Talanta 292 (2025) 127845

Contents lists available at ScienceDirect

Talanta

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

Comprehensive Raman spectroscopy analysis for differentiating toxic cyanobacteria through multichannel 1D-CNNs and SHAP-based explainability

María Gabriela Fernández-Manteca [a,b], Borja García García [a,b], Susana Deus Álvarez [c], Celia Gómez-Galdós [a,b], Andrea Pérez-Asensio [a,b], José Francisco Algorri [a,b,d], Agustín P. Monteoliva [c], José Miguel López-Higuera [a,b,d], Luis Rodríguez-Cobo [a,b,d,*], Alain A. Ocampo-Sosa [b,e,f,1], Adolfo Cobo [a,b,d,1]

[a] Photonics Engineering Group, Universidad de Cantabria, 39005, Santander, Spain
[b] Instituto de Investigación Sanitaria Valdecilla (IDIVAL), 39011, Santander, Spain
[c] Ecohydros S.L., 39600, Maliaño, Spain
[d] CIBER-BBN, Instituto de Salud Carlos III, 28029, Madrid, Spain
[e] Servicio de Microbiología, Hospital Universitario Marqués de Valdecilla, 39008, Santander, Spain
[f] CIBERINFEC, Instituto de Salud Carlos III, 28029, Madrid, Spain

ARTICLE INFO

Dataset link: https://doi.org/10.5281/zenodo.14727718, https://github.com/fmantecam/CyanoRamanDL

Keywords:
Cyanobacteria detection
Raman spectroscopy
One-dimensional convolutional neural networks
Harmful Algal Blooms
Water quality monitoring
Shapley Additive Explanations

ABSTRACT

Cyanobacterial blooms pose significant environmental and public health risks due to the production of toxins that contaminate water sources and disrupt aquatic ecosystems. Rapid and accurate identification of cyanobacterial species is crucial for effective monitoring and management strategies. In this study, we combined Raman spectroscopy with deep learning techniques to classify four toxic cyanobacterial species: *Dolichospermum crassum*, *Aphanizomenon* sp., *Planktothrix agardhii* and *Microcystis aeruginosa*. Spectral data were acquired using a confocal Raman microscope with a 532 nm excitation wavelength and subjected to preprocessing and filtering to enhance signal quality. We evaluated a multichannel one-dimensional convolutional neural network (1D-CNN) approach that incorporates raw spectra, baseline estimations, and preprocessed spectra. This multichannel approach improved overall classification accuracy, achieving 86% compared to 74% with a traditional single-channel 1D-CNN using only preprocessed spectra while maintaining low overfitting. Shapley Additive exPlanations (SHAP) were applied to identify critical spectral regions for classification to enhance interpretability. These findings highlight the potential of combining Raman spectroscopy with explainable deep learning methods as a powerful tool for water quality monitoring and the early detection of Harmful Algal Blooms (HABs).

1. Introduction

Cyanobacterial blooms in freshwater bodies have become a significant environmental and public health concern worldwide [1]. These microorganisms are among the oldest life forms on Earth and play a crucial role in global carbon and nitrogen cycles [2]. Under favorable conditions, such as high nutrient availability and warm temperatures, cyanobacteria can proliferate rapidly, leading to dense accumulations known as harmful algal blooms (HABs) [3]. These blooms are often characterized by the discoloration of water bodies, unpleasant odors, and the production of toxins that can harm aquatic life and human health [4].

Cyanobacteria produce various toxins, including microcystins, cylindrospermopsins, saxitoxins, and anatoxins, which can contaminate drinking water sources and recreational waters [5,6]. Exposure to these toxins may result in liver damage, neurotoxicity, and gastrointestinal illnesses in humans and animals [7]. For instance, microcystins are potent hepatotoxins and have been implicated in outbreaks of acute liver failure [8]. Moreover, cyanobacterial blooms can cause hypoxia in water bodies by depleting oxygen levels during decomposition, leading to fish death and biodiversity loss [9]. The economic impact is also significant, with increased costs for water treatment and losses in fisheries estimated to be billions of dollars annually [10].

* Corresponding author at: Photonics Engineering Group, Universidad de Cantabria, 39005, Santander, Spain.
E-mail address: luis.rodriguez@unican.es (L. Rodríguez-Cobo).
[1] These authors contributed equally to this work and share senior authorship.

https://doi.org/10.1016/j.talanta.2025.127845
Received 20 December 2024; Received in revised form 24 February 2025; Accepted 25 February 2025
Available online 7 March 2025

The increasing frequency of cyanobacterial blooms are often attributed to anthropogenic activities and climate change [11]. Nutrient enrichment from agricultural run-off and urbanization leads to eutrophication, providing an abundant supply of phosphorus and nitrogen that fuels cyanobacterial growth [12]. Climate change-induced water temperature and stratification increases further exacerbate the problem by creating optimal conditions for cyanobacterial proliferation [13]. Predictive models suggest that without intervention, the occurrence of HABs may increase by up to 20% in the next decade [14].

Rapid and accurate identification of cyanobacterial species is crucial for effective control strategies to protect public health and maintain ecological balance [5]. Traditional methods, such as manual sample collection followed by microscopy-based identification, are labor-intensive, time-consuming, and require analysis by highly qualified professionals [15]. Biochemical techniques like chlorophyll extraction and pigment analysis provide insights into photosynthetic activity but lack species-level specificity [16]. High-performance liquid chromatography (HPLC) can separate and quantify pigments like phycocyanin and chlorophyll; however, its reliability in distinguishing taxa is limited by overlapping pigment profiles and environmental variability, making it prone to misclassification [17]. Similarly, flow cytometry enables the rapid quantification of cyanobacterial populations by analyzing thousands of cells per minute; however, its high cost and complex setup limit its practical applicability for large-scale and field-based applications [18].

In this context, Raman spectroscopy has emerged as a promising tool for rapidly detecting and identifying microorganisms [19]. By measuring the inelastic scattering of monochromatic light, it generates a unique molecular fingerprint for each sample, enabling precise biochemical characterization [20]. Unlike other methods, Raman spectroscopy is non-destructive and requires minimal sample preparation. Furthermore, the development of portable Raman devices has extended its applicability, allowing for real-time, on-site analyses in field settings. Several studies have demonstrated the potential of Raman spectroscopy in identifying and classifying microorganisms. Heraud et al. utilized Raman spectroscopy to discriminate between different microalgal species, achieving a classification accuracy of over 90% [21]. Similarly, Schuster et al. reported an accuracy of 95% in distinguishing cyanobacterial species using Raman spectroscopy combined with multivariate analysis [22]. He et al. utilized confocal resonance Raman spectroscopy combined with Principal Component Analysis (PCA) and Discriminant Partial Least Squares (DPLS) analysis to identify unicellular algal genera, achieving a classification accuracy of over 90% [23] Additionally, Stöckel et al. used Raman spectroscopy to study the carotenoid content in cyanobacteria, providing insights into species differentiation based on pigment composition [24].

More recently, the integration of deep learning algorithms with Raman spectroscopy data has shown significant improvements in classification performance for microbial identification [25–27]. Ho et al. incorporated convolutional neural networks (CNNs) with Raman spectroscopy to achieve an accuracy of 98% in identifying pathogenic bacteria [28]. Similarly, Yu et al. developed a method combining Raman spectroscopy with long short-term memory (LSTM) neural networks, achieving over 94% accuracy in identifying marine pathogens [29]. Despite these advancements, challenges remain in processing and analyzing Raman spectroscopic data due to the complexity and variability of the signals [30]. Issues such as fluorescence background, instrumental noise, and overlapping spectral features can complicate spectral interpretation [31]. To overcome these limitations, robust preprocessing techniques—such as baseline correction, noise reduction, and normalization are essential for improving spectral quality and extracting meaningful information [32,33]. Furthermore, advanced algorithms and machine learning methods can further enhance the analysis by identifying subtle patterns and features within the data [34].

In this study, we present a methodology that combines Raman spectroscopy with a multichannel 1D-CNN to improve species-level identification of four toxic blue–green cyanobacteria. Filtering spectra was a critical step to ensure the quality of the training dataset, reducing noise and variability. The multichannel framework combines raw spectra, baseline estimations, and preprocessed data. This innovation increased classification accuracy from 74% when using only preprocessed spectra to 86% with the multichannel approach. Additionally, SHAP-based explanations were employed to highlight the most influential spectral regions, providing a more comprehensive understanding of the results. Overall, this methodology offers a more robust and transparent framework for distinguishing cyanobacterial species, ultimately supporting more effective management of HABs.

The four genera of toxic cyanobacteria selected for analysis were *Dolichospermum*, *Aphanizomenon*, *Planktothrix*, and *Microcystis*. These genera are among the most frequently reported in association with freshwater HABs worldwide, producing a variety of toxins with critical ecological and health impacts, including microcystins, saxitoxins, and cylindrospermopsins [35–39]. Despite their overall similarity, these cyanobacteria exhibit subtle differences in key traits [40]. These include pigment composition (e.g., carotenoids, phycobiliproteins, and chlorophyll), morphology (e.g., coccoid forms in Microcystis versus filamentous forms in *Dolichospermum*, *Planktothrix*, and *Aphanizomenon*), cell size (e.g., *Microcystis* cells typically measuring 4–6 μm in diameter, while *Dolichospermum* cells can reach up to 25 μm in length), and variations in cell wall composition (e.g., *Dolichospermum* and *Aphanizomenon* have thicker cell walls enriched with polysaccharides and glycolipids to support the formation of heterocysts, specialized cells for nitrogen fixation) [41,42]. These distinctions highlight the diversity found among cyanobacteria associated with HABs.

2. Materials and methods

2.1. Sample preparation

Four species of toxic cyanobacteria commonly found in reservoirs worldwide were selected for analysis: *Dolichospermum crassum*, *Aphanizomenon* sp., *Planktothrix agardhii*, and *Microcystis aeruginosa*. The isolates were provided by the Department of Biology at the Autonomous University of Madrid (UAM). Heterocyst-forming species were cultured in $BG11_0$ medium, which lacks combined nitrogen sources to promote heterocyst development. In contrast, non-heterocystous species were maintained in BG11 medium, a standard nutrient-rich medium for cyanobacterial cultivation. Cultures were incubated at 25 °C under continuous light at an illumination intensity between 70 and 130 μmol photons $m^{-2}s^{-1}$. Cultures were regularly manually agitated to ensure optimal growth conditions to prevent cell sedimentation and facilitate gas exchange.

Details of the analyzed samples are presented in Table 1, and bright-field images of each species are shown in Fig. 1.

2.2. Raman spectra acquisition

Raman spectra were obtained using a confocal Raman Spectroscopy microscope (XperRAM C Series, Nanobase) equipped with a 532 nm laser and a 20× objective lens (0.45 NA, MPlanFL N, Olympus). We set the laser power to 5 mW, and for each measurement, we conducted a single scan with a 500 ms acquisition time. The short acquisition time was selected to facilitate future measurements in a continuous-flow microfluidic system. The laser spot size was around 2–3 μm, smaller than cyanobacterial cells, enabling a detailed examination of specific cellular structures. Daily calibration was performed using a silicon wafer.

Samples were loaded into a microfluidic slide (Ibidi μ-Slide I Luer Glass Bottom) featuring a channel height of 250 μm and a reservoir volume of 62.5 μL. Automatic 2D Raman images were acquired on individual cells in a static state within the flow chamber. All measurements were performed with the glass side of the substrate facing the objective, and the laser was focused near this surface.

Table 1
Detailed information on the toxic cyanobacteria species used, including their respective codes, orders, genera, species, and culture media.

Code	Order	Genus	Species	Culture medium
UAM502	Nostocales	*Dolichospermum*	*D. crassum*	$BG11_0$
UAM588	Nostocales	*Aphanizomenon*	*Aphanizomenon* sp.[a]	$BG11_0$
UAM565	Oscillatoriales	*Planktothrix*	*P. agardhii*	BG11
UAM253	Chroococcales	*Microcystis*	*M. aeruginosa*	BG11

[a] Not identified.

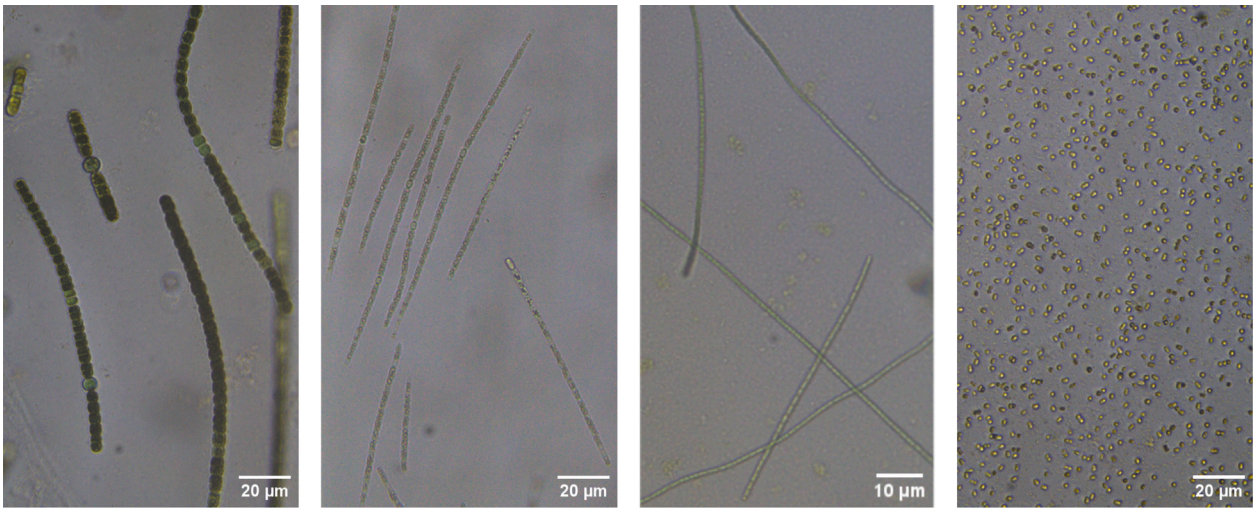

Fig. 1. Bright-field microscopy images of the toxic cyanobacterial species analyzed.

2.3. Preprocessing steps

The use of a 532 nm laser induces strong autofluorescence [43]. To ensure data consistency, we implemented an autofluorescence-based filtering criterion. We observed that when focusing the laser on cell bodies, fluorescence consistently saturated the detector in the 3200–3300 cm^{-1} region, which lies outside the Raman spectral range of interest (935–1575 cm^{-1}). We selected only the spectra that reached the detector's saturation threshold of 65,000 counts in this region. This threshold showed a strong correlation with Raman signal intensity, allowing for the selection of spectra with the most intense Raman responses. Only spectra meeting this criterion were considered valid, enabling us to exclude low-quality spectra, such as those collected from cell edges, and to minimize intraspecies spectral variability. Additionally, as dying cells progressively lose fluorescence [44], this approach ensured that analyzed spectra corresponded to cells in comparable physiological states.

The spectral data were cropped to the 800–1600 cm^{-1} range to focus on the most relevant Raman bands area. We excluded any spectra that exhibited saturation within this range from the analysis. Each remaining spectrum underwent smoothing with a Savitzky-Golay filter (21-point window, second-order polynomial), followed by baseline correction using Asymmetric Least Squares (ALS) with a smoothness parameter of 5×10^4 and an asymmetry parameter of 0.001. Subsequently, normalization was performed by scaling the total intensity of the spectra to 2000 counts, preserving the relative proportions between signals while ensuring a consistent scale. Finally, standardization was applied using Standard Normal Variate (SNV), resulting in spectra with a mean of 0 and a variance of 1.

Despite focusing the laser spot near the glass surface of the substrate, interference from the opposite side, composed of polymer, can occasionally occur. Residual Raman signals from the polymer were detected, as shown in the spectrum provided in Figure S1 of the supplemental material. A pronounced band at 910 cm^{-1} was observed, which could bias results if it appears faintly in cyanobacteria spectra. Additionally, fluorescence starts intensely around 1600 cm^{-1}, which may lead to improper baseline fitting in some cases. To address these issues, we implemented an additional cropping step that limits the spectral range to 935–1575 cm^{-1}. To further minimize polymer interference, the Non-Negative Least Squares (NNLS) algorithm was used to decompose each spectrum into a linear combination of reference spectra for cyanobacteria and polymer. Figure S1 in the Supplemental Material displays the reference spectra considered for the polymer substrate and cyanobacteria. Spectra showing more than 1% contribution from the polymer reference were excluded from further analysis.

A signal-to-noise ratio (SNR) filter was applied. Only spectra with an SNR greater than 30 between the regions of 800 cm^{-1} (without Raman signal) and 1002 cm^{-1} (CH_3 bonds, according to Table 3) were retained. We selected a high SNR threshold (SNR > 30) to ensure high-quality data and obtain the most representative spectra for each species. A lower threshold introduced greater variability, particularly in UAM565, which exhibits higher autofluorescence.

To improve data quality, we implemented an outlier detection procedure for all datasets. In the training and validation datasets, the Mahalanobis distance was calculated for each class individually, applying a fixed threshold of 40. This makes it particularly effective in identifying outliers in high-dimensional data where standard Euclidean distance may fail. For the test dataset, where data were unlabeled, we employed K-means clustering to automatically determine the optimal number of clusters using the elbow method. Within each cluster, spectra exceeding the 90th percentile of the Mahalanobis distance were classified as outliers and excluded from further analysis.

Figure S2 illustrates the outlier detection procedure, showing a K-means clustering plot and the Mahalanobis distance distribution for one of the classes. Additionally, Figure S3 provides two-dimensional Raman images highlighting the spectra removed as outliers. These images reveal no discernible spatial patterns among the excluded outliers, suggesting a random distribution of anomalous spectra throughout the dataset.

2.4. Deep learning pipeline

Raman spectroscopy produces complex, multicollinear, and high-dimensional spectra that challenge traditional machine learning methods, many of which assume feature independence. Our previous work [45] evaluated various machine learning algorithms for *Candida* species classification and found that a 1D-CNN achieved the best performance metrics and generalization. We later developed a simplified 1D-CNN with a single convolutional layer to differentiate capsular serotypes of *Klebsiella pneumoniae*, also obtaining high predictive performance [46]. In this study, we implement a multichannel 1D-CNN framework with the same architecture to extract relevant information from diverse spectral data.

The architecture of the 1D-CNN model is designed to process one-dimensional spectral data. The model begins with a convolutional layer of 64 filters of size 8 using 'same' padding to maintain input dimensions. A rectified linear unit (ReLU) activation function introduces non-linearity into the model. Following this, a MaxPooling1D layer with a pool size of 2 reduces the dimensionality and focuses on the most significant features. To prevent overfitting, a dropout layer with a rate of 0.25 is included. The output is flattened and passed through a dense layer with 128 units and another ReLU activation. Another dropout layer with the same rate precedes the final dense layer, which has units equal to the number of classes and employs a softmax activation function for multiclass classification [46].

We implemented a multichannel approach to expand the model's capabilities, introducing three independent input channels. Each channel processes a different version of the spectra: raw spectra, baseline estimations, and corrected spectra. The inclusion of baseline estimations as an independent channel provided the model with the relationship between raw and preprocessed spectra, making it easier to access information removed during correction. The corrected spectra undergo the entire preprocessing pipeline detailed in Section 2.3, while the raw and baseline datasets are subjected only to the standardization and normalization steps described in the same section. Each dataset follows the same convolutional pathway separately. Before reaching the dense layers, the outputs from these channels are concatenated, allowing the model to integrate information from all spectral types. Rather than exhaustively searching for an optimal preprocessing strategy, we prioritized integrating raw and preprocessed spectra in a multichannel framework to leverage their complementary information. The architecture of this multichannel CNN is depicted in Figure S4.

The dataset was split into 80% for training and 20% for validation. To address the class imbalance in the training-validation and test datasets, we randomly reduced the size of overrepresented classes, ensuring that no class had more than 20% additional samples compared to the minority class. The test dataset, measured from different cultures months later, was used to assess the model's performance on new data.

2.5. Shapley values for model interpretability

We employed SHapley Additive exPlanations (SHAP) values [47] to interpret the predictions of the multichannel 1D-CNN model. Shapley values, derived from cooperative game theory, quantify the contribution of individual features to a model's predictions by averaging their marginal impact across all possible subsets of features. This approach provides a rigorous framework for assessing feature importance in highly complex and nonlinear models.

In Raman spectroscopy, Shapley values enable the identification of spectral regions that influence the classification process, thus bridging the gap between deep learning predictions and the chemical properties of the sample. However, the exact computation of Shapley values is computationally infeasible for high-dimensional datasets, as it requires evaluating all possible feature combinations. To address this, the SHAP library provides efficient approximation methods tailored to specific models.

Table 2
Number of Raman spectra after preprocessing in the training/validation and test datasets for each cyanobacteria species.

Class	Train/val (balanced)	Test (balanced)
UAM502	9967	842
UAM588	9967	1010
UAM565	8690	1010
UAM253	8306	1010

For deep learning models, SHAP incorporates the DeepExplainer, a tool specifically designed to handle the internal structure of neural networks. Rather than evaluating every feature subset, DeepExplainer leverages carefully selected reference samples from the dataset and gradient-based information within the network to approximate each feature's contribution. It efficiently balances accuracy and computational cost by focusing on how the model's predictions change when inputs are perturbed relative to these reference samples.

3. Results and discussion

3.1. Raman spectra overview

Table 2 shows the number of Raman spectra remaining after the preprocessing filters described in Section 2.3. Additionally, Table S1 provides details of the spectra included in the train/val and test datasets, showing the number of spectra acquired per day for each species, the total spectra recorded per batch, and the proportion of each species within its batch.

As observed in Table S1, the limited number of culture batches and the uneven distribution of measurements within them required a random division of spectra in the train-validation dataset, as stratification by culture batch was not feasible.

In Fig. 2, normalized Raman spectra of the four cyanobacterial species are presented, showing (a) the raw spectra and (b) the baseline estimations, along with their respective standard deviations. For the baseline estimations, a broader range (50–3200 cm^{-1}) is used to highlight changes in autofluorescence curvature, which are more pronounced above 1600 cm^{-1}. The baseline estimation is used as an approximation of the samples' autofluorescence to assess its potential contribution to cyanobacterial classification.

As shown in Fig. 2, the estimated baselines reveal clear differences among the analyzed cyanobacterial species, directly linked to autofluorescence. UAM502, UAM588, and UAM253 exhibit similar baseline trends, characterized by a slight increase in intensity across the analyzed Raman range and a gradual rise in the 2500–3200 cm^{-1} region. In contrast, UAM565 presents a markedly different baseline profile, with a pronounced increase across the entire spectrum until stabilizing around 2500 cm^{-1}. This distinct behavior may be attributed to differences in photosynthetic pigment concentration, particularly chlorophylls and phycobiliproteins [48,49], as well as structural factors such as the presence of abundant gas vesicles, which are characteristic of*Planktothrix agardhii* and can affect light scattering and interactions with cellular components [50–52]. Additionally, UAM565 exhibits greater signal variability beyond 1300 cm^{-1}, as reflected in the standard deviation, which could further contribute to differences in spectral interpretation. Given that such variations may introduce biases in classification, assessing their impact on the corrected spectra is crucial for accurate spectral analysis and interpretation.

In Fig. 3, the preprocessed Raman spectra of the four cyanobacterial species are displayed after applying the methodology outlined in Section 2.2.

The fluorescence signal observed in the raw spectra, especially for UAM565, is significantly reduced after preprocessing. The processed signals become more consistent across species, with UAM565 showing comparable Raman features to the others.

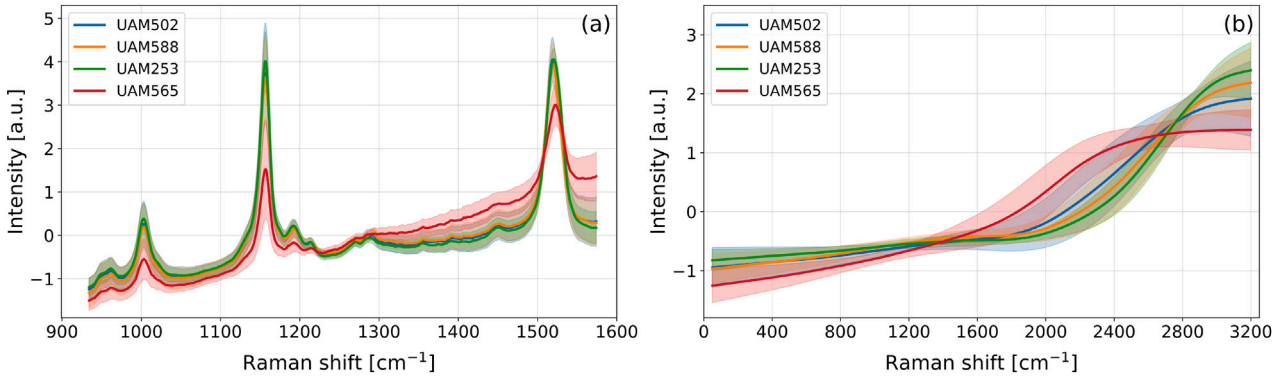

Fig. 2. Average normalized Raman spectra with standard deviation for each class: (a) raw spectra and (b) baseline estimations.

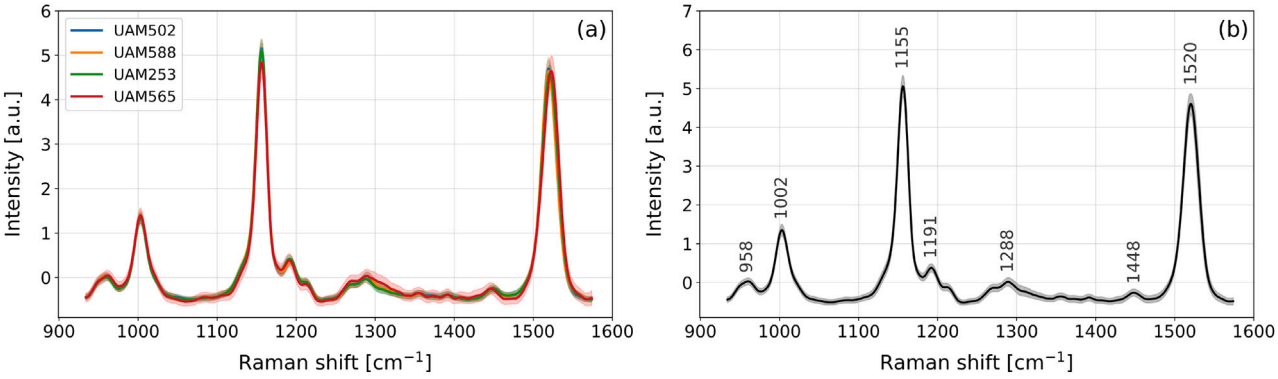

Fig. 3. (a) Average Raman spectra with standard deviation for the preprocessed data of each species. (b) Overall average Raman spectrum with standard deviation for the entire dataset, highlighting the most significant Raman bands.

Table 3
Main Raman band assignments of the four toxic cyanobacteria species analyzed, including Raman shift positions along with their associated bonds and substances [23,53–58].

Raman shift (cm^{-1})	Chemical bonds	Abbreviation	Substances
958	C–C str., CH out-of-plane (C=C), CH rock	ν(C–C), γ(CH)	Proteins, lipids, carotenoids
1002	Sym. C–CH_3 str., CH_3 bend	ν(C–CH_3), δ(CH_3)	Proteins, carotenoids
1155	C–C str., CH def.	ν(C–C), δ(CH)	Carotenoids, chlorophyll
1191	CH def.	δ(CH)	Carotenoids
1288	CH_2 def., Amide III (C–N, N–H bend)	δ(CH_2), ν(C–N), δ(N–H)	Lipids, proteins, chlorophyll, carbohydrates, carotenoids
1448	CH_2/CH_3 def.	δ(CH_2/CH_3)	Lipids, proteins, chlorophyll, carotenoids
1520	C=C str.	ν(C=C)	Carotenoids, chlorophyll

We propose including raw spectra and baseline estimations in the analysis pipeline, as this could enhance the robustness of automatic cyanobacterial species classification. It may also reveal whether preprocessing steps disadvantage certain species, impacting classification performance.

Table 3 provides a summary of the band assignments highlighted in Fig. 3b.

The Raman spectra presented in Fig. 3 and detailed in Table 3 reveal prominent bands at 1002, 1155, and 1520 cm^{-1}. The band at 1002 cm^{-1} is attributed to symmetric C–CH_3 stretching and CH_3 bending vibrations, commonly associated with proteins and carotenoids. The 1155 cm^{-1} band corresponds to C–C stretching and CH deformation modes, indicative of carotenoids and chlorophylls. The 1520 cm^{-1} band is linked to C=C stretching vibrations, characteristic of carotenoids and chlorophylls. These findings suggest that the dominant Raman signals in the spectra are primarily due to the vibrational modes of carotenoids and chlorophylls present in the samples.

3.2. Deep learning analysis

To evaluate the contribution of each spectral representation, we tested different input configurations. The highest test accuracy was

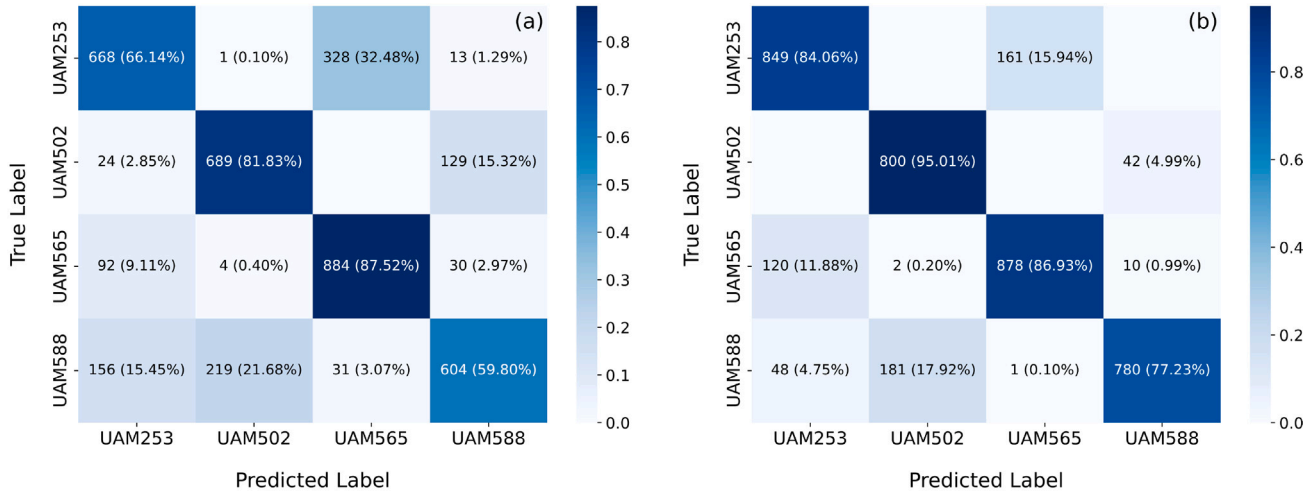

Fig. 4. Confusion matrices obtained using the 1D-CNN architectures described in Section 2.4: (a) 1D-CNN trained with preprocessed spectra and (b) 1D-CNN trained with a combination of raw spectra, baseline estimations, and preprocessed spectra. The overall accuracies obtained were (a) 74% and (b) 86%.

achieved by combining raw and preprocessed spectra (86%), outperforming only preprocessed (74%) or only raw (71%). This combination leverages their strengths: preprocessing removes noise and enhances spectral features, while raw spectra retain subtle details. Preprocessing also simplifies learning, helping the model focus on meaningful spectral patterns. We also trained a multichannel CNN using only raw and preprocessed spectra, omitting the baseline estimation channel, and obtained comparable performance (86%). We retain this channel to explore its individual contribution to spectral features, as its inclusion enables a more detailed analysis of autofluorescence.

Based on these findings, we compared the performance of the model trained with only preprocessed spectra, which is a common approach [28,46], to the multichannel CNN configuration incorporating raw, preprocessed, and baseline estimation spectra. The corresponding confusion matrices for these models are presented in Fig. 4.

The 1D-CNN trained on preprocessed spectra achieved an overall accuracy of 82% on both the training and validation datasets, dropping to 74% on the test dataset (Fig. 4a). In contrast, the multichannel 1D-CNN reached 95% accuracy on the training and validation datasets, respectively, and 86% on the test dataset (Fig. 4b), indicating that both models do not overfit. However, the test accuracies are lower, showing that generalization to unseen data remains imperfect. Since the test dataset was independently acquired, with cyanobacteria samples from different culture passages measured months apart, it represents novel samples to the models. Despite consistent preprocessing and measurement protocols, factors such as potential data leakage resulting from varying growth stages or genetic similarity between batches could contribute to the slightly reduced generalization.

The comparison between the confusion matrices in Fig. 4 highlights the significant improvement achieved with the multichannel 1D-CNN. The 1D-CNN trained solely on preprocessed spectra shows significant pairwise misclassifications, with UAM253 often confused with UAM565 and UAM588 with UAM502. Although the model achieves respectable accuracies for specific classes, such as 82% for UAM502 and 88% for UAM565, these systematic misclassifications introduce biases that compromise its overall reliability. In contrast, the multichannel 1D-CNN significantly reduces these confusions, although some misclassifications persist. By integrating raw spectra and baseline estimations alongside preprocessed spectra, the model leverages complementary information to reduce class overlap and eliminate the biases observed in the single-channel model.

The performance metrics of the multichannel 1D-CNN model are presented in Table 4.

The metrics presented in Table 4 demonstrate robust performance of the multichannel 1D-CNN model on the test data, with an overall accuracy of 86% and a macro-average F1-score of 86%, demonstrating a good balance between precision and recall. UAM502 stands out with a high Recall (95%), indicating that almost all its true instances are correctly identified. UAM588, with the highest precision (94%), shows reliable positive predictions, but its lower recall (77%) reveals that 18% of its true instances are classified as UAM502. Meanwhile, UAM565 and UAM253 exhibit balanced metrics, with recall of 87% and 84%, respectively, and precision of 84% and 83%, although there is some confusion between them. These results highlight that the main confusions occur in pairs: UAM502 with UAM588 and UAM565 with UAM253.

Table 4
Performance metrics for the multichannel 1D-CNN model trained on classifying four toxic cyanobacteria species. Metrics include accuracy, recall, precision, specificity, and F1-score for each class, along with the macro average and overall accuracy across all classes.

Cyanobacteria species	Accuracy	Recall	Precision	Specificity	F1-score
UAM502	0.94	0.95	0.81	0.94	0.88
UAM588	0.93	0.77	0.94	0.98	0.85
UAM565	0.92	0.87	0.84	0.94	0.86
UAM253	0.92	0.84	0.83	0.94	0.84
Macro average	0.93	0.86	0.86	0.95	0.86
Overall accuracy	0.86				

Consistently, Figure S5 in the supplemental material presents an LDA plot for the three datasets—raw spectra, baseline estimations, and preprocessed spectra—where the same pairwise class confusion is observed through the overlapping clusters of the species.

Fig. 4 shows that species prone to misclassification share the same growth medium, according to Table 1. To verify if the growth medium affects classification, spectra from UAM502 and UAM588 cultured in BG11 and $BG11_0$ were compared using 1500 preprocessed spectra per class. Figure S6 presents the average spectra and their standard deviations for these measurements, where the spectra appear highly similar under both culture conditions. Figure S7 illustrates two methods used to assess the impact of the growth medium on classification. In Figure S7a, the LDA visualization clearly separates the species regardless of the growth medium, indicating that the medium has little effect on the classification. Similarly, the 1D-CNN confusion matrix in Figure S7b shows that classification errors mainly occur between samples of the same species grown in different media. This suggests that the similarity observed between the cyanobacteria is not attributable to their shared growth medium but rather to their intrinsic molecular composition.

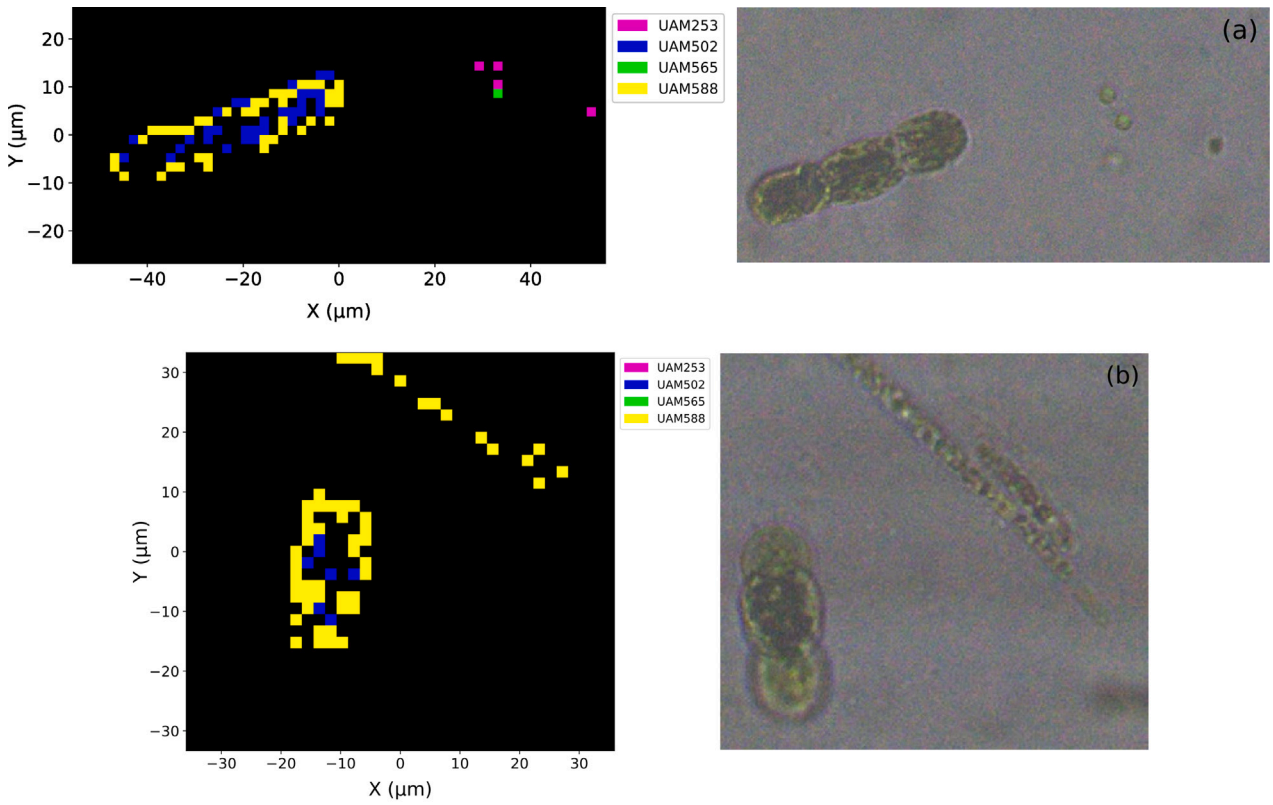

Fig. 5. Raman 2D images of cyanobacterial species classification in mixed samples using the multichannel model from Fig. 4b: (a) UAM502 and UAM253, (b) UAM502 and UAM588. Spectra in black did not pass the preprocessing filters, while other colors indicate the model's classification. Bright-field images provide a visual reference of the species' spatial distribution.

Additional measurements were performed under the same experimental and preprocessing conditions to better understand misclassifications between classes, focusing on mixed samples containing pairs of cyanobacterial species. Raman 2D images were generated using the automatic classification derived from predictions made by the multichannel model. These representations aim to identify patterns that explain the misclassifications observed among the species. Fig. 5 shows two representative images: a mixture of UAM502 and UAM253 and a mixture of UAM502 and UAM588, alongside bright-field images taken before the Raman measurements to provide a visual reference of the expected spatial distribution of the cyanobacteria species.

The results of the automatic classifications from mixed cyanobacteria samples did not yield particularly remarkable outcomes except for some specific cases, such as those shown in Fig. 5. The confusion observed between UAM502 and UAM588 often assigns the edges of UAM502 as UAM588. This may be related to pigment concentration differences in the cell membrane. On the other hand, the confusions between UAM253 and UAM565 seem more random, as no significant spatial patterns were identified in these cases.

It is worth noting that cyanobacterial membranes vary significantly in complexity, composition, and layer thickness—particularly in the peptidoglycan layer, which can exceed 700 nm in some species [59]. Such structural differences may influence the observed classification uncertainties, mainly because the measurements were performed using a confocal system. Since the laser penetrates only a few microns depending on the focal plane, subtle variations in membrane structure or pigment distribution can result in detectable differences in the Raman signal.

3.3. Interpretability from deep learning analysis

Fig. 6 shows the global absolute SHAP values for each class, representing the contribution of individual Raman shifts to the predictions made by the multichannel 1D CNN model. These results are shown independently for three channels: (a) raw spectra, (b) baseline estimations, and (c) preprocessed spectra.

According to Fig. 6, SHAP values indicate that the most relevant information for classification is directly associated with characteristic Raman bands. In contrast, baseline estimations (Fig. 6b) contribute minimally, suggesting that fluorescence signals, observed as a residual broad emission similar to the baseline estimation, have little impact on the model's classification performance. This is consistent with the LDA representation shown in Figure S5b, where the separation between clusters for baseline estimations is less evident compared to those for raw and corrected spectra.

Interestingly, many regions highlighted by SHAP values correspond to Raman bands that are not clearly visible in the average spectrum due to overlapping with more intense signals. This indicates that the model is capable of extracting subtle spectral information.

The SHAP-identified regions correspond to Raman bands commonly associated with carotenoids and chlorophylls, according to Table 3. Specifically, bands around 1520 cm^{-1} (C=C stretching), 1155 cm^{-1} (C–C stretching), 1002 cm^{-1} (symmetric C–CH_3 stretching and CH_3 bending) and 1191 cm^{-1} CH deformations align with well-characterized carotenoid signatures [57].

SHAP values also highlight regions where the presence of substances other than carotenoids is expected, such as chlorophylls, proteins and lipids. For example, the 1288 cm^{-1} band includes CH_2 deformations and amide III vibrations, indicating the presence of proteins and lipids, while the 1448 cm^{-1} band (CH_2/CH_3 deformations) suggests contributions from lipids and chlorophyll.

Considering this, precise molecular attribution remains challenging due to spectral overlap among these substances, but the strong correspondence between the most relevant Raman bands and known carotenoid and chlorophyll signatures reinforces their dominant role in cyanobacterial species differentiation.

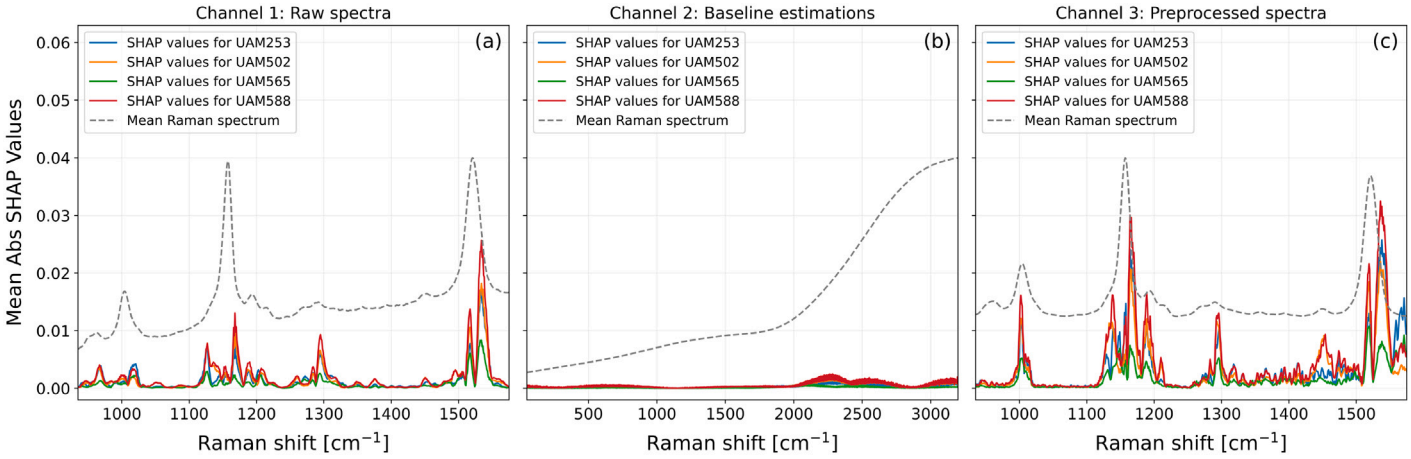

Fig. 6. Global absolute SHAP values for each class are shown for (a) raw spectra, (b) baseline estimations, and (c) preprocessed spectra obtained from the 1D-CNN multichannel model described in Fig. 4b. The scale for Mean Abs SHAP Values is consistent across the three channels. In the case of Raman spectra, the scale is expressed in arbitrary units.

Additionally, flat spectral regions, such as those between 1000–1100 cm^{-1}, show SHAP values close to zero, confirming that these areas, not associated with Raman bands, provide little to no relevant information for classification.

SHAP analysis demonstrates that raw and preprocessed spectra provide complementary insights. A comparison of SHAP values from raw (Fig. 6a) and preprocessed spectra (Fig. 6c) reveals greater consistency in regions around the Raman bands observed in the mean spectra after preprocessing Fig. 3. In raw spectra, SHAP values in these regions are less pronounced but remain well-defined, suggesting that noise and background signals negatively impact classification while still retaining key information for species differentiation. In contrast, preprocessed spectra more clearly distinguish key Raman bands, such as those near 1002, 1155, and 1520 cm^{-1}, confirming that preprocessing improves feature selection.

However, preprocessing can also suppress biologically relevant signals. For instance, the 958 cm^{-1} band, strongly highlighted in raw SHAP values, weakens after preprocessing, indicating potential information loss. Additionally, the 1500–1575 cm^{-1} region, located at the spectral range's edge, exhibits a progressive intensity increase in preprocessed spectra, likely due to artifacts introduced by baseline estimation, which may introduce classification biases.

Despite these limitations, the complementarity between raw and preprocessed spectra suggests that both contribute valuable information to the classification process. While the preprocessing applied may not be the optimal approach, it effectively enhances the model's focus on the most relevant spectral features, leading to improved classification performance.

4. Conclusions

This study presents a comprehensive methodology for the classification of four toxic cyanobacterial species (*Dolichospermum crassum*, *Aphanizomenon* sp., *Planktothrix agardhii* and *Microcystis aeruginosa*) by combining Raman spectroscopy with advanced deep learning techniques.

A preprocessing pipeline was carefully designed to address challenges in Raman spectroscopy and ensure that the selected spectra accurately represent each cyanobacterial species while minimizing external influences. Steps such as baseline correction, noise filtering, and signal normalization effectively tackled issues like polymer interference or spectral variability. These corrections ensured high-quality spectral data.

From a biological perspective, the data contained in the train-validation dataset do not differ significantly from what would be expected in a real environment, as cyanobacteria blooms are originated by successive regrowth of existing cells rather than independent populations. This continuous proliferation leads to populations with shared genetic backgrounds and similar physiological states, even when they arise from different culture batches.

However, a fluorescence-based filtering criterion was applied to reduce intraspecies variability and maintain comparable physiological states, promoting data consistency and reliability, even at the trade-off of excluding some valid spectra. These preprocessing steps are crucial for preventing biases and ensuring a representative dataset for classification. The slight performance reductions observed on the independent test dataset suggest further enhancements to improve model generalization, particularly by accounting for variations in sample growth stages or potential degradation states.

A multichannel 1D-CNN framework, incorporating raw spectra, baseline estimations and preprocessed spectral data, achieved significant improvements over traditional single-channel methods, reaching an overall accuracy of 86%. By integrating multiple spectral data, the multichannel approach effectively minimized systematic misclassifications commonly observed in single-channel models.

Despite the increased number of parameters in the multichannel approach, no evidence of overfitting was observed. Regularization techniques, including dropout and early stopping, were applied to control model complexity. Additionally, the multichannel CNN demonstrated good generalization to independent test data, indicating its robustness in classifying unseen samples.

The drop in accuracy from the training dataset (95%) to the test dataset (86%) could be attributed to the natural variability of the test set. Additionally, uncontrolled confounding variables may have contributed, as it has not been feasible to apply stratification strategies, for example, based on culture batches.

The integration of SHAP provided an interpretable framework for identifying the most relevant Raman spectral regions contributing to the classification. Key spectral bands, primarily associated with carotenoids and chlorophylls, were highlighted as critical biochemical markers. This improved interpretability not only enhances the model's reliability but also clarifies the molecular features that allow the differentiation of cyanobacterial species. Moreover, the complementarity between raw and preprocessed spectra underscores the importance of leveraging both datasets, as each captures distinct spectral details that contribute to a more accurate classification.

In addition, we have demonstrated that autofluorescence, represented by the baseline estimation channel, does not provide relevant information for classification. This finding confirms that the model predominantly relies on Raman features.

In future research, we will explore various directions. First, We will conduct validation in natural settings to assess the model's performance

under realistic conditions, specifically for *Microcystis aeruginosa*, which forms colonies in nature rather than existing as individual cells, as we observe in controlled cultures. While we expand our dataset with more cultures and a greater diversity of samples, we plan to further refine the model's generalization by increasing the training set and implementing stratification strategies on it. Second, efforts will focus on exploring dimensionality reduction techniques using SHAP values to improve computational efficiency and optimize spectral analysis. Third, work will focus on developing Raman spectroscopy integration into microfluidic systems for continuous-flow measurements and in situ monitoring, as no low-cost solutions currently exist to identify cyanobacterial species in real-time accurately. Finally, the dataset will be expanded to include a wider variety of cyanobacterial species and environmental conditions, improving the model's generalizability and robustness to variations in spectral data.

Our study underscores the potential of combining Raman spectroscopy with advanced deep learning techniques as a powerful tool for environmental monitoring. The proposed methodology improves the detection and classification of toxic cyanobacterial species, contributing to advancements in water quality management and ecosystem protection.

Funding

This work was supported by the R+D projects PREVAL23/05, INNVAL23/10, and INNVAL24/28, funded by Instituto de Investigación Marqués de Valdecilla (IDIVAL); J.F.A. acknowledges RYC2022-035279-I, funded by MCIN/AEI/10.13039/501100011033 and FSE+; TED2021-130378B-C21, funded by MCIN/AEI/10.13039/501100011033 and European Union NextGenerationEU/PRTR; PID2022-137269OB-C22, funded by MCIN/AEI/10.13039/501100011033 and FEDER, UE; Plan Nacional de I+D+i and Instituto de Salud Carlos III (ISCIII), Subdirección General de Redes *y* Centros de Investigación Cooperativa, Ministerio de Ciencia, Innovación *y* Universidades, through CIBER-BBN (CB16/01/00430) and CIBERINFEC (CB21/13/00068), co-financed by the European Regional Development Fund "A way to achieve Europe".

CRediT authorship contribution statement

María Gabriela Fernández-Manteca: Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Borja García García:** Writing – review & editing, Visualization, Methodology, Conceptualization. **Susana Deus Álvarez:** Writing – review & editing, Resources, Investigation, Conceptualization. **Celia Gómez-Galdós:** Writing – review & editing, Formal analysis, Conceptualization. **Andrea Pérez-Asensio:** Writing – review & editing, Validation. **José Francisco Algorri:** Writing – review & editing, Methodology, Funding acquisition. **Agustín P. Monteoliva:** Writing – review & editing, Resources, Conceptualization. **José Miguel López-Higuera:** Supervision, Project administration, Funding acquisition. **Luis Rodríguez-Cobo:** Writing – review & editing, Supervision, Methodology. **Alain A. Ocampo-Sosa:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Adolfo Cobo:** Writing – original draft, Supervision, Software, Methodology, Funding acquisition, 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

We thank the Biology Department of Universidad Autónoma de Madrid for their support and for providing the samples used in this study.

Abbreviations

The following abbreviations are used in this manuscript:

1D-CNN	One-Dimensional Convolutional Neural Network
HABs	Harmful Algal Blooms
PCA	Principal Component Analysis
DPLS	Discriminant Partial Least Squares
LSTM	Long Short-Term Memory
UAM	Universidad Autónoma de Madrid
NA	Numerical Aperture
ALS	Asymmetric Least Squares
SNV	Standard Normal Variate
NNLS	Non-Negative Least Squares
SNR	Signal-to-Noise Ratio
ReLU	Rectified Linear Unit
SHAP	SHapley Additive exPlanations

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.talanta.2025.127845.

Data availability

The Raman spectra datasets and the trained models used in this work are publicly available on Zenodo at https://doi.org/10.5281/zenodo.14727718. All Python scripts developed and used in this work are available in the GitHub repository https://github.com/fmantecam/CyanoRamanDL (accessed on 19 December 2024).

References

[1] H.W. Paerl, T.G. Otten, Harmful cyanobacterial blooms: causes, consequences and controls, Microb. Ecol. 65 (2013) 995–1010, http://dx.doi.org/10.1007/s00248-012-0159-y.

[2] B.A. Whitton, M. Potts (Eds.), Ecology of Cyanobacteria II: Their Diversity in Space and Time, Springer, Dordrecht, 2012, http://dx.doi.org/10.1007/978-94-007-3855-3.

[3] W.W. Carmichael, A world overview—One-hundred-twenty-seven years of research on toxic cyanobacteria—Where do we go from here? in: Advances in Experimental Medicine and Biology, vol. 619, Springer, New York, 2008, pp. 105–125, http://dx.doi.org/10.1007/978-0-387-75865-7_6.

[4] G.A. Codd, L.F. Morrison, J.S. Metcalf, Cyanobacterial toxins: risk management for health protection, Toxicol. Appl. Pharmacol. 203 (2005) 264–272, http://dx.doi.org/10.1016/j.taap.2004.02.016.

[5] I. Chorus, M. Welker (Eds.), Toxic Cyanobacteria in Water: A Guide To their Public Health Consequences, Monitoring and Management, second ed., CRC Press, Boca Raton, FL, USA, 2021.

[6] E. Dittmann, C. Wiegand, Cyanobacterial toxins—occurrence, biosynthesis and impact on human affairs, Mol. Nutr. & Food Res. 50 (2006) 7–17, http://dx.doi.org/10.1002/mnfr.200500162.

[7] I.R. Falconer, Cyanobacterial Toxins of Drinking Water Supplies: Cylindrospermopsins and Microcystins, CRC Press, Boca Raton, 2005.

[8] Q. Tan, H. Chu, J. Wei, S. Yan, X. Sun, J. Wang, L. Zhu, F. Yang, Astaxanthin alleviates hepatic lipid metabolic dysregulation induced by microcystin-LR, Toxins 16 (2024) 401, http://dx.doi.org/10.3390/toxins16090401.

[9] E. Hilborn, V. Beasley, One health and cyanobacteria in freshwater systems: Animal illnesses and deaths are sentinel events for human health risks, Toxins 7 (2015) 1374–1395, http://dx.doi.org/10.3390/toxins7041374.

[10] W.K. Dodds, et al., Eutrophication of U.S. freshwaters: analysis of potential economic damages, Environ. Sci. Technol. 43 (2009) 12–19, http://dx.doi.org/10.1021/es801217q.

[11] J. Huisman, et al., Cyanobacterial blooms, Nat. Rev. Microbiol. 16 (2018) 471–483, http://dx.doi.org/10.1038/s41579-018-0040-1.

[12] V.H. Smith, G.D. Tilman, J.C. Nekola, Eutrophication: impacts of excess nutrient inputs on freshwater, marine and terrestrial ecosystems, Environ. Pollut. 100 (1999) 179–196, http://dx.doi.org/10.1016/S0269-7491(99)00091-3.

[13] J.M. O'Neil, et al., The rise of harmful cyanobacteria blooms: The potential roles of eutrophication and climate change, Harmful Algae 14 (2012) 313–334, http://dx.doi.org/10.1016/j.hal.2011.10.027.

[14] S.C. Chapra, et al., Climate change impacts on harmful algal blooms in U.S. freshwaters: a screening-level assessment, Environ. Sci. Technol. 51 (2017) 8933–8943, http://dx.doi.org/10.1021/acs.est.7b01498.

[15] A. Zamyadi, C. Romanis, T. Mills, B. Neilan, F. Choo, L.A. Coral, D. Gale, G. Newcombe, N. Crosbie, R. Stuetz, R.K. Henderson, Diagnosing water treatment critical control points for cyanobacterial removal: Exploring benefits of combined microscopy, next-generation sequencing and cell integrity methods, Water Res. 152 (2019) 96–105, http://dx.doi.org/10.1016/j.watres.2019.01.020.

[16] M. Beutler, K.H. Wiltshire, B. Meyer, C. Moldaenke, C. Lüring, M. Meyerhöfer, U.P. Hansen, H. Dau, A fluorometric method for the differentiation of algal populations in vivo and in situ, Photosynth. Res. 72 (2018) 39–53, http://dx.doi.org/10.1023/A:1014870112121.

[17] L.J. Simmons, C.D. Sandgren, J.A. Berges, Problems and pitfalls in using HPLC pigment analysis to distinguish lake michigan phytoplankton taxa, J. Gt. Lakes Res. 42 (2016) 397–404, http://dx.doi.org/10.1016/j.jglr.2016.03.013.

[18] M. Cellamare, A. Rolland, S. Jacquet, J.F. Humbert, Flow cytometry sorting of freshwater phytoplankton, J. Appl. Phycol. 28 (2016) 279–297, http://dx.doi.org/10.1007/s10811-015-0547-4.

[19] M. Harz, P. Rösch, J. Popp, Vibrational spectroscopy—a powerful tool for the rapid identification of microbial cells at the single-cell level, Cytom. Part A 75 (2009) 104–113, http://dx.doi.org/10.1002/cyto.a.20685.

[20] J. Popp, W. Kiefer, J. Motz, Handbook of Biophotonics, Volume 2: Photonics for Health Care, Wiley-VCH, Weinheim, Germany, 2005.

[21] P. Heraud, B.R. Wood, J. Beardall, D. McNaughton, In vivo prediction of the nutrient status of individual microalgal cells using Raman microspectroscopy, FEMS Microbiol. Lett. 275 (2007) 24–30, http://dx.doi.org/10.1111/j.1574-6968.2007.00850.x.

[22] K.C. Schuster, I. Reese, E. Urlaub, J.R. Gapes, B. Lendl, Multidimensional information on the chemical composition of single bacterial cells by confocal Raman microspectroscopy, Anal. Chem. 72 (2000) 5529–5534, http://dx.doi.org/10.1021/ac000633s.

[23] S. He, W. Xie, P. Zhang, S. Fang, Z. Li, P. Tang, D. Wang, Preliminary identification of unicellular algal genus by using combined confocal resonance Raman spectroscopy with PCA and DPLS analysis, Spectrochim. Acta Part A: Mol. Biomol. Spectrosc. 190 (2018) 417–422, http://dx.doi.org/10.1016/j.saa.2017.09.036.

[24] S. Stöckel, S. Meisel, B. Lorenz, J. Popp, The application of Raman spectroscopy for the detection and identification of microorganisms, J. Raman Spectrosc. 47 (2016) 89–109, http://dx.doi.org/10.1002/jrs.4860.

[25] Y. Liu, Y. Gao, R. Niu, Z. Zhang, G.-W. Lu, H. Hu, T. Liu, Z. Cheng, Rapid and accurate bacteria identification through deep-learning-based two-dimensional Raman spectroscopy, Anal. Chim. Acta (2024) 343376, http://dx.doi.org/10.1016/j.aca.2024.343376.

[26] J. Xue, H. Yue, W. Lu, Y. Li, G. Huang, Y. Fu, Application of Raman spectroscopy and machine learning for *Candida auris* identification and characterization, Appl. Environ. Microbiol. 90 (2024) http://dx.doi.org/10.1128/aem.01025-24, e01025-24.

[27] S. Lu, Y. Huang, W.X. Shen, Y.L. Cao, M. Cai, Y. Chen, Y. Tan, Y.Y. Jiang, Y.Z. Chen, Raman spectroscopic deep learning with signal aggregated representations for enhanced cell phenotype and signature identification, PNAS Nexus 3 (2024) pgae268, http://dx.doi.org/10.1093/pnasnexus/pgae268.

[28] C.S. Ho, N. Jean, C.A. Hogan, L. Blackmon, S.S. Jeffrey, M. Holodniy, N. Banaei, A.A.E. Saleh, S. Ermon, J. Dionne, Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning, Nat. Commun. 10 (2019) 4927, http://dx.doi.org/10.1038/s41467-019-12898-9.

[29] H. Yu, D. Zhang, Z. Wang, J. Zhang, J. Xie, Analysis of Raman spectra by using deep learning methods in the identification of marine pathogens, Front. Microbiol. 12 (2021) 686239, http://dx.doi.org/10.3389/fmicb.2021.686239.

[30] I. Notingher, Raman spectroscopy cell-based biosensors, Sens. Actuators B: Chem. 51 (2006) 18–25, http://dx.doi.org/10.1016/j.snb.2005.02.033.

[31] J.M. Smulko, M.S. Wróbel, I. Barman, Noise in biological Raman spectroscopy, in: Proceedings of the 2015 International Conference on Noise and Fluctuations (ICNF), Xiamen, China, vol. 2–6, 2015, pp. 1–6.

[32] L. Ashton, K. Lau, C.L. Winder, R. Goodacre, Raman spectroscopy: lighting up the future of microbial identification, Futur. Microbiol. 6 (2011) 941–943, http://dx.doi.org/10.2217/fmb.11.80.

[33] R.T. Vulchi, V. Morgunov, R. Junjuri, T. Bocklitz, Artifacts and anomalies in Raman spectroscopy: A review on origins and correction procedures, Molecules 29 (2024) 4748, http://dx.doi.org/10.3390/molecules29194748.

[34] A.V. Karmenyan, D.A.Sr. Vrazhnov, E.A. Sandykova, E.V. Perevedentseva, A.S. Krivokharchenko, V.A. Nadtochenko, C.-L. Cheng, T.V. Kabanova, T.E. Malakhova, Informative feature selection method for Raman micro-spectroscopy data, in: Proceedings of the XV International Conference on Pulsed Lasers and Laser Applications, vol. 12086, Russian Federation, Tomsk, 2021, 120861H, http://dx.doi.org/10.1117/12.2613966.

[35] Centers for Disease Control and Prevention (CDC), Summary Report – One Health Harmful Algal Bloom System (OHHABS), United States, 2020, U.S. Department of Health and Human Services, CDC, Atlanta, Georgia, 2022.

[36] R. Vieira-Lanero, S. Barca, M.C. Cobo, F. Cobo, Occurrence of freshwater cyanobacteria and bloom records in spanish reservoirs (1981–2017), Hydrobiology 1 (2022) 122–136, http://dx.doi.org/10.3390/hydrobiology1010009.

[37] Z. Svirčev, D. Lalić, G.Bojadžija. Savić, N. Tokodi, D. Drobac, L. Chen, J. Meriluoto, Global geographical and historical overview of cyanotoxin distribution and cyanobacterial poisonings, Arch. Toxicol. 93 (2019) 2429–2481, http://dx.doi.org/10.1007/s00204-019-02524-4.

[38] H. Li, M. Barber, J. Lu, R. Goel, Microbial community successions and their dynamic functions during harmful cyanobacterial blooms in a freshwater lake, Water Res. 185 (2020) 116292, http://dx.doi.org/10.1016/j.watres.2020.116292.

[39] A.D. Turner, M. Dhanji-Rapkova, A. O'Neill, L. Coates, A. Lewis, K. Lewis, Analysis of microcystins in cyanobacterial blooms from freshwater bodies in England, Toxins 10 (2018) 39, http://dx.doi.org/10.3390/toxins10010039.

[40] A. Stüken, R.J. Campbell, A. Quesada, A. Sukenik, P.K. Dadheech, C. Wiedner, Genetic and morphologic characterization of four putative cylindrospermopsin producing species of the cyanobacterial genera Anabaena and Aphanizomenon, J. Plankton Res. 31 (2009) 465–480, http://dx.doi.org/10.1093/plankt/fbp011.

[41] S. Cirés, A. Quesada, Catálogo de cianobacterias planctónicas potencialmente tóxicas de las aguas continentales españolas, Ministerio de Medio Ambiente y Medio Rural y Marino, Madrid, Spain, ISBN: 978-84-491-1072-6, 2011.

[42] K. Kumar, R.A. Mella-Herrera, J.W. Golden, Cyanobacterial heterocysts, Cold Spring Harb. Perspect. Biol. 2 (2010) a000315, http://dx.doi.org/10.1101/cshperspect.a000315.

[43] H. Wei, D. Brune, D. Anderson, J. Shi, An integrative Raman microscopy-based workflow for rapid *in situ* analysis of microalgal lipid bodies, Biotechnol. Biofuels 8 (2015) 164, http://dx.doi.org/10.1186/s13068-015-0349-1.

[44] K. Schulze, D.A. López, U.M. Tillich, F. Dürr, M. Frohme, A simple viability analysis for unicellular cyanobacteria using a new autofluorescence assay, automated microscopy, and ImageJ, BMC Biotechnol. 11 (2011) 118, http://dx.doi.org/10.1186/1472-6750-11-118.

[45] M.G. Fernández-Manteca, A.A. Ocampo-Sosa, C. Ruiz de Alegría Puig, M.P. Roiz, J. Rodríguez-Grande, F. Madrazo, J. Calvo, L. Rodríguez-Cobo, J.M. López-Higuera, M.C. Fariñas, A. Cobo, Clasificación automática de especies de Candida utilizando espectroscopia Raman y aprendizaje automático, Spectrochim. Acta Part A: Mol. Biomol. Spectrosc. 290 (2023) 122270, http://dx.doi.org/10.1016/j.saa.2022.122270.

[46] M.G. Fernández-Manteca, A.A. Ocampo-Sosa, D.Fernandez. Vecilla, M.Siller. Ruiz, M.P. Roiz, F. Madrazo, J. Rodríguez-Grande, J. Calvo-Montes, L. Rodríguez-Cobo, J.M. López-Higuera, M.C. Fariñas, A. Cobo, Identification of hypermucoviscous *Klebsiella pneumoniae* K1, K2, K54 and K57 capsular serotypes by Raman spectroscopy, Spectrochim. Acta Part A: Mol. Biomol. Spectrosc. 301 (2024) 124533, http://dx.doi.org/10.1016/j.saa.2024.124533.

[47] SHAP Documentation. Available online: https://shap.readthedocs.io/en/latest/index.html, (Accessed on 25 2024).

[48] D.K. Saini, S. Pabbi, P. Shukla, Cyanobacterial pigments: Perspectives and biotechnological approaches, Food Chem. Toxicol. 120 (2018) 616–624, http://dx.doi.org/10.1016/j.fct.2018.08.002.

[49] M. Roldán, C. Ascaso, J. Wierzchos, Fluorescent fingerprints of endolithic phototrophic cyanobacteria living within halite rocks in the Atacama Desert, Appl. Environ. Microbiol. 80 (2014) 4851–4859, http://dx.doi.org/10.1128/AEM.03428-13.

[50] Y. Kolodny, Y. Avrahami, H. Zer, M.J. Frada, Y. Paltiel, N. Keren, Phycobilisome light-harvesting efficiency in natural populations of the marine cyanobacteria Synechococcus increases with depth, Commun. Biol. 5 (2022) 727, http://dx.doi.org/10.1038/s42003-022-03677-2.

[51] C. Pancrace, M.-A. Barny, R. Ueoka, A. Calteau, T. Scalvenzi, J. Pédron, V. Barbe, J. Piel, J.-F. Humbert, M. Gugger, Insights into the Planktothrix genus: Genomic and metabolic comparison of benthic and planktic strains, Sci. Rep. 7 (2017) 41181, http://dx.doi.org/10.1038/srep41181.

[52] C. Djediat, K. Feilke, A. Brochard, L. Caramelle, S. Kim Tiam, P. Sétif, T. Gauvrit, C. Yéprémian, A. Wilson, L. Talbot, B. Marie, D. Kirilovsky, C. Bernard, Light stress in green and red Planktothrix strains: The orange carotenoid protein and its related photoprotective mechanism, Biochim. et Biophys. Acta (BBA) - Bioenerg. 1861 (2020) 148037, http://dx.doi.org/10.1016/j.bbabio.2019.06.009.

[53] N.I. Novikova, H. Matthews, I. Williams, M.A. Sewell, M.K. Nieuwoudt, M.C. Simpson, N.G.R. Broderick, Detecting phytoplankton cell viability using NIR Raman spectroscopy and PCA, ACS Omega 7 (2022) 5962–5971, http://dx.doi.org/10.1021/acsomega.1c06262.

[54] R.E. Barletta, J.W. Krause, T. Goodie, H. El Sabae, The direct measurement of intracellular pigments in phytoplankton using resonance Raman spectroscopy, Mar. Chem. 176 (2015) 164–173, http://dx.doi.org/10.1016/j.marchem.2015.09.005.

[55] V.E. de Oliveira, M.A.C. Neves Miranda, M.C.S. Soares, H.G.M. Edwards, L.F.C. de Oliveira, Study of carotenoids in cyanobacteria by Raman spectroscopy, Spectrochim. Acta. Part A, Mol. Biomol. Spectrosc. 150 (2015) 373–380, http://dx.doi.org/10.1016/j.saa.2015.05.044.

[56] E. Perevedentseva, N. Melnik, E. Muronets, A. Averyushkin, A. Karmenyan, I. Elanskaya, Raman spectroscopy with near IR excitation for study of structural components of cyanobacterial phycobilisomes, J. Lumin. 120224 (2023) 120224, http://dx.doi.org/10.1016/j.jlumin.2023.120224.

[57] Y.D. Winters, T.K. Lowenstein, M.N. Timofeeff, Identification of carotenoids in ancient salt from Death Valley, Saline Valley and Searles Lake, California, using laser Raman spectroscopy, Astrobiology 13 (2013) 1065–1080, http://dx.doi.org/10.1089/ast.2012.0952.

[58] W. Liu, Z. Wang, Z. Zheng, L. Jiang, Y. Yang, L. Zhao, W. Su, Density functional theoretical analysis of the molecular structural effects on Raman spectra of β-carotene and lycopene, Chin. J. Chem. 30 (2012) 2573–2580, http://dx.doi.org/10.1002/cjoc.201200661.

[59] M. Mehdizadeh Allaf, H. Peerhossaini, Cyanobacteria: Model microorganisms and beyond, Microorganisms 10 (2022) 696, http://dx.doi.org/10.3390/microorganisms10040696.