Science of the Total Environment 938 (2024) 173546

Contents lists available at ScienceDirect

Science of the Total Environment

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

Review

Recent advances in algal bloom detection and prediction technology using machine learning

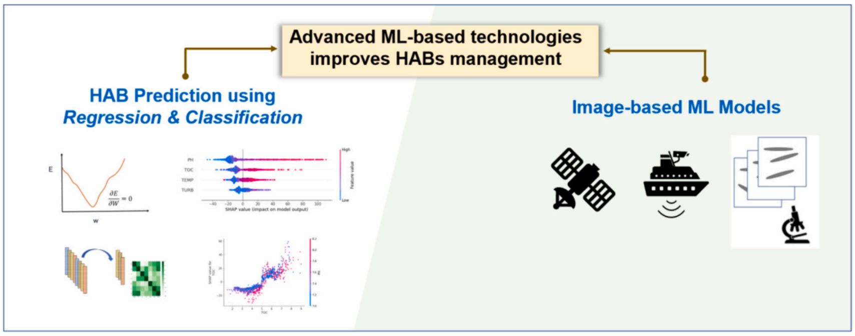
Jungsu Park [a], Keval Patel [b], Woo Hyoung Lee [b,*]

[a] Department of Civil and Environmental Engineering, Hanbat National University,125, Dongseo-daero, Yuseong-gu, Daejeon 34158, Republic of Korea
[b] Department of Civil, Environmental and Construction Engineering, University of Central Florida, 12800 Pegasus Dr., Orlando, FL 32816, United States

HIGHLIGHTS

- Machine learning-based models lead to improved harmful algal bloom prediction.
- Explainable artificial intelligence improves better understanding of the developed models.
- Image processing using machine learning enhances algal detection and monitoring strategies.

GRAPHICAL ABSTRACT


Advanced ML-based technologies improves HABs management
HAB Prediction using Regression & Classification
Image-based ML Models

ARTICLE INFO

Editor: Ouyang Wei

Keywords:
Algal bloom detection
Algal bloom prediction
Harmful algal bloom
Image-based machine learning
Machine learning

ABSTRACT

Harmful algal blooms (HAB) including red tides and cyanobacteria are a significant environmental issue that can have harmful effects on aquatic ecosystems and human health. Traditional methods of detecting and managing algal blooms have been limited by their reliance on manual observation and analysis, which can be time-consuming and costly. Recent advances in machine learning (ML) technology have shown promise in improving the accuracy and efficiency of algal bloom detection and prediction. This paper provides an overview of the latest developments in using ML for algal bloom detection and prediction using various water quality parameters and environmental factors. First, we introduced ML for algal bloom prediction using regression and classification models. Then we explored image-based ML for algae detection by utilizing satellite images, surveillance cameras, and microscopic images. This study also highlights several real-world examples of successful implementation of ML for algal bloom detection and prediction. These examples show how ML can enhance the accuracy and efficiency of detecting and predicting algal blooms, contributing to the protection of aquatic ecosystems and human health. The study also outlines recent efforts to enhance the field applicability of ML models and suggests future research directions. A recent interest in explainable artificial intelligence (XAI) was discussed in an effort to understand the most influencing environmental factors on algal blooms. XAI facilitates interpretations of ML model results, thereby enhancing the models' usability for decision-making in field management and improving their overall applicability in real-world settings. We also emphasize the significance of obtaining high-quality, field-representative data to enhance the efficiency of ML applications. The effectiveness of ML models in detecting and predicting algal blooms can be improved through management strategies for data

* Corresponding author at: 12800 Pegasus Dr. Suite 211, Orlando, FL 32816-2450, United States.
E-mail addresses: parkjs@hanbat.ac.kr (J. Park), Keval.Patel@ucf.edu (K. Patel), woohyoung.lee@ucf.edu (W.H. Lee).

https://doi.org/10.1016/j.scitotenv.2024.173546
Received 17 November 2023; Received in revised form 18 May 2024; Accepted 24 May 2024
Available online 27 May 2024

quality, such as pre-treating missing data and integrating diverse datasets into a unified database. Overall, this paper presents a comprehensive review of the latest advancements in managing algal blooms using ML technology and proposes future research directions to enhance the utilization of ML techniques.

1. Introduction

Algae offer many advantages due to their ability to generate oxygen and biomass (Shao et al., 2021), the latter of which can be processed to produce sustainable energy such as feedstocks and biofuels such as biodiesel. Nonetheless, due to the increasing water pollution resulting from the excessive release of nitrogen (N) and phosphorus (P) into water bodies, there is a potential for the substantial proliferation of algae. Algal blooms, which pose threats to both drinking water supply systems and the ecological health of water resources, have been a major issue in water quality management in many countries around the world, including Australia, China, the European Commission, South Korea, and the United States of America (USA), over the past few decades (Herath, 1997; Wang et al., 2022a; West et al., 2021; Wurtsbaugh et al., 2019). In the case of Lake Erie, managing algal blooms has been a critical issue in water quality management over the last two decades, notably when a significant bloom in 2014 resulted in a three-day ban on tap water usage in Toledo, Ohio, USA (Ho and Michalak, 2015). There has been growing attention towards managing eutrophication and algal blooms in the Lake Taihu area, the Chinese largest freshwater lake, due to escalating pollution inputs since the 1960s (Wang et al., 2022a). Specific types of algae (e.g., cyanobacteria) have the capability to generate neurotoxins, and the overgrowth of these noxious algae is referred to as harmful algal blooms (HABs) (Erdner et al., 2008). Exposure to these toxins can lead to various adverse reactions in humans (Hill et al., 2020), such as gastrointestinal toxicity from short-term exposure to possibly promoting cancers and liver disease from long-term exposure (Erdner et al., 2008). Furthermore, the economic impacts of HABs on the fishing and tourism industry have been estimated to cost the USA $82 million per year (Hill et al., 2020). Unless well managed, water bodies with elevated levels of N and P can undergo HABs, rendering them unsafe for both consumption and recreation. Therefore, it is critical to develop and utilize suitable techniques for detecting algae prior to the escalation of HABs. By monitoring physico-biochemical patterns within the algae ecosystem such as monitoring water quality parameters, and identifying and quantifying algal species, timely warnings (e.g., HAB alert with recreational advisory guidance levels) (Gong et al., 2023) and prediction of HABs (like weather forecasts) can be provided to ensure the protection of human health (Hill et al., 2020).

Recent research has been dedicated to utilizing rapidly advancing machine learning (ML) models for predicting algal blooms and enhancing monitoring efficiency. Traditional approaches to HAB monitoring and prediction have typically involved direct microscopic observation of algae or indirect analysis using chlorophyll-a (Chl-*a*) concentrations in a laboratory setting. However, both methods require specialized analytical expertise and are time-intensive and laborious. In contrast, advanced ML algorithms for object detection are emerging as efficient and cost-effective alternatives to algal bloom monitoring. Traditional methods for predicting algal blooms, such as mechanistic models, require consideration of various physical, biological, and chemical factors that influence them. To identify such factors (e.g., algal growth rate), time-consuming and labor-intensive experiments are required. However, ML models offer the advantage of building efficient models with reduced dependence on experimentally determined factors, and they can produce models with good performance if sufficient high-quality data is available. Various data-driven ML models, including artificial neural networks (ANN), deep learning, and ensemble ML models, exhibiting excellent performance in future prediction, continue to be explored for algal bloom prediction (Gupta et al., 2023; Wen et al., 2022).

The primary objectives of this review article are to comprehensively explore recent studies on ML algorithms including image processing in predicting HABs and detecting algae (Fig. 1). This study aims to systematically present the theoretical concepts and characteristics of various ML models that are increasingly utilized for the detection and prediction of algal blooms. Furthermore, recent studies on explainable artificial intelligence (XAI), a novel method for quantitative interpretation of black-box based ML models, were reviewed and presented directions to increase the practical application of machine learning models.

We analyzed the current status and characteristics of various ML models used in field algal bloom management, and proposes data quality management strategies to enhance the performance and utility of data-driven models such as ML. The article offers insights into the application of both regression and classification models for HAB prediction (Fig. 1), highlighting the distinction between models that predict quantitative algal bloom status and those that categorize bloom levels and can be used for issuing an algal bloom alert. Furthermore, the review discussed the growing use of image classification algorithms for algal detection.

Algal bloom management is a crucial issue in water environmental management, and efforts are ongoing to apply advanced ML models for efficient algal bloom prediction and monitoring. This paper consolidates recent research trends in efficient algal prediction and monitoring, and by summarizing the current research landscape, this paper provides insights into future research directions necessary for effective algal bloom management using ML-based technology.

The following sections of this research are organized as follows: Section 2 presents a critical review of recent research on algal bloom prediction using ML algorithms (regression vs. classification), and Section 3 reviews recent ML-based technologies for algal bloom detection. Additionally, the application of novel XAI algorithms to overcome the limitations of ML is included in Section 4. Considerations for effective ML-driven HAB management strategies are presented in Section 5. Lastly, Section 6 provides a comprehensive conclusion of this study.

2. Development of ML models for algal bloom prediction

2.1. Model development overview

In the recent decade, various ML models have been used to predict algal blooms in freshwater bodies such as rivers and reservoirs. The ML models have general advantages in that they are suitable for non-linear data while they have high complexity and computational cost compared to traditional statistical methods (Cruz et al., 2021). Various ML algorithms such as ANN, deep learning, and support vector machine (SVM) have been used consistently for algal bloom prediction over the past decade (Vilas et al., 2014). Ensemble models such as random forest (RF) and gradient boosting decision tree (GBDT) algorithms have been widely used from the mid-2010s until recently. Deep learning models such as Long Short-Term Memory (LSTM), gated recurrent units (GRU), and Transformer have also been increasingly used for algal bloom prediction (Qian et al., 2023; Rostam et al., 2021; Vilas et al., 2014).

Recently, automated machine learning (AutoML) has been developed, which automates the ML model development process, including data pretreatment, model selection, and hyperparameter optimization (Madni et al., 2023; Prasad et al., 2021).

In general, ML models are categorized into regression models, aimed at predicting continuous outcomes, and classification models, designed to classify data points into distinct categories or classes. The independent variables used for model development include various water quality parameters, hydrological parameters, and meteorological parameters, while Chl-*a* and algal cell numbers are commonly used as target variables for algal prediction. Regression models predict the actual values of Chl-*a* or algal cell numbers, while classification models estimate the alert levels necessary for algal bloom management alerts. Model performance is evaluated using various indices, including coefficient of determination, root mean squared errors (RMSE), and mean absolute error (MAE) for regression models, and accuracy, recall, and precision for classification models.

2.2. *Theoretical base of ML models*

The ML algorithms used for predicting algal blooms can be commonly categorized into models based on neural networks, such as ANN and deep learning; tree-based algorithms; ensemble models, which improve prediction through a combination of multiple models; and SVM, known for their effectiveness in classification and regression tasks. This section presents the concepts and theoretical bases of these models.

ANN is a well-known ML algorithm, consisting of three layers: an input layer, hidden layer(s), and an output layer. The hidden layer is composed of multiple nodes, and the model is optimized by identifying the optimal values for the weights and biases of each node. These optimal values are determined to minimize the loss between observed and predicted model outputs during the training process, commonly employing a gradient descent algorithm. ANN-based models are effective in modeling complex non-linear relationships, and their computational capacity can be enhanced by increasing the number of hidden layers. The efficiency of ANN model training has been significantly improved by the back-propagation (BP) method (Rumelhart et al., 1986). However, training deep networks has been a challenge, leading to various efforts to overcome this limitation. On such effort is the development of the Boltzmann machine, a stochastic neural network where each node is fully connected to every other node.

The Restricted Boltzmann Machine (RBM) simplifies the learning process by removing the connections between visible-visible and hidden-hidden units (Hinton, 2012; Hinton et al., 2006). Deep Belief Networks (DBN), which stack RBMs, are considered an early form of deep learning model (Hinton et al., 2006). However, increasing the number of hidden layers led to various computational problems, such as the vanishing gradient issue that arises during the BP process used to update the weights of the neural network model. The rectified linear unit (ReLU) has been instrumental in addressing the vanishing gradient problem. The use of ReLU as an activation function in a neural network model, instead of the traditional sigmoid function, mitigates the vanishing gradient issue during the BP process (Nair and Hinton, 2010). Dropout is another critical technique that enhances the performance of deep learning models. During the training process, dropout works by randomly "dropping out" a certain number of outputs from the layers of the hidden layers. This helps prevent overfitting of the model to the training data and improves the model's performance on unseen data (Srivastava et al., 2014).

Recurrent Neural Networks (RNNs) are a class of neural networks designed specifically to recognize patterns in sequential data, making them widely used for various data types, including text and time series data. LSTM, a type of RNN, mitigates the vanishing gradient problem often encountered in traditional RNNs, by introducing a structure known as the "cell state" (Hochreiter and Schmidhuber, 1997). The LSTM structure is composed of the cell state and three gates: forget gate, input gate, and output gate. The cell state carries information from one time step to the next, with the three gates selectively controlling the flow of information through time. In the first step, the forget gate determines which information to discard from the cell state. Next, at the input gate, new information is selected to be stored in the cell state. Finally, at the output gate, the LSTM decides what information should be output at the current time step. The output gate in an LSTM uses the current input and the previous hidden state to compute its activation, applying learned weights and a sigmoid function. Separately, the current cell state is processed through a tanh function to scale its values between −1 and 1. The LSTM then computes its output at the current timestep, known as the current hidden state, by multiplying the output gate activation and the transformed cell state.

GRU is a type of RNN architecture that also mitigate vanishing gradient issues like LSTMs, but in a simpler way (Cho et al., 2014). GRU has a hidden state and two gates, namely an update gate and a reset gate, which regulate the flow of information within the hidden state. The reset gate decides the extent of past information to discard, while the update gate determines the amount of information from the previous hidden state that should be carried forward to the current state. By applying learned weights to the current input and the previous hidden state, then passing them through a sigmoid function, the update gate generates values between 0 and 1 to serve as coefficients for controlling this information flow.

SVM is a representative supervised learning algorithm that was widely used until recently (Boser et al., 1992; Cortes and Vapnik, 1995). The SVM model is trained to find a hyperplane or decision boundary in an N-dimensional space (where N is the number of features) that maximizes the distance between data points. The distance between the decision boundary is referred to as "margin" and the data closest to the decision boundary is called "support vector" which determines the position and orientation of the decision boundary. The SVM is widely used both for classification and regression models.

Ensemble ML has become a widely used approach in recent years. RF

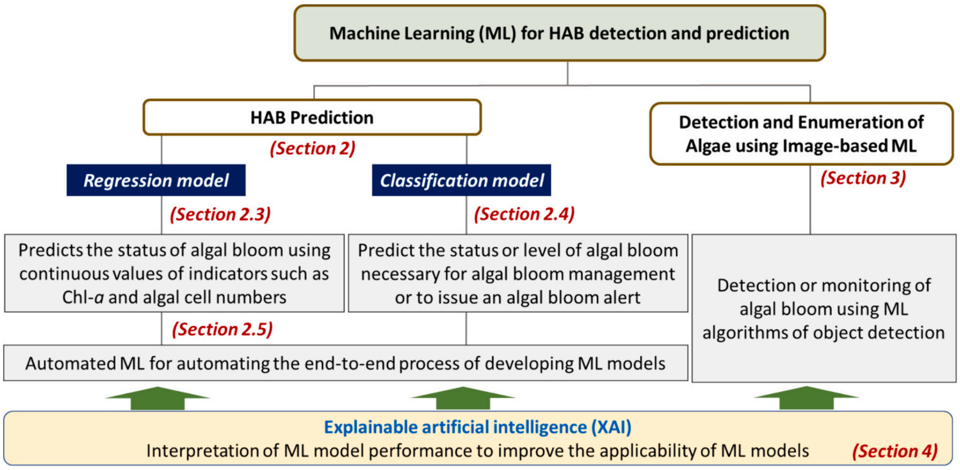

Machine Learning (ML) for HAB detection and prediction
HAB Prediction
(Section 2)
Detection and Enumeration of Algae using Image-based ML
(Section 3)
Regression model
(Section 2.3)
Classification model
(Section 2.4)
Predicts the status of algal bloom using continuous values of indicators such as Chl-a and algal cell numbers
Predict the status or level of algal bloom necessary for algal bloom management or to issue an algal bloom alert
Detection or monitoring of algal bloom using ML algorithms of object detection
(Section 2.5)
Automated ML for automating the end-to-end process of developing ML models
Explainable artificial intelligence (XAI)
Interpretation of ML model performance to improve the applicability of ML models
(Section 4)

Fig. 1. An overview of applying machine learning models for algal bloom management.

algorithm is a prime example of an ensemble learning method. The RF algorithm generates multiple individual decision tree models, and the final model's outcome is determined by either voting (for classification models) or averaging (for regression models) the results of these individual models (Breiman, 2001). Each decision tree is trained on a subset of the input features, which are randomly selected using a technique known as bagging. This process encourages diversity among the individual models, thus enhancing the robustness and generalization of the overall RF model. GBDT is one of the prominent ensemble ML algorithms (Friedman, 2001). In GBDT, multiple decision tree models, often referred to as "weak learners," are generated sequentially. The performance of the model is enhanced by utilizing the information from the previous weak learner to inform the creation of the next tree model. The model is trained with the goal of minimizing the residual error at each step. XGBoost (Chen and Guestrin, 2016) is one of the most popular GBDT algorithms, and light gradient-boosting machine (LGBM) is another innovative GBDT algorithm known for its computational speed and suitability for large datasets. LGBM employs the gradient-based one-side sampling method and exclusive feature bundling algorithm to selectively determine the number of input data for the model training strategy (Ke et al., 2017). It also utilizes a leaf-wise tree growth strategy. These algorithms are designed to improve computational efficiency and accelerate the model simulation speed of LGBM models.

Research into applying variously developed ML models for algal bloom prediction has been actively pursued in recent years. Due to the nature of data-driven models, the performance of predictions is influenced by the characteristics of the input data used to build the model. Even using the same model, performance can vary depending on the

Table 1
A summary of regression ML models to predict algal blooms.

Main model	Input variables	Target variables	Performance evaluation	Ref.
Generalized regression neural network (GRNN), SVM	Monthly/biweekly water quality and daily meteorological data including wind speed and solar radiation records	Chl-*a*	R^2, RMSE, and MAE values of 0.819, 5.436, and 3.167, respectively	(Li et al., 2014)
SVM	Weekly upwelling indices, temperature, salinity, occurrence of a bloom in the previous week (Bloom-1w) or two weeks before the sampling (Bloom-w2), day of the year, ria code	Bloom status (presence/ below low detection limit(P/ BD)) of Pseudo-nitzschia spp.	The best overall accuracy of P/BD was 78.57	(Vilas et al., 2014)
LSTM, ordinary least square (OLS), MLP, RNN	Weekly water temperature, pH, BOD, COD, DO, cyanobacteria cell number, water level, and pondage	Chl-*a*	RMSE of LSTM 16.09 and for OLS, MLP, and RNN were 17.75, 16.42, and 16.13, respectively	(Lee and Lee, 2018)
Multivariate Timing-Random Deep Belief Net (MT-RDBN) model, which combines multi-factor time series analysis and deep belief net	pH, NH_4^+-N, and water temperature with 10 samples per day	Chl-*a*	RMSE 3.88 % for testing data	(Wang et al., 2019).
M5P (a tree algorithm), extreme learning machine (ELM)	Daily water temperature, rainfall, solar radiation, total nitrogen, total phosphorus, N/P ratio, and Chl-*a*	Chl-*a*	The R^2 of M5P and ELM were 0.83, 0.46, 0.44, 0.39 and 0.87, 0.59, 0.48, 0.40 after 1, 3, 5 and 7 d, respectively	(Yi et al., 2019)
MLR, SVM, ANN	EC, DO, water temperature, TN, TP, BOD, COD, TSS, Chl-*a*, precipitation, transparency (Monthly)	Chl-*a* and transparency	MAE, RMSE, R^2/R^2 of SVM for Chl-*a* prediction varied from 0.56 to 0.80 for various sites and seasons	(Mamun et al., 2019)
SVM, DT, RF, ANN, MLR, TSP, RNN, DNN, LSTM	Salinity, DO, turbidity, pH, Secchi Disk Depth, SS, water temperature, Total Inorganic Nitrogen, Ammonia Nitrogen, PO_4^{3-}-P, TP, TN, NO_2^--N, NO_3^--N, Silica (1556 instances during 1986–2018)	Chl-*a*	MAE: 0.0256–0.5607 RMSE: 0.0360–0.6359 MSE: 0.0013–0.4044 LSTM showed the best performance	(Rostam et al., 2021)
RF, SVM, MLP	water and environmental parameters including Secchi depth, salinity, water temperature, NO_3^--N, PO_4^{3-}-P, Chl-*a*, Zooplankton abundance, sunlight duration, wind speed with daily datasets obtained from interpolation	Chl-*a*	RMSE, R^2 and adjusted R^2/adjusted R^2 were 0.77, 0.74, 0.76 for SVM, RF and MLP, respectively	(Amorim et al., 2021)
LSTM, CNN	MODIS-Aqua level 3 Chl-*a* data	Chl-*a*	RMSE for LSTM 3.402142 and CNN 4.361724/R (correlation coefficient) for LSTM 0.338385 and CNN 0.111790, respectively	(Yussof et al., 2021)
SVM	Biweekly/monthly total inorganic nitrogen (TIN), PO_4^{3-}-P, DO, water temperature, secchi-disc depth	Chl-*a*	RMSE 0.660 and correlation coefficient 0.984	(Deng et al., 2021)
AdaBoost, ANN, GBDT, KNN, SVM.	Weekly Chl-*a*, nitrate, and phosphate	Harmful algal bloom	MSE ranges 0.031–0.61 and R^2 ranges 0.939–0.956 AdaBoost showed the best R^2 as 0.956	(Yu et al., 2021)
CART, RF, LR	15 min interval pH, EC, water temperature, and system battery	Chl-*a*	MAE ranges 4.40–6.22	(Mozo et al., 2022)
LSTM	Marine hydrological multidepth environmental data including NH_4^+-N, PO_4^{3-}-P, Chl-*a*, depth, pressure etc.	Harmful algal bloom	MAE, RMSE, Sum of squared error, Mean absolute percentage error, fitting degree(R) with highest R is 82.1 %	(Wen et al., 2022)
CNN	8 water quality variables (water temperature, pH, EC, DO, TOC, TN, TP, Chl-*a*) and four weather variables as input variables	Chl-*a*	R^2 and RMSE of the optimal model were 0.934 and 5.463	(Lee et al., 2022)
GBR, LSTM	Meteorological data including air temperature and nutrients data (daily, 1–2 weeks)	Chl-*a*	MAE, RMSE, R^2 LSTM shows the best R^2 as 0.2	(Lin et al., 2023)
RF	Physicochemical, hydrological, meteorological observation	Biomass composition of phytoplankton community	$R^2 > 0.74$ for total biomass simulation	(Liu et al., 2023)
Transformer	TP, PO_4^{3-}-P, TN, NO_3^--N, NH_4^+-N, COD, TOC	Chl-*a*	R^2 0.85, RMSE 0.35	(Qian et al., 2023)

data used for model development, and more advanced and complicated models do not always yield better results. Consequently, there has been vigorous research into not only using advanced algorithms like ensemble ML and LSTM but also applying a range of ML models including earlier developed ones such as ANN and SVM, independently or in combination, to analyze their performance and application characteristics.

2.3. Regression models for algal bloom prediction

Various ML algorithms have been used for the development of regression models to predict Chl-*a* or algal cell numbers, the two most widely used indicators for the quantitative representation of algal blooms. Higher values of Chl-*a* or algal cell numbers indicate more severe algal blooms. The main model, input variables and target variable used for the algal bloom prediction regression models were summarized in Table 1.

The commonly used algorithms include various types of models such as ANN, SVM, tree-based ensemble models (e.g., RF and GBDT), and deep learning models. Many studies have employed multiple models instead of a single specific model, as summarized in Table 1. The input variables comprise various factors representing basic water quality elements (e.g., pH, dissolved oxygen (DO), temperature), organic materials (e.g., biochemical oxygen demand (BOD), chemical oxygen demand (COD)), and nutrients (e.g., total nitrogen (TN), total phosphorus (TP)), with Chl-*a* being the most frequently utilized item for the target-dependent variable. The model performances were evaluated using diverse indices such as RMSE and MAE. Additionally, the coefficient of determination (R^2) is widely employed as an index to enable the comparison of multiple model performances, given that a high R^2 indicates a strong agreement between the model and observations.

SVM is an ML model that has been widely used since the early stages of research for predicting algae. Li et al. (Li et al., 2014) developed prediction models for algal blooms using various ML methods and observation data from Tolo Harbour in Hong Kong. Three types of models were created using BP neural network, generalized regression neural network (GRNN), and SVM. Monthly/biweekly water quality and daily meteorological data (e.g., wind speed and solar radiation records) from January 1997 to December 2004 were used for model development, with Chl-*a* used as the target variable. The experimental results revealed that the SVM model exhibited the best performance, with R^2, RMSE, and MAE values of 0.819, 5.436, and 3.167, respectively, for the testing dataset. Vilas et al. (Vilas et al., 2014) developed and validated SVM models for the prediction of *Pseudo-nitzschia* spp. using eight years of data collected in the coastal embayments (rias) in the NW part of Spain. The models accurately identified presence/below low detection limit (P/BD) and bloom/no bloom conditions of *Pseudo-nitzschia* spp. and predicted blooms in the coastal systems of the Galician rias.

Since the late 2010s, research has continued to apply various ML models, including various deep learning models, for algal prediction. LSTM is one of the most popularly used algorithms, and various studies have used LSTM with multiple ML models for algal prediction. Lee and Lee (2018) utilized three deep learning models, namely Multilayer Perceptron (MLP), RNN, and LSTM, for algal bloom prediction. This study used raw data obtained from 16 field monitoring stations in South Korea with different observation frequencies, such as daily and weekly. The data were standardized to weekly frequencies for model development. The model employed water temperature, pH, BOD, COD, DO, cyanobacteria cell number, water level, and pondage as independent variables to predict Chl-*a* concentration as the target variable. The results indicated that the LSTM model exhibited the best performance, with an average RMSE of 16.09 for the 16 sites. Wang et al. (Wang et al., 2019) proposed a model called the Multivariate Timing-Random Deep Belief Net (MT-RDBN), which combines multi-factor time series analysis and deep belief nets. The MT-RDBN model utilizes autoregressive and multivariate regression models to describe the relationships between the characterization factor at current and previous times, as well as between the characterization factor and the influencing factors. Rostam et al. (Rostam et al., 2021) presented a complete framework methodology for predicting algal growth that includes sensor assembly and integration, data acquisition, and predictive modeling using data-driven approaches such as ML and deep learning. Various models including SVM, DT, RF, ANN, multilinear regression (MLR), RNN, DNN, and LSTM were used for model development. Model performance was evaluated using MAE, RMSE, and MSE, with values varying from MSE of 0.0256–0.5607, RMSE of 0.0360–0.6359, and MSE of 0.0013–0.4044, respectively. LSTM showed the best performance with MAE of 0.0256, RMSE of 0.0360, and MSE of 0.0013, respectively. The study demonstrates that using time series with deep learning algorithms, specifically LSTM, is the best fit for accurately predicting algal growth. Saboe et al. (2021) hypothesized that temporal microbial potentiometric sensor (MPS) signal patterns can predict changes in water quality parameters using AI/ML tools. The proof of concept was first tested by correlating MPS signals with high algae concentrations in an algal cultivation pond. The study then expanded to predict multiple water quality parameters in real surface waters, like irrigation canals. Data from the MPS system was used to train LSTM algorithms, which predicted parameters such as turbidity, conductivity, Chl-*a*, blue-green algae, DO, and pH. Real-time observations over 9 months with a 30-min observation frequency were used for model development. Results demonstrated the usefulness of MPSs and AI/ML tools in predicting key surface water quality parameters through a single composite signal, offering a novel and cost-effective approach for water quality monitoring.

LSTM has shown excellent predictive performance on time-series data and is widely utilized. However, similar to other ML models, the performance of the model is significantly influenced by the characteristics of the input data. Therefore, there is a variety of ongoing research aimed at improving LSTM's performance by enhancing the composition of input data. Yussof et al. (Yussof et al., 2021) applied LSTM and Convolutional Neural Network (CNN) methods to predict HAB events on the West Coast of Sabah, Malaysia. Satellite time-series data was used, with Chl-*a* as an HAB indicator. The dataset covered eight-day intervals from January 2003 to December 2018. In this study, the LSTM model proved more accurate than the CNN model based on RMSE and correlation coefficient criteria. Wen et al. (Wen et al., 2022) proposed a local spatiotemporal HABs forecasting model (STHFM) based on maritime station monitoring (MSM) data. The model uses principal component analysis (PCA) to select main environmental factors (MEFs) related to HABs and determines multiple warning levels based on algae growth rate. An improved LSTM network incorporating MEFs time series information from the Autoregressive Integrated Moving Average (ARIMA) model is used for forecasting. Tested on NOAA's public dataset, the model achieves a prediction accuracy of 82.1 % and a small prediction error, demonstrating good HABs monitoring performance. Lin et al. (Lin et al., 2023) applied two ML models, gradient boost regressor (GBR) and LSTM network, to predict Chl-*a* concentrations in a mesotrophic lake. The input variables for model development include daily meteorological data (e.g., air temperature and wind speed) and nutrient data (e.g., NO_x, O_2, PO_4^{3-}, and TP). They tested three predictive workflows: one using only available measurements of daily meteorological data and nutrient data with 1–2 weeks observation frequency and the other two using a two-step approach with pre-generated environmental factors such as daily nutrients and hydrodynamic data from process-based models. Observations between 2004–2016 and 2017–2020 were used for training and testing the model, respectively. The ML models outperformed process-based models in predicting Chl-*a* concentrations, and the hybrid model improved predictions of algal bloom timing and magnitude.

The Transformer stands as a state-of-the-art deep learning algorithm that surpasses the constraints of the RNN model by adeptly capturing extensive dependencies across sequences and enabling parallel processing. The self-attention mechanism within the Transformer enables the model to concentrate on important information. Qian et al. (Qian et al., 2023) developed a Transformer model, named Bloomformer-1, to identify the drivers of algal growth in freshwater without requiring extensive prior knowledge or experiments. Four traditional ML models, including Extra Trees Regression (ETR), Gradient Boosting Regression Tree (GBRT), SVR, and MLR, were used to compare the model performance with Bloomformer-1. The results show that Bloomformer-1 exhibited the best performance with an R^2 value of 0.85 and an RMSE of 0.35. The data used for model development were collected from the Henan and Hebei sections of China between August 2018 to August 2022, and included water parameters such as TP, PO_4^{3-}-P, TN, NO_3^--N, NH_4^+-N, COD, and total organic carbon (TOC). The target variable used was Chl-*a*, which served as an indicator of phytoplankton biomass.

Deep learning models have a high level of complexity and demonstrate strong performance across various fields, leading to their widespread use. However, it's not always the case that higher complexity models exhibit superior performance. Results can vary based on the characteristics of each region and the data being used. Therefore, research involving a variety of models such as ANN, SVM, and ensemble models, in addition to deep learning models, continues to be conducted for algal prediction up to the present. Yi et al. (Yi et al., 2019) developed two models, M5P, a tree-based model, and extreme learning machine (ELM), to predict short-term algal bloom in the Youngsan River, South Korea. The models were developed using a dataset that included daily measurements of water temperature, rainfall, solar radiation, TN, TP, N/P ratio, and Chl-*a* from January 2013 to December 2016. The models predicted Chl-*a* levels after 1, 3, 5, and 7 days. The M5P model showed the highest performance in predicting Chl-*a* after one day, while the ELM model demonstrated better capability for Chl-*a* prediction spanning 1–7 days. In a period of rapidly increasing algal blooms, the ELM model showed higher accuracy than the M5P model. The R^2 values for the M5P and ELM models were 0.83, 0.46, 0.44, 0.39 and 0.87, 0.59, 0.48, and 0.40 after 1, 3, 5, and 7 days, respectively.

Mamun et al. (Mamun et al., 2019) developed ML models to predict algal Chl-*a* and water clarity in reservoirs during 2000–2017, influenced by the Asian monsoon, using MLR, SVM, and ANN models. Monthly observations of electrical conductivity (EC), DO, water temperature, TN, TP, BOD, COD, total suspended solids (TSS), Chl-*a*, precipitation, and transparency were used for model development. The SVM model performs better than the MLR and ANN models in predicting the values of Chl-*a* and transparency. The model performance was evaluated using RMSE, $R^{2,}$ and MAE for three zones: riverine, transitional, and lacustrine zone, where the R^2 of SVM were 0.75, 0.73, and 0.80, respectively, in three sites for validation data. The model accuracy was also compared in three seasons: pre-monsoon (January to June), monsoon (July to August), and post-monsoon (September to December), where the R^2 of SVM was 0.56, 0.63, 0.80 for three seasons for validation data, respectively. The analysis of the relative importance of input variables presents that water temperature, TP, TN, nutrient ratios (e.g., N/P), and rainfall are important in predicting Chl-*a* and transparency in the reservoir. Amorim et al. (Amorim et al., 2021) developed a model to predict Chl-*a* concentrations using measured water and environmental parameters, including Secchi depth, salinity, water temperature, NO_3, PO_4, Chl-*a*, zooplankton abundance, sunlight duration, and wind speed. However, the field observations had many missing values and thus interpolated daily dataset of 13 years from November 2001 to April 2015 was obtained and used for modeling. Three machine learning algorithms, namely RF, support vector regressor (SVR), and neural networks MLP regressor, were used, and SVR showed the best performance among the three algorithms. Deng et al. (Deng et al., 2021) developed models to predict algal growth and eutrophication in Tolo Harbour, Hong Kong, using ANN and SVM. The results showed that both methods were effective, with ANN providing faster results and SVM offering greater accuracy with longer training times. Water quality indicators, including total inorganic nitrogen (TIN, mg/L), phosphorus (PO_4, mg/L), Chl-*a* (μg/L), DO (mg/L), water temperature (°C), and Secchi depth (m), were used for model development to predict Chl-*a* concentration. The model performance was evaluated using RMSE, and the correlation coefficient was used to measure the goodness of fit between observation and model prediction. Yu et al. (Yu et al., 2021) developed an ML-based method to predict algal blooms using environmental parameters. Five algorithms were used for model development, including Adaptive Boosting (AdaBoost), ANN, GBDT, K-nearest Neighbor (KNN), and SVM. The method performance was validated on real datasets from two locations in the US and China. Results show that the developed ML method effectively predicts short-term concentrations by selecting appropriate features, providing insight into crucial factors for HAB outbreaks. The R^2 ranged from 0.939 to 0.956 for the model developed from weekly-basis water quality data. AdaBoost demonstrated the best model performance with an R^2 of 0.956.

Mozo et al. (Mozo et al., 2022) utilized three ML models, namely RF, Linear Regression (LR), and Classification and Regression Trees (CART), to develop the Chl-*a* soft-sensor. The models were trained and tested using various data aggregation techniques to enhance their inference performance. The soft-sensors were designed using compact and energy-efficient ML algorithms to infer Chl-*a* fluorescence with low-cost input variables that can be deployed on buoys with limited battery and hardware resources. The model was built using field observations collected over three years at 15-min intervals from two different areas of As Conchas freshwater reservoir in northwest Spain, where the four variables (i.e., pH, EC, water temperature, and system battery) were used as independent variables while Chl-*a* was used as the target variable for prediction. Liu et al. (Liu et al., 2023) utilized RF to predict phytoplankton community shifts based on multi-source environmental factors. The RF models accurately predicted algal communities in Lake Mjosa, Norway's largest lake, with hydro-meteorological variables being the most influential factors. The input variables used in the model included multi-source environmental factors such as physicochemical (e.g., TN, TP, and water temperature), hydrological (e.g., input and output discharge, and discharge difference), meteorological (e.g., precipitation, air temperature, and sunshine duration), and spatial factors (e.g., latitude and longitude). The target variables were the composition and biomass of phytoplankton communities. The analysis of feature importance on model performance revealed that antecedent hydro-meteorological factors were the most important factors.

Over the past decade, various machine learning models such as ANN, RF, GBDT, and SVM, along with deep learning models such as LSTM, have been utilized for predicting algal blooms and various water quality parameters. Despite the recent surge in the use of various deep learning models like LSTM, it is observed that relatively simpler models such as ANN, SVM, and RF remain prevalent. In practice, the acquisition of all necessary data for model construction is limited, and the available data that can be collected in the field are typically utilized for building ML models. Given this reality, selecting the most suitable model that aligns with the characteristics of the input data can improve the efficiency of ML models in the field. Additionally, continuous efforts to acquire high-quality field data over extended periods are essential for enhancing the efficiency of algal bloom management using ML models.

2.4. *Classification models for algal bloom prediction*

Prediction of algal bloom status or level is also important for algal bloom management strategy including early warning, and thus various ML models were also used for the development of classification model to forecast algal bloom status or level (Table 2).

SVM and tree-based ensemble models such as RF and GBDT are popular algorithms used for the development of models for algal bloom classification, where the status or class of algal bloom is determined by the occurrence level of Chl-*a* or algal cell numbers. Xia et al. (Xia et al., 2020) used a gradient boosting machine (GBM) model to predict algal blooms in a large river in China, using various water quality parameters (e.g., water temperature, TN, and TP) and hydrologic variables (e.g.,

Table 2
A summary of classification ML models to predict algal blooms.

Model	Input variables	Target variables	Performance evaluation	Ref.
GBM	10-day water quality data including TN, TP, water temperature, and daily hydrological data including water level, flow velocity	Algal bloom level (binary)	Median Kappa of 0.9 for the best GBM model	(Xia et al., 2020)
ANN, SVM	A total of 14 water quality, metalogical, and hydrological variables(weekly)	Algal alert level for early warning of blooms	Accuracy, sensitivity, specificity, precision were 0.81, 0.86, 0.79, 0.72 for ANN, and 0.73, 0.86, 0.64, 0.62 for SVM	(Park et al., 2021)
ANN, SVM	Meteorological data (air temperature, accumulated precipitation), hydrodynamic data (inflow, discharge, water level), and water quality data (TDN, NO$_3^-$-N, NH$_4^+$-N, TDP, PO$_4^{3-}$-P and conductivity) where observation frequency ranges 4–29 days with an average of 7.4 days	Algal alert level	Precision, recall, accuracy Precision ranges 45.2–92.9 The best precision was observed using ANN with synthetic-added dataset.	(Kim et al., 2021)
SVM, RF, MLP (multilayer perceptron)	Water temperature, transparency, water color, DO, conductivity, turbidity, pH, SS, COD, TN, TP, Chl-*a* etc. with observation frequency of four times per year.	Occurrence of *Microcystis* blooms	Accuracy of SVM 0.950, RF 0.924, MLP 0.792 and F-measure of SVM 0.863, RF 0.677, MLP 0.538	(Mori et al., 2022)
RF, GBDT, Naïve Bayes (NB), fusion of RF-GBT	physical and chemical properties of the microalgae, including cell count, biomass weight, pH, ORP, temperature, CO$_2$ concentration in air, and dissolved oxygen	Classification of three microalgae varieties	Accuracy, recall, specificity, and precision with accuracy ranges from 86.11 to 93.11 % The best accuracy using the fusion of RF-GBT model	(Koc et al., 2023)

water level and stream flow discharge) for model development. Algal bloom is defined by algal density into two groups ($\leq 10^7$ cells/ml and $>10^7$ cells/ml). For model development, 10-day water quality data (e.g., TN, TP, algal density, and Chl-*a*) at three sections and daily hydrological data of water level, flow velocity, and streamflow rates from 2003 to 2014 were used. Two GBM models were developed, using explanatory variables from the current 10-day (GBMc model) or previous 10-day period (GBMp model). The model performance was evaluated using accuracy, Cohen's Kappa statistic where the results showed that GBMp showed higher accuracy with a median Kappa of 0.9. Park et al. (Park et al., 2021) developed ML models for an early warning system to predict HABs in a freshwater reservoir to protect the aquatic ecosystem and human health. ANN and SVM models were used to predict algae alert levels based on intensive water quality, hydrodynamic, and meteorological data. The study applied sensitivity analyses for the input variables and optimized the parameters of the models. The results indicated that the ANN model performed better than the SVM model and determined 6- and 7-day sampling intervals as efficient early-warning periods. The model was developed using 14 input variables of water quality, meteorological data, and hydrodynamic data including total dissolved nitrogen (TDN), NO$_3^-$-N, NH$_4^+$-N, total dissolved phosphorus (TDP), PO$_4^{3-}$-P, conductivity, sampling interval, water level of reservoir, inflow and discharge of lake, discharge for hydropower, precipitation, air temperature, and wind speed. The algal alert level was considered as the target variable for prediction. Water quality data was generally collected weakly from 2013 to 2019, and weekly averaged values of metalogical and hydrological data were used for model development. For the evaluation of the model performance, accuracy, sensitivity, specificity, and precision were 0.81, 0.86, 0.79, and 0.72 for ANN, and 0.73, 0.86, 0.64, and 0.62 for SVM, respectively.

Recently, class imbalance in data has been recognized as one of the factors that can degrade model performance. Research on improving model performance by addressing this issue has been consistently conducted (Kim and Park, 2023; Kim et al., 2021). Kim et al. (Kim et al., 2021) also developed an early warning system using ANN and SVM models based on meteorological (e.g., air temperature and accumulated precipitation), hydrodynamic (e.g., inflow, discharge, and water level), and water quality data (TDN, NO$_3^-$-N, NH$_4^+$-N, TDP, PO$_4^{3-}$-P, and conductivity). Due to an imbalance in alert level data, the adaptive synthetic (ADASYN) sampling method was employed to enhance prediction performance. The study showed that combining original and synthetic data improved the model performance in predicting critical alert levels. The model precision ranged from 45.2 to 92.9, with the best precision observed using ANN with a synthetic-added dataset. The improved models can aid in designing management practices to mitigate algal blooms within reservoirs.

In recent years, new approaches have emerged to enhance the performance of ML models, such as selecting input variables with a higher relative effect on model performance or fusing multiple models (Mori et al., 2022; Park et al., 2022b). Mori et al. (Mori et al., 2022) developed an ML model for predicting the occurrence of *Microcystis* in water reservoirs using water quality data. The data observed between 2004 and 2020, with an observation frequency of four times per year, was used for model development. The model uses feature engineering and selection to improve accuracy, and the input independent variables include various water quality parameters (e.g., temperature, transparency, water color, DO, conductivity, turbidity, pH, SS, COD, TN, TP, and Chl-*a*). The target variable is the occurrence of *Microcystis* blooms. SVM, RF, and MLP were used for prediction, with accuracy and F-measure as evaluation indices. The results show that the model performance was improved by feature engineering and feature selection of input variables. Koc et al. (Koc et al., 2023) cultivated three different microalgae species (i.e., *Chlorella kessleri*, *Botryococcus braunii* and *Synechococcus leopoliensis*) using various light sources and collected data on essential cultivation parameters. A 10-fold-cross validation method was employed to partition the algal growth dataset, and three machine learning algorithms—RF, Gradient Boosted Trees (GBT), and Naïve Bayes (NB), —were utilized. The model was developed using the physical and chemical properties of the microalgae, including cell count, biomass weight, pH, ORP, temperature, CO$_2$ concentration in air, and DO as independent input variables to classify microalgae varieties. The researchers also computed an RF-GBT fusion to enhance accuracy. The best prediction, with a 93.11 % accuracy, was achieved using the RF-GBT fusion-based algorithm, while RF, GBT, and NB had accuracies of 93.06 %, 90.28 %, and 86.11 % respectively.

Similar to regression models, classification models have also been developed using a variety of algorithms from relatively simpler algorithms like ANN and SVM to deep learning. Algal bloom levels were often used as the target variable of these ML models, making classification models particularly suitable for such applications. The performance of classification models is greatly influenced by the characteristics of the input data, especially considering the challenges of monitoring natural phenomena where acquiring data of desired concentrations is limited. Thus, future improvements in model performance could be achieved through research efforts aimed at improving the composition of input data, such as addressing data imbalance through additional preprocessing algorithms.

2.5. Automated ML for algal bloom prediction

Automated ML (AutoML) is a field of study that focuses on creating algorithms and tools that automate the end-to-end process of developing

ML models, from data preprocessing to model selection, optimization of hyperparameters, and model evaluation. Neural architecture search (NAS) is a representative AutoML approach that searches an optimal structure of neural networks (Zoph and Le, 2016). Recently, various open-source libraries have been used for the development of ML models for the prediction of water quality. Prasad et al. (Prasad et al., 2021) developed two AutoML models using the “mljar-supervised” algorithm and the Tree-Based Pipeline Optimization Tool (TPOT). The models were created using 9 parameters and around 5000 records of field observations collected between 2009 and 2019. The 9 parameters included TDS, turbidity, pH, COD, iron, phosphate, sodium, chloride, and nitrate, and were used to calculate the water quality index. The dataset was balanced between classes using the Synthetic Minority Oversampling Technique (SMOTE). The results showed that the AutoML and TPOT models achieve higher accuracy of 1.4 % and 0.5 %, respectively, compared to conventional ML techniques for binary and multi-class water data. TPOT was also found to be 0.6 % more accurate than conventional ML techniques for multi-class water data. Auto H2O and Auto-sklearn are also representative open-source auto ML libraries (Feurer et al., 2020; LeDell and Poirier, 2020). Both libraries provide detailed reports on the specific performance of individual models selected in automated ML, as well as the weights of individual models within the final auto ML model developed from the ensemble of those individual models. Fig. 2 presents an example of a variable importance heatmap of input variables for a model to predict Chl-*a* concentration. The heatmap (Fig. 2) is a part of the result report in the auto H2O model, where the relative importance of input variables for each individual model included in the auto H2O model is presented. The model was developed using the open-source auto H2O library (LeDell and Poirier, 2020). Field observation data from the Miho River monitoring station, covering the period from April 1, 2016, to December 31, 2021, and reported in the Water Environmental Information System, which is managed by the National Institute of Environmental Research of Korea, were used for model development. The color scale bar in Fig. 2 indicates the relative variable importance. A variable with higher importance in the heatmap (e.g., red color with higher number) has a greater impact on model performance.

AutoML improves the usability of ML models, allowing non-expert developers to use ML models with relatively simple preprocessing. It is believed that continuous research aimed at enhancing the usability of ML models, similar to AutoML, will be necessary to enable the broader use of advanced ML models in algae bloom management.

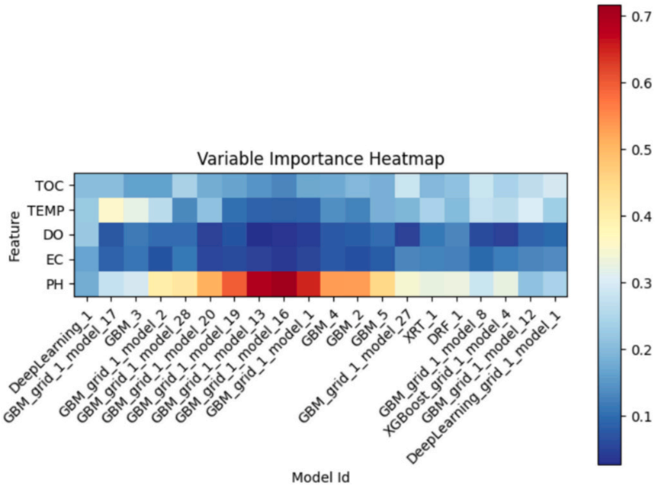


Fig. 2. An example of variable importance heatmap of auto H2O model.

3. Detection and enumeration of algae using image-based ML

3.1. Image-based algal detection overview

Given the diversity in harmfulness among different algal species, quantitative analysis of the types and quantities of algal bloom species is crucial for effective algal bloom management. Conventionally, for image-based algal detection, a well-trained algae taxonomist classifies and counts the algae species found in microscopic images to investigate the algae populations at a given time and geographic location. However, there is a shortage of such experts and their classification accuracy is around 67–83 % due to the large diversity of algae species (there are over 30,000 species) (Xu et al., 2022). Furthermore, manual counting and classification are very time-intensive. Thus, there is a growing effort to replace this conventional method with automated ML models using different types of imaging strategies. Furthermore, image-based ML technology is expected to continue developing in the future. By updating various field images, continuous improvement and increased accuracy in field algal detection technology are also expected.

Over the past decade, object detection technology using ML has shown significant advancements. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) has been a great contributor to the development of ML models for image classification. As a CNN architecture of deep learning models used for computer vision application, Residual Neural Network (ResNet), which was the winner of the 2015 ILSVRC competition, demonstrated the practical application of ML models for image classification by achieving remarkable performance with a top-5 error rate of <5 % (Alyafeai and Ghouti, 2020). Until recently, CNN was one of the most fundamental algorithms used for object detection. CNN extracts features from input data through two processes: convolution and pooling (Krizhevsky et al., 2017; LeCun et al., 1998; Sultana et al., 2020; Zeiler and Fergus, 2014). The first step in a CNN is the convolutional process, where the features of the input image are extracted using a convolution kernel that slides along the input feature matrix. The output of the convolutional process is then reduced in dimension through a subsequent pooling process.

Earlier microscopy-based studies explored how ML algorithms can be used to distinguish and classify different species of algae, and count their number of cells accurately (Chen et al., 2020; Qian et al., 2020; Ruiz-Santaquiteria et al., 2020; Suh et al., 2021). After establishing that different ML algorithms (e.g., CNN, R-CNN, LR, and SVM) are effective for accurate counting and classification of algae in microscopic images, subsequent research endeavors focused on optimizing the code (e.g., YOLO, AlgaeFiner, and deep CNN) to reduce the computational time and improve the capability to differentiate among various genera (Abdullah et al., 2022; Gong et al., 2023; Liu et al., 2022; Park et al., 2022a; Xu et al., 2022; Zhou et al., 2023). Emphasis was also placed on improving the practicality of the system. For instance, one study created an algorithm from the YOLO model that can better recognize algae species from low magnification images which reduces costs as high magnification microscopes are expensive. Other studies made the code more lightweight through the implementation of better algorithms by developing an algal self-organized detection system, which reduces the analytical time and computational power required (Gong et al., 2023). Furthermore, several studies aimed to expand the range of genera that can be precisely classified (Gong et al., 2023; Park et al., 2022a; Qian et al., 2020; Xu et al., 2022).

There are also a few studies that use ML to automatically analyze non-microscopic images of algae, such as from ship and coastal surveillance cameras (Wang et al., 2022b; Zou et al., 2022), unmanned aerial vehicles (Yang et al., 2022), or satellites (Hill et al., 2020). These studies can detect macro-algal populations (e.g. clumps of floating algae) from the collected images/videos. Table 3 provides an overview of the image-based ML algae detection studies that will be discussed in this section.

Overall, the benefits of using various image-based ML algorithms for

Table 3
A summary of algal detection using image-based ML.

HABs detection method	Algorithm(s) used	Number of genera/ species classified	Average precision	Image acquisition method	Reference
Algae classification and counting from microscope images	SegNet, Mask-RCNN[a]	10 species (diatoms)	85 %	Microscopy (60× magnification)	Ruiz-Santaquiteria et al. (2020)
	LR[b], SVM[c], and XGBoost[d] combined	9 species (red tide algae)	96 %	Microscopy (40× magnification)	Chen et al. (2020)
	Faster RCNN[a]	27 genera	74.64 %	Microscopy (high magnification)	Qian et al. (2020)
	Weighted mask RCNN	1 genus (*Microcystis*)	92.5 %	Microscopy (200× and 400× magnification)	Suh et al. (2021)
	YOLO[e]	2 genera (*Chlorella* and *Isochrysis*)	Up to 80 %	Microscopy (low magnification – 10×)	Liu et al. (2022)
	YOLO	27 genera	89.8 %	Microscopy (high magnification)	Park et al. (2022a)
	YOLO	4 genera	91.0 %	Microscopy (high magnification)	Abdullah et al. (2022)
	Modified CNN	13 genera	93 %	Publicly available microscopy pictures (i.e., online algae databases)	Xu et al. (2022)
	TOOD[f], YOLO, RCNN tested	6 genera	82.6 % (YOLO model)	Microscopy (40× magnification)	Zhou et al. (2023)
	YOLO	54 genera	70.6 %	Microscope images using automatic image acquisition equipment: Algae-Hub (20× and 40× magnification)	Gong et al. (2023)
Algae communities from naval, aerial, and satellite images	CNN	1 species (*K. brevis*)	86 % prediction accuracy up to 8 days in future	Satellite imaging + historical records of HABs	Hill et al. (2020)
	AlgaeFiner	2 species (*Ulva prolifera* and *Sargassum*)	45–49 %	Offshore and ship surveillance cameras	Zou et al. (2022)
	AlgaeMask	2 species (*Ulva prolifera* and *Sargassum*)	45 %	Surveillance videos	Wang et al. (2022b)
	RecepNet (semantic segmentation)	Blue-green algae	82 %	Unmanned aerial vehicle (UAV) images	Yang et al. (2022)

[a] RCNN or R-CNN: Region-based Convolutional Neural Network.
[b] LR: Logistic Regression.
[c] SVM: Support Vector Machine.
[d] XGBoost: Extreme Gradient Boosting.
[e] YOLO: You Only Look Once.
[f] TOOD: Task-aligned one-stage object detection.

detection and enumeration of algae are a significant reduction in time and improved accuracy across diversely populated samples. This leads to significant cost reduction in sample analysis and enables faster data collection and processing. Being able to obtain more data can allow for better predictions and management of HABs. While current research indicates promising results for the future, some improvements such as increasing the number of genera detection and the accuracy of detection can still be made before practical use.

3.2. *Segmentation for automatic enumeration and identification of single algal cells from microscopic images*

To classify and enumerate algae cells from microscope images, ML algorithms must first identify single cells using a process called segmentation (Ruiz-Santaquiteria et al., 2020). This allows the software to automatically detect objects (e.g., single cells) in a microscopic image of a water sample by drawing a boundary around each object to select it. The shape created by the boundary, which contains all the pixels inside, is called a Region of Interest (RoI) (Ruiz-Santaquiteria et al., 2020). Once the individual cells are selected, features can be extracted from each cell such as texture, shape, and size (Chen et al., 2020). The process of segmentation facilitates cell counting and even cells aggregated in clusters can be accurately enumerated, while feature extraction can classify the cells by species.

There are two segmentation techniques used for algal monitoring: instance segmentation and semantic segmentation. In instance segmentation, a deep neural network is trained to recognize and distinguish between various objects, while in semantic segmentation, a segmentation mask is used to isolate a specific portion of an image from the rest of an image using parameters specified by the user such as object size range and pixel intensity range. A mask is a file or variable that has the locations of the RoIs (Ruiz-Santaquiteria et al., 2020). Ruiz-Santaquiteria et al. (2020) showed that instance segmentation is more accurate than semantic segmentation in picking out and counting individual diatom cells (Ruiz-Santaquiteria et al., 2020). They used the SegNet model as the semantic segmentation model and the Mask Region-based Convolutional Neural Network (Mask R-CNN) as the instance segmentation model. They found that although their instance segmentation model has greater average precision in classifying algae species, it has lower detection ability and thus misses some cells in the image compared to their semantic segmentation model (Ruiz-Santaquiteria et al., 2020). Using instance segmentation, (Suh et al., 2021) used weighted Mask R-CNN to improve boundary distinguishment between a collection of objects (i.e., objects with touching boundaries). However, this technique required obtaining an optimal value for a parameter and they only managed with one genus of algae: *Microcystis*.

Through the integration of various ML algorithms, classification accuracies can be improved (Chen et al., 2020; Qian et al., 2020). In Chen et al. (2020), three algorithms consisting of Logistic Regression (LR), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost) were used to classify algae images based on extracted features. Their model achieved over 95 % segmentation efficiency and 96 % classification accuracy with 200 test images of red tide algae species (Chen et al., 2020).

3.3. *Improving region-based CNN (R-CNN) algorithm for faster detection*

R-CNN is considered one of the early deep learning algorithms for object detection with a two-stage process (Girshick et al., 2014). In the first stage of R-CNN, the model proposes regions where the target object is located using a selective search algorithm. A CNN is then used to extract the features of the proposed regions, and these features are classified using an SVM. In R-CNN, each proposed region is individually passed through a CNN to extract features, which is computationally expensive and time-consuming.

Fast R-CNN addresses this issue by using a RoI pooling layer to extract features from the entire feature map, rather than applying the CNN to each proposed region individually (Girshick, 2015). The RoI pooling layer takes the proposed regions as inputs and extracts a fixed-size feature map for each region, allowing the CNN to be applied only once to the entire input image. This reduces the number of computations and makes Fast R-CNN significantly faster than R-CNN. Faster R-CNN introduces a Region Proposal Network (RPN), which shares convolutional features with the object detection network (Ren et al., 2015). The RPN generates object proposals much faster than the selective search algorithm used in R-CNN and Fast R-CNN, resulting in faster training and testing times. This allows Faster R-CNN to achieve state-of-the-art performance on object detection tasks with improved speed.

Qian et al. (2020) applied Faster R-CNN to a large dataset of colored microscopic images which contained 27 genera of algae including cyanobacteria with a mean average precision of 74.6 % (Qian et al., 2020). Although the average precision was lower than Chen et al. (2020) (discussed in the previous section), being able to differentiate between many genera of algae is a significant improvement as field samples tend to have a great diversity of algae species.

3.4. *YOLO models for more efficient single algal cell detection*

YOLO models are proposed to be better than CNN models for identifying and classifying objects in microscopic images of microalgae (Abdullah et al., 2022; Park et al., 2022a). The YOLO models were recently developed in 2016 for algal image detection to improve classification accuracy and inference time (Redmon et al., 2016). Over time, different versions have been developed, which include a tiny version of each model using a smaller number of convolutional layers to reduce processing time compared to standard models (Bochkovskiy et al., 2020; Jiang et al., 2020; Park et al., 2022a).

The YOLO algorithm is considered to be the first one-stage algorithm to include region proposal and object classification in a single stage (Redmon et al., 2016). The first version of YOLO, YOLO v1, reduced the inference time for object detection by processing the region proposal and object detection in a single stage, but it showed relatively low accuracy compared to other two-stage models. Since then, the YOLO models have been improved with versions 2 and 3 (Redmon and Farhadi, 2017; Redmon and Farhadi, 2018). To assess object detection models (e.g., R-CNN and YOLO), the mean average precision (mAP) is used, and a higher score indicates greater accuracy in the model detections. YOLO v3 showed comparable accuracy with a mAP of intersection over a union threshold of 0.5 (mAP-50) ranging from 51.5 to 57.9. This level of accuracy is similar to other two-stage models while having several times faster inference time than other two-stage models (Redmon and Farhadi, 2018). YOLO is a representative one-stage object detection model, and since then, new versions have been continuously developed and used in various fields including algal detection.

Recently, several independent studies used YOLO models to classify and count algal cells from microscope images. Fig. 3 shows the representative workflow of using YOLO models to analyze images of microalgae. First, the microscope images are acquired and labeled by hand to facilitate the training of various YOLO models. Then, additional microscope images are utilized to evaluate the ability of the diverse models in classifying and counting the algal cells present.

Liu et al. (Liu et al., 2022) used YOLO to classify and count ocean microalgae from low-magnification images (10×). As high-magnification images are more expensive to acquire and contain fewer algae cells per image, being able to accurately classify and count algae cells from a lower magnification image is attractive. The authors used an improved algae-YOLO object detection approach to automatically count single algae cells from low magnification with an 82.3 % reduced parameter space size without loss of accuracy (Liu et al., 2022). However, they only analyzed images of lab-grown cultures containing one species of algae (either C*hlorella* sp. or *Isochrysis* sp.) which limits their model for field application as natural water samples will contain diverse types of algae species. Despite this, being able to accurately detect cells at lower magnification and significantly reduce the parameter space can be built upon in future studies to help bring automated detection of algae one step closer to practical use (e.g., cost reduction).

Park et al. (Park et al., 2022a) were able to accurately classify and count 30 algae genera using 4 different versions of YOLO (YOLO.v3, YOLO.v3 tiny, YOLO.v4, and YOLO.v4 tiny). The tiny versions are built off the standard version to be faster but less accurate by decreasing the number of convolutional layers. However, the study found that YOLO.v4 tiny achieved the highest mAP at 89.8 %, while also being able to analyze the images the quickest (4 fps). Abdullah et al. (2022) also tested YOLO.v3, YOLO.v4, and YOLO.v5 with 4 algal species (*Cosmarium*, *Closterium*, *Scenedesmus*, and *Spirogyra*) and obtained a slightly higher mAP (90.1 %) using YOLO.v5 compared to the similar study (Park et al., 2022a).

Recently, Gong et al. (Gong et al., 2023) compared the performance of various versions of YOLO models, including versions 5 to 7, for the classification of 53 algal genera. The mAP-50 of the YOLO models ranged from 56.1 % to 70.6 %, with YOLO v7 showing the best model performance with a mAP-50 of 70.6 %.

Zhou et al. (Zhou et al., 2023) also used YOLO algorithms to classify and count algae genera even when they are in different physiological states (e.g., bleaching, translating, or normal) and compared with various other algorithms (i.e., TOOD and RCNN) to find the YOLO model performed best at 82.6 % precision. Various authors have been continuously releasing new versions of the YOLO model, up until YOLO V8 in early 2023.

3.5. *Designing ML models for practical use in microscopic detection of algae*

Manual algal classification from microscopic images often faces

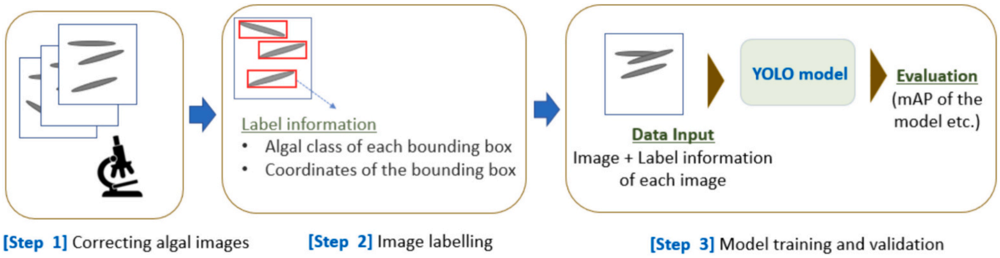

Label information
Algal class of each bounding box
Coordinates of the bounding box
Data Input
Image + Label information of each image
YOLO model
Evaluation
(mAP of the model etc.)
[Step 1] Correcting algal images
[Step 2] Image labelling
[Step 3] Model training and validation

Fig. 3. A procedure for algae classification and counting from microscopic images using YOLO models.

challenges in achieving high accuracies (Xu et al., 2022). Clever design of ML models can allow for high accuracy algae classification and enumeration without extensive model training.

Xu et al. (Xu et al., 2022) utilized deep CNN's ability to differentiate between objects using publicly available microscopic images of different algae species to train their algorithms. Approximately 800 algal images were used along with 400 of these images for testing and 13 different algal genera were classified (Xu et al., 2022). They demonstrated that efficient and accurate algal classification was possible even with a small number of algal images for training a CNN due to a series of technologies that were applied for efficient feature extraction. This has the potential for a national-scale application by collecting public databases (e.g., algal images) without the need for extensive sampling events (Xu et al., 2022).

Gong et al. (Gong et al., 2023) used a self-organized algorithm that does not require a set of training images but rather learns over time as more images are fed. Using this method, they were able to create a dataset that ultimately had 28,329 images with 562,512 single-cell images covering 54 genera. Furthermore, their ML program accurately classified and counted the algae on a 2 cm × 2 cm field image (which corresponds to a 100 μL sample) within 5 min. They also included an interface to upload their results to the cloud to help send warnings about potential algal blooms.

3.6. *Detection of floating algae using ML*

While most studies focus on improving automated counting and detection in microscopic images, ML-based image analysis can also be applied to the images of macroscopic floating algae communities. This method offers the advantage of automatically detecting algae using simple images collected by surveillance cameras on ships, eliminating the need for microscopic images of collected samples. However, this approach has its limitations, as it can only detect species of floating algae, and requires high concentrations to form visible floating communities. Discussed below are two studies that use ML for analyzing images collected from ship and shoreline surveillance cameras.

Zou et al. (Zou et al., 2022) proposed a new instance-segmentation network named AlgaeFiner to monitor floating-macroalgae (specifically *Ulva prolifera* and *Sargassum*), which have recently caused outbreaks within relatively short periods in China. They analyzed RGB images collected from surveillance cameras aboard ships and other facilities along the shores. Mask Transfiner network was added to the AlgaeFiner to enhance the quality of floating-algae segmentation even with images taken in various atmospheric and water conditions (e.g., cloudy, foggy, windy, wavy, sunny/glary, and rainy) with 45–49 % detection precision.

Wang et al. (Wang et al., 2022b) proposed a new algorithm called AlgaeMask which also uses instance segmentation to detect floating algae from ship and coastline surveillance videos. Their algorithm was based on the CenterMask algorithm and was able to achieve up to ~45 % detection precision of *Ulva prolifera* and *Sargassum* as opposed to up to 15 % with CenterMask.

3.7. *Detecting spatiotemporal HABs trends using remote sensing imaging*

In recent years, there have been many remote sensing studies on HABs detection and monitoring including satellite-derived methods. Khan et al. (2021) provides a thorough meta-analysis review on remote sensing studies for HAB detection and monitoring that were published up to 2020. As shown by their meta-analysis, there have not been many studies that utilize ML on remote sensing data (Khan et al., 2021). Using remote sensing imaging is advantageous in that vast areas of water can be monitored. However, similar to images of floating macroalgae from surveillance cameras (discussed in Section 3.6), remote sensing data is limited by its inability to fully characterize the diversity and numbers of algal species in a body of water. In this section, we summarized the ML modeling using remote sensing imaging.

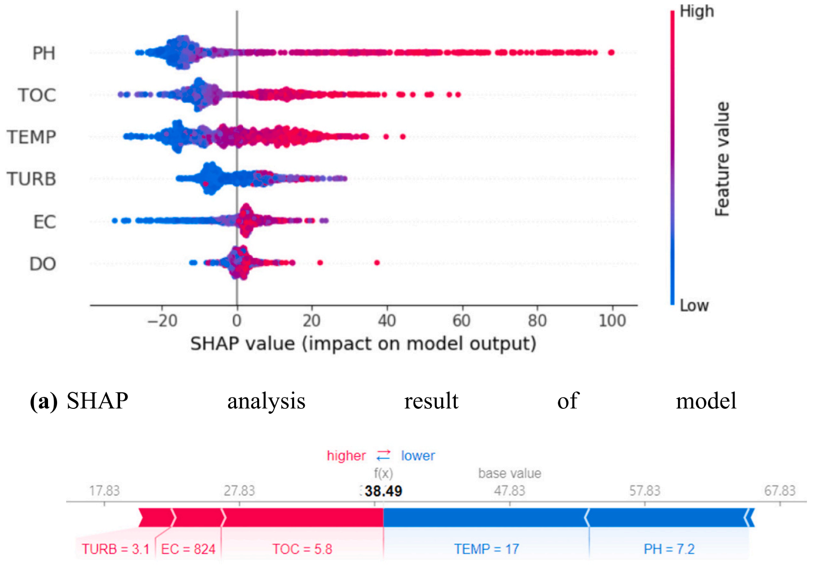
The most common method of HAB detection is currently based on Chl-*a* (Hill et al., 2020). Hill et al. (Hill et al., 2020) used ML to incorporate historical data on HABs outbreaks and reflectance band data from satellite images to detect *Karenia brevis* algae (*K. brevis*) HAB events. Because HAB outbreaks follow well-defined spatiotemporal patterns, those patterns can be used to inform ML algorithms to predict future outbreaks. Specifically, in this study, images taken by satellites measured Chl-*a* from backscattered light of waters to quantify algae concentration. Using this data of backscattered light (i.e., Chl-*a* concentration), a CNN algorithm was trained to compare the spatiotemporal patterns in the data to historical patterns to forecast future HABs. They achieved 86 % prediction accuracy up to 8 days ahead in Floridan coastal waters, which is significantly improved compared to previous models, such as the HAB Operational Forecast System in Florida, which can only predict respiratory-related HAB outbreaks up to 4 days in advance.

Although instance segmentation networks are quite popular due to their accuracy, semantic segmentation networks are less complex and can be designed to reduce computational complexity (Yang et al., 2022). Using a large dataset of unmanned aerial vehicles images, Yang et al. (Yang et al., 2022) developed a real-time semantic segmentation network, RecepNet, based on a bilateral segmentation network (BiSeNetV2) and were able to achieve 82 % mean intersection over union (mIoU) for blue-green algae detection.

4. Explainable artificial intelligence (XAI)

ML models are often referred to as black-box models. Due to the inherent nature of black-box models, interpreting the outcomes of ML models can be challenging. This is regarded as a significant limitation when employing ML models in practical field management and decision-making processes. To address these limitations, various algorithms are used for interpreting model simulation results. Ensemble models such as RF, XGBoost, and LGBM often have internal algorithms to quantify the relative importance of input variables. In recent years, XAI is increasingly used for the interpretation of various factors on target variables in ML models (Adadi and Berrada, 2018; Arrieta et al., 2020; Park et al., 2023). Shapley (SHAP) analysis is a popular XAI algorithm (Lundberg et al., 2018; Lundberg and Lee, 2017).

Fig. 4 is an example of the SHAP analysis result for an XGBoost model to predict Chl-*a* concentration using the field observation data from the Miho River monitoring station, the same data used in Section 2.5 for an example of auto H2O model. The python open source libraries XGBoost (Chen and Guestrin, 2016; XGBoost), SHAP (Lundberg et al., 2018) and Scikit-learn (Pedregosa et al., 2011) were used for development and visualization of the model in Fig. 4. Fig. 4(a) visualizes the SHAP values of input variables used for model training. The y-axis of the graph is determined by sorting variables based on their influence on the model results, with the most influential variable at the top, followed by others in descending order of influence. In Fig. 4(a), each dot represents the SHAP value of an individual observation, where the color of the dots corresponds to the actual observation values. Higher measurement values are represented in red, while lower values are indicated in blue. A positive SHAP value indicates that the measurement of the corresponding variable contributes to an increase in the predicted target value of the model. Conversely, a negative SHAP value means that the variable has an impact that leads to a decrease in the target value from model prediction. For example, as shown in Fig. 4(a), the variable PH has the most significant impact on the model's performance, indicating that higher values of PH tend to contribute to an increase in the predicted values of Chl-*a*. SHAP analysis also allows for detailed interpretation of individual measurements. Fig. 4(b) is an example of SHAP analysis for a certain date of measurement, wherein the recorded TEMP value on the date was 17 °C, resulting in a predicted Chl-*a* concentration of 38.49 mg/m^3. The largest bar scale (i.e., TEMP) indicates the most



(a) SHAP analysis result of model training

(b) SHAP value analysis for an individual observation day.

Fig. 4. An example of SHAP value analysis.
(a) SHAP analysis result of model training.
(b) SHAP value analysis for an individual observation day.

significant impact on the model performance. Additionally, the color of the variable indicates its tendency to either decrease (blue) or increase (red) the predicted Chl-*a* concentration.

Recent studies used SHAP to understand and quantify the effect of various environmental factors on algal blooms (Lee et al., 2022; Park et al., 2022b). Park et al. (Park et al., 2022b) calculated three indices SHAP, feature importance (FI) and variance inflation factor (VIF) to quantify the relative importance of input variables on an XGBoost model to predict Chl-*a* concentration. The water quality monitoring data collected in three parallelly located field stations in Geum River, South Korea between October 2017 and March 2021 were used for model development. The results showed that the model performed most stably when the priority of input variables was determined by SHAP.

The factors influencing algal blooms are diverse and include nutrient levels, weather conditions, and more. These characteristics can vary depending on the overall pollution load and the specific characteristics of the target area. Existing studies also propose various influencing factors on pollution. TN and TP are among the significant contributors, and their impact can vary based on whether concentrations exceed a certain threshold or not. Mamun et al. (2019) analyzed the relative importance of input variables in predicting Chl-*a* using three ML algorithms: MLR, SVM, and ANN. The effect of input variables on model performance varied among different models. Water temperature was found to be the most important variable for MLR and SVM, followed by TN and TP. On the other hand, TSS and BOD were found to be more important for ANN. Xia et al. (2020) analyzed the relative importance of input variables on a GBM model to predict algal blooms using the Varlmp function in a caret package and presented that water level and water temperature were more important than nutrient concentration since the concentration of TN and TP were usually above thresholds and not limiting algal blooms. Ly et al. (Ly et al., 2021) used ML algorithms to predict algal blooms in the Han River, South Korea, using monthly data collected from 40 field stations between 2011 and 2020. Eight different ML algorithms, including RNN, LSTM, GRU, SVM, and decision tree regression (DTR), were compared for their suitability to predict the Trophic State Index (TSI) values based on Chl-*a*. The ML algorithms helped identify the most important water quality parameters contributing to algal bloom prediction, showing that eutrophication and algal proliferation were influenced by the interplay between nutrients, organic contaminants, and environmental factors. Qian et al. (Qian et al., 2023) used an attention mechanism to analyze the driving factors of algal growth in their deep learning-based Transformer model, and the results indicated that TP had the highest effect on algal growth in the Henan section, China, whereas TN had the highest effect on algal growth in the Hebei section, China.

Water temperature is another factor that is consistently identified as one of the key factors influencing algal growth. Baek et al. (2021) utilized a numerical model and ML to identify environmental factors influencing *Alexandrium catenella* blooms through intensive monitoring and DT methods. *A. catenella* is an algal species responsible for red tide, causing paralytic shellfish poisoning. The study found that water temperature was the primary driving factor for *A. catenella* blooms, followed by phosphate concentration and retention time. The DT model revealed that water temperature below 17.2 °C, higher phosphate levels, and increased retention time were key factors in the algal species' growth. These findings can help predict *A. catenella* blooms and inform mitigation strategies. The combination of ML and numerical simulation could be an effective approach for managing *A. catenella* blooms. The total classification accuracy of *A. catenella* bloom levels was 82.25 % for the training set and 75.0 % for the test set (Baek et al., 2021).

Tamvakis et al. (Tamvakis et al., 2021) used ML techniques to predict the presence of 18 potentially harmful marine microalgae at the genus level, based on a small set of abiotic variables identified as drivers of blooms. The RF algorithm accurately identified the presence of most genera, with a mean accuracy of 89.2 % across all samples. Analysis of the input variable importance revealed that temperature had the most significant effect on the presence of genera, where this effect varied among different genera. Lee et al. (Lee et al., 2022) developed a CNN model to predict the concentration of Chl-*a* using eight water quality variables (water temperature, pH, EC, DO, TOC, TN, TP, and Chl-*a*) and four weather variables (e.g., average wind speed) as input variables. Daily water quality data were collected from 40 automatic water quality monitoring stations between April 2015 and December 2018 from four major rivers in Korea. The optimal CNN model showed an R^2 value of 0.934 and an RMSE value of 5.463, respectively. The SHAP analysis was performed to quantify the relative importance of input variables on the model performance, which showed that Chl-*a* in the previous time period had the most significant impact on the prediction, with water temperature, DO, and TP identified as major factors affecting Chl-*a* prediction. Jung et al. (Jung et al., 2023) applied ML to identify the main factors influencing the occurrence of blue-green algae in a stagnant river area. They used an RF model and evaluated its accuracy using validation data. The input independent variables included water quality parameters such as pH, EC, DO, BOD, COD, TP, and Chl-*a*, hydraulic data including outflow and hydraulic retention time (HRT), and meteorological data such as temperature and precipitation. The target variable was the occurrence of blue-green algae. The model was trained on data from 2015 to 2019 and tested on data from 2020 to 2022. The researchers used the mean decrease in Gini to evaluate the importance of each variable in the model. The results showed that overall temperature is the most important factor affecting the occurrence of blue-green algae. The evaluation result also showed that the RF model had high accuracy in predicting the occurrence of blue-green algae in stagnant rivers (Jung et al., 2023).

Overall, prior studies have highlighted that various environmental factors such as water quality and weather conditions from upstream sites can serve as input variables to construct predictive models for algal blooms. Through quantitative analysis facilitated by XAI, the impact of diverse upstream influencing factors can be assessed, shedding light on the underlying causes of algal bloom occurrences including pollution sources. Continuous exploration of XAI for conducting scientific and quantitative analyses of environmental factors affecting water quality, and leveraging these results for decision-making in water quality improvement, holds the potential to expand the applicability of ML-driven models for algal bloom management.

5. Consideration for effective ML-driven HAB management

Here, we demonstrated the current development and utilization of various ML models for HAB prediction and algae detection based on time series data and image analysis using object detection technology. Data-driven models, such as ML, have the advantage of relatively easy initial construction as they do not require the estimation of parameters based on experiments. However, in real-world scenarios, the quality and characteristics of input data, such as the measurement frequency of input data, selection of input variables, and the correlation between measurement items, can significantly impact the performance of the model (Park et al., 2022b). Therefore, first of all, obtaining high-quality and field-representative data is essential for developing desirable models and optimizing overall model performance. Given the intricate internal algorithms of ML models, acquiring a sufficient quantity of high-quality data is necessary to achieve optimal model performance. Real-time monitoring data collected from on-site, field deployable sensors can be highly valuable for ML models. Additionally, regular accuracy management is required to prevent errors or biases in field sensor data. In field monitoring stations, long-term missing data over several months is often observed. Thus, efforts should be made to minimize missing data through proper interpolation of mission data or measures using duplicate measurement devices to enhance data quality for model construction. Furthermore, during the planning and site selection stages of establishing a new field monitoring plan, determining appropriate measurement frequencies, measurement parameters, and measurement locations, while considering the application of ML models can significantly contribute to enhancing the efficiency of water quality management with the use of rapidly evolving ML tools.

In addition, by establishing an integrated database that manages various data on water quantity and quality together, the effectiveness of utilizing ML models can be improved. To predict the occurrence of algal blooms, it is essential to leverage not only water quality data but also a range of measurements including water quantity, weather conditions, and watershed environment. It is believed that the efficiency of data acquisition and quality control on water management can be significantly enhanced through the establishment of a system that integrates and manages data from these different domains.

Exploring XAI presents a promising avenue for expanding the application of ML algorithms to algae bloom management. For example, when constructing an ML model to predict the occurrence of algal blooms, XAI can be employed to perform a quantitative assessment of the factors influencing the increase in algal blooms. This approach facilitates the analysis of the necessary measures to reduce the occurrence of algal blooms. Such analysis results can contribute to enhancing the efficiency of decision-making for algae bloom management, such as prioritizing pollution reduction projects needed to mitigate algal bloom occurrences. Ongoing research into applying XAI for algae bloom management indicates a burgeoning field. Through continuous research on utilizing XAI in the future, it will be possible to advance and improve the efficiency of algal bloom management techniques.

6. Conclusions

The public health and ecological concerns towards HABs have propelled research to better manage HABs and monitor algae populations. For HABs management, it is crucial to analyze the current status and predict future occurrences of algae.

Research into using advanced ML models for algal bloom detection and prediction is relatively recent and has evolved alongside the development history of ML models. As new models are developed, research typically focuses on applying them to the detection and prediction of algal blooms. Initially, early-developed ML models such as ANN and SVM were used in algal bloom prediction. Over time, more advanced models with superior performance, such as ensemble models and deep learning models, have been developed and used for algal prediction. In terms of object detection models, those based on the CNN algorithm have continuously improved in both performance and detection speed. The YOLO model, in particular, has demonstrated excellent performance while reducing inference time, making it one of the most widely used object detection algorithms for algal cell detection to date.

One of the advantages of ML models over traditional mechanistic models in predicting algal blooms is that they eliminate the need to identify physico-chemical-biological factors affecting algal growth through time-consuming and labor-intensive experiments. The application of ML-based image detection and prediction models with excellent performance can reduce the time, manpower, and costs required for field management, enhancing real-time responsiveness. Ultimately, the integration of all available data (e.g., satellite images and on-site real-time sensor data) into ML algorithms can improve algal bloom prediction significantly and efficiently with high accuracy. As discussed, algae blooms follow certain patterns though they may occur randomly and behave differently in geographical areas. One can imagine that the use of ML along with the collection of high quality and high quantities of data on algal populations and the incorporation of data with spatiotemporal patterns (e.g., satellite images) and historical records can significantly improve algal bloom prediction, maybe even to a month or two in advance. It may also be able to combine and incorporate geographic information into the HAB prediction to provide practical implications for the prevention and control of HAB. For example, is the surrounding area of HAB events highly agricultural versus commercial? Are there other geographical considerations (e.g., temperature and altitude) that can influence algal blooms and their dynamics? Advancements in ML algorithms can provide the answers to these questions, leading to the development of a timely and highly accurate prediction tool for HAB management.

As observed in previous studies, ML-based algorithms are rapidly advancing. By reviewing past studies, we can identify directions to further enhance the effectiveness of ML technology. The success of ML models largely depends on the quality of the data used in model construction and how this data is preprocessed and selected for model construction. However, more complex models do not necessarily mean better performance. Therefore, a wide variety of ML models continue to be actively used and research often involves applying multiple ML models together to compare their performances. The key is to choose models that align with the specific characteristics of the input data. AutoML emerges as an innovative approach that facilitates the construction of optimal models suited to the characteristics of the data. As research continues in this field, we can expect both the performance of ML models and their ease of use to improve.

Generally, ML models require sufficient data for training the models, and on-site real-time sensor data from existing monitoring systems is a useful tool for obtaining the necessary data for applying ML models. Securing high-quality data can maximize the effectiveness of advanced ML models. Considering the ongoing increase in the utilization of ML models, it is essential to conduct research and considerations regarding the types of data and acquisition methods (e.g., measurement frequency, locations, and parameters to be measured) that can enhance the effectiveness of model development for field applications. Additionally, the establishment of a systematic platform for the integrated management of data acquired from various fields can enhance the usability of the data and the efficiency of quality control. Further, XAI provides a quantitative and scientific interpretation for the results of ML models, increasing the applicability of ML-based technology in policy development and decision-making for mitigating algal bloom.

CRediT authorship contribution statement

Jungsu Park: Writing – review & editing, Writing – original draft, Methodology, Investigation, Conceptualization. **Keval Patel:** Writing – original draft, Investigation. **Woo Hyoung Lee:** Writing – review & editing, Writing – original draft, Methodology, Investigation, 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.

Data availability

Data will be made available on request.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1F1A1065518) (50 %). This study was supported by a grant from the Ministry of Land, Infrastructure and Transport of the Korean government (22UGCP-B157945-03) (50 %).

References

Abdullah, Ali S., Khan, Z., Hussain, A., Athar, A., Kim, H.-C., 2022. Computer vision based deep learning approach for the detection and classification of algae species using microscopic images. Water 14, 2219.

Adadi, A., Berrada, M., 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6, 52138–52160.

Alyafeai, Z., Ghouti, L., 2020. A fully-automated deep learning pipeline for cervical cancer classification. Expert Syst. Appl. 141, 112951.

Amorim, F.L.L., Rick, J., Lohmann, G., Wiltshire, K.H., 2021. Evaluation of machine learning predictions of a highly resolved time series of chlorophyll-a concentration. Appl. Sci. 11, 7208.

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Moling, D., Benjamins, R., Chatila, R., Herrera, F., 2020. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion 58, 82–115.

Baek, S.-S., Kwon, Y.S., Pyo, J., Choi, J., Kim, Y.O., Cho, K.H., 2021. Identification of influencing factors of A. catenella bloom using machine learning and numerical simulation. Harmful Algae 103, 102007.

Bochkovskiy A, Wang C-Y, Liao H-YM. Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 2020.

Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pp. 144–152.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Chen T, Guestrin C. Xgboost: a scalable tree boosting system. Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.

Chen S, Shan S, Zhang W, Wang X, Tong M. Automated red tide algae recognition by the color microscopic image. 2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2020, pp. 852–861.

Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 2014.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297.

Cruz, R.C., Reis Costa, P., Vinga, S., Krippahl, L., Lopes, M.B., 2021. A review of recent machine learning advances for forecasting harmful algal blooms and shellfish contamination. J. Mar. Sci. Eng. 9, 283.

Deng, T., Chau, K.-W., Duan, H.-F., 2021. Machine learning based marine water quality prediction for coastal hydro-environment management. J. Environ. Manag. 284, 112051.

Erdner, D.L., Dyble, J., Parsons, M.L., Stevens, R.C., Hubbard, K.A., Wrabel, M.L., et al., 2008. Centers for oceans and human health: a unified approach to the challenge of harmful algal blooms. Environ. Health 7, S2.

Feurer M, Eggensperger K, Falkner S, Lindauer M, Hutter F. Auto-sklearn 2.0: The next generation. arXiv preprint arXiv:2007.04074 2020; 24.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 1189–1232.

Girshick, R., 2015. Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587.

Gong, X., Ma, C., Sun, B., Zhang, J., 2023. An efficient self-organized detection system for algae. Sensors 23, 1609.

Gupta, A., Hantush, M.M., Govindaraju, R.S., 2023. Sub-monthly time scale forecasting of harmful algal blooms intensity in Lake Erie using remote sensing and machine learning. Sci. Total Environ. 900, 165781.

Herath, G., 1997. Freshwater algal blooms and their control: comparison of the European and Australian experience. J. Environ. Manag. 51, 217–227.

Hill, P.R., Kumar, A., Temimi, M., Bull, D.R., 2020. HABNet: machine learning, remote sensing-based detection of harmful algal blooms. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 3229–3239.

Hinton, G.E., 2012. A practical guide to training restricted Boltzmann machines. In: Neural Networks: Tricks of the Trade: Second Edition, pp. 599–619.

Hinton, G.E., Osindero, S., Teh, Y.-W., 2006. A fast learning algorithm for deep belief nets. Neural Comput. 18, 1527–1554.

Ho, J.C., Michalak, A.M., 2015. Challenges in tracking harmful algal blooms: a synthesis of evidence from Lake Erie. J. Great Lakes Res. 41, 317–325.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9, 1735–1780.

Jiang Z, Zhao L, Li S, Jia Y. Real-time object detection method based on improved YOLOv4-tiny. arXiv preprint arXiv:2011.04244 2020.

Jung, W.S., Jo, B.G., Kim, Y.D., 2023. A study on the occurrence characteristics of harmful blue-green algae in stagnant rivers using machine learning. Appl. Sci. 13, 3699.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al., 2017. Lightgbm: a highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems, 30.

Khan, R.M., Salehi, B., Mahdianpari, M., Mohammadimanesh, F., Mountrakis, G., Quackenbush, L.J., 2021. A meta-analysis on harmful algal bloom (HAB) detection and monitoring: a remote sensing perspective. Remote Sens. 13, 4347.

Kim, J., Park, J., 2023. Evaluation of multi-classification model performance for algal bloom prediction using CatBoost. J. Korean Soc. Water Environ. 39, 1–8.

Kim, J.H., Shin, J.-K., Lee, H., Lee, D.H., Kang, J.-H., Cho, K.H., et al., 2021. Improving the performance of machine learning models for early warning of harmful algal blooms using an adaptive synthetic sampling method. Water Res. 207, 117821.

Koc, D.G., Koc, C., Ekinci, K., 2023. Fusion-based machine learning approach for classification of algae varieties exposed to different light sources in the growth stage. Algal Res. 71, 103087.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2017. Imagenet classification with deep convolutional neural networks. Commun. ACM 60, 84–90.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86, 2278–2324.

LeDell, E., Poirier, S., 2020. H2o automl: scalable automatic machine learning. In: Proceedings of the 7th ICML Workshop on Automated Machine Learning (AutoML). 2020. ICML.

Lee, S., Lee, D., 2018. Improved prediction of harmful algal blooms in four Major South Korea's Rivers using deep learning models. Int. J. Environ. Res. Public Health 15, 1322.

Lee, D., Kim, M., Lee, B., Chae, S., Kwon, S., Kang, S., 2022. Integrated explainable deep learning prediction of harmful algal blooms. Technol. Forecast. Soc. Chang. 185, 122046.

Li, X., Yu, J., Jia, Z., Song, J., 2014. Harmful algal blooms prediction with machine learning models in Tolo Harbour. In: 2014 International Conference on Smart Computing. IEEE, pp. 245–250.

Lin, S., Pierson, D.C., Mesman, J.P., 2023. Prediction of algal blooms via data-driven machine learning models: an evaluation using data from a well-monitored mesotrophic lake. Geosci. Model Dev. 16, 35–46.

Liu, D., Wang, P., Cheng, Y., Bi, H., 2022. An improved algae-YOLO model based on deep learning for object detection of ocean microalgae considering aquacultural lightweight deployment. Front. Mar. Sci. 9.

Liu, M., Huang, Y., Hu, J., He, J., Xiao, X., 2023. Algal community structure prediction by machine learning. Environ. Sci. Ecotechnol. 14, 100233.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Proces. Syst. 30.

Lundberg SM, Erion GG, Lee S-I. Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888 2018.

Ly, Q.V., Nguyen, X.C., Lê, N.C., Truong, T.-D., Hoang, T.-H.T., Park, T.J., et al., 2021. Application of machine learning for eutrophication analysis and algal bloom prediction in an urban river: a 10-year study of the Han River, South Korea. Sci. Total Environ. 797, 149040.

Madni, H.A., Umer, M., Ishaq, A., Abuzinadah, N., Saidani, O., Alsubai, S., et al., 2023. Water-quality prediction based on H2O AutoML and explainable AI techniques. Water 15, 475.

Mamun, M., Kim, J.-J., Alam, M.A., An, K.-G., 2019. Prediction of algal chlorophyll-a and water clarity in monsoon-region reservoir using machine learning approaches. Water 12, 30.

Mori, M., Flores, R.G., Suzuki, Y., Nukazawa, K., Hiraoka, T., Nonaka, H., 2022. Prediction of Microcystis occurrences and analysis using machine learning in high-dimension, low-sample-size and imbalanced water quality data. Harmful Algae 117, 102273.

Mozo, A., Morón-López, J., Vakaruk, S., Pompa-Pernía, Á.G., González-Prieto, Á., Aguilar, J.A.P., et al., 2022. Chlorophyll soft-sensor based on machine learning models for algal bloom predictions. Sci. Rep. 12, 13529.

Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. Proceedings of the 27th International Conference on Machine Learning (ICML-10), 2010, pp. 807–814.

Park, Y., Lee, H.K., Shin, J.-K., Chon, K., Kim, S., Cho, K.H., et al., 2021. A machine learning approach for early warning of cyanobacterial bloom outbreaks in a freshwater reservoir. J. Environ. Manag. 288, 112415.

Park, J., Baek, J., Kim, J., You, K., Kim, K., 2022a. Deep learning-based algal detection model development considering field application. Water 14, 1275.

Park, J., Lee, W.H., Kim, K.T., Park, C.Y., Lee, S., Heo, T.-Y., 2022b. Interpretation of ensemble learning to predict water quality using explainable artificial intelligence. Sci. Total Environ. 832, 155070.

Park, J., Joo, J.C., Kang, I., Lee, W.H., 2023. The use of explainable artificial intelligence for interpreting the effect of flow phase and hysteresis on turbidity prediction. Environ. Earth Sci. 82, 375.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Prasad, D.V.V., Kumar, P.S., Venkataramana, L.Y., Prasannamedha, G., Harshana, S., Srividya, S.J., et al., 2021. Automating water quality analysis using ML and auto ML techniques. Environ. Res. 202, 111720.

Qian P, Zhao Z, Liu H, Wang Y, Peng Y, Hu S, et al. Multi-target deep learning for algal detection and classification. 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2020, pp. 1954-1957.

Qian, J., Pu, N., Qian, L., Xue, X., Bi, Y., Norra, S., 2023. Identification of driving factors of algal growth in the South-to-North Water Diversion Project by Transformer-based deep learning. Water Biol. Secur. 100184.

Redmon, J., Farhadi, A., 2017. YOLO9000: better, faster, stronger. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 7263–7271.

Redmon J, Farhadi A. Yolov3: An incremental improvement. arXiv preprint arXiv: 1804.02767 2018.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: unified, real-time object detection. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 779–788.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, p. 28.

Rostam, N.A.P., Malim, N.H.A.H., Abdullah, R., Ahmad, A.L., Ooi, B.S., Chan, D.J.C., 2021. A complete proposed framework for coastal water quality monitoring system with algae predictive model. IEEE Access 9, 108249–108265.

Ruiz-Santaquiteria, J., Bueno, G., Deniz, O., Vallez, N., Cristobal, G., 2020. Semantic versus instance segmentation in microscopic algae detection. Eng. Appl. Artif. Intell. 87, 103271.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature 323, 533–536.

Saboe, D., Ghasemi, H., Gao, M.M., Samardzic, M., Hristovski, K.D., Boscovic, D., et al., 2021. Real-time monitoring and prediction of water quality parameters and algae concentrations using microbial potentiometric sensor signals and machine learning tools. Sci. Total Environ. 764, 142876.

Shao, H., Kiyomoto, S., Kawauchi, Y., Kadota, T., Nakagawa, M., Yoshimura, T., et al., 2021. Classification of various algae canopy, algae turf, and barren seafloor types using a scientific echosounder and machine learning analysis. Estuar. Coast. Shelf Sci. 255, 107362.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15, 1929–1958.

Suh, S., Park, Y., Ko, K., Yang, S., Ahn, J., Shin, J.-K., et al., 2021. Weighted mask R-CNN for improving adjacent boundary segmentation. J. Sens. 2021, 8872947.

Sultana, F., Sufian, A., Dutta, P., 2020. A review of object detection models based on convolutional neural network. In: Intelligent Computing: Image Processing Based Applications, pp. 1–16.

Tamvakis, A., Tsirtsis, G., Karydis, M., Patsidis, K., Kokkoris, G.D., 2021. Drivers of harmful algal blooms in coastal areas of Eastern Mediterranean: a machine learning methodological approach. Math. Biosci. Eng. 18, 6484–6505.

Vilas, L.G., Spyrakos, E., Palenzuela, J.M.T., Pazos, Y., 2014. Support vector machine-based method for predicting Pseudo-nitzschia spp. blooms in coastal waters (Galician rias, NW Spain). Prog. Oceanogr. 124, 66–77.

Wang, L., Zhang, T., Wang, X., Jin, X., Xu, J., Yu, J., et al., 2019. An approach of improved multivariate timing-random deep belief net modelling for algal bloom prediction. Biosyst. Eng. 177, 130–138.

Wang, J., Zhou, Y., Bai, X., Li, W., 2022a. Effect of algal blooms outbreak and decline on phosphorus migration in Lake Taihu, China. Environ. Pollut. 296, 118761.

Wang, X., Wang, L., Chen, L., Zhang, F., Chen, K., Zhang, Z., et al., 2022b. AlgaeMask: an instance segmentation network for floating algae detection. J. Mar. Sci. Eng. 10, 1099.

Wen, J., Yang, J., Li, Y., Gao, L., 2022. Harmful algal bloom warning based on machine learning in maritime site monitoring. Knowl.-Based Syst. 245, 108569.

West, J.J., Järnberg, L., Berdalet, E., Cusack, C., 2021. Understanding and managing harmful algal bloom risks in a changing climate: lessons from the European CoCliME Project. Front. Clim. 3, 636723.

Wurtsbaugh, W.A., Paerl, H.W., Dodds, W.K., 2019. Nutrients, eutrophication and harmful algal blooms along the freshwater to marine continuum. Wiley Interdiscip. Rev. Water 6, e1373.

XGBoost. n.d. Available online: https://pypi.org/project/xgboost/.

Xia, R., Wang, G., Zhang, Y., Yang, P., Yang, Z., Ding, S., et al., 2020. River algal blooms are well predicted by antecedent environmental conditions. Water Res. 185, 116221.

Xu, L., Xu, L., Chen, Y., Zhang, Y., Yang, J., 2022. Accurate classification of algae using deep convolutional neural network with a small database. ACS ES&T Water 2, 1921–1928.

Yang, K., Wang, Z., Yang, Z., Zheng, P., Yao, S., Zhu, X., et al., 2022. RecepNet: network with large receptive field for real-time semantic segmentation and application for blue-green algae. Remote Sens. 14, 5315.

Yi, H.-S., Lee, B., Park, S., Kwak, K.-C., An, K.-G., 2019. Prediction of short-term algal bloom using the M5P model-tree and extreme learning machine. Environ. Eng. Res. 24, 404–411.

Yu, P., Gao, R., Zhang, D., Liu, Z.-P., 2021. Predicting coastal algal blooms with environmental factors by machine learning methods. Ecol. Indic. 123, 107334.

Yussof, F.N., Maan, N., Md Reba, M.N., 2021. LSTM networks to improve the prediction of harmful algal blooms in the West Coast of Sabah. Int. J. Environ. Res. Public Health 18, 7650.

Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13. Springer, pp. 818–833.

Zhou, S., Jiang, J., Hong, X., Fu, P., Yan, H., 2023. Vision meets algae: a novel way for microalgae recognization and health monitor. Front. Mar. Sci. 10, 1105545.

Zoph B, Le QV. Neural architecture search with reinforcement learning. arXiv preprint arXiv:1611.01578 2016.

Zou, Y., Wang, X., Wang, L., Chen, K., Ge, Y., Zhao, L., 2022. A high-quality instance-segmentation network for floating-algae detection using RGB images. Remote Sens. 14.