Enhancing chlorophyll-a predictions using optimal machine learning models and field spectral reflectance

Allam, Mona; Zhang, Lifu; Sun, Xuejian; Kisi, Ozgur; Kushwaha, Nand Lal; Yongxin, Liu; Menenti, Massimo; Heddam, Salim

DOI
10.1007/s12145-024-01658-z

Publication date
2025

Document Version
Final published version

Published in
Earth Science Informatics

Citation (APA)
Allam, M., Zhang, L., Sun, X., Kisi, O., Kushwaha, N. L., Yongxin, L., Menenti, M., & Heddam, S. (2025). Enhancing chlorophyll-a predictions using optimal machine learning models and field spectral reflectance. *Earth Science Informatics*, *18*(2), Article 384. https://doi.org/10.1007/s12145-024-01658-z

Earth Science Informatics (2025) 18:384
https://doi.org/10.1007/s12145-024-01658-z

RESEARCH

Enhancing chlorophyll-a predictions using optimal machine learning models and field spectral reflectance

Mona Allam[1,2] · Lifu Zhang[1] · Xuejian Sun[1] · Ozgur Kisi[3,4,5] · Nand Lal Kushwaha[6,7] · Liu Yongxin[1] · Massimo Menenti[1,8] · Salim Heddam[9]

Received: 20 September 2024 / Accepted: 8 December 2024

Abstract

This study evaluates the effectiveness of hyperspectral data to retrieve chlorophyll a (Chl-a) concentrations using various Machine Learning (ML) methods, specifically to determine whether spectral reflectance can provide accurate estimations of Chl-a. The study aims to address the gap in understanding how hyperspectral measurements correlate with Chl-a concentrations and to explore the potential for improving water quality assessment by accurately estimating Chl-a concentrations, which is essential for environmental monitoring, especially in aquatic ecosystems. The method proposed is evaluated using different Chl-a concentrations defined by the experiment design using Rhodamine B. The main reason for preparing pre-defined solutions of Chl-a is to verify the sensitivity of spectral measurements to Chl-a concentrations. In this paper, we aim to measure the pure signature of the Chl-a in which spectral reflectance of each Chl-a concentration is measured with 10 replicates by the spectrometer HS-1000WFL3. Six ML methods were investigated; (i) the multilayer perceptron artificial neural network (MLPNN), (ii) the support vector regression (SVR), (iii) the random forest regression (RFR), (iv) the Gaussian process regression (GPR), (v) Relevance Vector Machine (RVM) and (vi) Extreme Gradient Boosting (XGboost). 70 % of the data is used in training the models and 30 % of the data was used for their validation. We applied two bands 446 nm and 595 nm that are highly correlated with Chl-a. The models are evaluated using coefficient of determination (R2), Nash-Sutcliffe efficiency (NSE), root-mean-square error (RMSE), and mean absolute error (MAE). The results for the input variable, band 595 nm achieved the best predictive accuracy using the MLPNN method with R2, NSE, RMSE and MAE of approximately ≈0.859, ≈0.853, ≈26.722 and ≈19.05, respectively. The research also aims to lay the groundwork for future studies in water quality monitoring and management, using hyperspectral data and ML to improve our understanding of aquatic environments.

Keywords Chl-a · MLPNN · SVR · RFR · GPR · XGboost

Communicated by: H. Babaie

✉ Lifu Zhang
zhanglifu@aircas.ac.cn

[1] Department of National Engineering Research Center for Satellite Remote Sensing Applications, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China

[2] Environment Climate Changes Research Institute, National Water Research Centre, El Qanater EI Khairiya 13621/5, Egypt

[3] Department of Civil Engineering, Luebeck University of Applied Sciences, 23562 Lübeck, Germany

[4] Department of Civil Engineering, Ilia State University, 0162 Tbilisi, Georgia

[5] School of Civil, Environmental and Architectural Engineering, Korea University, Seoul 02841, South Korea

[6] Department of Soil and Water Engineering, Punjab Agricultural University, Ludhiana, Punjab 141004, India

[7] ICAR-Indian Agricultural Research Institute, Pusa Campus, New Delhi 110012, India

[8] Geosciences and Remote Sensing Department, Delft University of Technology, Stevinweg 12628 CN, Delft, the Netherlands

[9] Faculty of Science, Agronomy Department, Hydraulics Division, University 20 Août 1955 Skikda, Route El Hadaik, BP 26, Skikda, Algeria

Published online: 28 April 2025

Introduction

Chlorophyll-a, a photosynthetic pigment found in algae and plants, plays a pivotal role in aquatic ecosystems (Shin et al. 2020). Its concentration in water bodies serves as a key indicator of water quality, ecosystem health, and the potential for harmful algal blooms. Monitoring chlorophyll-a levels is vital for understanding and managing water resources, as it directly correlates with the overall state of aquatic environments, nutrient levels, and the presence of pollutants (Poddar et al. 2019; Hang et al. 2022).. Estimation of Chl-a in aquatic systems presents several challenges, which traditional methods struggle to address efficiently. Firstly, the manual collection and processing of samples are time-consuming and expensive, making it difficult to achieve high spatial and temporal resolution in monitoring. Moreover, these traditional methods are often constrained to specific points of measurement, limiting their capacity to capture the dynamic and spatially heterogeneous nature of Chl-a distributions within water bodies. Furthermore, water quality assessment demands the monitoring of several other environmental variables, such as turbidity, temperature, and dissolved organic matter (Kushwaha et al. 2023).

The past few decades have witnessed remarkable advancements in remote sensing technology, particularly the application of field spectral reflectance, to overcome these challenges and act as promising tools for cost-effective, rapid, and large-scale Chl-a concentration estimation (Silveira Kupssinskü et al. 2020; Shin et al. 2020; Hang et al. 2022). Spectrometers, such as the ones mounted on satellites, drones and other platforms, offer high spectral and spatial resolution, allowing for detailed and widespread data collection. These spectrometers measure the reflected sunlight in various wavelength ranges, providing spectral signatures that contain valuable information about water quality parameters, including Chl-a (Silveira Kupssinskü et al. 2020).

The spectroradiometers on satellite for chlorophyll a monitoring commonly belongs to passive remote sensing. These instruments use sunlight as light source and is easily affected by atmospheric interference. The spectrophotometer in the laboratory uses internal Xenon light, with the wavelength signal from 200 – 1000nm. It has the capability to capture spectral features in the ultraviolet band compared to the spectroradiometers on satellites. The light field conditions are also different. The spectroradiometers on satellite is easily affected by solar condition and atmospheric interference. The light field of lab spectroradiometers is much more stable and controllable. It enables better quality water spectra. In situ hyperspectral monitoring offers a transformative approach to water quality assessment by addressing the limitations of traditional laboratory methods. Once established, it can significantly reduce the need for frequent, labor-intensive testing across multiple locations. By enabling direct field data collection, hyperspectral systems allow researchers and environmental managers to minimize the time and costs associated with laboratory analyses. High-resolution spectral data on chlorophyll-a (Chl-a) provides the foundation for developing more accurate predictive models for ecosystem health and water quality (Keller et al. 2018). Hyperspectral measurements enable continuous monitoring across extensive areas and over time, capturing dynamic changes in chlorophyll-a concentrations that laboratory methods, limited by discrete sampling, might miss. While initial setup costs for hyperspectral equipment can be high, the ability to monitor large areas without extensive fieldwork can reduce overall expenses compared to the recurring costs of laboratory analyses (Pandey et al. 2024). Combining hyperspectral data with machine learning algorithms enhances the accuracy of Chl-a estimation. This integration allows for more effective monitoring and management of water quality (An et al. 2020) . Pahlevan et al. (2022) applied Mixture Density Networks (MDNs) to the inverse problem of simultaneously retrieving water quality indicators, including chlorophyll-a (Chla) use in situ measurements to train and optimize the developed models for the relevant spectral measurements (400–800 nm) of the Operational Land Imager (OLI), MultiSpectral Instrument (MSI), and Ocean and Land Color Instrument (OLCI) aboard the Landsat-8, Sentinel-2, and Sentinel-3 . Cao et al. (2020) employed a machine learning approach termed the extreme gradient boosting tree (BST) to develop an algorithm for Chl*a* estimation from OLI in turbid lakes. Kolluru and Tiwari (2022) proposed a novel approach to derive Chl-a by using multi-layer perceptron Neural Network (MLPNN) with Resilient backpropagation method based on the four ocean color bands existent in most of the ocean color sensors. Hu et al. (2021) develop a machine learning approach to reduce the impact of spectral noise and improve algorithm performance at the global scale for multiple satellite sensors.

Over the past five years, a novel spectrometer product capable of buoyancy on water has been developed (Aguzzi et al. 2020). An instance of a buoy spectrometer, the HS-VN1000WF3, has been developed by Tianjin Progoo Information Technology Co., Ltd. in China. This spectrometer is capable of gathering spectral reflectance data at a stationary location on a river and also in the Laboratories. Gathering water samples near the buoy spectrometer can help to develop a more precise model of water quality parameters, by identify the highly correlated spectra with the measured water parameters , then build relationship between the water parameters and the measured spectra (Zhang et al. 2022).

Machine learning (ML) models have become increasingly important (Singhal et al. 2019; Chusnah and Chu 2022). The integration of machine learning (ML) algorithms with field spectral reflectance data has garnered considerable attention in the domain of Chl-a estimation. ML, a subfield of artificial intelligence, has demonstrated its potential to enhance predictive accuracy and reduce data-processing time. ML algorithms are designed to process vast datasets and recognize complex patterns within the data, making them well-suited for extracting meaningful information from spectral reflectance measurements (Singhal et al. 2019). Machine learning techniques, such as regression, Random forest (RF), support vector regression (SVR), and artificial neural networks (ANN), have demonstrated their potential in modeling the relationship between spectral data and Chl-a concentrations. Park et al., (2015) demonstrated the application of ANN and support vector machine (SVM) machine learning approach for the estimation of Chl-a concentrations in the Juam Reservoir and Yeongsan Reservoir Korea. (He et al. 2020) examined eight distinct machine learning techniques, which encompassed SVR, ANN, gradient boosting machine, Random forests (RF), standard (ocean chl-a three-band algorithm for MODIS (moderate-resolution imaging spectroradiometer)) CI-OC3M, multiple linear regression (MLR), generalized additive regression, and principal component regression to estimate the distribution of Chl-a within the Gulf of St. Lawrence, located in Canada. The results of this analysis indicated that among these techniques, SVR demonstrated the most favorable performance in accurately estimating Chl-a concentrations. (Shin et al. 2020) explored the Recurrent Neural Network (RNN) model showed superior performance over SVR, Bagging, RF, Extreme Gradient Boosting (XGBoost), and Long–Short-Term Memory (LSTM) in the prediction of Chl-a concentrations in the Nakdong River, Korea.The recent literature reviewed here clearly indicates the growing interest and advancements in the integration of spectral data and machine learning for chlorophyll-a estimation. These studies underscore the potential of ML algorithms to enhance the accuracy of Chl-a predictions and offer cost-effective, large-scale monitoring solutions.

The cooperative use of field spectral reflectance data in conjunction with sophisticated machine learning (ML) models is used to achieve increased accuracy, promptness, and economic efficiency in forecasting abilities. The implications of this paper go beyond academic research, which is helpful to environmental regulatory agencies, water resource management authorities, and academics interested in protecting and improving water quality. In this paper, we investigate how state-of-the-art ML algorithms combine with field spectral reflectance data improve predictions of Chl-a. Our study investigates the potential of machine learning models to estimate Chl-a concentration using field spectral reflectance data, a relatively novel approach in environmental monitoring. We test the hypothesis that these models can yield accurate estimations by evaluating various machine learning algorithms for their ability to predict Chl-a concentration. The article offers a thorough data analysis and assessing different machine-learning techniques using hyperspectral data measurement. This detailed examination underscores the scientific rigor and systematic methodology of the study. Additionally, the article highlights the addressed issues and the significance of the research, emphasizing its potential impact on hyperspectral data and ML for Chl_a prediction. This work advances environmental modeling by demonstrating the feasibility of predicting Chl-a concentration, paving the way for future studies to explore innovative water quality assessment and management methods.

The aims of the study are to evaluate the effectiveness of hyperspectral data to retrieve chlorophyll a (Chl-a) concentrations using various Machine Learning (ML) methods, specifically to determine whether spectral reflectance can provide accurate estimations of Chl-a. The research aims to address the gap in understanding how hyperspectral measurements correlate with Chl-a concentrations and to explore the potential for improving water quality assessment by accurately estimating Chl-a levels.

Methodology

Determination of chlorophyll-a concentrations

In our study the Chl-a concentrations were prepared in the laboratory of Tianjin Progoo information technology Co by using Rhodamine B 2.5% as indicator to Chl-a concentrations (Koli and Sharma 2021), according to (Luck et al. 2012). Rhodamine B is a fluorescent dye that can absorb and re-emit light in a way that enhances the detection of chlorophyll a. In spectrophotometric or fluorometric studies, Rhodamine B can assist in calibrating the detection system for chlorophyll a, helping to achieve accurate quantitative analysis in water quality and environmental studies. A recent study calibrated in-situ Chla fluorescence data, this study use Rhodamine B as indicator to Chla, the results of this study showed that ,Fluorescence in-situ sensors are particularly useful to detect and quantify sudden phytoplankton biomass variations through high frequency measurements (Levi et al. 2024).The Chl-a concentrations were contained in a black shade bucket to avoid the interference of environmental stray light during the spectra measurement. The fluorescence signal of 0.625 mg/L Rhodamine B corresponds to the 66.0 µg/L chlorophyll-a in the 22℃ water, which can be used to construct relationship between Rhodamine B and Chl-a. The fluorescence of Rhodamine B is influenced by the

water temperature. So, the water temperature was measured after the spectrum was measured. The influence of the water temperature was corrected according to Xylem Inc. (2020). The prepared concentration chlorophyll-a standard solution ranged from 11- 63 µg/L. Flowchart of the approach used for Chl-a prediction Fig. 1.

Spectra measurement

Progoo information technology Co has developed a buoy spectrometer water quality detection system that has the capability to monitor various water quality parameters and also can measure the spectra of the water parameters. The

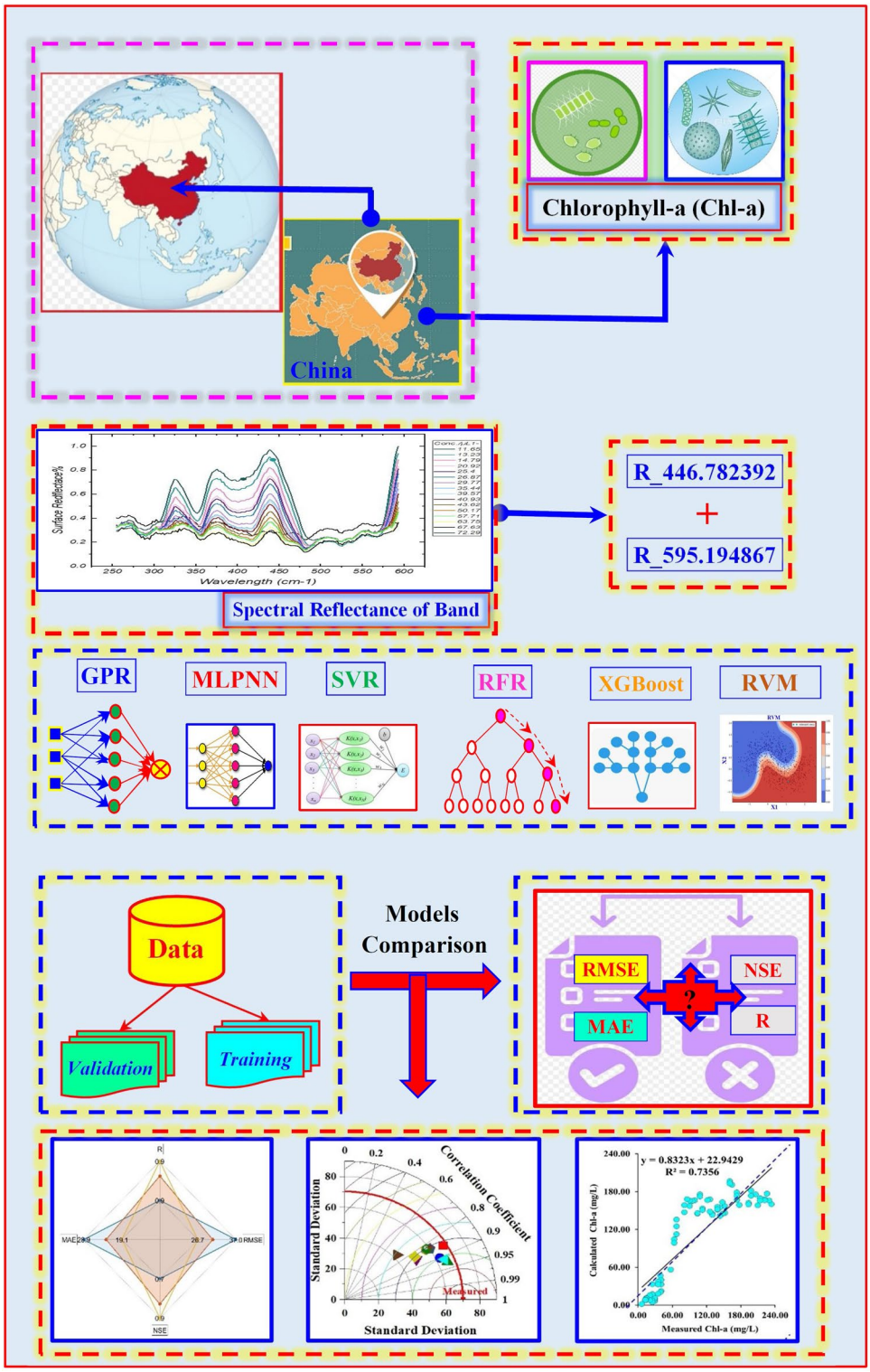

Fig. 1 Flowchart of the approach used for Chl-a prediction

spectral range of buoy spectrometer is from 400 nm to 1000 nm. In our study, we use the second version of the buoy spectrometer to measure the spectra of the prepared Chl-a concentrations in the progoo laboratory (Zhang et al. 2022). The spectra measurement of chlorophyll-a was divided into the following steps. First, the probe of the spectrometer was cleaned with a soft bristle brush and then rinsed with deionized water. Second, the spectrometer was calibrated using deionized water then reference spectra of pure water was measured. To examine the effectiveness of the calibration, we measured the spectra of pure water again, the spectra of both water samples supposed to be same. If the mean coefficient of variation of the measured spectra is less than 5%, the calibration is qualified. Third, the prepared concentrations were measured by the calibrated spectrometer. Each concentration measurement was repeated 3 times, with a total of 15 spectra obtained for one sample. The absorption coefficient at chlorophyll-specific wavelengths (e.g., 440 nm and 675 nm) are used as indicators to retrieve the chlorophyll a concentration. Algorithms are applied to absorption data to relate spectral features to chlorophyll a concentrations, e.g.,: Empirical Models: use band ratios to estimate concentrations. Semi-Analytical Models: Incorporate the specific absorption coefficients of chlorophyll a.

Spectral characteristics of chlorophyll-a

Figure 2 shows the surface reflectance at different Chl-a concentrations. Only wavelengths between 300 and 600 nm are shown on the x-axis to reduce interference from instrument noise. Surface reflectance is shown via the y-axis. Every labeled or color-coded data point represents a particular concentration of chlorophyll. The spectra reflect correctly well-known absorption features of Chl-a. According to the overall tendency, Chl-a concentrations often have increased spectral reflectance at specific wavelengths. For example, a concentration of 11.65 ppm of chlorophyll shows comparatively greater reflectance at 324 nm. Likewise, 13.23 and 14.79 concentrations exhibit greater reflectance at 373 nm and 438 nm. These findings align with previous studies on chlorophyll and how it interacts with light. According to (Markwell et al. 1995), the blue and red portions of the spectrum are where chlorophyll absorbs light the strongest, with absorption maxima occurring at 430 nm and 665 nm. Since the absorption range includes 324 nm, 373 nm, and 438 nm, this explains why the reflectance at these wavelengths is greater. The low reflectance seen in the 488–567 nm region, which corresponds to the chlorophyll absorption peak at 665 nm, can be explained by the fact that reflectance is higher at wavelengths where absorption is lower (Gitelson et al. 2003). Reflectance rises for all chlorophyll concentrations above 567 nm, even if there is no discernible difference between them, because of scattering effects that occur beyond the primary absorption peaks (Gitelson et al. 2003). The summary statistic of the Chl-a concentration and the two spectral characteristics are represented in (Table 1), which showed that band 446 has the highest skewness and kurtosis (leptokurtic) while the Chl-a has negative kurtosis higher than one and it is platykurtic. In the table, we provide the mean, maximal,

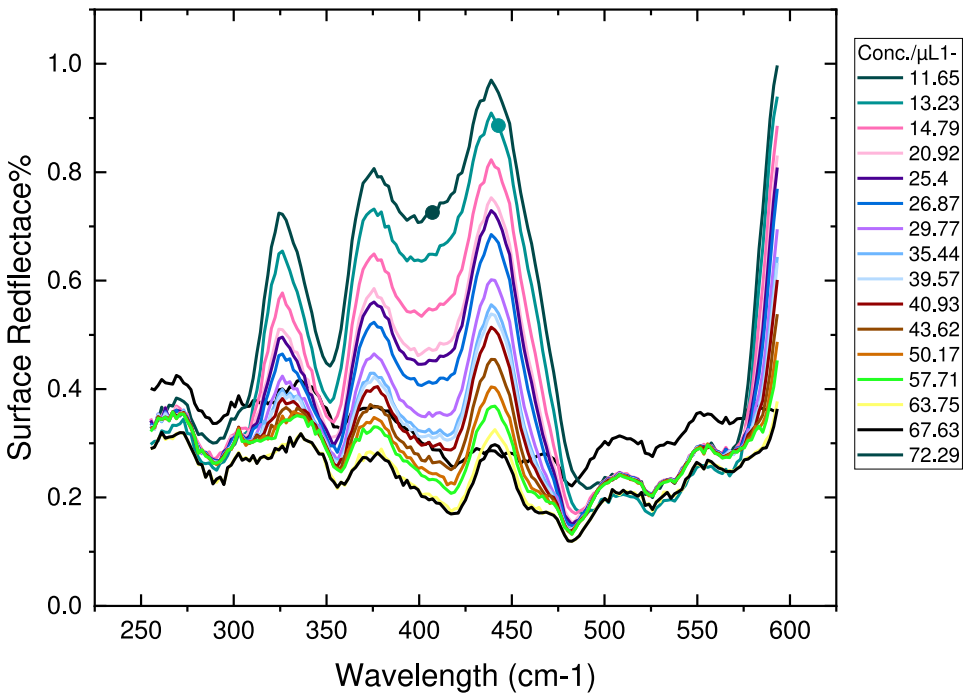

Fig.2 Surface spectral reflectance at different Chl-a concentration

Table 1 Summary statistics of raw water quality variables and Chl-a

Variables	Subset	Unit	X_{mean}	X_{max}	X_{min}	S_x	C_v	R	Skweness	Kurtosis
Chl-a	Training	mg/L	115.788	234	6.84	72.33	0.625	1	−0.061	−1.314
	Validation	mg/L	108.156	234	6.84	70.12	0.648	1		
	All data	mg/L	113.502	234	6.84	71.36	0.629	1		
Band_446	Training	//	0.352	1.109	0.166	0.244	0.693	−0.77	1.931	3.314
	Validation	//	0.365	1.097	0.162	0.243	0.665	0.777		
	All data	//	0.356	1.109	0.162	0.24	0.674	0.772		
Band_595	Training	//	0.496	1.199	0.252	0.27	0.545	0.803	1.483	1.03
	validation		0.51	1.18	0.254	0.277	0.543	0.798		
	All data		0.5	1.199	0.252	0.269	0.538	0.801		

[Abbreviations: X_{mean} mean, X_{max} maximum, X_{min} minimum, S_x standard deviation, C_v coefficient of variation, R coefficient of correlation with Chl-a]

minimal, standard deviation, the coefficient of variation, and the correlation coefficient. More details about the statistical description of can be found in (Özbayrak et al. 2023). Before applying the six machine learning models, all data were standardized using the Z-score method. This method helps in avoiding the problem of outlier. However, according to the calculated statistical indices reported in Table 1, there is no outlier in the dataset. The machine leaning models were developed using various parameters and by adopting the trail and error for finding the best models' parameters. For example, varying the number of hidden neurons, the number of trees, the parameters of the gaussian function, and the kernel parameters of the SVR models. The number of training data was 187 and 80 for validation. In the present study, machine learning models were developed using MATLAB, and the XGBoost using Python. The procedures of calculation of chlorophyll a, step 1: prepare chlorophyll a concentration ranging from 11.65 to 72.29 µg/L , step 2: measure spectra using buoy spectroradiometer (200 - 1000 nm), step 3: correlation analysis identifies highly correlated bands with chlorophyll a : 446 nm & 595 nm. Step 5: 446 nm & 595 nm used as input to evaluate and compare results of 6 ML methods, MLPNN,SVR, RFR, iv, GPR, RVM and XGboost. Step 6: Final Result : MLPNN achieves the highest predictive Chlorophyll a accuracy as shown in Figure 3.

Machine learning models for prediction of Chl-a in aquatic environments

The machine learning models developed in the present study namely, (*i*) the multilayer perceptron artificial neural network (MLPNN), (*ii*) the support vector regression (SVR), (*iii*) the random forest regression, (*iv*) the Gaussian process regression (GPR), (*v*) the relevance vector machine (RVM) and (*vi*) the XGBoost regression models. More details about the models can be found in (Ekmekcioğlu et al. 2022; Citakoglu and Coşkun 2022; Demir and Citakoglu 2024).

Multilayer perceptron neural network (MLPNN)

Multilayer perceptron neural networks (MLPNNs) are widely adopted feedforward neural networks known for their rapid execution, ease of implementation, and modest data requirements (Ali et al. 2017; Hakim et al. 2021). The architecture of an MLPNN typically comprises three sequential layers: the input layer, the hidden layer, and the output layer Fig.4. The hidden layer plays a pivotal role in processing and transmitting input data to the output layer. However, it's crucial to strike the right balance when determining the number of neurons in the hidden layer.

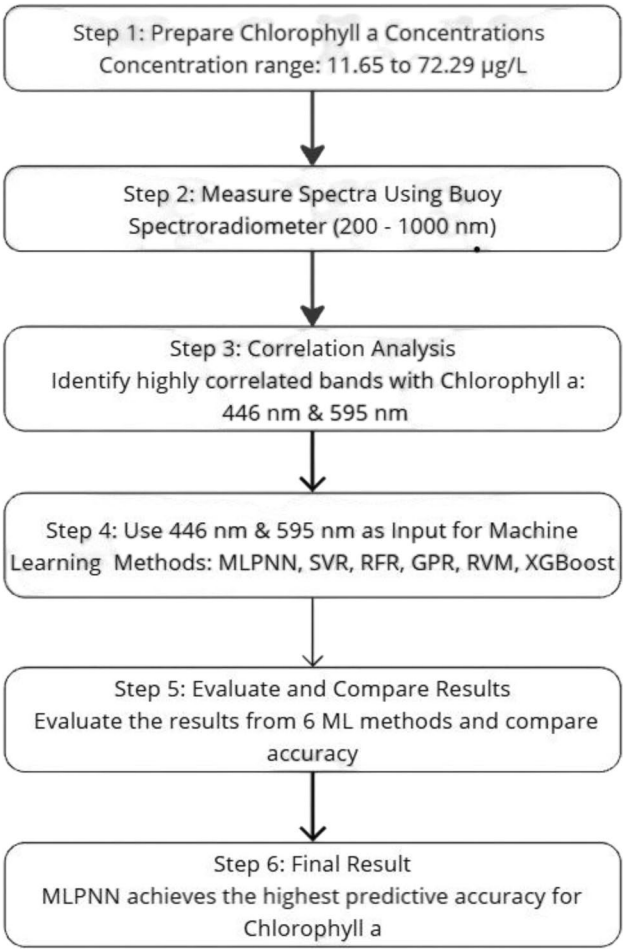

Fig.3 Procedure for calculation of chlorophyll a using spectral data and machine learning methods

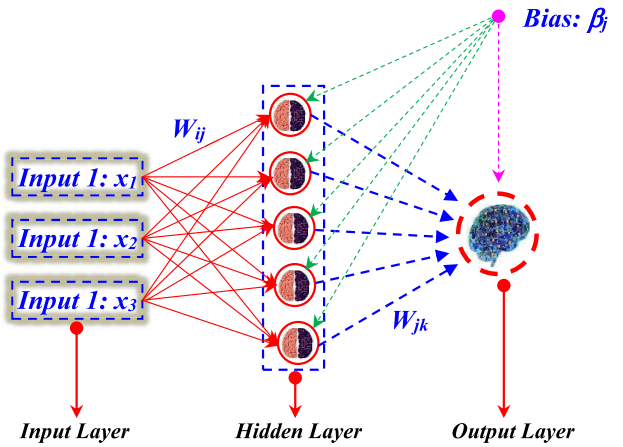

Fig.4 Schematic diagram of the multilayer perceptron neural network (MLPNN)

An insufficient or excessive number of neurons can lead to challenges in generalization and result in overfitting issues. Every neuron 'j' within the hidden layer computes the sum of its input signals 'x_i,' which are weighted by their respective connection weights 'w_{ji}.' The output of each neuron is defined as:

$$Y_j = f\left(\sum w_{ji} x_i\right) \tag{1}$$

Here, 'f' denotes an activation function that utilizes the weighted input summations. Common choices for activation functions include simple threshold functions, sigmoid functions, or hyperbolic tangent functions.

Random forest regression (RFR)

Random Forest (RFR) represents an ensemble technique introduced by Breiman (2001), widely applied across various studies. RF combines multiple decision trees in parallel Fig.5, employing a bagging (bootstrapping and aggregation) strategy. Bootstrapping involves training individual decision trees on distinct subsets of the input training data, effectively reducing model variance and yielding precise outcomes. The final decision is made by aggregating the decisions of individual trees, leading to improved generalization (Misra and Li 2020). The generalization accuracy of a random forest hinges on the quality of individual trees and the extent of correlation among them. Random forest models have consistently demonstrated their robust predictive capabilities and versatility, effectively addressing classification and regression tasks, particularly in scenarios with limited sample sizes and high-dimensional data (Biau and Scornet 2015; Kushwaha et al. 2024b).

Support vector regression (SVR)

Support Vector Machines (SVM) have been a significant advancement in the field of machine learning, with a history dating back to their introduction by (Boser et al. 1992). Since their inception, SVMs have found applications in a wide range of fields, such as machine learning, optimization, statistics, and functional analysis (Vishwakarma et al. 2023; Abd-Elaty et al. 2023). The SVM concept can be generalized to become applicable to regression problems (Kushwaha et al. 2021). The SVR, one of the applications of the SVM, finds hyperplanes that minimize the errors and maximize the margins of continuous data. A schematic diagram of SVR is displayed in Fig.6. The equation for linear SVM can be written as follows:

$$x_1, y_1 \ldots\ldots\ldots\ldots x_n, y_n \tag{2}$$

Where y_i is either 1 or −1, depending on the class to which the point is assigned. Each represents an

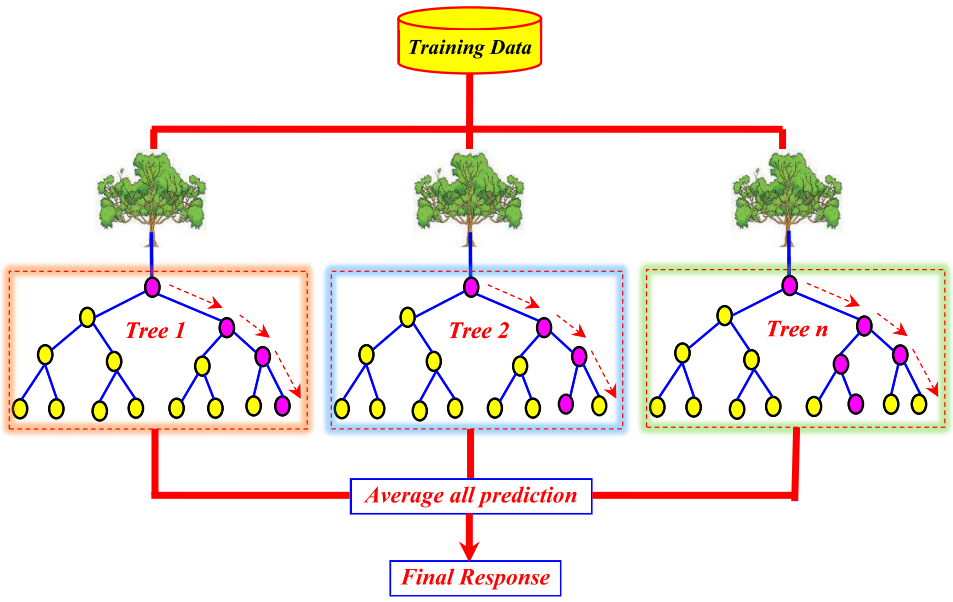

Fig.5 Schematic diagram of the random forest regression (RFR)

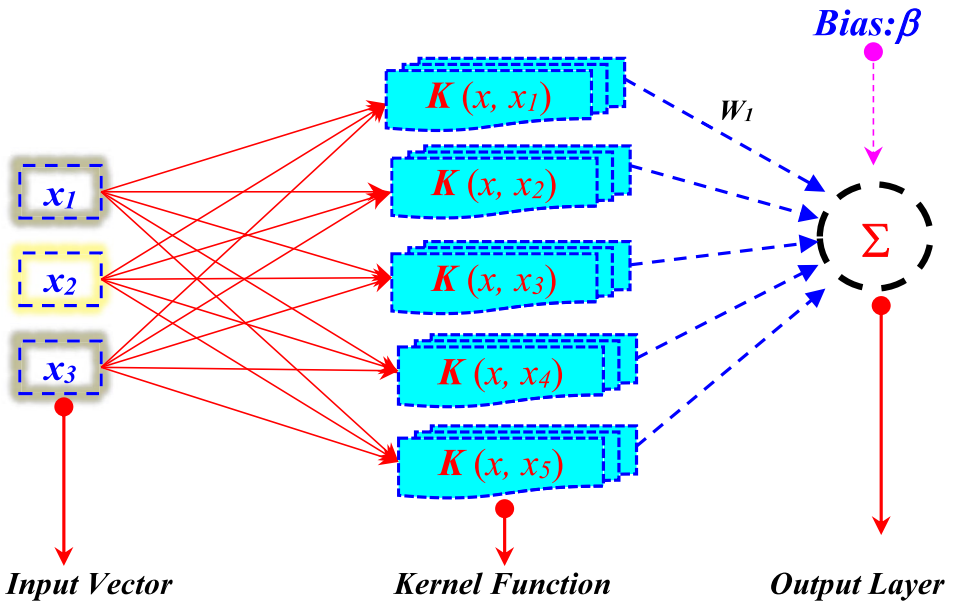

Fig. 6 Schematic diagram of the support vector regression (SVR)

n-dimensional real vector. The maximum-margin hyperplane that divides the group of points, when from the group of points when which is determined to maximize the distance among different points from either group. The hyperplane which satisfies the following equation for a set of points can be written as below:

$$w^T x - b = 0 \tag{3}$$

Where w is the normal vector to the hyperplane. The parameter denotes the offset of the hyperplane from the origin along the normal vector.

Gaussian process regression (GPR)

A Gaussian process (GP) is a stochastic process that generates a sequence $\{Xt\}t \in \tau$ over time, such that its impact on the finiteness of a linear combination of Xt (or, more generally, any linear function derived from the sample function Xt) is negligible Fig. 7. This linear combination typically follows a normal distribution (Daemi et al. 2019; Li et al. 2020). Consider a training dataset $T = \{(x_i, y_i)/i = 1, 2,\ldots, N\}$ that describes the Gaussian process's behavior. When employing Gaussian processes for regression tasks, often referred to as kriging, a fundamental

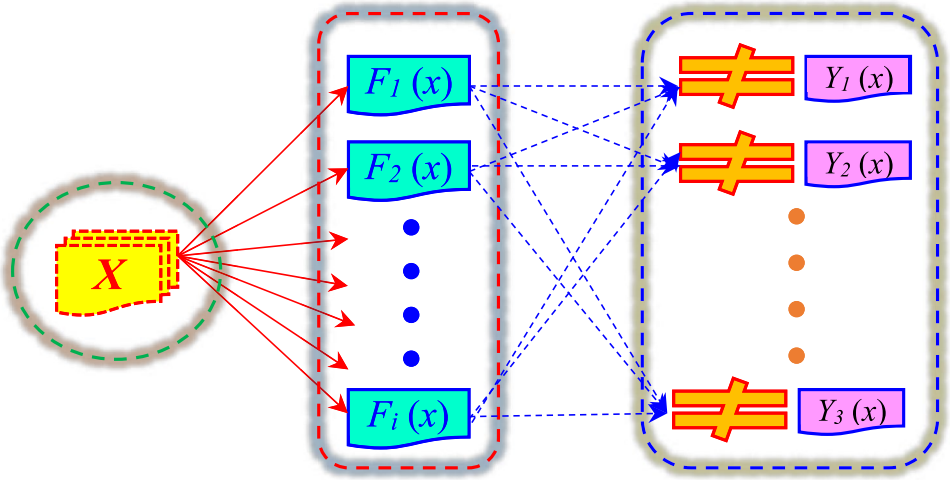

Fig. 7 Schematic diagram of the Gaussian Process Regression (SVR)

assumption is made. For a Gaussian process 'f' observed at the 'x' coordinates, the vector of values f(x) is essentially a sample from a multivariate Gaussian distribution, with the dimension matching the number of observed coordinates 'n'. Notably, Gaussian processes can be fully characterized based on their second-order statistics. Hence, assuming a Gaussian process with a zero mean, the definition of the covariance matrix 'K' (a positive definite kernel) completely determines the Gaussian process's behavior. This covariance matrix plays a pivotal role in defining essential properties of Gaussian processes, such as isotropy, stationarity, smoothness, and periodicity (T).

Relevance vector machine (RVM)

The Relevance Vector Machine (RVM) is a sophisticated machine learning method employed for predicting Chl-a concentrations in aquatic environments (Zou et al. 2020). Unlike traditional support vector machines, RVM offers a Bayesian framework that allows for probabilistic regression, enabling the model to provide uncertainty estimates along with predictions. RVM selects a subset of relevant data points, known as relevance vectors, to construct a sparse model, reducing computational complexity while maintaining predictive accuracy (Bowd et al. 2005). By effectively identifying relevant features and optimizing model parameters through Bayesian inference, RVM can accurately forecast Chl-a concentrations based on diverse environmental variables. Its ability to handle uncertainty and sparsity makes RVM a valuable tool for Chl-a prediction, contributing to improved understanding and management of water quality in aquatic ecosystems (Camps-Valls et al. 2006).

Extreme gradient boosting (XGboost)

The Extreme Gradient Boosting (XGBoost) is a powerful machine learning technique used for predicting Chl-a concentrations in aquatic ecosystems. It employs an ensemble learning framework that iteratively improves the predictive accuracy of a model by combining multiple weak learners, typically decision trees, into a strong predictive model (Ahn et al. 2023). XGBoost optimizes the model's performance by minimizing a loss function and employing gradient boosting techniques, which involve adjusting the weights of misclassified instances in subsequent iterations. This iterative approach allows XGBoost to effectively capture complex relationships and patterns in the data, making it particularly well-suited for predicting Chla concentrations based on various environmental factors and input variables. Additionally, XGBoost offers flexibility in model tuning and regularization parameters, allowing for fine-tuning to optimize predictive performance (Chen et al. 2015). Overall, XGBoost has demonstrated promising results in accurately forecasting Chla concentrations, contributing to enhanced understanding and management of water quality in aquatic ecosystems (Niroumand-Jadidi and Bovolo 2022).

Models structures

In the present study, the six machine learning models, i.e., the MLPNN, RFR, GPR, SVR, RVM and XGBoost were compared according to three different scenarios in Table 2. All models were first calibrated during the training stage and later

Table 2 The input combinations of different machine learning models

Models' configurations	Input variables	Output
Scenario 01, i.e., MLPNN1, SVR1, RFR1…	Band 446, Band 595	Chl-a
Scenario 02 i.e., MLPNN2, SVR2, RFR2…	Band 446	Chl-a
Scenario 03 i.e., MLPNN3, SVR3, RFR3…	Band 595	Chl-a

validated using the validation dataset. As we have only two input variables, and in total only three possible combinations, we have not applied linear or nonlinear techniques

Performance assessment of the models

Four performances metrics namely root-mean-square error (RMSE), mean absolute error (MAE), Nash-Sutcliffe efficiency (NSE), the determination coefficient (R2), and the mean absolute percentage error (*MAPE*) were used for models' performances evaluation (Kushwaha et al. 2022, 2024a). They are calculated as follow:

$$MAE = \frac{\sum_{i=1}^{N} |Chl_{pre,i} - Chl_{obs,i}|}{N} \quad (4)$$

$$MAPE = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{Chl_{obs,i} - Chl_{pre,i}}{Chl_{obs,i}} \right| (\%) \quad (5)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Chl_{obs,i} - Chl_{pre,i})^2}{N}} \quad (6)$$

$$NSE = 1 - \left[\frac{\sum_{i=1}^{N} (Chl_{obs,i} - Chl_{pre,i})^2}{\sum_{i=1}^{N} (Chl_{obs,i} - \overline{Chl_{obs}})^2} \right] \quad (7)$$

$$R^2 = \left(\frac{\sum_{i=1}^{N} \left(Chl_{obs,i} - \overline{Chl_{obs}} \right) \left(Chl_{pre,i} - \overline{Chl_{pre}} \right)}{\sqrt{\sum_{i=1}^{N} (Chl_{obs,i} - \overline{Chl_{obs}})^2 \sum_{i=1}^{N} (Chl_{pre,i} - \overline{Chl_{pre}})^2}} \right)^2 \quad (8)$$

$\overline{Chl}_{obs}$ and $\overline{Chl}_{pre}$ are the mean measured, and mean predicted Chl-a, respectively,Chl_{obs} and Chl_{pre} specifies the observed and predicted Chl-a, and N shows the number of data points.

Results and discussion

Models evaluation and comparison

Training and testing outcomes of the machine learning methods are compared in (Table 3) with respect to R2, NSE, RMSE and MAE criteria. In the training period, the MLPNN2 has the highest R2 and NSE accuracies, indicating that it performs the best in approximating the relationship between the input features and Chl-a concentrations, it is closely followed by MLPNN2, and MLPNN1 has the lowest R2 and NSE accuracy. In terms of RMSE and MAE, lower values are better. The MLPNN2 has the lowest RMSE, indicating that it has the smallest errors during training. The MLPNN3 follows, and the MLPNN1 has the highest RMSE and MAE, meaning it has the largest errors. Based on the provided training performances, the MLPNN2 generally outperforms the other two models

Table 3 Performances of different machine learning models for Chl-a prediction

Training						Validation				
	R^2	NSE	RMSE	MAE	MAPE	R^2	NSE	RMSE	MAE	MAPE
MLPNN1	0.766	0.762	34.990	25.91	26.40	0.736	0.718	36.989	28.91	35.20
MLPNN2	0.867	0.865	26.413	18.49	17.90	0.812	0.805	30.747	23.12	30.00
MLPNN3	0.846	0.846	28.132	17.91	15.30	0.859	0.853	26.722	19.05	19.70
SVR1	0.721	0.717	38.154	26.36	25.00	0.699	0.698	38.279	26.85	23.90
SVR2	0.706	0.705	38.991	27.95	25.20	0.687	0.686	39.071	28.17	25.90
SVR3	0.714	0.711	38.553	26.56	24.70	0.687	0.686	39.043	27.95	27.70
GPR1	0.974	0.974	11.516	6.745	6.70	0.533	0.436	52.309	36.33	53.00
GPR2	0.712	0.713	38.463	27.51	25.20	0.689	0.687	38.950	28.69	26.90
GPR3	0.716	0.716	38.256	27.23	25.30	0.684	0.683	39.219	28.54	26.80
RFR1	0.978	0.979	10.480	6.259	6.40	0.728	0.622	42.852	26.68	20.90
RFR2	0.980	0.979	10.388	5.787	6.00	0.689	0.577	45.335	28.32	22.50
RFR3	0.925	0.925	19.658	10.70	10.00	0.837	0.829	28.806	17.83	18.80
RVM1	0.832	0.692	39.852	30.126	30.00	0.668	0.666	40.282	30.769	31.40
RVM2	0.825	0.681	40.529	31.819	38.10	0.662	0.657	40.833	32.339	39.00
RVM3	0.826	0.682	40.481	31.742	37.60	0.660	0.656	40.856	32.169	37.90
XGBoost1	0.999	0.999	0.131	0.079	0.10	0.666	0.525	48.043	29.009	21.80
XGBoost2	0.999	0.999	0.371	0.204	0.20	0.655	0.507	48.907	29.906	22.40
XGBoost3	0.999	0.999	0.402	0.246	0.20	0.669	0.641	41.739	22.603	24.10

(MLPNN1and MLPNN3) in approximating Chl-a concentrations. Comparing the training performances of the SVR models reveals that the SVR1 has the highest R2 and NSE accuracy, followed closely by the SVR3, and the SVR2 has the lowest R2 and NSE accuracy. Furthermore, the SVR1 has the lowest RMSE and MAE during training, it is followed by the SVR3 and the SVR2 has the highest RMSE and MAE. The SVR1 generally outperforms the other two models (SVR2 and SVR3) in approximating Chl-a concentrations.

Comparing the training performances of three GPR models indicates that the GPR1 has the highest R2 and NSE accuracy, indicating that it performs exceptionally well in approximating the relationship between the input features (wavelengths lights) and Chl-a concentrations. The GPR2 and GPR3 have lower R2 and NSE accuracies. The GPR1 has the lowest RMSE and MAE, suggesting that it has the smallest errors during training by a wide margin. The GPR3 has a slightly lower RMSE and MAE than the GPR2, indicating that GPR3 performs slightly better in terms of RMSE and MAE. Based on the provided training performances, the GPR1 outperforms the other two models (GPR2 and GPR3) in approximating Chl-a concentration. Comparing the training performances of the RFR models shows that the RFR2 has the highest R2 and NSE accuracies closely followed by the RFR1 and the RFR3 has lower R2 and NSE accuracies. Moreover, the RFR2 has demonstrated the most outstanding training performance, boasting the highest R2 the lowest RMSE and MAE and it is closely followed by the RFR1, and the RFR3 has the highest RMSE and MAE, suggesting it has larger errors. The RFR2 and RFR1 perform exceptionally well in approximating Chl-a concentrations. The RFR2 has the highest R2 accuracy, closely followed by the RFR1 and GPR1. The MLPNN2 also performs well in terms of R2 accuracy. The SVR models have relatively lower R2 accuracy. The RFR1 and RFR2 have the highest NSE accuracies, followed by GPR1. The MLPNN and SVR models have relatively lower NSE accuracies. The RFR2 and RFR1 have the lowest RMSE values, indicating that they have the smallest errors in training. The GPR1 also performs well in terms of RMSE. SVR and MLP models have higher RMSE values. The RFR2 and RFR1 have the lowest MAE values, indicating they have the smallest absolute errors during training. The GPR1 also performs well in terms of MAE. The SVR models have the highest MAE values. The RFR2 and RFR1 tend to perform the best in approximating the Chl-a concentrations based on the provided training performance metrics, with low RMSE and MAE values and high R2 and NSE accuracies. The GPR1 also performs well in terms of RMSE and MAE, and the MLP2 performs well in terms of R2 accuracy and RMSE in the training period. The SVR models generally have lower performance compared to the other three methods.

Among the three MLPNN models, the MLPNN3 has the highest R2 and NSE accuracies on the validation set, indicating a strong correlation between predicted and actual values. The MLPNN2 follows with the second-highest R2 and NSE accuracies, and the MLPNN1 has the lowest R2 and NSE accuracies. In terms of RMSE and MAE, the MLPNN3 has the lowest RMSE and MAE on the validation set, indicating it has the smallest errors during validation. The MLPNN2 follows with the second-lowest RMSE and MAE, and the MLPNN1 has the highest RMSE and MAE. Based on the provided validation performance metrics, the MLPNN3 tends to perform the best in predicting Chl-a concentrations. It has the highest accuracy for R2 and NSE, as well as the lowest RMSE and MAE on the validation data. The MLPNN2 also performs well, especially in terms of R2 and NSE accuracy, and is followed by the MLPNN1, which has the highest RMSE and MAE, indicating higher errors in predicting Chl-a concentrations on the validation set. The SVR1 has the highest R2 and NSE accuracies on the validation set while the SVR2 and SVR3 have similar R2 and NSE accuracies but are slightly lower than the SVR1. The SVR1 has the lowest RMSE and MAE on the validation set, indicating it has the smallest errors during validation. The SVR3 follows with a slightly higher RMSE and MAE, and the SVR2 has the highest RMSE and MAE. Based on the provided validation performance metrics, the SVR1 tends to perform the best in predicting Chl-a concentrations among the three SVR models. It has the highest accuracy for R2 and NSE, as well as the lowest RMSE and MAE on the validation data. The SVR3 performs slightly better than the SVR2 in terms of RMSE and MAE, but all three SVR models have relatively similar NSE and R2 accuracy on the validation set.

GPR2 has the highest R2 and NSE accuracies on the validation set and the GPR3 closely follows with a slightly lower R2 and NSE accuracy. The GPR1 has the lowest R2 and NSE accuracies. In terms of RMSE and MAE, GPR2 has the lowest values on the validation set. The GPR3 follows with a slightly higher RMSE and MAE, and the GPR1 has the highest RMSE and MAE, suggesting it has the largest errors. In summary, based on the provided validation performance metrics, the GPR2 tends to perform the best in predicting Chl-a concentrations among the three GPR models. It has the highest accuracy for R2 and NSE, as well as the lowest RMSE and MAE on the validation data. The GPR3 also performs well, with slightly lower R2, NSE and higher RMSE, and MAE values than the GPR2. The GPR1 has the lowest R2 and NSE accuracy, as well as the highest RMSE and MAE during validation and is the least accurate among the three GPR models. The RFR3 has the highest R2 and NSE accuracy on the validation set and the RFR1 follows with the second-highest R2 and NSE accuracy. The RFR2 has the lowest accuracy. In terms of RMSE and MAE, the RFR3 has the lowest values on the validation set and

the RFR1 follows with a slightly higher RMSE. The RFR2 has the highest RMSE and MAE. In summary, based on the provided validation performance metrics, the RFR3 offered the best accuracy in predicting Chl-a concentrations among the three RFR models. It has the highest accuracy for R2 and NSE, as well as the lowest RMSE and MAE on the validation data. RFR1 also performs well, especially in terms of R2 and NSE accuracy, and is followed by RFR2, which has the highest RMSE and MAE, indicating higher errors in predicting Chl-a concentrations on the validation set. Among the models, the RFR2 achieved the best overall training performance with the highest R2 and NSE accuracies and the lowest RMSE and MAE. MLPNN3 performed best on the validation data with the highest R2 and NSE accuracies and the lowest RMSE and MAE

The prediction accuracies of three methods are compared on the scatterplot in Fig. 8 for the validation period. It can be observed from the graphs, the MLPNN3 model has less scattered predictions and followed by the RFR3 justifying the statistics given in Table 3. Comparison of the fit line equations clearly indicates the slope (0.8678) and bias (8.8494) coefficients of the MLPNN3 model are closer to the 1 and 0 (ideal line, y = x) which proves the success of this model in predicting Chl-a concentration. It has the highest R2 which means that it can explain the 86% variance in modeling chl-a concentration. The models are further compared in violin, boxplot in Figs.8-9.

Discussion

In the previous section, we have provided the results obtained using six machine-learning models, i.e., the MLPNN, SVR, RFR, GPR, RVM, and the XGBoost. All models were first calibrated and in a second stage validated using validation dataset. The performances of each ML model are mainly governed by the best selection of the hyperparameters; however, any hyperparameter can have its own effect on the model response. In other word, we only hope to reach the smallest errors indices and the biggest fitting capability (i.e., the high R2 and NSE values), but it is hard to easily find the suitable hyperparameters. Furthermore, one can observe that, even after achieving the calibration stage by reaching the maximal training epoch, the best RMSE and MAE are recorded, and similar results with negligible difference can be obtained by slightly changing the fixed hyperparameters. More precisely, for the MLPNN model, having only one single hidden layer with sigmoidal function, we only varied the number of hidden neurons and the best performances were obtained with 08 neurons without the needs for increasing this number beyond 10 neurons. For the RFR, we only varied the number of tree (NumTrees) and the best performances were obtained using 10 trees. For the RVM, we applied the model with various width and bias (α,β), and we find the best width value to be approximately equal to three. For the GPR model, we varied the parameters of the squared exponential kernel function. For the SVR, we optimize the parameters of the Gaussian radial basis function by trial and error. Finally, for the XGBoost, we changed the learning rate, tree depth, and regularization parameters.

In this research, we investigated a range of machine-learning models to estimate Chl-a concentration using light from two distinct wavelengths. Training and testing outcomes of three machine learning methods were compared with respect to R2, NSE, RMSE and MAE criteria.

In the training period, the MLPNN2 had the highest R2 and NSE accuracies, indicating that it performs best in approximating the relationship between the input features and Chl-a concentrations, it was closely followed by the MLPNN3, and the MLPNN1 had the lowest R2 accuracy. Meanwhile, when it came to the validation data, the MLPNN3 outperformed the other models, showcasing the highest R2 and NSE accuracies along with the lowest RMSE and MAE. The RMSE values of MLPNN1, MLPNN2, SVR1, SVR2, SVR3, GPR1, GPR2, GPR3, RFR1, RFR2 and RFR3 in validation were approximately 38.41%, 15.05%, 43.17%, 46.16%, 45.98%, 4.23%, 45.73%, 46.76%, 60.34%, 69.65% and 7.80% higher than that of the MLPNN3, respectively. Similarly, the MAE values of MLPNN1, MLPNN2, SVR1, SVR2, SVR3, GPR1, GPR2, GPR3, RFR1, RFR2 and RFR3 in validation were approximately 51.87%, 21.34%, 41.05%, 47.86%, 46.60%, 90.62%,50.55%, 49.74, 40.03%, 48.63% and 6.40% higher than that of MLPNN3, respectively. In the validation period, the second-best model after MLPNN3 is RFR3. The MLPNN3 was the best model closely followed by the RFR3 in terms of predictive accuracy for Chl-a concentrations during the validation period.

The overall results indicated that the MLPNN is more successful in predicting Chl-a concentration compared to SVR, GPR and RFR methods. MLPNN's outperformance compared to other algorithms can be attributed to several factors. Firstly, MLPNN's ability to model non-linear relationships is crucial in environmental applications like Chl-a prediction, where spectral reflectance and chlorophyll dynamics often display complex, non-linear patterns. Neural networks like MLPNN are inherently capable of learning these complex relationships, especially compared to SVR and GPR models, which may be limited to more linear mappings or constrained in capturing non-linearity. Another reason for MLPNN's superior performance is its flexible architecture. In this study, the MLPNN was optimized by adjusting the number of hidden neurons, allowing it to capture nuances within the spectral data better than simpler models. This flexibility made MLPNN particularly adept at approximating the relationship between spectral bands and Chl-a concentrations, which was evident in its lower RMSE and MAE values. Moreover, MLPNN's robustness against

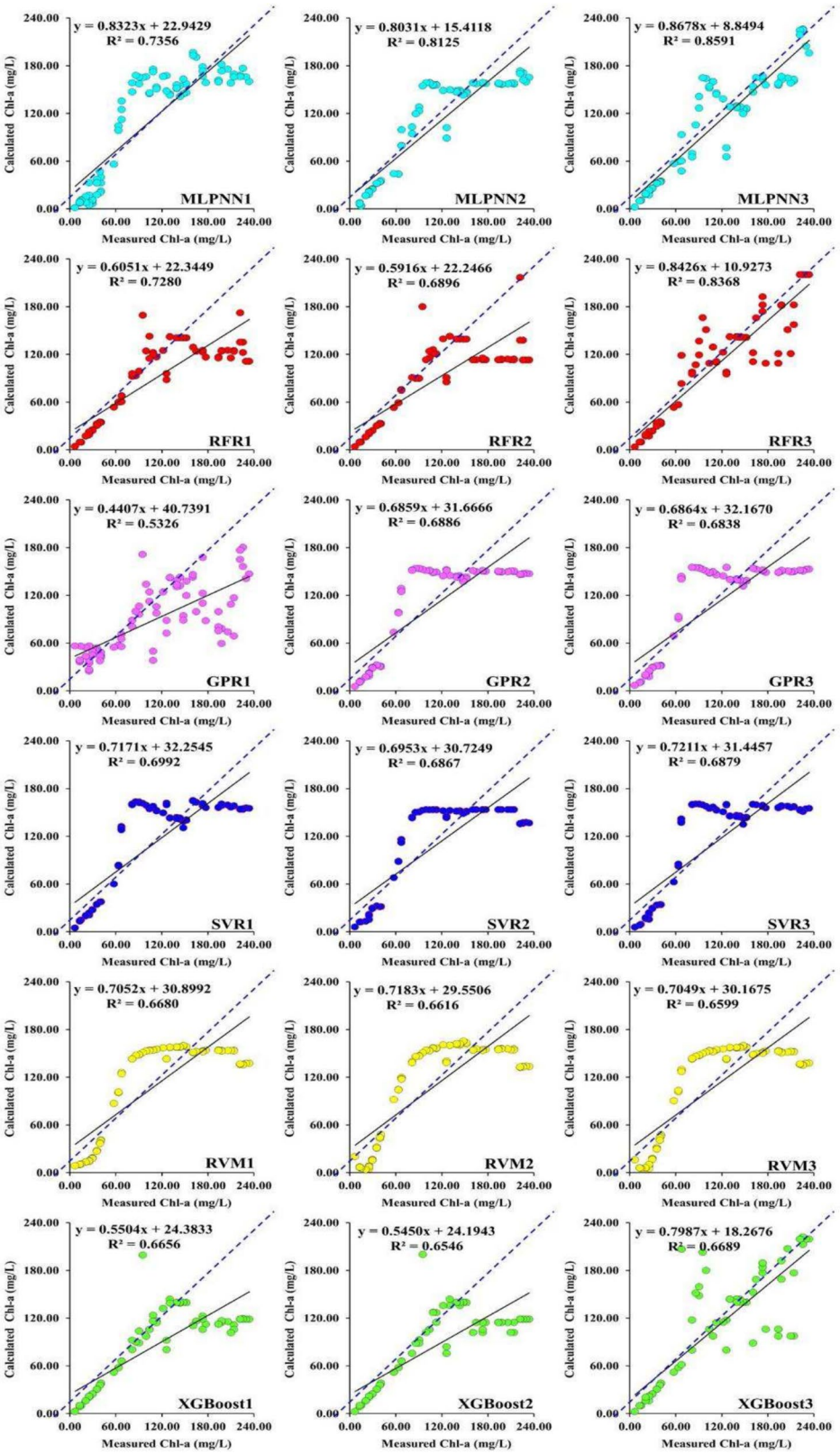

Fig. 8 Scatterplot of measured and predicted Chl-a using machine learning models

overfitting likely contributed to its strong validation performance. By fine-tuning the architecture, MLPNN balanced complexity and generalization, which was reflected in its lower error metrics on validation data compared to models like SVR. Unlike RFR, which relies on aggregated decision trees that may miss finer spectral variations, MLPNN's

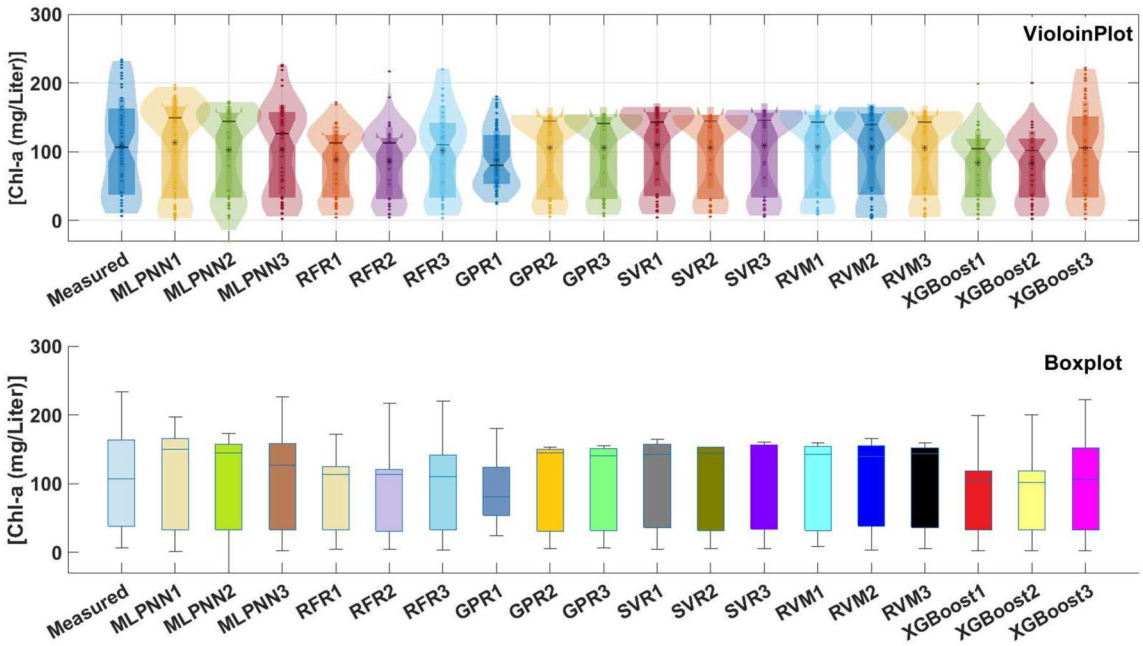

Fig. 9 Comparison between measured and predicted Chl-a: Boxplot (upper panel) and Violinplot (lower panel)

multi-layer approach enabled it to achieve higher accuracy by effectively capturing subtle spectral data patterns.

The scatterplots illustrated that none of the models were able to capture certain peaks; instead, they tended to underestimate them. Furthermore, the models encountered difficulty in accurately representing extreme fDOM values, both low and peak. This limitation can be attributed to the insufficient presence of extreme values within the dataset. Consequently, the data-driven techniques employed in this study failed to fully comprehend the dynamics of extremes, a phenomenon similarly noted by Akhlaq et al. (2023). On the other hand, it has been noted that increasing the number of input variables may lead to a decrease in model accuracy. For instance, the models in the 3rd case have a better accuracy compared to 1st case having more inputs, the addition of band 446 nm or band 595nm input resulted in a decline in the prediction accuracy of the models. This observation is consistent with findings from prior studies conducted by researchers such as Shi et al. (2012) and Zhang et al. (2022). These studies consistently indicated that augmenting the number of inputs does not necessarily enhance prediction accuracy; rather, it may have an adverse effect on variance. Consequently, this could lead to the development of more complex models with diminished prediction performance.

Kim and Ahn, (2022) recently applied random forest in prediction of Chl-a concentration of the Han River basin, China and compared with multilayer perceptron ANN and support vector machine. They used total nitrogen, total organic carbon, potential of hydrogen (pH), water temperature, electrical conductivity, total phosphorus, dissolved oxygen, mean AT minimum and maximum AT as model inputs. They obtained R^2 of 0.747 for the best random forest model. Baek et al., (2022) used two deep learning models, Long-Short Term Memory (LSTM) and Convolution Neural Networks (CNNs) for predicting Chl-a concentration of upper course of the Nakdong River, South Korea. They utilized minimum and maximum AT, precipitation, evapotranspiration, wind speed, solar radiation, daylight hours, streamflow and water temperature in the stream as inputs. Their best models provided R2 of 0.87, 0.83, 0.85 and 0.50 for four different sites. In our study, the best model produced a R2 of 0.86, indicating its success in predicting Chl-a concentration compared to previous literature. In our next step we will assess the effect of dissolved substances on the reflectance of Chl-a rich inland waters by direct measurements of Chl-a in samples of lake water.

The results of this study demonstrate the value of hyperspectral data combined with machine learning for large-scale, real-time monitoring of Chl-a concentrations. Unlike laboratory methods, which are precise but costly and time-intensive, this approach enables continuous and extensive monitoring, capturing dynamic spatial and temporal changes in aquatic ecosystems. While initial setup costs for hyperspectral equipment may be high, the ability to reduce manual sample collection and recurring laboratory expenses makes this approach more economical over time. The integration of field spectral reflectance data with ML models, as demonstrated in this study, represents a scalable, efficient alternative for monitoring Chl-a and managing water quality in aquatic environments.

While the initial investment in hyperspectral equipment can be high, the long-term savings in fieldwork, labor, and logistical costs make it a cost-effective solution for large-scale studies. Unlike traditional methods, which require extensive field sampling, laboratory analysis, and logistical coordination—steps that become increasingly costly as the study area expands—hyperspectral sensors enable continuous, automated data collection across vast regions. This capability reduces the need for repeated field visits and provides high-resolution temporal and spatial data, making it ideal for monitoring dynamic parameters like chlorophyll-a (Pandey et al. 2024). Traditional approaches, as noted by Randolph et al. (2008), can take several days to weeks to process, which is neither cost-effective nor timely, particularly for short-lived events like algal blooms. In contrast, hyperspectral remote sensing, combined with machine learning, allows for rapid and accurate chlorophyll-a estimation. Furthermore, advancements such as automated hyperspectral radiometer networks enhance scalability by providing continuous, real-time validation of satellite-derived data Ruddick et al. (2024). The high temporal and spatial resolution of hyperspectral data enables comprehensive monitoring of chlorophyll-a trends over time, providing critical insights that traditional methods often miss. This continuous monitoring not only reduces costs but also improves the effectiveness of environmental management efforts. Based on our findings, integrating hyperspectral data with machine learning algorithms represents a cost-efficient approach for large-scale chlorophyll-a monitoring, particularly for applications requiring rapid and frequent updates over expansive areas.

Conclusion

Based on the published literature, the findings in various investigations suggest that the concentration of Chl-a in freshwater is closely associated with various wavelength light, which constitute the major motivation of the present study. In this study, we explore various machine-learning models for retrieving Chl-a concentration from two different wavelength lights. Thus, a modelling framework is developed, and thereby the Chl-a was accurately quantified with high degree of precision. Among the four machine learning proposed in the present study, the MLPNN model produced systematically better estimates of Chl-a than the RFR, GRP and SVR models. The ability of MLPNN to flexibly adjust the number of layers and neurons allowed it to better approximate relationships in spectral data, enhancing its predictive power. Furthermore, MLPNN's multi-layer structure likely provided it with an edge over decision-tree-based methods like RFR, as it could more effectively model subtle variations in the spectral inputs. In conclusion, this study demonstrates that MLPNN is a powerful approach for Chl-a concentration prediction, offering a significant improvement in accuracy and reliability over traditional ML models when applied to complex environmental data. Taking into account that the effect of two wavelength was investigated, we can conclude that the results varied significantly from one model to another. The MLPNN and the RFR were already reported as being the exclusive algorithms for which the combination of the two wavelength lights (i.e., band 446 and band 595) have led to tangible improvements in the models performances, with an accuracy of R2≈0.927 and NSE≈0.853 for the MLPNN, and an accuracy of R2≈0.915 and NSE≈0.829 for the RFR, respectively. Finally, it is important to note that, the lowest accuracy was obtained using the GPR lower than all other models. Together with the in situ measured data, our proposed approach will contribute to the improvement of our monitoring and assessment of water quality.

Future research should investigate advanced neural network architectures, such as convolutional or recurrent neural networks, to further enhance predictive accuracy by capturing spatial or temporal patterns in hyperspectral data. MLPNN's success in this study suggests that neural network models are promising tools for accurate, large-scale water quality assessment.

Author contributions M.A.: conceptualization, methodology, data curation, investigation, software, visualization, formal analysis, validation, writing-original draft, writing, reviewing and editing. L.Z.: conceptualization, methodology, data curation, investigation, writing-original draft, writing, reviewing and editing. X.S.: conceptualization, methodology, data curation, investigation, writing-original draft, writing, reviewing and editing. O.K.: conceptualization, supervision, validation, software, visualization, formal analysis, validation, writing-original draft, writing, reviewing, editing. N.L.: Investigation, visualization, formal analysis, validation, writing-original draft. L.Y.: visualization, formal analysis, validation, writing-original draft. M.M.: formal analysis, validation, writing, reviewing and editing. S.H.: conceptualization, visualization, formal analysis, validation, writing-original draft, software, revision, writing-original draft.

Funding This work was supported by the National Natural Science Foundation of China (Grant No. 41830108)

Data Availability The data presented in this study are available under request

Declarations Ethical Approval

We declare that this manuscript has complied with all the ethical requirements of the journal.

Consent to Participate All authors of this manuscript have agreed to participate in the writing of the manuscript.

Consent for Publication All the authors of this manuscript consented to its publication.

Competing Interests The authors declare no competing interests.

References

Abd-Elaty I, Kushwaha NL, Patel A (2023) Novel Hybrid Machine Learning Algorithms for Lakes Evaporation and Power Production using Floating Semitransparent Polymer Solar Cells. Water Resour Manag 37:4639–4661. https://doi.org/10.1007/s11269-023-03565-2

Aguzzi J, Albiez J, Flögel S et al (2020) A Flexible Autonomous Robotic Observatory Infrastructure for Bentho-Pelagic Monitoring. Sensors 20:1614. https://doi.org/10.3390/s20061614

Ahn JM, Kim J, Kim K (2023) Ensemble Machine Learning of Gradient Boosting (XGBoost, LightGBM, CatBoost) and Attention-Based CNN-LSTM for Harmful Algal Blooms Forecasting. Toxins 15:608. https://doi.org/10.3390/toxins15100608

Akhlaq A, Shaik Dawood MSI, Jaffar Syed MA, Sulaeman E (2023) A study on the effect of piezoelectric nonlinearity on the bending behaviour of smart laminated composite beam. Materials 16:2839. https://doi.org/10.3390/ma16072839

Ali Z, Hussain I, Faisal M et al (2017) Forecasting Drought Using Multilayer Perceptron Artificial Neural Network Model. Adv Meteorol 2017:e5681308. https://doi.org/10.1155/2017/5681308

An G, Xing M, He B et al (2020) Using Machine Learning for Estimating Rice Chlorophyll Content from In Situ Hyperspectral Data. Remote Sens 12:3104. https://doi.org/10.3390/rs12183104

Baek S, Abbas A, Park M (2023) Cho KH (2022) Deep learning-based algorithms for long-term prediction of chlorophyll-a in catchment stream. J Hydrol 626:130240

Biau G, Scornet E (2015) A Random Forest Guided Tour. ArXiv151105741 Math Stat

Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on Computational learning theory. Association for Computing Machinery, New York, NY, USA, pp 144–152

Bowd C, Medeiros FA, Zhang Z et al (2005) Relevance Vector Machine and Support Vector Machine Classifier Analysis of Scanning Laser Polarimetry Retinal Nerve Fiber Layer Measurements. Invest Ophthalmol Vis Sci 46:1322–1329. https://doi.org/10.1167/iovs.04-1122

Breiman L (2001) Random Forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324

Camps-Valls G, Gómez-Chova L, Muñoz-Marí J et al (2006) Retrieval of oceanic chlorophyll concentration with relevance vector machines. Remote Sens Environ 105:23–33. https://doi.org/10.1016/j.rse.2006.06.004

Cao Z, Ma R, Duan H et al (2020) A machine learning approach to estimate chlorophyll-*a* from Landsat-8 measurements in inland lakes. Remote Sens Environ 248:111974. https://doi.org/10.1016/j.rse.2020.111974

Chen S, McKinney GJ, Nichols KM, Colbourne JK, Sepúlveda MS (2015) Novel Cadmium Responsive MicroRNAs in Daphnia pulex. Environ Sci Technol 49:14605–14613. https://doi.org/10.1021/acs.est.5b03988

Chusnah WN, Chu H-J (2022) Estimating chlorophyll-a concentrations in tropical reservoirs from band-ratio machine learning models. Remote Sens Appl Soc Environ 25:100678. https://doi.org/10.1016/j.rsase.2021.100678

Citakoglu H, Coşkun Ö (2022) Comparison of hybrid machine learning methods for the prediction of short-term meteorological droughts of Sakarya Meteorological Station in Turkey. Environ Sci Pollut Res 29:1–25

Daemi A, Kodamana H, Huang B (2019) Gaussian process modelling with Gaussian mixture likelihood. J Process Control 81:209–220. https://doi.org/10.1016/j.jprocont.2019.06.007

Demir V, Çıtakoğlu H (2024) Comparison of multiple machine learning methods for estimating digital elevation points. In: Çiner A, Ergüler ZA, Bezzeghoud M et al (eds) Recent Research on Geotechnical Engineering, Remote Sensing, Geophysics and Earthquake Seismology. Springer Nature Switzerland, Cham, pp 155–158

Ekmekcioğlu Ö, Başakın EE, Özger M (2022) Developing metaheuristic optimization based ensemble machine learning algorithms for hydraulic efficiency assessment of storm water grate inlets. Urban Water J 19:1093–1108. https://doi.org/10.1080/1573062X.2022.2134806

Gitelson AA, Gritz Y, Merzlyak MN (2003) Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. J Plant Physiol 160:271–282. https://doi.org/10.1078/0176-1617-00887

Hakim AMY, Baja S, Rampisela DA, Arif S (2021) Modelling land use/land cover changes prediction using multi-layer perceptron neural network (MLPNN): a case study in Makassar City, Indonesia. Int J Environ Stud 78:301–318. https://doi.org/10.1080/00207233.2020.1804730

Hang X, Li Y, Li X et al (2022) Estimation of Chlorophyll-a Concentration in Lake Taihu from Gaofen-1 Wide-Field-of-View Data through a Machine Learning Trained Algorithm. J Meteorol Res 36:208–226. https://doi.org/10.1007/s13351-022-1146-y

He J, Chen Y, Wu J et al (2020) Space-time chlorophyll-a retrieval in optically complex waters that accounts for remote sensing and modeling uncertainties and improves remote estimation accuracy. Water Res 171:115403. https://doi.org/10.1016/j.watres.2019.115403

Hu C, Feng L, Guan Q (2021) A Machine Learning Approach to Estimate Surface Chlorophyll a Concentrations in Global Oceans From Satellite Measurements. IEEE Trans Geosci Remote Sens 59:4590–4607. https://doi.org/10.1109/TGRS.2020.3016473

Keller S, Maier PM, Riese FM et al (2018) Hyperspectral Data and Machine Learning for Estimating CDOM, Chlorophyll a, Diatoms, Green Algae and Turbidity. Int J Environ Res Public Health 15:1881. https://doi.org/10.3390/ijerph15091881

Kim K-M, Ahn J-H (2022) Machine learning predictions of chlorophyll-a in the Han river basin. Korea. J Environ Manage 318:115636. https://doi.org/10.1016/j.jenvman.2022.115636

Koli P, Sharma U (2021) Use of pigments present in the crude aqueous extract of the spinach for the simultaneous solar power and storage at natural sun intensity. Adv Energy Sustain Res 2:2100079. https://doi.org/10.1002/aesr.202100079

Kolluru S, Tiwari SP (2022) Modeling ocean surface chlorophyll-a concentration from ocean color remote sensing reflectance in global waters using machine learning. Sci Total Environ 844:157191. https://doi.org/10.1016/j.scitotenv.2022.157191

Kushwaha NL, Rajput J, Elbeltagi A et al (2021) Data Intelligence Model and Meta-Heuristic Algorithms-Based Pan Evaporation Modelling in Two Different Agro-Climatic Zones: A Case Study from Northern India. Atmosphere 12:1654. https://doi.org/10.3390/atmos12121654

Kushwaha NL, Rajput J, Sena DR et al (2022) Evaluation of Data-driven Hybrid Machine Learning Algorithms for Modelling Daily Reference Evapotranspiration. Atmosphere-Ocean 60:519–540. https://doi.org/10.1080/07055900.2022.2087589

Kushwaha NL, Rajput J, Suna T et al (2023) Metaheuristic approaches for prediction of water quality indices with relief algorithm-based feature selection. Ecol Inform 75:102122. https://doi.org/10.1016/j.ecoinf.2023.102122

Kushwaha NL, Kudnar NS, Vishwakarma DK et al (2024) Stacked hybridization to enhance the performance of artificial neural networks (ANN) for prediction of water quality index in the Bagh river basin India. Heliyon 10:e31085. https://doi.org/10.1016/j.heliyon.2024.e31085

Kushwaha NL, Sushanth K, Patel A et al (2024) Beach nourishment for coastal aquifers impacted by climate change and population growth using machine learning approaches. J Environ Manage 370:122535. https://doi.org/10.1016/j.jenvman.2024.122535

Levi EE, Jeppesen E, Nejstgaard JC, Davidson TA (2024) Chlorophyll-a determinations in mesocosms under varying nutrient and temperature treatments: in-situ fluorescence sensors versus in-vitro measurements. Open Res Eur 4:69. https://doi.org/10.12688/openreseurope.17146.1

Li X, Yuan C, Li X, Wang Z (2020) State of health estimation for Li-Ion battery using incremental capacity analysis and Gaussian process regression. Energy 190:116467. https://doi.org/10.1016/j.energy.2019.116467

Luck J, Shearer SA, Luck BD, Payne F (2012) Evaluation of a rhodamine-WT dye/glycerin mixture as a tracer for testing direct injection systems for agricultural sprayers. Biol Syst Eng Pap Publ

Markwell J, Osterman JC, Mitchell JL (1995) Calibration of the Minolta SPAD-502 leaf chlorophyll meter. Photosynth Res 46:467–472. https://doi.org/10.1007/BF00032301

Misra S, Li H (2020) Chapter 9 - Noninvasive fracture characterization based on the classification of sonic wave travel times. In: Misra S, Li H, He J (eds) Machine Learning for Subsurface Characterization. Gulf Professional Publishing, pp 243–287

Niroumand-Jadidi M, Bovolo F (2022) Extreme gradient boosting machine learning for total suspended matter (TSM) retrieval from Sentinel-2 imagery. In: Remote Sensing of the Ocean, Sea Ice, Coastal Waters, and Large Water Regions 2022. SPIE, pp 30–37

Pahlevan N, Smith B, Alikas K et al (2022) Simultaneous retrieval of selected optical water quality indicators from Landsat-8, Sentinel-2, and Sentinel-3. Remote Sens Environ 270:112860. https://doi.org/10.1016/j.rse.2021.112860

Pandey A, Pandey P, Garg V, et al (2024) Remotely Sensed Hyperspectral Data to Determine Chlorophyll-a in River Water. In: Agarwal A, Yadav B, Nema M, et al. (eds) Towards Water Circular Economy. Springer Nature Switzerland, Cham, pp 176–187

Park Y, Cho KH, Park J et al (2015) Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea. Sci Total Environ 502:31–41. https://doi.org/10.1016/j.scitotenv.2014.09.005

Poddar S, Chacko N, Swain D (2019) Estimation of Chlorophyll-a in Northern Coastal Bay of Bengal Using Landsat-8 OLI and Sentinel-2 MSI Sensors. Front Mar Sci 6:598

Özbayrak A, Ali MK, Çıtakoğlu H (2023) Buckling load estimation using multiple linear regression analysis and multigene genetic programming method in cantilever beams with transverse stiffeners. Arab J Sci Eng 48:5347–5370. https://doi.org/10.1007/s13369-022-07445-6

Randolph K, Wilson J, Tedesco L et al (2008) Hyperspectral remote sensing of cyanobacteria in turbid productive water using optically active pigments, chlorophyll *a* and phycocyanin. Remote Sens Environ 112:4009–4019. https://doi.org/10.1016/j.rse.2008.06.002

Ruddick KG, Brando VE, Corizzi A et al (2024) WATERHYPERNET: a prototype network of automated in situ measurements of hyperspectral water reflectance for satellite validation and water quality monitoring. Front Remote Sens 5:1347520. https://doi.org/10.3389/frsen.2024.1347520

Shin Y, Kim T, Hong S et al (2020) Prediction of Chlorophyll-a Concentrations in the Nakdong River Using Machine Learning Methods. Water 12:1822. https://doi.org/10.3390/w12061822

Silveira Kupssinskü L, Thomassim Guimarães T, Menezes de Souza E et al (2020) A Method for Chlorophyll-a and Suspended Solids Prediction through Remote Sensing and Machine Learning. Sensors 20:2125. https://doi.org/10.3390/s20072125

Singhal G, Bansod B, Mathew L et al (2019) Chlorophyll estimation using multi-spectral unmanned aerial system based on machine learning techniques. Remote Sens Appl Soc Environ 15:100235. https://doi.org/10.1016/j.rsase.2019.100235

Vishwakarma DK, Kuriqi A, Abed SA et al (2023) Forecasting of stage-discharge in a non-perennial river using machine learning with gamma test. Heliyon 9:e16290. https://doi.org/10.1016/j.heliyon.2023.e16290

Xylem Inc (2020) EXO user manual: advanced water quality monitoring platform (Revision K). Xylem Inc. Retrieved from YSI.com/EXO

Zhang D, Zhang L, Sun X, Gao Y, Lan Z, Wang Y, Zhai H, Li J, Wang W, Chen M, Li X, Hou L, Li H (2022) A new method for calculating water quality parameters by integrating space–ground hyperspectral data and spectral-in situ assay data. Remote Sens 14(15):3652. https://doi.org/10.3390/rs14153652

Zou D, Tong L, Wang J et al (2020) A Logical Framework of the Evidence Function Approximation Associated with Relevance Vector Machine. Math Probl Eng 2020:e2548310. https://doi.org/10.1155/2020/2548310