



A portable and low-cost optical device for pigment-based taxonomic classification of microalgae using machine learning

Vitor Magalhães^a, Vânia Pinto^{a,c}, Paulo Sousa^{a,c}, José A. Afonso^{a,c}, Luís Gonçalves^{a,c}, Emilio Fernández^b, Graça Minas^{a,c,*}

^a CMEMS-UMinho, University of Minho, Campus de Azurém, Guimarães 4800-058, Portugal

^b Centro de Investigación Marina. Universidade de Vigo, Vigo 36310, Spain

^c LABBELS –Associate Laboratory, Braga, Guimarães, Portugal

ARTICLE INFO

Keywords:

Microalgae identification
HABs
Fluorometry
Machine Learning
Portable device

ABSTRACT

The proliferation of certain phytoplankton species may lead to harmful algal blooms (HABs) that can affect living resources and human health. Therefore, an accurate identification of phytoplankton populations is essential for the sustainable management of some activities relevant for the blue economy, such as aquaculture, being also relevant for environmental monitoring and marine research purposes. Microalgae taxonomic discrimination, based on their pigment composition, is a versatile and promising technique to detect and identify potential HABs. In this work, a portable and low-cost device for taxonomic identification of microalgae, based on the pigment composition of 16 species belonging to 6 different phyla, was developed. It uses the fluorescence intensity signal emitted by each species at three wavelengths (575 nm, 680 nm and 730 nm) when excited at five wavelengths (405 nm, 450 nm, 500 nm, 520 nm and 623 nm) to create a fluorescence signature for each species. Furthermore, several machine learning classifiers were studied using this fluorescence signature as features to train and classify each species according to their respective taxonomic group. The Extreme Gradient Boosting (XGBoost) classifier was able to correctly predict microalgae monocultures with 97 % accuracy at the phylum level and 92 % accuracy at the order level. The obtained results confirm the potential of this technique for fast, accurate and low-cost identification of microalgae.

1. Introduction

Microalgae play a crucial role in assessing water quality, as well as marine and public health. They serve as important indicators of the overall health and ecological balance of aquatic ecosystems [1]. However, they can also cause environmental problems when some species proliferate exponentially, a natural phenomenon known as harmful algal blooms (HABs). Some HABs are constituted by species capable of producing toxins, others are non-toxic high biomass proliferations that can lead to water discoloration and oxygen depletion [2]. These events can take place in large or small areas depending on the species and external conditions. The efficient and timely monitoring, with large spatial resolution, of these risk situations is essential to safeguard public health and blue economy, particularly shellfish and finfish aquaculture. However, the implementation of these monitoring programmes is currently complex, expensive and demands skilled personnel.

Phytoplankton is composed of microscopic photosynthetic

unicellular organisms where sizes range from less than 1 µm to colonies larger than 500 µm, and comprises diverse evolutionary lineages, including eukaryotes and cyanobacteria, that show differences in cell morphology, ornamentation, photosynthetic pigments (chlorophylls, carotenoids and phycobilins) and other biochemical markers. All these features allow discrimination between species [3–5].

Standard methods for phytoplankton bloom identification and quantification rely mostly on morphological analysis and/or pigment composition [6,7]. The most commonly used method is image-based analysis, which is usually done through microscopy and requires well-trained personnel. Despite being very robust in detecting and quantifying target species, it is time-demanding and restricted to lab-based operations. Pigment composition-based methods are able to distinguish microalgae due to their spectral properties using fluorometric techniques. Several research works have substantiated the viability of employing fluorescence techniques for the classification of microalgae [4,7–10]. However, existing devices, such as FlowCAM [11],

* Corresponding author at: CMEMS-UMinho, University of Minho, Campus de Azurém, Guimarães 4800-058, Portugal.

E-mail addresses: vpinto@dei.uminho.pt (V. Pinto), gminas@dei.uminho.pt (G. Minas).

Imaging Flow Cytobot [12], CytoBuoy [13] and Laser Optical Plankton Counter [14], while accurate, are often bulky and expensive, which might hinder their application for real-time, on-site detection of harmful algae over large spatial scales, especially in remote or resource-constrained environments.

In addition to these devices, there is an ongoing effort to analyze phytoplankton and thus predict HABs, using satellite ocean color imaging [15] and meteorological forecasts [16], and while useful for broad monitoring, lack the resolution and specificity needed to identify diverse phytoplankton species in coastal areas where HABs are most prevalent.

Unmanned aerial vehicles (UAVs) have been also used to monitor possible HAB outbreaks with high spatial resolution and cost-effectiveness, but battery power consumption and bad weather are limiting factors for extended sample collection and data quality when using this approach [17,18].

The integration of sensors and satellite data with machine learning (ML) and deep learning (DL) is a hot topic in the research field, which allowed further expanding the capability for accurate, faster, cheaper and widespread monitoring of phytoplankton, by combining data from satellites and/or autonomous sensors capable of in situ monitoring featuring also the capability for ML and DL for real time analysis of the collected data set [19–23]. There is already extensive research showing the feasibility of combining fluorescence-based optical sensors with ML and DL with promising results [24–27].

As the need for efficient and reliable in situ phytoplankton identification has grown, the development of portable and autonomous devices has become increasingly important. Despite some progress in this area, current portable devices still face challenges regarding cost, portability, and real-time identification accuracy. There is an urgent need for a compact, affordable, and highly sensitive device that can be massively deployed for systematic in-situ monitoring of both toxic and non-toxic phytoplankton. Such a tool would have a significant impact on understanding HABs dynamics, thereby providing critical data to predict, for instance, shellfish and finfish toxification episodes, enabling sustainable harvesting decisions and safeguarding public health.

Thus, in this work, a portable and low-cost (estimated less than 250 EUR) device for taxonomic identification of microalgae species based on fluorometry and ML was developed. Our device addresses the limitations of existing methods by offering a compact, affordable, and portable solution, enabling on-site classification of microalgae.

By making this technology accessible, even in developing countries, the device provides early detection of important microalgal populations, such as toxic species, offering a transformative approach to managing HABs and their environmental and economic impacts, bridging the gap between current technology and the need for real-time, accessible monitoring. The compact design and low power requirements ensure the device can be widely used, even in remote or resource-limited regions, where traditional bulky lab equipment cannot be deployed.

The device uses the fluorescence signal of several species emitted at three wavelengths regions (575 nm, 680 nm, 730 nm) when excited at five wavelengths (405 nm, 450 nm, 500 nm, 520 nm and 623 nm). These excitation and emission wavelengths were selected based on fluorometry analysis, using a commercial equipment, scanning the excitation spectra and acquiring the emission spectra for each species. The different patterns of microalgae fluorescence that arise from the combination of various excitation sources at different emission regions allow for classification of microalgae through ML. Several ML classifiers were tested and optimized in order to find the best model that produced the most accurate results.

2. Materials and methods

2.1. Phytoplankton cultures

A total of 16 species belonging to 6 phyla were selected to conduct the study. The phytoplankton monocultures were collected from the

Toralla Marine Science Station (ECIMAT), Vigo, Spain. A description of their taxonomic classification (Table S1) as well as their main characteristics, including morphological variables, pigments and toxins (Table S2) are described in the [supplementary material](#), section S1.

2.2. Fluorescence characterization

The emission and excitation spectra of monocultures species were obtained using the Shimadzu RF-5301PC spectrophotometer at the Toralla Marine Science Station (ECIMAT). Knowledge on the excitation spectra of each species was key to select the light-emitting diodes (LEDs) sources that allowed the best potential discrimination between different taxonomic groups.

2.3. Portable device for fluorescence measurements

Based on a fluorescence spectral discrimination approach, a portable device for detecting multiple fluorescence signals was implemented. This method involves exciting microalgae pigments with light sources at different excitation regions and quantifying the subsequent emission of light in regions of interest. The device includes four main subsystems: (1) an excitation source; (2) a photodetection system; (3) an electronics system; and (4) a power source consisting of batteries. A 3D representation of the optical apparatus of the prototype is shown in [Fig. S1 of supplementary material](#).

The excitation source comprises a set of high-intensity LEDs (Polychromatic LuxiGen Multi-Color LED 897-LZ7A4M2PD0000 plus a Bivar UV3TZ-405–30) at wavelengths that cover the range of interest (405 nm, 450 nm, 500 nm, 520 nm and 623 nm, outputted from [Section 3.1](#)) ([Fig. S2 of supplementary material](#)); they are arranged in a compact area ($3.4 \times 3.4 \text{ mm}^2$) facilitating its integration in miniaturized devices. Each LED is activated for 1 second to limit the decrease in fluorescence intensity due to the photochemical quenching effect [28].

The photodetection system includes a silicon photodiode (Hamamatsu, S1336–8BK) placed below the cuvette holder for capturing the fluorescence intensity, with a range of sensitivity from ultraviolet to infrared. Its responsivity, high-quantum efficiency, non-uniformity, non-linearity and low-noise makes it widely used in optical measurement equipment. Above the photodiode there is a mobile drawer with three optical bandpass filters (centred wavelength outputted from [Section 3.1](#)). One centred at 575 nm with a FWHM of 10 nm (575BP10 from Laser Components); other centred at 680 nm with a FWHM of 10 nm (FB680–10 from Thorlabs); and the third centred at 730 nm with a FWHM of 13 nm (730BP15 from Laser Components) ([Fig. S3 of supplementary material](#)). They are used to suppress the excitation signals and enhance sensitivity by significantly reducing background noise. Additionally, the excitation and detection subsystems are assembled in a 90° configuration using a 3D-printed support to ensure the correct alignment and positioning of the optical components.

The signal acquisition is susceptible to ambient noise interference, electronic components and circuit interactions sources, hampering the signal-to-noise ratio. So, a lock-in amplifier [29] was used for detection and measurement of low amplitude signals, even when the signal of interest is completely embedded in noise. It uses a phase-sensitive detection technique, which isolates the measured signal through a reference signal with a certain frequency and phase. In this way, the noise present at frequencies different from the ones of the reference signal will be rejected. Here, this is achieved by modulating the excitation signal at 1 kHz (pulsing the excitation LEDs at 1 kHz). The photocurrent outputted by the photodiode, after converted to voltage, is filtered by a 1 kHz band-pass filter isolating the desired signal from unwanted noise frequencies. To restore the original signal, a synchronous demodulator, operating at 1 kHz, translates the modulated signal into direct current (DC), effectively eliminating un-synchronized signals [29,30]. Finally, the useful signal will be processed by a computer through a microcontroller.

The developed device for microalgae fluorescence detection is shown in Fig. 1. Their compact design and small size provides portability allowing their operation outside the laboratory.

2.4. Machine learning algorithm

For the ML implementation, first, the microalgae species were associated with their respective taxonomic phylum. For instance, the *Alexandrium tamarense* and *Alexandrium minutum* species were associated to the Miozoa phylum they belong to (see Table S2 of supplementary material). The two species of cyanobacteria despite belonging to the same phylum, were considered as two distinct groups (Cyanobacteria 1 and Cyanobacteria 2) due to their very distinct pigment composition.

Fluorescence intensity ratios were calculated between the excitation and emission wavelengths that provided the greatest distinction, creating a unique spectral signature for each species (more details described in Section 3). These signatures were then analyzed by ML models for taxonomic classification.

To avoid any possibility of data leakage and ensure reliable model evaluation, the dataset was randomly split into training (80 %) and testing (20 %) sets. This resulted in 148 samples for training and 38 samples for testing. The random splitting was done at the species level, ensuring that the same species did not appear in both the training and test sets to avoid overfitting or data leakage. To further ensure the robustness of the model, a 3-fold cross-validation was employed on the training set. This means the training set was split into three equal parts, with two parts used for training and one part for validation, cycling through all combinations. Further details of the tested ML classifiers are in Section S6 of the Supplementary Information.

Due to the logistical challenge of culturing 16 different microalgae species at a quasi-synchronous growth phase, the training set had a relatively small number of samples. To overcome this limitation, a diverse array of ML classifiers with varying hyperparameters was employed to optimize model performance.

3. Results and discussion

3.1. Fluorescence spectra

Fig. 2 shows the fluorescence emission measured at 575 nm, 680 nm and 730 nm, respectively Fig. 2a, b and c, over a range of excitation wavelengths. The emission at 575 nm corresponds to the phycoerythrin pigment emission, an important marker pigment for cryptophytes (RI) and cyanobacteria (Syn034). The 680 nm and 730 nm correspond

mostly to the chlorophyll an emission. Only one species representative of each phylum (with exception of cyanobacteria) is represented in Fig. 2 to avoid confusing visualization due to excessive overlapping of curves.

The overall fluorescence of microalgae is a complex interplay of various pigments and their interactions within the photosynthetic apparatus. Due to variations in their pigment composition, different algal groups exhibit distinct patterns of photon absorption depending on the excitation region in the spectrum. As a result, they emit fluorescence with varying patterns allowing exploration of the relationships between emitted fluorescence from different excitation sources, thereby identifying discriminant features between algal groups.

Species like *Rhodomonas lens* (RI) and *Synechococcus* sp. (Syn033 and Syn034) have very distinct fluorescence properties because of their unique pigment composition, allowing easy distinction between them. For instance, *Rhodomonas lens* (RI) and *Synechococcus* (Syn034) both have the pigment phycoerythrin, and, for that reason, they emit fluorescence at 575 nm when excited at 500–540 nm. Another *Synechococcus* species (Syn033) has the pigment allophycocyanin with an excitation peak at 620 nm, which also sets it apart from other species. The remaining species have excitation peaks much more similar and fluorescence emission only around 680–730 nm region, making discrimination challenging.

We selected five specific regions within the spectrum where the variations and magnitude in fluorescence intensity were most pronounced among the species (see Fig. 2). The chosen excitation wavelengths were centred at 405 nm, 450 nm, 500 nm, 520 nm, and 623 nm. Fig. 3 represents the emission spectra for 8 species belonging to 7 different phyla, for the 5 selected wavelengths. Discrimination between *Rhodomonas lens* (RI) and the two *Synechococcus* species is quite straightforward. Their unique pigment composition gives rise to very distinct fluorescence patterns when excited at the 5 selected wavelengths. The more challenging part is the discrimination between species that only emit in the 680–730 nm region, which is the case for most species in this study. Despite having fluorescence emission at similar regions, their different pigment composition affects how each species will absorb each different wavelength, and, as a result, the intensity of emitted fluorescence at this specific wavelength will be slightly different. These differences can provide distinguishing features to ML classifiers that can be used to identify different taxonomic groups for microalgae where fluorescence emission overlaps.

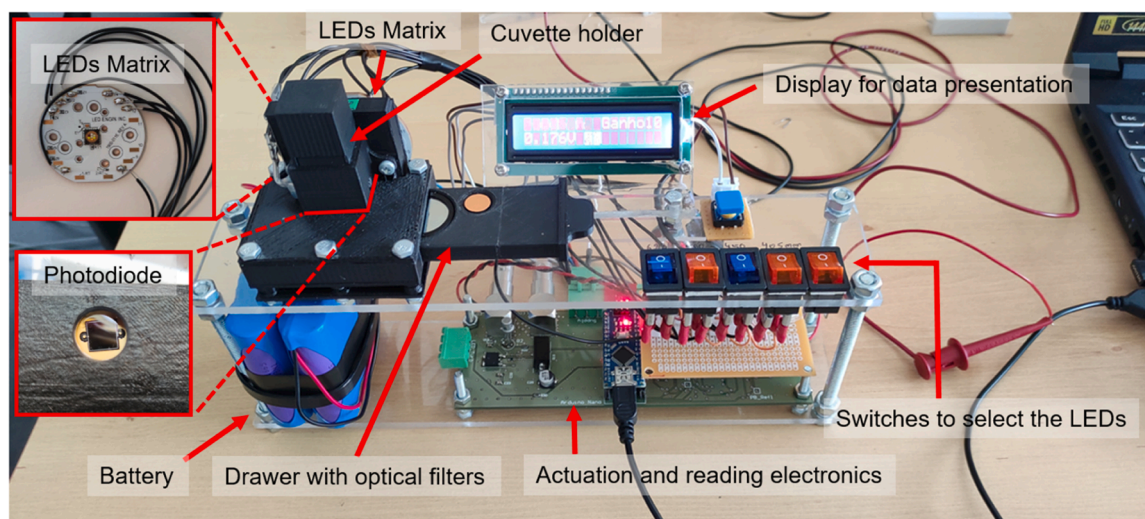


Fig. 1. Portable device to detect multiple fluorescence signals.

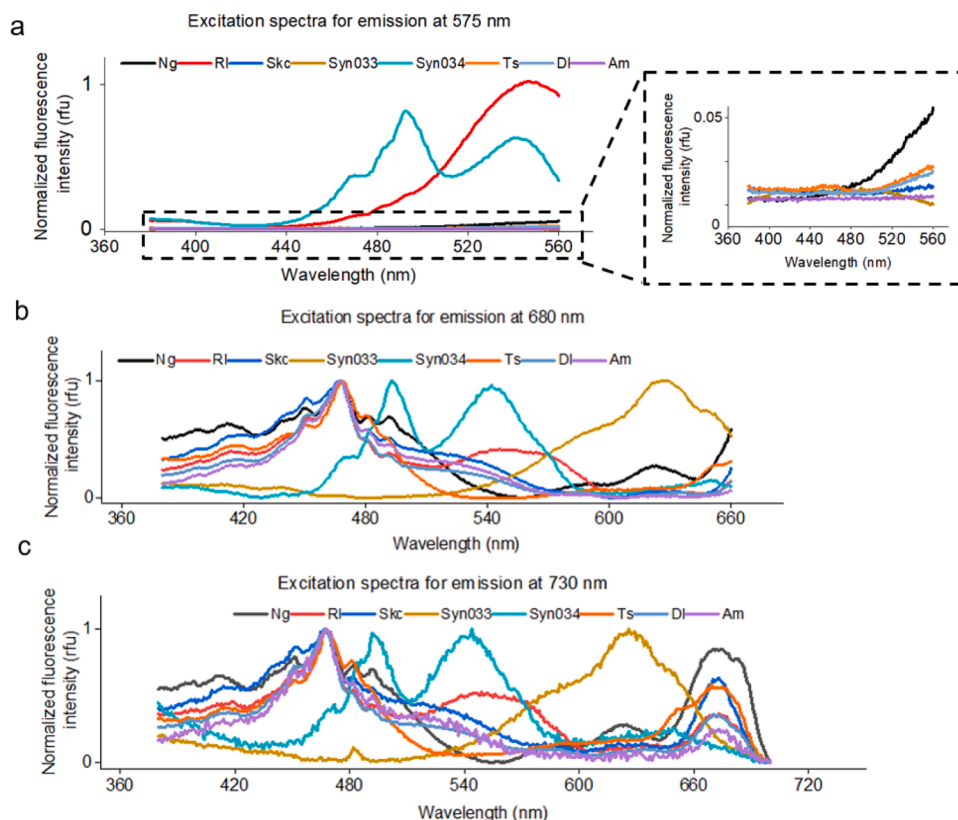


Fig. 2. Excitation spectra for fluorescence emission at a) 575 nm (dashed lines correspond to a zoomed section of the plot), b) 680 nm, and c) 730 nm. Each line represents the excitation spectrum of a different species, with abbreviations corresponding to those listed in Table S2 of the supplementary material.

3.2. Device validation

After an initial calibration (procedures detailed in section S3 of supplementary material), the developed device was used for measuring the emission signals of different monocultures at the wavelengths previously stated. The samples were collected at the stationary phase of growth, although due to the different growth rates of the tested species, dilutions were performed to maintain similar concentrations. The detected fluorescence signal was validated by comparing the obtained measurements with the excitation spectra of the microalgae. Correction factors, using the measured mean value of the filtered seawater samples ($n = 8$), were applied to the microalgae fluorescence signal, yielding a final value of 1 (for correcting possible inherent biases). These correction factors, specific for each combined LED and emission filter, are detailed in Table S3 of the supplementary material. When conducting measurements, these correction factors are applied to the photodiode output voltage, resulting in an adjusted fluorescence emission signal.

Fig. 4 shows the fluorescence pattern recorded by the device compared with the respective excitation spectra for a selection of microalgae representative of each phylum. In general, the fluorescence patterns are in good agreement with the excitation curves. *Rhodomonas lens* (Fig. 4a) and *Synechococcus* species (Fig. 4b c) exhibit distinct fluorescence patterns. However, subtle differences were found in the fluorescence patterns of the other species. Focusing only on the fluorescence emission in the regions of 680 nm and 730 nm, both *Alexandrium minutum* (Fig. 4d), *Diatrypa lutheri* (Fig. 4e) and *Skeletonema costatum* (Fig. 4f) exhibits significantly lower fluorescence emission when excited at 623 nm (red bar) compared with excitation at 520 nm (green bar). By contrast, *Nannochloropsis gaditana* (Fig. 4g) and *Tetraselmis suecica* (Fig. 4h) showed fluorescence emission values at 623 nm either similar or even higher compared to those at 520 nm. *Alexandrium minutum* (Fig. 4d) shows slightly lower fluorescence emission when

excited at 405 nm (purple bar) when compared to 500 nm excitation (cyan bar). In contrast, *Nannochloropsis gaditana* (Fig. 4g), *Tetraselmis suecica* (Fig. 4h) and *Diatrypa lutheri* (Fig. 4e) display higher fluorescence emission at an excitation of 405 nm compared to at 500 nm, while *Skeletonema costatum* (Fig. 4f) presents similar intensity between both excitations. These results align well with the emission spectra depicted in Fig. 3 and allowed assigning a fluorescence signature to each species, based on the fluorescence patterns measured by the device.

3.3. Fluorescence signature

A fluorescence signature was fixed employing a ratiometric approach based on ratios between fluorescence signals obtained using the five different excitation LEDs for each emission region (basically, derived from Fig. 4). Specifically, for emissions at 680 nm and 730 nm, we derived a total of ten ratios for each emission. In the case of the 575 nm emission, a single ratio was determined, involving the 520 nm and 405 nm LEDs. Each abbreviation letter corresponds to a specific ratio between two excitation LEDs (see table S4 of supplementary material) and the number, 575 nm, 680 nm or 730 nm, corresponds to the emission wavelength used (Fig. 5, x-axis). This means that each species is characterized by a total of 21 ratios (10 ratios for the emission region at 680 nm, 10 ratios for the emission region at 730 nm, and one ratio for the 575 nm emission region), each ratio corresponding to a point of the fluorescence signature of each species. Fig. 5 shows the fluorescence signature of all the species tested, calculated from the average of 12–20 measurements for each species. As an example, the ratio C575 (the first in the x-axis) corresponds to the ratio of emitted fluorescence intensities at 575 nm when excited with the 520 nm and the 405 nm LEDs. Since only *Rhodomonas lens* (RI) and *Synechococcus* (Syn034) have phycoerythrin, they can be easily distinguished from other species by using that ratio. *Synechococcus* (Syn033) shows a high signal at 623 nm

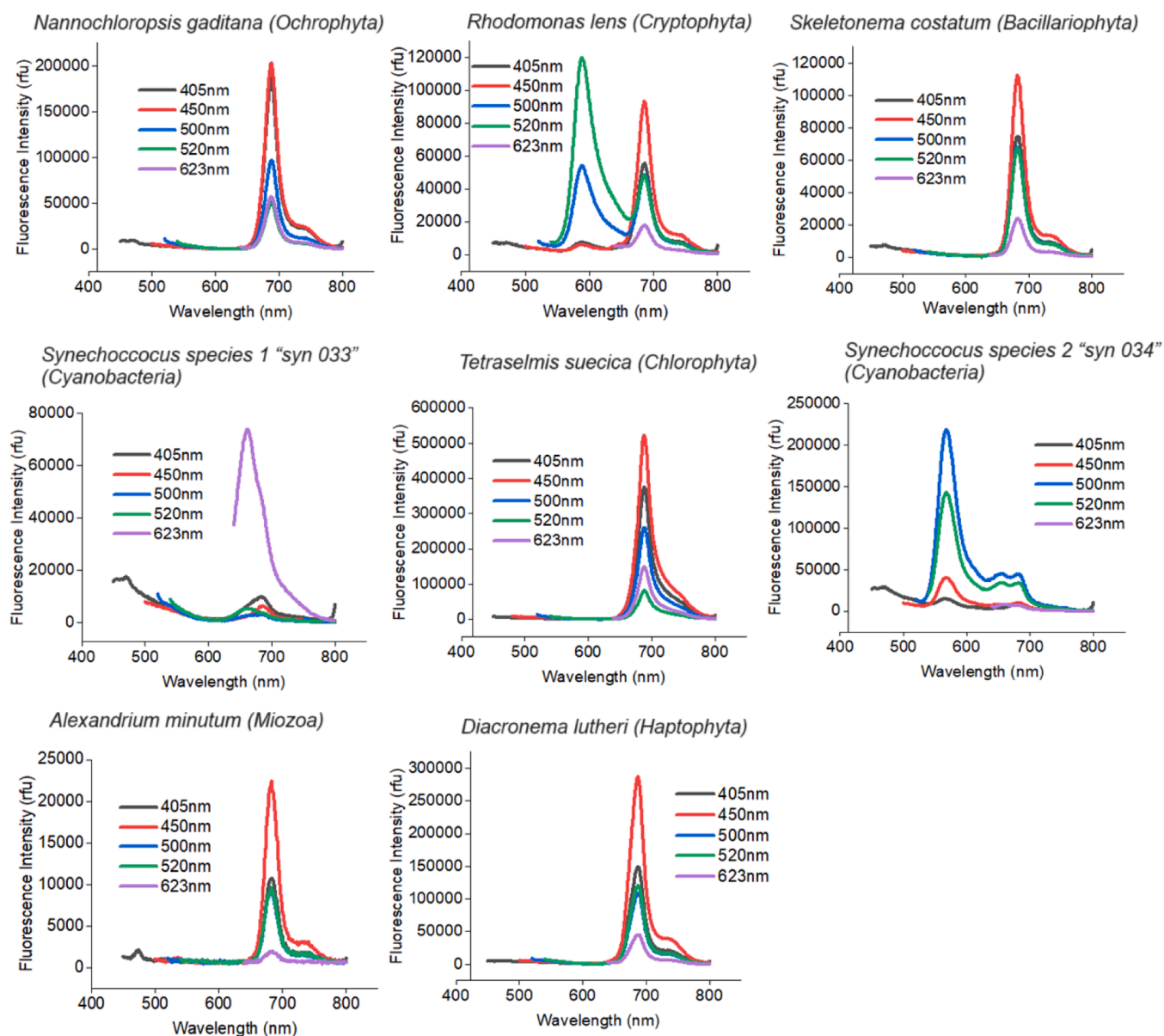


Fig. 3. Emission spectra for the excitation at 405 nm, 450 nm, 500 nm, 520 nm and 623 nm, for 8 phytoplankton species representative of 7 different phyla.

because of the presence of phycocyanin, and for that reason, their I and J ratios at 680 nm and 730 nm emissions are clearly distinguished from those of other species. However, for most ratios, identifying distinguishing patterns between different species is more complex, and this is the case where ML assumes an important role, as these ratios (fluorescence signatures) can be used as features in the ML models. Leveraging these features (ratios), the ML models will strive to differentiate and categorize microalgae, with the ultimate goal of maximizing accuracy in assigning data points to their appropriate classes. The adopted methodology is thus designed to identify the presence of a specific taxonomic group in a given sample.

3.4. Machine learning classification results

A total of 12 ML classifiers were tested to find the one that outputted the best performing in the dataset. In the training set, the model constructs the classifier based on the labeled observations and tries to separate the predefined classes as best as it can. In the testing set, the model predicts the class of unlabeled observations.

The process is divided into two parts. First, each model was trained

and validated using only the training set with a cross-validation of 3. This means that the training set was split in a way that 1/3 represented a validation set and 2/3 the training set. This was repeated 3 times, allowing all training samples to be included in the validation set at some point. Every classifier was also optimized by iteration of several hyperparameters and the ones yielding the best score were saved. This cross-validation technique along with the iteration of several hyperparameters maximized the robustness and accuracy of the results, mitigating the impact of the limited training data. In the second part, the optimized classifiers with the best accuracy were trained with the whole training set, and their classification performance was evaluated using the testing set. Accuracy was the main metric used to evaluate the performance of the ML models. Accuracy is defined as the ratio between the number of correctly predicted samples and the total number of samples. It measures the proportion of correct predictions made by the model overall predictions, both true positives and true negatives. However, other metrics such as precision and recall were also used to provide a more comprehensive performance evaluation, especially when handling potential imbalances in the dataset. These additional metrics help assess the model's performance in classifying less-represented

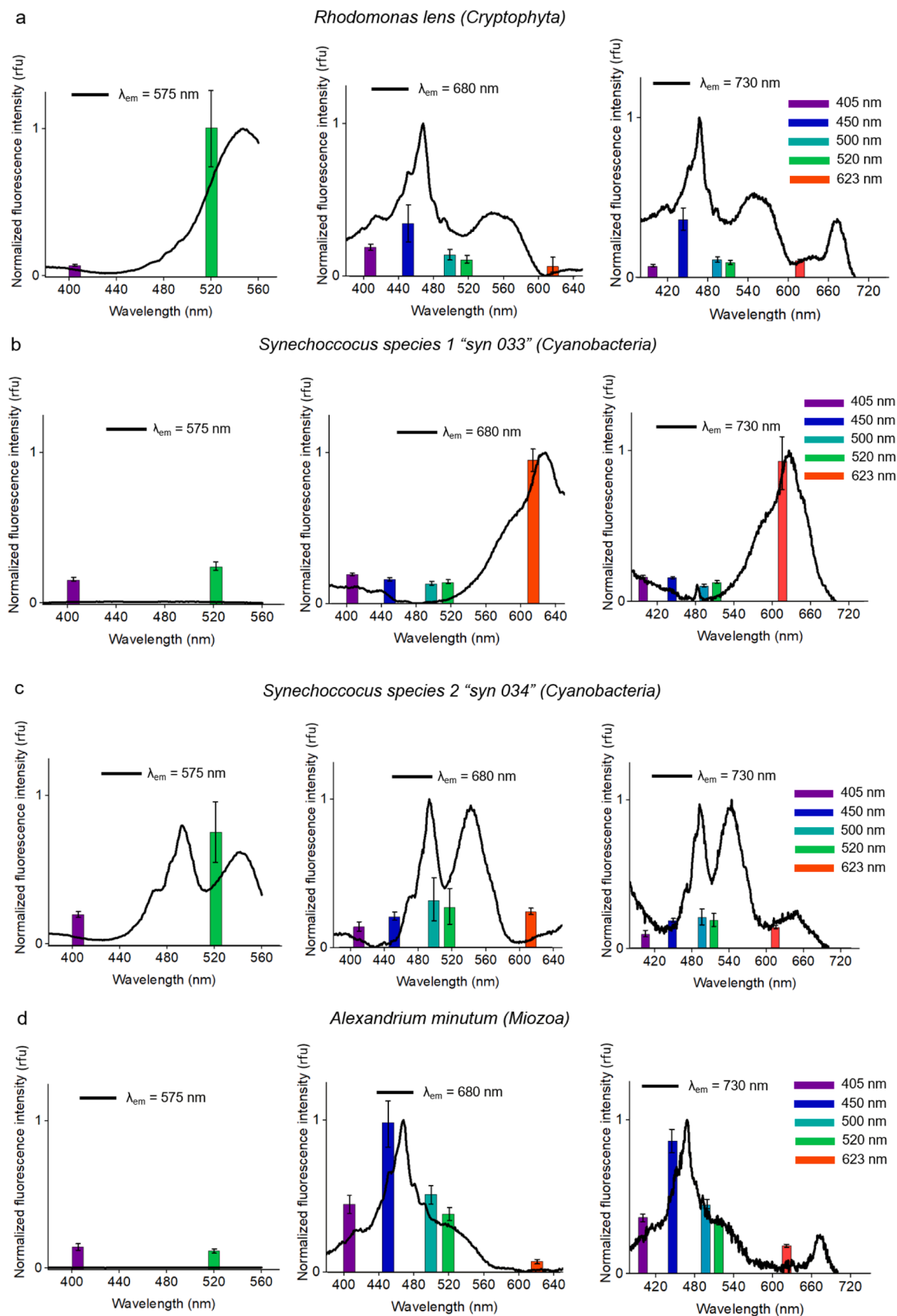


Fig. 4. Comparison of the fluorescence pattern recorded by the device (bar plots) with the fluorescence excitation spectrum for a) *Rhodomonas lens* (n=5), b and c) *Synechococcus* species (n=23), d) *Alexandrium minutum* (n=14), f) *Skeletonema costatum* (n=13), g) *Nannochloropsis gaditana* (n=11), and h) *Tetraselmis suecica* (n=12). Error bars represent the standard deviation in n samples measured.

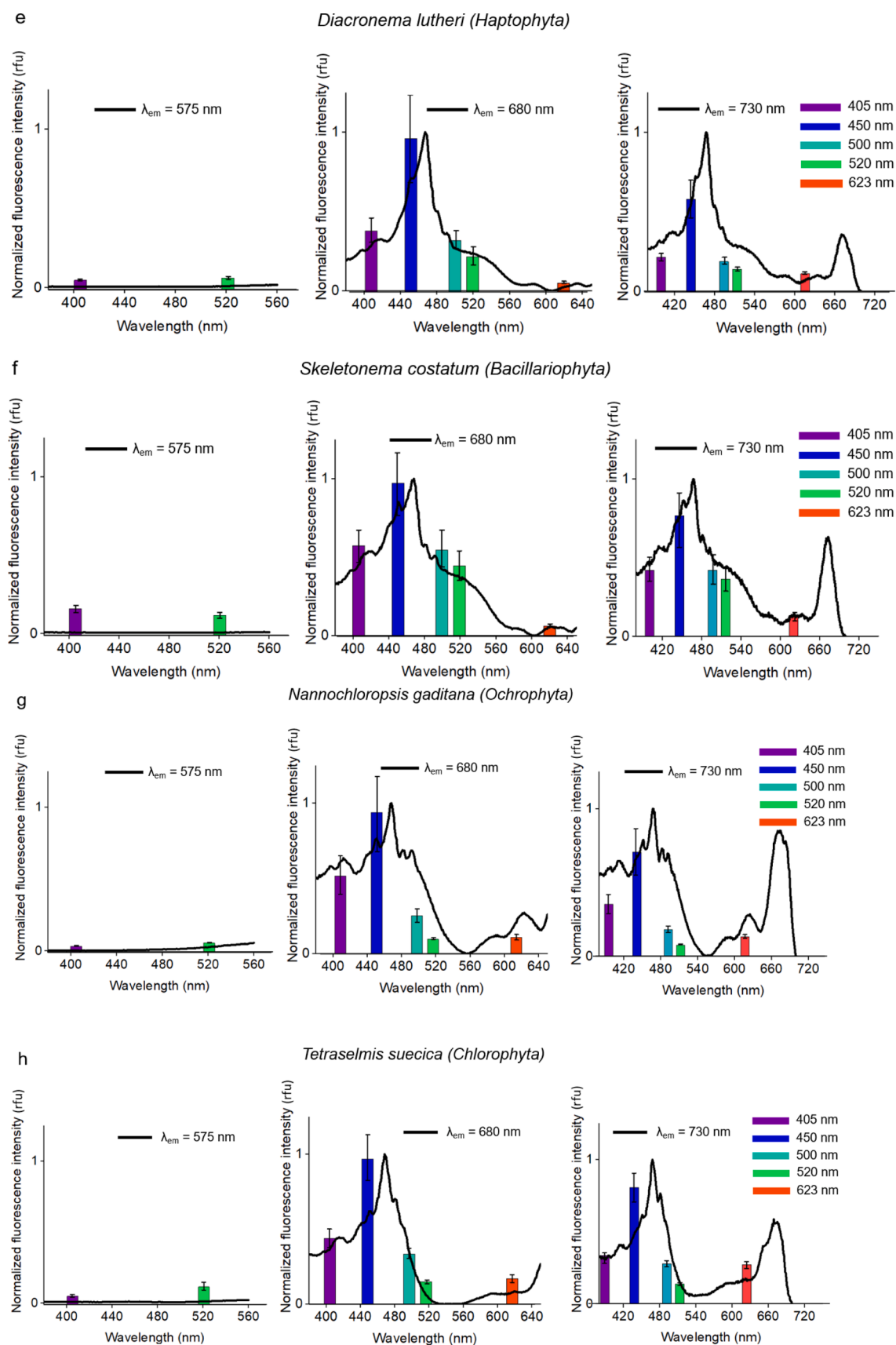


Fig. 4. (continued).

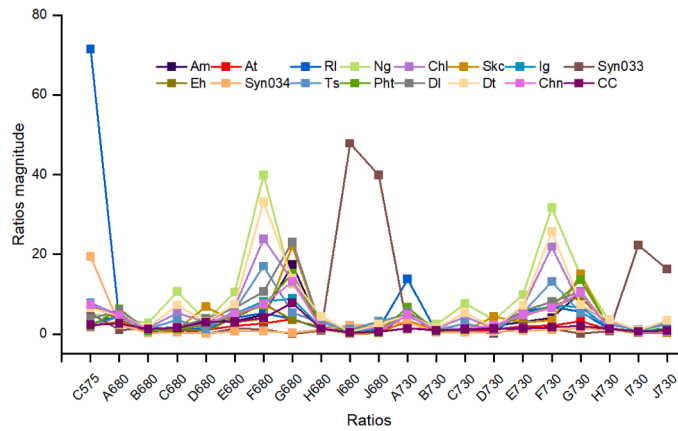


Fig. 5. Fluorescence signature composed of 21 ratios for each species.

Table 1
Results for training and validation of several ML classifiers at the phylum level.

Classifier	Training time (s)	Prediction time (s)	Best parameters	Accuracy for training dataset
Nearest Neighbors	18.0059	0.0171	'n_neighbors': 2	0.94
Decision Tree	0.1747	0.0003	'max_depth': 6	0.87
Random Forest	70.5316	0.0136	'max_depth': 2, 'max_features': 4, 'min_samples_leaf': 25, 'n_estimators': 100	0.29
Logistic Regression	1.2498	0.0003	'C': 1e-05, 'penalty': 'none'	0.93
Support Vector Machine	0.2005	0.0015	'C': 10	0.93
Neural Network	190.5183	0.0033	'activation': 'tanh', 'alpha': 1, 'hidden_layer_sizes': (100, 100, 100), 'learning_rate': 'constant'	0.96
AdaBoost	5.3509	0.0120	'learning_rate': 1, 'n_estimators': 50	0.93
Naive Bayes	0.0693	0.0021	'var_smoothing': 1e-09	0.87
Linear Discriminant Analysis	0.7927	0.0002	'shrinkage': 0, 'solver': 'lsqr', 'tol': 0.0001	0.92
Quadratic Discriminant Analysis	0.3432	0.0007	'reg_param': 0.01, 'tol': 0.0001	0.81
Gradient Boosting	277.7135	0.0130	'learning_rate': 0.01, 'max_depth': 3, 'n_estimators': 500	0.90
Extreme Gradient Boosting	34.6022	0.0020	'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 50	0.92

classes, mitigating the risk of bias toward majority classes.

Table 1 shows the selected classifiers, their respective training time and prediction time, and the accuracy, which represent the accuracy of the models on the training data for the best hyperparameters. More detailed information about the classifiers and respective hyperparameters is presented in [supplementary material S6](#). The classifiers with accuracy equal to or above 0.90 were selected to classify the test data. The accuracy of the selected classifiers for the testing data are presented in Table 2.

XGBoost is an ensemble ML algorithm widely popular for its exceptional performance on structured data, consistently achieving state-of-

Table 2

Accuracy of the classifiers for the test data.

Classifier	Accuracy for test dataset
Nearest Neighbors	0.95
Logistic Regression	0.87
Support Vector Machine	0.89
Neural Network	0.87
AdaBoost	0.87
Linear Discriminant Analysis	0.82
Gradient Boosting	0.87
Extreme Gradient Boosting	0.97

The Extreme Gradient Boosting (XGBoost) stood out with an accuracy in the test data of 0.97. This means that it correctly classified 37 of the 38 testing samples (97%).

the-art results across a variety of ML tasks [31]. It builds multiple decision trees sequentially, each one improving upon the errors made by the previous trees. XGBoost's key strength lies in its ability to correct residual errors iteratively, making it particularly effective for structured datasets and robust against small dataset size and imbalances, which would otherwise hinder models such as Random Forest or Neural Networks. While Neural Networks require large datasets to avoid overfitting, XGBoost is more resistant to overfitting due to its use of gradient boosting and regularization techniques, an important advantage when working with smaller datasets like the one in this study. Additionally, XGBoost employs parallelization and efficient memory usage, making it relatively faster and more efficient than many other classifiers [32,33].

Analyzing the selected hyperparameters for XGBoost, a learning rate of 0.1 indicates that the model learns gradually, with each tree contributing 10 % to the ensemble's prediction. The trees have a max depth of 3, which helps prevent overfitting and keeps the model simpler. With 50 estimators, XGBoost was able to capture patterns in the data without becoming overly complex. This combination of parameters allows XGBoost to strike a balance between model complexity and accuracy.

The classification matrix for the XGBoost, trained with the whole training set and respective testing set, is represented in Fig. 6, showing the correctly and incorrectly classified phyla. This matrix provides a detailed breakdown of the predictions made by the model (predicted label) and how they compare against the real taxonomic classification of

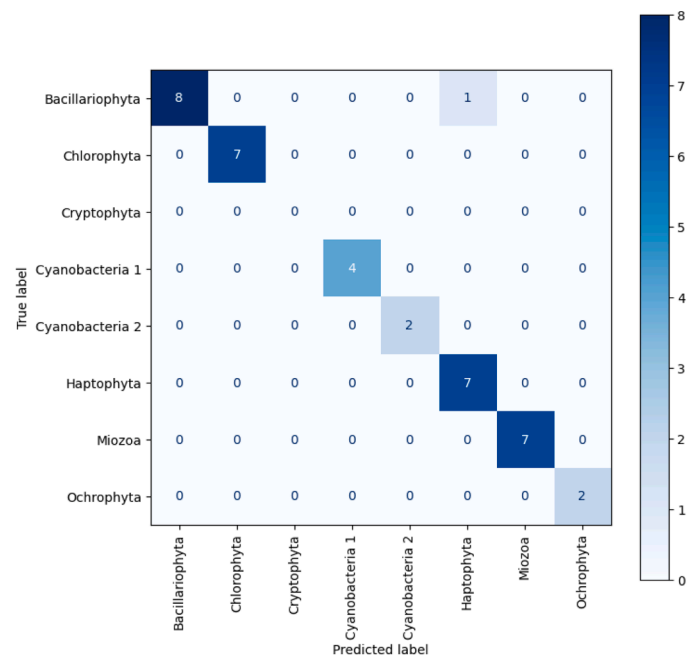


Fig. 6. Test data classification matrix for XGBoost at the phylum level.

microalgae (true label). The model was able to correctly classify almost all testing data to the correct phylum, with the exception of one sample belonging to the Bacillariophyta phylum which was incorrectly classified as Haptophyta.

The next step was to investigate the capability of XGBoost to correctly classify the microalgae at the order level. For that, each species was associated with one of the 14 taxonomic orders. The XGBoost was trained for the order level and its classification performance was evaluated using the testing set. For the order level, the XGBoost classifier showed an accuracy of 0.85 for the training data and an accuracy score of 0.92 for the testing data. The confusion matrix for the XGBoost trained with the whole training set and respective testing set is represented in Fig. 7, showing the correctly and incorrectly classified taxonomic order. Here, the model misclassified three testing samples, a Chlorellales sample wrongly predicted as a Chlamydomonadales and vice versa, and a Bacillariales incorrectly classified as a Thalassiosirales, making it a total of 35 (out of 38) correctly classified samples.

A summary of the main classification metrics for the XGBoost classifier at the phylum and order level is presented in Table 3. In addition to accuracy, other metrics such as precision, recall and F1-score, as well as the macro/micro-average scores, were used to evaluate the performance of the model through a set of different metrics and to circumvent the sometimes-misleading character of accuracy, as this metric may be high due to the model's ability to accurately predict the majority class, while performing poorly on the minority class. At the phylum level, all metrics show high values for the test data, showcasing the potential of this model to classify each phylum correctly, with a weighted average of 0.97 and 0.98 for recall and precision, respectively, showcasing the model's ability to accurately classify microalgae across diverse taxonomic groups. High values for all metrics were obtained except for the Chlamydomonadales, Chlorellales, Thalassiosirales and Bacillariales orders, which showed lower precision and/or recall. Despite that, the overall performance of the model is satisfactory, with a weighted average of 0.92 and 0.93 for recall and precision, respectively. XGBoost demonstrated excellent performance not only in accuracy but also in robustness against overfitting and the ability to generalize from a relatively small dataset, making it particularly suited for applications where data collection may be limited, but high classification accuracy is still required.

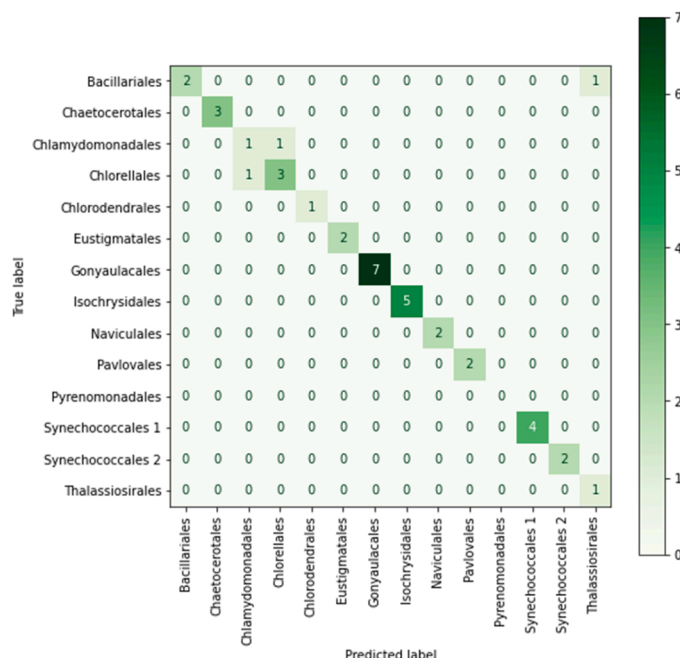


Fig. 7. Test data classification matrices for XGBoost at the order level.

Microalgae classification, employing a range of methods based on fluorescence analysis has undergone a substantial evolution and fine-tuning with the integration of ML techniques. The combination of conventional fluorescence methods with the capabilities of AI has opened up new possibilities and significantly enhanced precision and efficiency in distinguishing and characterizing various algae species. In this regard, Young-Ho et al. [34] demonstrated the distinction and the quantification of green algae and cyanobacteria and Persichetti et al. [9] employed a ratiometric approach distinguishing green algae, diatoms, and numerous cyanobacteria species. In our research, we focus on seawater species aiming to differentiate between representatives of seven different phyla (16 species). We used a smaller number of emission regions of interest (three regions), when compared with those in literature, but we employed multiple wavelength excitations (five LEDs) along with a photodiode equipped with optical filters to detect the fluorescence signals, enabling a compact and fully portable device for classifying microalgae. Additionally, our model achieves accuracy levels (97 % accuracy for 7 phyla and 92 % accuracy for 14 taxonomic orders) similar to those reported in previous works using similar pigment-based techniques with ML [7,35]. Nevertheless, the herein reported model advantage the added feature of a compact and portable device, suitable for use beyond the laboratory environment. As far as the authors' knowledge is concerned, there are no similar systems in the literature that use fluorometry and ML in a compact and portable device for phytoplankton taxonomic classification.

The future integration of this technology with specific sorting and concentration of microalgae of interest could enable a higher accuracy of the taxonomic discrimination of phytoplankton, especially in seawater samples with high-diversity phytoplankton communities. This could be addressed through strategic integration with a spiral micro-channel device recently developed by our group [36] as a pre-step process, which showcased the effectiveness of inertial microfluidics in sorting, isolating, and concentrating microalgae of interest based on size. The isolation of microalgae before examination narrows down the analysis to a more confined range of species. Additionally, the ability to enhance cell concentration prior to measurements is valuable to increase the fluorescence signal. Also noteworthy is the capability to isolate harmful species and significantly concentrate them increasing the chance of early detection of toxic species during the early phase of bloom development.

Autonomous microalgae identification and in situ monitoring will allow a massification of data collection which is a key factor in monitoring programs and can complement information from satellite images and modelling techniques [20,37–40]. While satellite images offer large-scale coverage and can capture spatial patterns, in situ detection focuses on local-scale details and provides sea-truth data for validation and calibration of satellite observations and models. By autonomously identifying microalgae in their natural environment, more precise insights into the composition and abundance of microalgae species in specific locations are gained. Additionally, real-time monitoring capabilities enable to swiftly detect and respond to sudden changes in microalgae populations. This integrated approach would likely enhance the effectiveness and efficiency of water monitoring and management, ultimately leading to improved environmental protection and optimized resource utilization.

4. Conclusions

The use of spectrofluorometric techniques coupled with ML capabilities demonstrated to be a promising tool for fast and accurate discrimination of phytoplankton species. This work confirms the potential of this approach as a robust, low-cost, and portable device implemented method for rapid taxonomic classification of microalgae growing in monocultures, with up to 97 % accuracy at the phylum level and 92 % accuracy at the order level. Future research will prioritize miniaturization and automation of the device for in situ testing while

Table 3
Classification report for the XGBoost model at the phylum and order level.

Phylum	Order	Precision (phylum/order)	Recall (phylum/order)	F1-score (phylum/order)	Support
Bacillariophyta		1.00	0.89	0.94	9
	Bacillariales	1.00	0.67	0.80	3
	Chaetocerotales	1.00	1.00	1.00	3
	Thalassiosirales	0.50	1.00	0.67	1
	Naviculales	1.00	1.00	1.00	2
Chlorophyta		1.00	1.00	1.00	7
	Chlamydomonadales	0.50	0.50	0.50	2
	Chlorellales	0.75	0.75	0.75	4
	Chlorodendrales	1.00	1.00	1.00	1
Ochrophyta		1.00	1.00	1.00	2
Miozoa	Eustigmatales	1.00	1.00	1.00	2
	Gonyaulacales	1.00	1.00	1.00	7
Haptophyta		0.88	1.00	0.93	7
	Isochrysidales	1.00	1.00	1.00	5
	Pavlovaes	1.00	1.00	1.00	2
Cryptophyta		0.00	0.00	0.00	0
	Pyrenomonadales	0.00	0.00	0.00	0
Cyanobacteria 1		1.00	1.00	1.00	4
	Synechococcales 1	1.00	1.00	1.00	4
Cyanobacteria 2		1.00	1.00	1.00	2
	Synechococcales 2	1.00	1.00	1.00	2
	micro avg	0.97/0.92	0.97/0.92	0.97/0.92	38
	macro avg	0.86/0.84	0.86/0.85	0.86/0.84	38
	weighted avg	0.98/0.93	0.97/0.92	0.97/0.92	38

expanding the database to enhance ML classification accuracy. The low-cost nature cost (estimated less than 250 EUR) of the presented device coupled with ML can potentially enable real-time monitoring of HABs, crucial for a fast response and effective mitigation measures. We are aware that the implemented device alone cannot substitute laboratory analysis, but it might be a valuable tool to help increasing the sampling spatial coverage, particularly in coastal areas, and providing reliable and efficient data for long-term water monitoring.

Author contributions

The manuscript was written through contributions of all authors. All authors have given approval for the final version of the manuscript.

CRediT authorship contribution statement

G. Minas: Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Formal analysis. **José A. Afonso:** Writing – review & editing. **Paulo Sousa:** Writing – original draft, Investigation, Conceptualization. **Emilio Fernandez:** Writing – review & editing, Validation, Supervision, Methodology, Formal analysis, Data curation. **Luís Goncalves:** Writing – review & editing, Validation, Formal analysis. **Vitor Magalhães:** Writing – original draft, Methodology, Investigation, Data curation, Conceptualization. **Vânia Pinto:** Writing – original draft, Supervision, Methodology, Investigation, Funding acquisition, Data curation, 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.

Acknowledgements

This work has been supported by the projects Eye-on-shell (EXPL/EAM-OCE/1155/2021), through National Funds (OE), within the scope of the Scientific Research and Technological Development Projects (IC&DT) program in all scientific domains (PTDC), through the

Foundation for Science and Technology, I.P. (FCT, I.P), and Next generation imaging Technologies to probe structure and function of biological specimen across scales in their natural environment (Project 101094250 – IMAGINE – Horizon-Infra 2022 Tech-01), and partially supported from the European Union’s Horizon 2020 Research and Innovation Programme via ASSEMBLE Plus under grant agreement No 730984. The authors are grateful for the support given by ECIMAT in accessing phytoplankton cultures as well as analysis equipment. Furthermore, the authors would like to thank María Huete-Ortega, Damián Costas, Tania Tato, Arantxa Martínez and Mercedes Olmedo for their help with phytoplankton analysis. Vitor Magalhães thanks the FCT for the PD/BD/150581/2020 grant. Vânia Pinto and Paulo Sousa thanks FCT for their contracts funding provided through 2021.01087. CEECIND and 2021.01086.CEECIND, respectively. The authors also acknowledge the partial financial support within the R&D Unit Project Scope: UIDB/04436/2020.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at [doi:10.1016/j.snb.2024.136819](https://doi.org/10.1016/j.snb.2024.136819).

Data availability

No data was used for the research described in the article.

References

[1] A. Zaghloul, M. Saber, S. Gadow, F. Awad, Biological indicators for pollution detection in terrestrial and aquatic ecosystems, *Bull. Natl. Res. Cent.* 44 (1) (2020) 1–11.
[2] P. Lassus, N. Chomérat, P. Hess, E. Nézan, Toxic and harmful microalgae of the World Ocean. Denmark, International Society for the Study of Harmful Algae/ Intergovernmental Oceanographic Commission of UNESCO. IOC Manuals and Guides, 68 (Bilingual English/French). Harmful Algae 63 (2016) 203.
[3] S.W. Jeffrey, S.W. Wright, M. Zapata, Microalgal classes and their signature pigments, in: *Chemotaxonomy and Applications in Oceanography*, Cambridge University Press, United Kingdom, 2011, pp. 3–77.
[4] D.F. Millie, O.M. Schofield, G.J. Kirkpatrick, G. Johnsen, T.J. Evens, Using absorbance and fluorescence spectra to discriminate microalgae, *Eur. J. Phycol.* 37 (3) (2002) 313–322.
[5] S. Promdaen, P. Wattuya, N. Sanevas, Automated microalgae image classification, *Procedia Comput. Sci.* 29 (2014) 1981–1992.

- [6] N. Hashemi, J.S. Erickson, J.P. Golden, F.S. Ligler, Optofluidic characterization of marine algae using a micro-flow cytometer, *Biomicrofluidics* 5 (3) (2011) 032009.
- [7] S.E. Zieger, S. Seoane, A. Laza-Martínez, A. Knaus, G. Mistlberger, I. Klimant, Spectral characterization of eight marine phytoplankton phyla and assessing a pigment-based taxonomic discriminant analysis for the in situ classification of phytoplankton blooms, *Environ. Sci. Technol.* 52 (24) (2018) 14266–14274.
- [8] H.L. MacIntyre, E. Lawrenz, T.L. Richardson, Taxonomic discrimination of phytoplankton by spectral fluorescence, *Chlorophyll A Fluoresc. Aquat. Sci.: Methods Appl.* (2010) 129–169.
- [9] G. Persichetti, E. Viaggiu, G. Testa, R. Congestri, R. Bernini, Spectral discrimination of planktonic cyanobacteria and microalgae based on deep UV fluorescence, *Sens. Actuators B: Chem.* 284 (2019) 228–235.
- [10] P. Alexander, G. Evgeniy, V. Sergey, M. Zhanna, O. Tatiana, The study of fluorescence features of microalgae from the genus *Pseudo-nitzschia* and the possibility of their detection in water, *Algal. Research* 64 (2022) 102662.
- [11] N.J. Poulton, FlowCam: quantification and classification of phytoplankton by imaging flow cytometry, *Imaging Flow. Cytom.: Methods Protoc.* (2016) 237–247.
- [12] R.J. Olson, A. Shalapyonok, D.J. Kalb, S.W. Graves, H.M. Sosik, Imaging FlowCytobot modified for high throughput by in-line acoustic focusing of sample particles, *Limnol. Oceanogr.: Methods* 15 (10) (2017) 867–874.
- [13] G.B.J. Dubelaar, P.L. Gerritzen, CytoBuoy: a step forward towards using flow cytometry in operational oceanography, *Sci. Mar.* 64 (2) (2000) 255–265.
- [14] S. Schultes, R.M. Lopes, Laser Optical Plankton Counter and Zooscan intercomparison in tropical and subtropical marine ecosystems, *Limnol. Oceanogr.: Methods* 7 (11) (2009) 771–784.
- [15] L. Shen, H. Xu, X. Guo, Satellite remote sensing of harmful algal blooms (HABs) and a potential synthesized framework, *Sensors* 12 (6) (2012) 7778–7803.
- [16] V. Asnaghi, D. Pecorino, E. Ottaviani, A. Pedroncini, R.M. Bertolotto, M. Chiantore, A novel application of an adaptable modeling approach to the management of toxic microalgal bloom events in coastal areas, *Harmful Algae* 63 (2017) 184–192.
- [17] I.C. Robbins, G.J. Kirkpatrick, S.M. Blackwell, J. Hillier, C.A. Knight, M.A. Moline, Improved monitoring of HABs using autonomous underwater vehicles (AUV), *Harmful Algae* 5 (6) (2006) 749–761.
- [18] D. Wu, R. Li, F. Zhang, J. Liu, A review on drone-based harmful algae blooms monitoring, *Environ. Monit. Assess.* 191 (2019) 1–11.
- [19] M. Izadi, M. Sultan, R.E. Kadiri, A. Ghannadi, K. Abdelmohsen, A remote sensing and machine learning-based approach to forecast the onset of harmful algal bloom, *Remote Sens.* 13 (19) (2021) 3863.
- [20] A. Mozo, J. Morón-López, S. Vakuruk, Á.G. Pompa-Pernía, Á. González-Prieto, J.A. P. Aguilar, J.M. Ortiz, Chlorophyll soft-sensor based on machine learning models for algal bloom predictions, *Sci. Rep.* 12 (1) (2022) 13529.
- [21] P. Otálora, J.L. Guzmán, F.G. Acien, M. Berenguel, A. Reul, Microalgae classification based on machine learning techniques, *Algal Res.* 55 (2021) 102256.
- [22] R. Reimann, B. Zeng, M. Jakopek, M. Burdukiewicz, I. Petrick, P. Schierack, S. Rödriger, Classification of dead and living microalgae *Chlorella vulgaris* by bioimage informatics and machine learning, *Algal Res.* 48 (2020) 101908.
- [23] J. Wen, J. Yang, Y. Li, L. Gao, Harmful algal bloom warning based on machine learning in maritime site monitor-ing, *Knowl.-Based Syst.* 245 (2022) 108569.
- [24] Y.S. Kwon, Monitoring coastal chlorophyll-a concentrations in coastal areas using machine learning models, *Water* 10 (8) (2018) 1020.
- [25] W. Xu, J. Niu, W. Gan, S. Gou, S. Zhang, H. Qiu, T. Jiang, Identification of paralytic shellfish toxin-producing microalgae using machine learning and deep learning methods, *J. Oceanol. Limnol.* (2022) 1–16.
- [26] S. Shan, L. Xu, K. Chen, M. Tong, X. Wang, A rapid fluorescence approach on differentiation of typical dinoflagellate of East China Sea, *Spectrochim. Acta Part A: Mol. Biomol. Spectrosc.* 276 (2022) 121216.
- [27] S.E. Zieger, G. Mistlberger, L. Troi, A. Lang, F. Confalonieri, I. Klimant, Compact and low-cost fluorescence based flow-through analyzer for early-stage classification of potentially toxic algae and in situ semiquantification, *Environ. Sci. Technol.* 52 (13) (2018) 7399–7408.
- [28] Y.-H. Shin, J.Z. Barnett, E. Song, M.T. Gutierrez-Wing, K.A. Rusch, J.-W. Choi, A portable fluorescent sensor for on-site detection of microalgae, *Microelectron. Eng.* 144 (2015) 6–11, <https://doi.org/10.1016/j.mee.2015.01.005>.
- [29] L. Orozco, Synchronous detectors facilitate precision, low-level measurements, *Analog Dialog-* 48 (11) (2014) 1–5.
- [30] Sá, L.R., Pinto, V.C., Sousa, P.J., Magalhães, V.H., Carvalho, D.A., Afonso, J.A., ... & Gonçalves, L.M. (2023, June). Portable and Low-Cost Fluorometer for Phytoplankton Monitoring. In 2023 IEEE 7th Portuguese Meeting on Bioengineering (ENBENG) (pp. 96–99). IEEE.
- [31] Brownlee, J. (2020, August 14). A gentle introduction to the gradient boosting algorithm for machine learning. Retrieved from MachineLearningMastery: <https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/>.
- [32] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, *J. Comput. Syst. Sci.* 55 (1) (1997) 119–139.
- [33] Li, M., & Baxter, J. (1999). Boosting algorithms as gradient descent in function space. *NIPS: New Orleans, LA, USA*.
- [34] Y.-H. Shin, J.Z. Barnett, M.T. Gutierrez-Wing, K.A. Rusch, J.-W. Choi, A hand-held fluorescent sensor platform for selectively estimating green algae and cyanobacteria biomass, *Sens. Actuators B Chem.* 262 (2018) 938–946, <https://doi.org/10.1016/j.snb.2018.02.045>.
- [35] J.L. Deglint, C. Jin, A. Wong, Investigating the automatic classification of algae using the spectral and morphological characteristics via deep residual learning, in: F.C. Karray (Ed.), *Image Analysis and Recognition. ICIAR (2019) Lecture Notes in Computer Science*, 11663, Springer, Cham, 2019 <https://doi.org/10.1007>.
- [36] V. Magalhães, V. Pinto, P. Sousa, L. Gonçalves, E. Fernández, G. Minas, Spiral inertial microfluidics for separation and concentration of phytoplankton, *Algal Res.* 103317 (2023).
- [37] S. Groom, S. Sathyendranath, Y. Ban, S. Bernard, R. Brewin, V. Brotas, M. Wang, Satellite ocean colour: current status and future perspective, *Frontiers in Marine Science* 6 (2019) 485.
- [38] F. Chai, K.S. Johnson, H. Claustre, X. Xing, Y. Wang, E. Boss, A. Sutton, Monitoring ocean biogeochemistry with autonomous platforms, *Nat. Rev. Earth Environ.* 1 (6) (2020) 315–326.
- [39] W. Guan, M. Bao, X. Lou, Z. Zhou, K. Yin, Monitoring, modeling and projection of harmful algal blooms in China, *Harmful Algae* 111 (2022) 102164.
- [40] A. Bracher, H.A. Bouman, R.J. Brewin, A. Bricaud, V. Brotas, A.M. Ciotti, A. Wolanin, Obtaining phytoplankton diversity from ocean color: a scientific roadmap for future development, *Front. Mar. Sci.* 4 (2017) 55.

Vitor Magalhães holds a degree in Bioengineering and completed his Master's degree in Micro/Nano Technologies at the University of Minho in 2019. In 2020, he served as an invited lecturer at the University of Minho, focusing on modelling and simulations for microsystems. Currently, he is pursuing a doctorate in Marine Science, Technology, and Management. His research interests include the development of microfluidic systems, optical detection, numerical simulations, and marine monitoring devices.

Vânia Pinto is graduated in Biotechnology Engineering with master's in Micro/Nano Technologies. In 2017, she received the PhD degree in Biomedical Engineering from the University of Minho, Portugal. In 2021, she was awarded with a FCT competitive 6-year junior research contract at CMEMS and is an Invited Assistant Professor, lecturing in the areas of biosensors, microtechnologies and bionanotechnologies. Her current research interests include the development of biomedical microsystems, focused on microfluidics, microfabrication, immunosensors and lab-on-chip (LOC) devices.

Paulo Sousa is graduated in Biotechnology Engineering with master's in Micro/Nano Technologies. In 2017, he received the PhD degree in Biomedical Engineering from the University of Minho, Portugal. He was awarded with a Junior Researcher Contract - Stimulus of Scientific Employment, Individual Support (CEECIND), reference 2021.01086. CEECIND - 4th Edition (2021), funded by Fundação para a Ciência e Tecnologia (FCT). His current research interests include the development of sensors and actuators for organ-on-a-chip devices. His main research areas are on biomedical microsystems, focused on lab-on-a-chip devices, microfluidics, microfabrication techniques, biosensors, immunosensors, micro-pressure sensors, optical detection and organ-on-a-chip.

José A. Afonso graduated in Electronic Engineering from Federal University of Rio de Janeiro (UFRJ), Brazil, in 1993. He received the MSc degree in Informatics, in 1997, and the PhD degree in Industrial Electronics from University of Minho, Portugal, in 2005. Currently, he is an Assistant Professor in the Department of Industrial Electronics of University of Minho and a researcher at CMEMS (Center for MicroElectroMechanical Systems) and the LABBELS Associate Laboratory. His main research interests are in the areas of wireless sensor networks, Internet of Things, body area networks, mobile phone sensing, quality of service, and wireless communication and protocols.

L.M. Goncalves graduated in 1993 and received his MSc degree in 1999, both in Industrial Electronics Engineering, University of Minho, Portugal. His PhD thesis was on thermoelectric microsystems for on-chip cooling and energy harvesting, Cardiff/Minho (2007). From 1993–2002 he researched on embedded systems and electronics and since 2002, he has been lecturing at Electronics Department, University of Minho. He has experience in microelectronics, micro-fabrication, thin-film deposition, characterization and patterning, particularly in thermoelectric materials and micro-devices for energy harvesting and sensors. Currently, he is an Associated Professor in the Department of Industrial Electronics of University of Minho and a researcher at CMEMS (Center for Micro-ElectroMechanical Systems) and at the LABBELS Associate Laboratory. His current interests focuses on microelectronic sensors for oceanic applications.

Emilio Fernández received a PhD in Biology from the University of Oviedo in 1990. He was a post-doctoral research fellow at Plymouth Marine Laboratory (UK) and the Rosenstiel School for Marine and Atmospheric Sciences of the University of Miami and at the Institute for Marine Research of the Spanish Research Council (CSIC). He took a position as lecturer of Ecology at the University of Vigo in 1995. Since 2011, he is full professor of Ecology at the same university. His general research interest is the assessment of the functional response of ecosystems to global change. In particular, he is focused on studying biologically-mediated fluxes of matter and energy in planktonic marine systems in relation to environmental changes. He is the coordinator of the Biological Oceanography Group at the University of Vigo.

Graça Minas received the Ph.D. degree in 2004 from University of Minho, Portugal. Her thesis work was in cooperation with the Laboratory for Electronic Instrumentation, Delft University of Technology, The Netherlands, and dealt with lab-on-a-chips for biological fluids analysis. Presently, she is an Associate Professor with Habilitation at the Department of Industrial Electronics and a researcher at CMEMS (Center for MicroElectroMechanical Systems) and at the LABBELS Associate Laboratory, both at University of Minho, where she is involved in lab-on-a-chip devices research. Her research is basically focussed on microfluidics and on-chip integration of electronic circuits, optical filters, (bio)sensors and microactuators.