represents only a small deviation from the real biomass concentration value, but it becomes a much larger discrepancy when applied to smaller biomass values. Note in Fig. 4a that for biomass values between 400 and 800 mg/L, the model's predictions are highly accurate. However, for values outside this range, the model tends to deviate significantly, highlighting the challenges of accurately predicting lower or higher biomass values. Additionally, it is important to note that the dataset contains fewer samples with biomass values greater than 800 mg/L, which may contribute to the model's decreased accuracy for these higher values.

A second model refinement was included by focusing on a multispectral version of the model so to allow the use of a more cost-efficient image acquisition sensor system. This approach probed promising predictive accuracy, probing a hyperspectral system might not be necessary. In this model refinement only wavelengths: 439, 620 and 650 nm were considered (Lorenzo et al., 2022), altogether with pH and temperature data. Prediction indexes highlight that this refinement does not induce better predictions, in fact both indexes perform worst, with a prediction error around 250 mg/L (p.e. 20 %) for both types of data (standardized and normalized).

Learning rounds were also conducted using the FCN algorithm. However, the performance metrics for FCN were generally worse than those obtained with the SVR algorithm (see supplementary materials). Notably, the results from the 3rd learning round (see supplementary materials) highlight the poor performance of the FCN model. Since the 3rd learning round results were better using the SVR algorithm, the normalization refinement was not tested on the FCN hyperspectral based model. Instead, the multispectral refinement was applied directly to both standardized and normalized SNV data (see supplementary materials). The multispectral refinement on normalized SNV data (Fig. 4b) emerged as the best data type, achieving a MAE of 37 mg/L (p.e. below 4 %). Interestingly, no significant differences in prediction accuracy were observed between lower and higher biomass values, indicating that the model's performance was consistent across the range of biomass concentrations.

Notably, the SVR model exhibited lower MAE and MSE when using hyperspectral data, but its performance deteriorated when the multispectral refinement was applied. When the input data was treated as multispectral and limited to the three most relevant wavelengths, as identified in the literature, the model's performance improved significantly, achieving a prediction error of just 37 mg/L (p.e. 3.7 %). This suggests that for the FCN model, reducing the number of input variables can enhance model accuracy, indicating that including all spectral data might introduce unnecessary noise that hampers predictions. However, the opposite behaviour is depicted by the SVR model, the reduction of spectral data when using the multispectral processing led to a significant decline in prediction accuracy, with the error increasing from 59 mg/L (p.e. 5 %) to 253 mg/L (p.e. 20 %).

The CNN model directly uses the multispectral version of the hyperspectral images, since the use of the hyperspectral images require very high computational capabilities, thus selecting only 3 out of the total 300 wavelengths (Lorenzo et al., 2022). Additionally, this model processes only the RAW data, consisting of the pixel values from the images. The model was trained with the data as described in Table 2. In

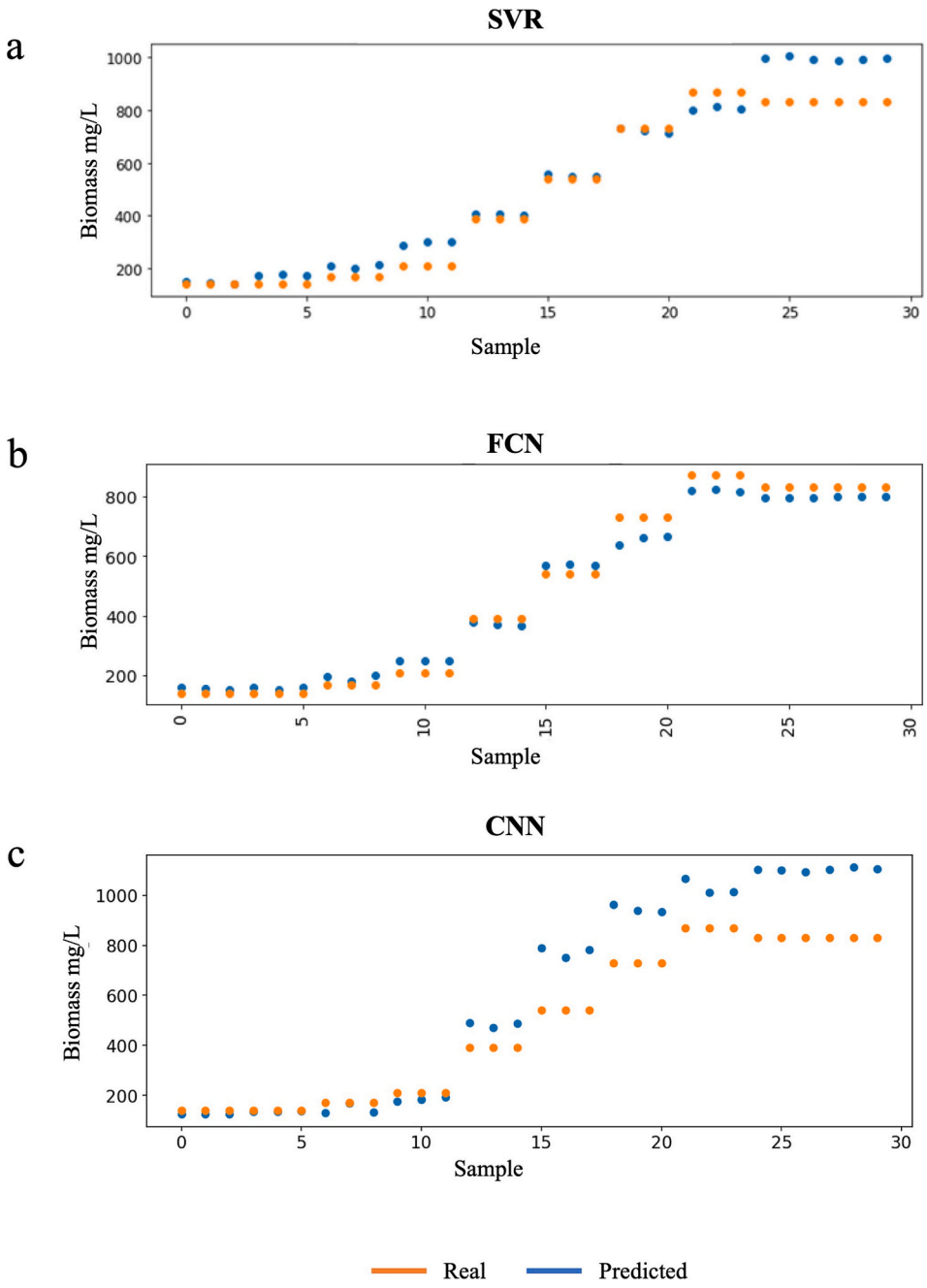


Fig. 4. Best Biomass prediction results for each ML algorithm a) SNV data type in learning 3, showing a prediction error of 59 mg/L). b) Multispectral normalised SNV data in round 3.2, with a prediction error of 37 mg/L). c) Refined multispectral model in round 3.2, with a prediction error of 137 mg/L.

the 3rd learning round, the model's performance was not highly accurate (see supplementary materials), with errors slightly above a quarter of the range of possible values. However, the model was still able to differentiate between significant changes in biomass (see supplementary materials).

The image model has also been refined to work more efficiently with a cost-efficient sensory solution. The pre-processing of the solution was simplified. Instead of working with the whole image, the refined version works with a 58x58 pixels crop. The sensory solution will capture 58x58 pixel areas of the sample; thus, no pre-processing will be required. This refinement led to improved performance, with slightly worse errors during training, but a notable improvement in testing errors, reducing the prediction error to 137 mg/L (p.e. below 12 %). Fig. 4c illustrates that, while the model performs better overall, predictions for biomass values greater than 800 mg/L remain less accurate.

The performance of the machine learning models was assessed across different dataset acquisition to evaluate whether the relationship between hyperspectral image features and biomass concentrations was maintained over time. Final models were trained using the complete dataset and their performances were compared to understand the effect of data characteristics and model architecture.

Among the models tested, Support Vector Regression (SVR) showed the best performance when trained on full hyperspectral data. However, to explore more practical and cost-effective alternatives, a reduced multispectral approach using only three wavelengths was also evaluated. In this case, the FCN outperformed the SVR, whose performance dropped significantly, indicating that model architecture plays a critical role in adapting to limited input features. This highlights the potential of FCNs for deployment in resource-constrained monitoring systems using simplified spectral setups.

Across both spectral and image-based approaches, model performance consistently decreased at biomass concentrations above 800 mg/L. This limitation can be attributed to the lack of sufficient training data in that range, resulting in reduced model generalization. These findings indicate the need for further data acquisition focused on high biomass concentrations to develop models that are robust across the full operational range.

Image-based models, specifically Convolutional Neural Networks (CNNs), performed worse overall in this study. While these models serve as a promising proof of concept for leveraging spatial features in spectral data, they require further training on larger (ideally $>$ 1,000 samples) (Li et al., 2022), more balanced datasets to achieve reliable biomass estimation. Their current performance emphasizes the importance of dataset size and structure in training high-capacity models such as CNNs.

Several factors contributed to prediction inaccuracies, including the limited sample size in certain biomass intervals and potential variability in image acquisition—such as changes in illumination, sensor noise, and slight inconsistencies in setup. Standardizing acquisition protocols and expanding the dataset would therefore be essential steps toward improving both model accuracy and robustness.

Interestingly, the multispectral models using just three selected wavelengths achieved accuracy levels that were comparable to those of models trained on the full hyperspectral dataset. This finding underscores the feasibility of using simplified and more economical spectral systems for biomass monitoring, especially in practical settings where full hyperspectral equipment may be impractical.

Lastly, prediction trends varied across training rounds. While the final models tended to slightly overestimate biomass, earlier rounds showed underestimation, revealing evolving model biases depending on dataset composition and training scope. These variations reinforce the importance of diverse and well-balanced datasets in building generalizable predictive models.

4. Conclusions

To develop an accurate biomass estimator model, first there is the need to construct a large dataset connecting hyperspectral data with biomass concentrations. Three different models have been developed to establish this links: two that utilized spectral information from the hyperspectral images, plus info about the bioreactor pH and temperature; and one directly using the hyperspectral images. Standard Normal Variate (SNV) preprocessing was applied to correct data collection issues and improve the performance of the spectral model. Both spectral and image based approached found the models performance is worst for biomass values higher than 800 mg/L, indicating the need for additional data representing higher biomass concentrations.

Furthermore, a refined version of the models was also trained so the models could be used as part of a cost-effective sensory system. The comparison between the best refined performance spectral-based solution and the refined image-based solution showed that the spectral-based testing errors are smaller. The FCN model provided the best biomass prediction with a mean absolute error of 37.337 mg/L (p.e. 4 %). Moreover, this low prediction error demonstrates the success of the dataset size and model training approach, validating the robustness of the developed biomass estimation model.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used ChatGPT to improve language and readability. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

CRediT authorship contribution statement

J.M. Fernández Montenegro: Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **I. Pérez Couñago:** Investigation. **S. Iglesias Moreira:** Writing – review & editing, Investigation. **F. Rodríguez Lorenzo:** Writing – review & editing, Investigation, Conceptualization. **P. Villar Sola:** Investigation. **E. Pancorbo González:** Investigation. **B. Altamira-Algarra:** Writing – review & editing, Investigation. **J. Garcia:** Writing – review & editing, Funding acquisition. **S. Muíños Lanín:** Funding acquisition. **E. Gonzalez-Flo:** Writing – original draft, Conceptualization.

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.

Acknowledgments

This research was supported by the European Union's Horizon 2020 research and innovation programme under the grant agreement No 101000733 (project PROMICON). B.A.A. thanks the Agency for Management of University and Research (AGAUR) for her grant [FIA-GAUR_2021]. J.G. acknowledges the support provided by the ICREA Academia program.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.biortech.2025.132943.

Data availability

Data will be made available on request.

References

Altamira-Algarra, B., Lage, A., García, J., Gonzalez-Flo, E., 2024a. Photosynthetic species composition determines bioplastics production in microbiomes: strategy to enrich cyanobacteria PHB-producers. Algal Res. 79, 103469. https://doi.org/10.1016/j.algal.2024.103469.

Altamira-Algarra, B., Lage, A., Meléndez, A.L., Arnau, M., Gonzalez-Flo, E., García, J., 2024b. Bioplastic production by harnessing cyanobacteria-rich microbiomes for long-term synthesis. Sci. Total Environ. 176136. https://doi.org/10.1016/j.scitotenv.2024.176136.

Altamira-Algarra, B., Rueda, E., Lage, A., San León, D., Martínez-Blanch, J.F., Nogales, J., García, J., Gonzalez-Flo, E., 2023. New strategy for bioplastic and exopolysaccharides production: Enrichment of field microbiomes with cyanobacteria. N. Biotechnol. 78, 141–149. https://doi.org/10.1016/j.nbt.2023.10.008.

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data 2021 8:1*, *8*(1), 1–74. https://doi.org/10.1186/S40537-021-00444-8.

Apha, 2005. Standard Methods for the Examination of Water and Wastewater. . *American Public Health Association/American Water Works Association/Water Environment Federation*, 21st Edition.

Barnes, R.J., Dhanoa, M.S., Lister, S.J., 1989. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. Appl. Spectrosc. 43 (5), 772–777. https://doi.org/10.1366/0003702894202201.

Cabello-Solorzano, K., Ortigosa de Araujo, I., Peña, M., Correia, L., Tallón-Ballesteros, J., A., 2023. The Impact of Data Normalization on the Accuracy of Machine Learning Algorithms: a Comparative Analysis. In: *Lecture Notes in Networks and Systems*. https://doi.org/10.1007/978-3-031-42536-3_33.

Gao, B.C., Montes, M.J., Davis, C.O., Goetz, A.F.H., 2009. Atmospheric correction algorithms for hyperspectral remote sensing data of land and ocean. Remote Sens. Environ. 113 (SUPPL. 1), S17–S24. https://doi.org/10.1016/J.RSE.2007.12.015.

G.C. Holst T.S. Lomheim CMOS/CCD Sensors and Camera Systems Second Edition 2018 CMOS/CCD Sensors and Camera Systems, Second Edition 10.1117/3.2524677.

Huang, L., Qin, J., Zhou, Y., Zhu, F., Liu, L., Shao, L., 2023. Normalization Techniques in Training DNNs: Methodology, Analysis and Application. IEEE Trans. Pattern Anal. Mach. Intell. 45 (8), 10173–10196. https://doi.org/10.1109/TPAMI.2023.3250241.

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.

Li, Z., Yoon, J., Zhang, R., Rajabipour, F., Srubar, W. V., Dabo, I., & Radlińska, A. (2022). Machine learning in concrete science: applications, challenges, and best practices. *Npj Computational Materials 2022 8:1*, *8*(1), 1–17. https://doi.org/10.1038/s41524-022-00810-x.

Lorenzo, F.R., Lorenzo, M.P., Castilla, L.H., Álvarez Rodríguez, J.A., Iglesias, S., Gómez, S., Fernández Montenegro, J.M., Rueda, E., Diez-Montero, R., Garcia, J., Gonzalez-Flo, E., 2022. Monitoring PHB production in Synechocystis sp. with hyperspectral images. Water Sci. Technol. 86 (1), 211–226. https://doi.org/10.2166/WST.2022.194.

Murphy, T.E., Macon, K., Berberoglu, H., 2014. Rapid algal culture diagnostics for open ponds using multispectral image analysis. Biotechnol. Prog. 30 (1), 233–240. https://doi.org/10.1002/btpr.1843.

Ogashawara, I., Mishra, D. R., Mishra, S., Curtarelli, M. P., & Stech, J. L. (2013). A Performance Review of Reflectance Based Algorithms for Predicting Phycocyanin Concentrations in Inland Waters. *Remote Sensing 2013, Vol. 5, Pages 4774-4798*, *5* (10), 4774–4798. https://doi.org/10.3390/RS5104774.

Pääkkönen, S., Pölönen, I., Raita-Hakola, A.M., Carneiro, M., Cardoso, H., Mauricio, D., Rodrigues, A.M.C., Salmi, P., 2024. Non-invasive monitoring of microalgae cultivations using hyperspectral imager. J. Appl. Phycol. 36 (4), 1653–1665. https://doi.org/10.1007/s10811-024-03256-4.

Raj, R., Walker, J.P., Vinod, V., Pingale, R., Naik, B., Jagarlapudi, A., 2021. Leaf water content estimation using top-of-canopy airborne hyperspectral data. Int. J. Appl. Earth Obs. Geoinf. 102, 102393. https://doi.org/10.1016/J.JAG.2021.102393.

Rivas-Perea, P., Cota-Ruiz, J., Chaparro, D.G., Venzor, J.A.P., Carreón, A.Q., Rosiles, J. G., Rivas-Perea, P., Cota-Ruiz, J., Chaparro, D.G., Venzor, J.A.P., Carreón, A.Q., Rosiles, J.G., 2012. Support Vector Machines for Regression: a Succinct Review of Large-Scale and Linear programming Formulations. International Journal of Intelligence Science 3 (1), 5–14. https://doi.org/10.4236/IJIS.2013.31002.

Sagrado, S., María, E. B., Medina, J., & Martín, Y. (2005). *Enfoque ISO 17025 (2 edición)*. www.aenor.es.

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 (3), 1565–1575. https://doi.org/10.1007/S10811-022-02735-W/TABLES/3.

Senatore, V., Rueda, E., Bellver, M., Díez-Montero, R., Ferrer, I., Zarra, T., Naddeo, V., García, J., 2023. Production of phycobiliproteins, bioplastics and lipids by the cyanobacteria Synechocystis sp. treating secondary effluent in a biorefinery approach. Sci. Total Environ. 857, 159343. https://doi.org/10.1016/J.SCITOTENV.2022.159343.

Shahrimie, M.A.M., Mishra, P., Mertens, S., Dhondt, S., Wuyts, N., Scheunders, P., 2016. Modeling effects of illumination and plant geometry on leaf reflectance spectra in close-range hyperspectral imaging. Workshop on Hyperspectral Image and Signal Processing, Evolution in Remote Sensing. https://doi.org/10.1109/WHISPERS.2016.8071753.

Su, Y., Liu, Y., Huang, X., 2012. Circle Detection based on Voting for Maximum Compatibility. IEICE Trans. Inf. Syst. E95.D(6), 1636–1645. https://doi.org/10.1587/TRANSINF.E95.D.1636.

Xu, Z., Jiang, Y., Ji, J., Forsberg, E., Li, Y., He, S., 2020. Classification, identification, and growth stage estimation of microalgae based on transmission hyperspectral microscopic imaging and machine learning. Opt. Express 28 (21), 30686. https://doi.org/10.1364/oe.406036.