


## Article

# Estimating Chlorophyll-*a* Concentration from Hyperspectral Data Using Various Machine Learning Techniques: A Case Study at Paldang Dam, South Korea

GwangMuk Im <sup>1,†</sup>, Dohyun Lee <sup>1,†</sup>, Sanghun Lee <sup>1,†</sup>, Jongsu Lee <sup>2</sup>, Sungjong Lee <sup>3,\*</sup>, Jungsu Park <sup>4,\*</sup> and Tae-Young Heo <sup>1,\*</sup> 

<sup>1</sup> Department of Information and Statistics, Chungbuk National University, Chungbuk 28644, Republic of Korea

<sup>2</sup> Solution Development Team, Korea Water Resources Corporation, Daejeon 34430, Republic of Korea

<sup>3</sup> Department of Environmental Science, Hankuk University of Foreign Studies, Gyeonggi 17035, Republic of Korea

<sup>4</sup> Department of Civil and Environmental Engineering, Hanbat National University, 125 Dongseo-daero, Yuseong-gu, Daejeon 34158, Republic of Korea

\* Correspondence: sjlee80@hufs.ac.kr (S.L.); parkjs@hanbat.ac.kr (J.P.); theo@cbnu.ac.kr (T.-Y.H.); Tel.: +82-43-261-3741 (T.-Y.H.)

† These authors contributed equally to this work.

**Abstract:** Algal blooms have been observed worldwide and have had a serious impact on industries that use water resources, which is a problem for people and the environment. For this reason, an algae warning system is used to count the number of cyanobacterial cells and the concentration of chlorophyll-*a*. Several studies using multispectral or hyperspectral data to estimate chlorophyll concentration have recently been carried out. In the present study, a comparative approach was applied to estimate the concentration of chlorophyll-*a* at Paldang Dam, South Korea using hyperspectral data. We developed a framework for estimating chlorophyll-*a* using dimension reduction methods, such as principal component analysis and partial least squares, and various machine learning algorithms. We analyzed hyperspectral data collected during a field survey to locate peaks in the chlorophyll-*a* spectrum. The framework that used support vector regression achieved the highest  $R^2$  of 0.99, a mean square error (MSE) of 1.299  $\mu\text{g}/\text{cm}^3$ , and showed a small discrepancy between observed and real values relative to other frameworks. These findings suggest that by combining hyperspectral data with dimension reduction and a machine learning algorithm, it is possible to provide an accurate estimation of chlorophyll-*a*. Using this, chlorophyll-*a* can be obtained in real time through hyperspectral sensor data input from drones or unmanned aerial vehicles using the learned machine learning algorithm.

**Keywords:** hyperspectral; chlorophyll-*a*; machine learning



**Citation:** Im, G.; Lee, D.; Lee, S.; Lee, J.; Lee, S.; Park, J.; Heo, T.-Y.

Estimating Chlorophyll-*a* Concentration from Hyperspectral Data Using Various Machine Learning Techniques: A Case Study at Paldang Dam, South Korea. *Water* **2022**, *14*, 4080. <https://doi.org/10.3390/w14244080>

Academic Editor: Jian Liu

Received: 31 October 2022

Accepted: 6 December 2022

Published: 14 December 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Water is a vital resource for ecosystems and humans alike. Therefore, damage to water resources can cause significant risks in the aquatic ecosystems living nearby, not only humans. In particular, Harmful algal blooms (HABs) have been reported for a long time [1], but it is hard to say that there is still a perfect countermeasure. Abnormal phenomena, such as global warming and weather variability [2], human natural destruction, such as aquaculture industry, and the resulting eutrophication, were suggested as the cause of HAB. Damage caused by HABs has been reported in many lakes and rivers around the world, including in Korea [3–5], owing to climate change caused by global warming [6].

Algal bloom is a phenomenon in which floating algae proliferate in large quantities in eutrophic lakes or in slow-flowing rivers and seas, turning the water green. When this phenomenon is severe, the algae block the sunlight from entering the water, which prevents the photosynthesis of water plants [7]. Additionally, these HABs produce a high

level of biomass and can negatively affect the environment and human activity through seawater discoloration, anoxia, and mucilage induction [8]. Furthermore, when algal bloom is left untreated, it becomes a critical threat to inland water and estuarine environments. HABs negatively affect aquatic ecosystems and public service systems; they cause the mass mortality of fish, animal death, human health hazard, and difficulties in managing food and water supply [9].

Algal blooms that are potentially harmful to the ecosystem occur as a result of environmental pollution in dams, lakes, reservoirs, and rivers all over the world. Algal blooms in South Korea have resulted in the mass growth of cyanobacteria, which threatens the steady and safe supply of drinking water and wreaks havoc on the ecosystem. Although it is imperative to detect and fix the influx of phosphorus, which is the primary cause of algal bloom, there is a growing need to quickly identify the appearance of algal bloom and to remove phosphorus and floating waste. The concentration of chlorophyll-*a* is a commonly used indicator that quantitatively represents the status of algal bloom. The concentration of chlorophyll-*a*, which is one of the pigments required by cyanobacterial for photosynthesis, is an important parameter in investigations on algal bloom [10,11].

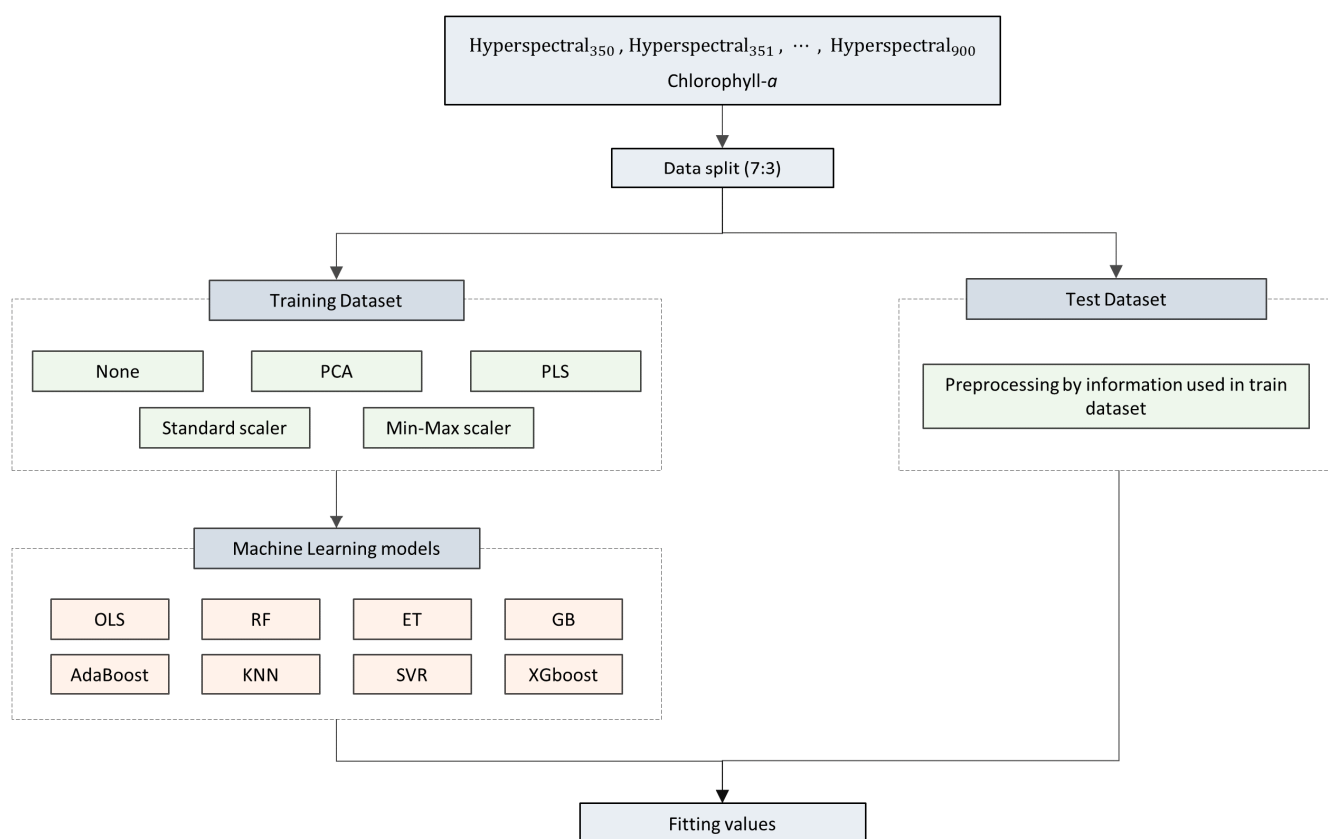
There is a growing body of research that explores the estimation of chlorophyll-*a*. It has been confirmed that existing chlorophyll-*a* estimation studies differ in terms of the type of data and analysis method used, which include mathematical analysis using second-order derivatives of spectra [12] and the use of a regression formula for river data [13]. In terms of the data used, water quality and weather conditions have been used [14], as well as hyperspectral data [15]. In particular, hyperspectral sensors have been widely used for estimating chlorophyll-*a* [16] because of their efficiency in the early estimation of chlorophyll-*a* [17].

Methods that combine machine learning and measured hyperspectral data have been recently developed for estimating chlorophyll-*a*. Sina Keller [18] estimated chlorophyll-*a* applying machine learning techniques, such as principal component analysis (PCA) and extreme tree [19], to hyperspectral sensors and obtained an  $R^2$  of 0.941. We studied how to estimate the actual chlorophyll-*a* using only hyperspectral sensors without using water and weather sensors to reduce costs. Since chlorophyll-*a* is estimated using only hyperspectral sensors, it is easy to measure algae in a larger area using airplanes or drones [20].

In this study, we attempted to develop a model that provides optimal performance in fitting chlorophyll-*a* to hyperspectral data, and we present a framework for estimating chlorophyll-*a* with high performance by applying various machine learning models to hyperspectral data and comparing them from various perspectives.

## 2. Materials and Methods

The data collection, preprocessing, and machine learning methods employed in this study are discussed in this section. Figure 1 shows the workflow, which divides the entire process of this study into several steps. First, the obtained hyperspectral data were separated into training and testing data to estimate and evaluate chlorophyll-*a*. Subsequently, eight different machine learning models were used to learn raw data. Four different methods (standard scaler, min-max scaler, PCA, and partial least squares (PLS) [21]) were used for preprocessing and we could figure out the most appropriate model by comparing each model's evaluation metrics. This section includes data collection methods, data descriptions, a preprocessing method, and a model definition.



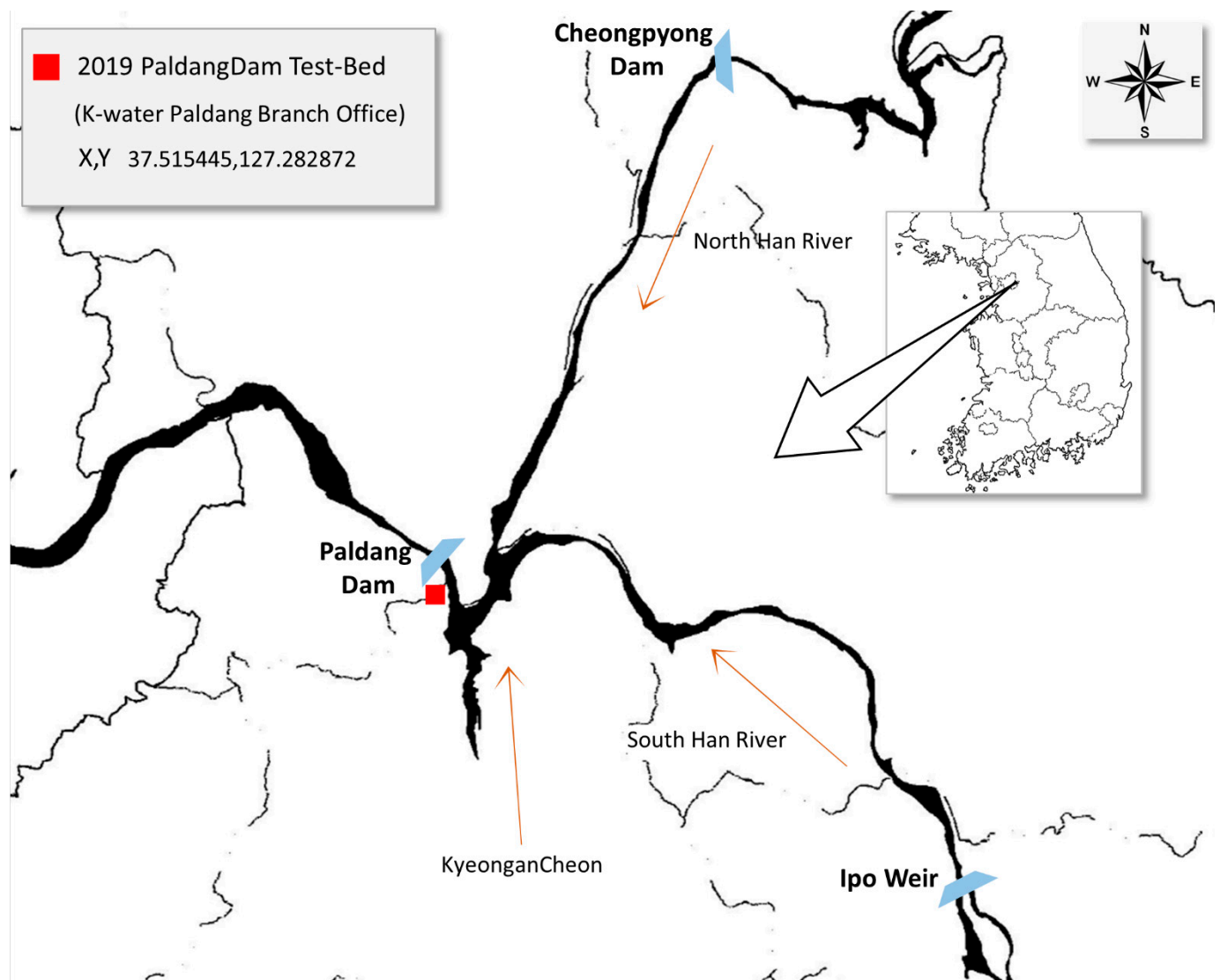
**Figure 1.** Machine learning framework for chlorophyll-*a* estimation. Hyperspectral<sub>*i*</sub> denotes an *i* nm hyperspectral wavelength (*i* = 350 nm, . . . , 900 nm). The abbreviation in Figure 1 is described as follows. OLS (ordinary least squares), RF (random forest), ET (extra trees), GB (Gradient Boosting), AdaBoost (Adaptive Boosting), KNN (k-nearest neighbor), SVR (support vector regression), XGboost (Extreme Gradient Boosting).

### 2.1. Hyperspectral Datasets

The data for this study were collected from Paldang Dam in Hanam, Gyeonggi-do, Korea using chlorophyll-*a* measurement sensors and hyperspectral sensors. This is an important location for the management of drinking water quality in Seoul, the capital of the Republic of Korea. In light of this, Paldang Dam is continuously conducting research on chlorophyll-*a* [22,23]. Figure 2 presents a map that shows the location of Paldang Dam.

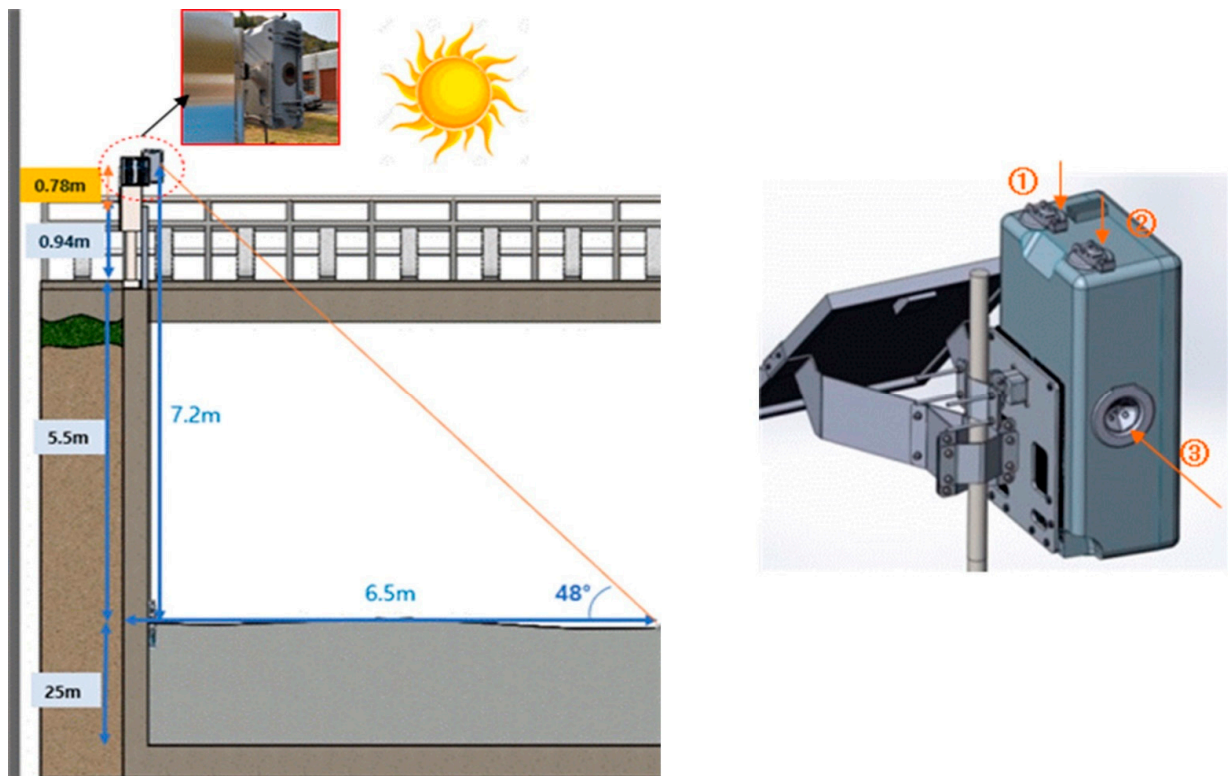
### 2.2. Sampling Chlorophyll-*a*, Hyperspectral Sensor

A hyperspectral sensor (Water insight spectrometer station; WISP station) [24] and a water quality sensor (YSI 6600EDS) were installed in the inflow waterway at Paldang Dam, where an observatory is operated by a government agency. Hyperspectral data and water quality data were collected periodically. The hyperspectral sensor measures 551 wavelengths between 350 nm and 900 nm at frequencies at intervals of 15 min from 01:00 to 8:00 every day. The water quality sensor measures chlorophyll-*a* with the same time frequency. The measured wavelength is used as independent variables to predict the measured chlorophyll-*a*. Figure 3 shows the technical principles of the hyperspectral sensor.

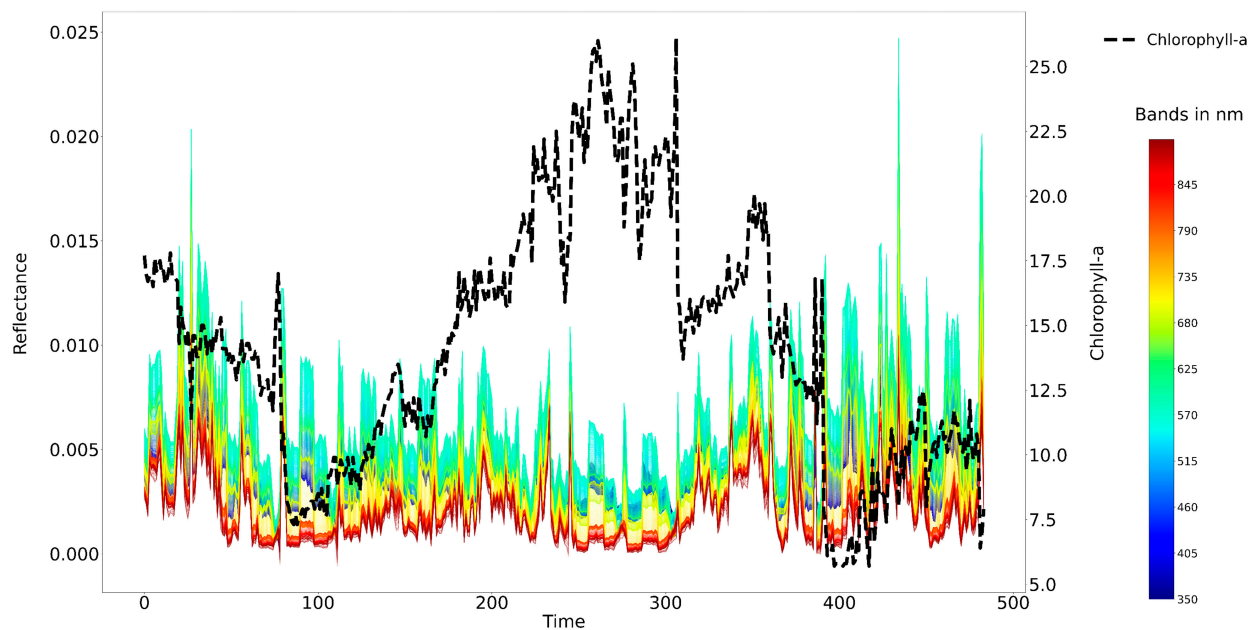


**Figure 2.** Map of the location where the data were collected.

Data were collected from 10 June 2019 to 6 November 2019, but due to missing values for chlorophyll-*a*, the data used for analysis were collected from 9 September 2019 at 02:45 to 7 October 2019 at 01:15. A total of 483 observations were used. The shape of the data observations is shown in Figure 4. Because estimating the dependent variable (chlorophyll-*a*) when an explanatory variable (hyperspectral wavelength data) is given is an estimation problem, rather than a prediction problem, time dependence is ignored in this study.



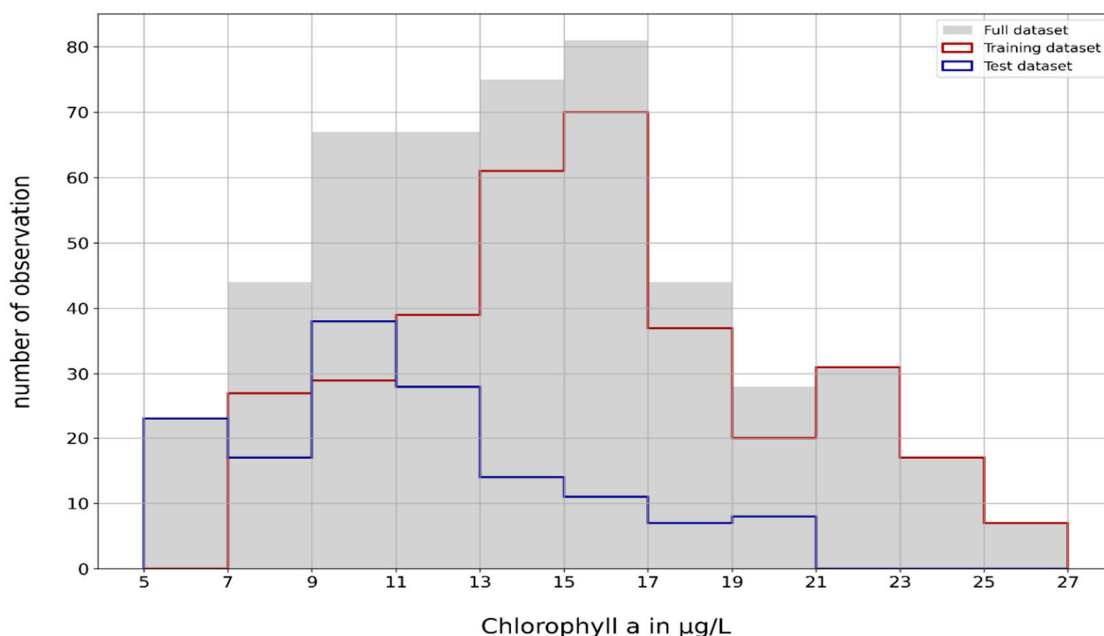
**Figure 3.** Technological principle of hyperspectral sensor WISP station. Upper lens 2 point (①,②): Light must enter through the two lenses during measurement, not under overcast or rainy conditions. Side lens 1 point (③): The distance between the measurement point and the point to be measured is calculated; it is then installed such that the reflected light reaches the side lens point at a  $48^\circ$  angle, as illustrated on the left. This is the spectroscopic principle by which the reflected light is measured using a single wavelength band.



**Figure 4.** Chlorophyll-*a* and hyperspectral data observed through the sensor. The black dotted line is the chlorophyll-*a* observed at each time point, and the spectrum is the hyperspectral data for the bands, 350 nm to 900 nm, observed at each time point.

### 2.3. Data Preprocessing

To evaluate the estimated chlorophyll-*a*, preprocessing was performed by dividing the training and testing data into a 7:3 ratios. For the total of 483 observations, training data and testing data observations numbered 338 and 146, respectively, and the data division is shown in Figure 5.



**Figure 5.** Chlorophyll-*a* distribution. Each dataset (gray bars) is divided into training (red) and testing (blue) subsets in chronological order.

Four alternative preprocessing approaches were used to improve the performance of the model on hyperspectral data, which is high-dimensional data. We compared and fitted eight different machine learning models to each approach to obtain the best model. To begin, we used the standard scaler method, which is a method for transforming data into a normal distribution by setting the mean to 0 and the variance to 1. The min-max scaler can be used to scale the values of observations using lower and upper limits that are predefined. Pretreatment was carried out in this investigation by setting the lower limit to 0 and the upper limit to 1.

For high-dimensional data, algorithms such as PLS and PCA are often utilized for dimensionality reduction to improve prediction performance and speed [25]. PCA is a method that employs the main components that maximize the variance of a linear combination of explanatory variables, whereas PLS uses the main components that maximize the linear combination of explanatory variables and the covariance of dependent variables.

#### 2.3.1. PCA

PCA is an unsupervised feature extraction technique that transforms high-dimensional data into low-dimensional data. The principal component  $PC_I$  ( $i = 1, 2, \dots, k$ ), which is a new variable created by PCA, can be generated as many times as the value of  $p$ , which is the maximum number of original variables. Principal components that explain more than 70–80% of the variation in the original variable are generally chosen and employed. PCA can cause data loss because it reduces the dimensionality of the data.

#### 2.3.2. PLS

PLS is a method for reducing the dimensions of a variable by creating a new variable through a linear combination of variables. The value of  $p$ , which is the maximum number of original variables, is used to generate a latent variable, which is a variable extracted from

a linear combination of variables. The mean square error (MSE) is used to determine the ideal number of latent variables.

In this study, only five principal components were used in PCA and PLS, and they accounted for 99.9% of a total of 551 hyperspectral bands ranging from 350 nm to 900 nm.

#### 2.4. Machine Learning Algorithm

To estimate the chlorophyll-*a*, ordinary least squares (OLS), support vector regression (SVR) [26], the k-nearest neighbor (KNN) [27], the bagging model, and the boosting model were used. Furthermore, optimal hyperparameters were obtained using grid search for cross validation, and the optimized model was determined by preventing overfitting.

##### 2.4.1. SVR

SVR is a generalized support vector machine (SVM) algorithm. Unlike SVM, SVR is a method for finding a regression line that optimizes flatness. SVM finds the hyperplane that maximizes the distance between each data point adjacent to various groups.

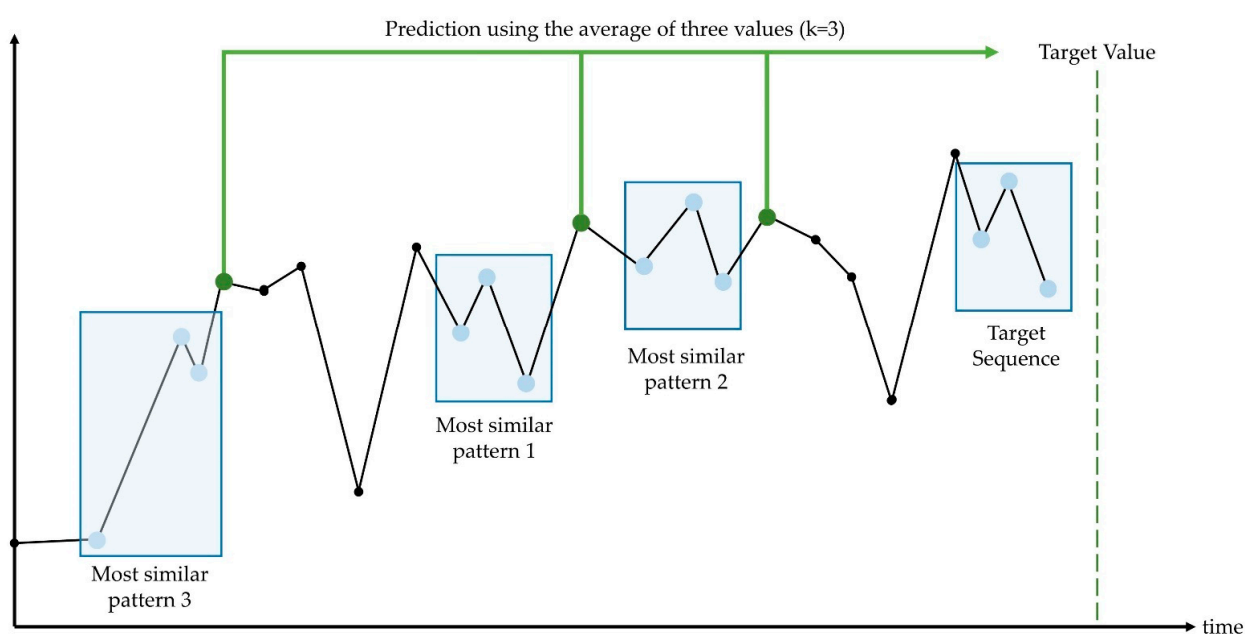
$$L_{SVR} = \min \frac{1}{2} ||\omega||^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*)$$

$\xi$  is the distance between the regression line and the regression equation's upper boundary, and  $\xi^*$  is the distance between the regression line and the regression equation's lower boundary.

##### 2.4.2. KNN

KNN is a learning method that uses the  $k$  most similar objects to identify unlabeled objects. It can be extended to categorical dependent variables and to regression and time series prediction.

In univariate time series data, KNN predicts a value using information for the pattern most similar to the previous pattern of the value to be predicted. The parameters used are  $w$  and  $k$ , similarity is calculated using the Euclidean distance of the previous  $w$  points, and the most similar  $k$  patterns are used. For  $m$  multivariate time series data observations, similarity is calculated using the  $m$ -dimensional Euclidean distance calculation value. Figure 6 is a summary of KNN's method.

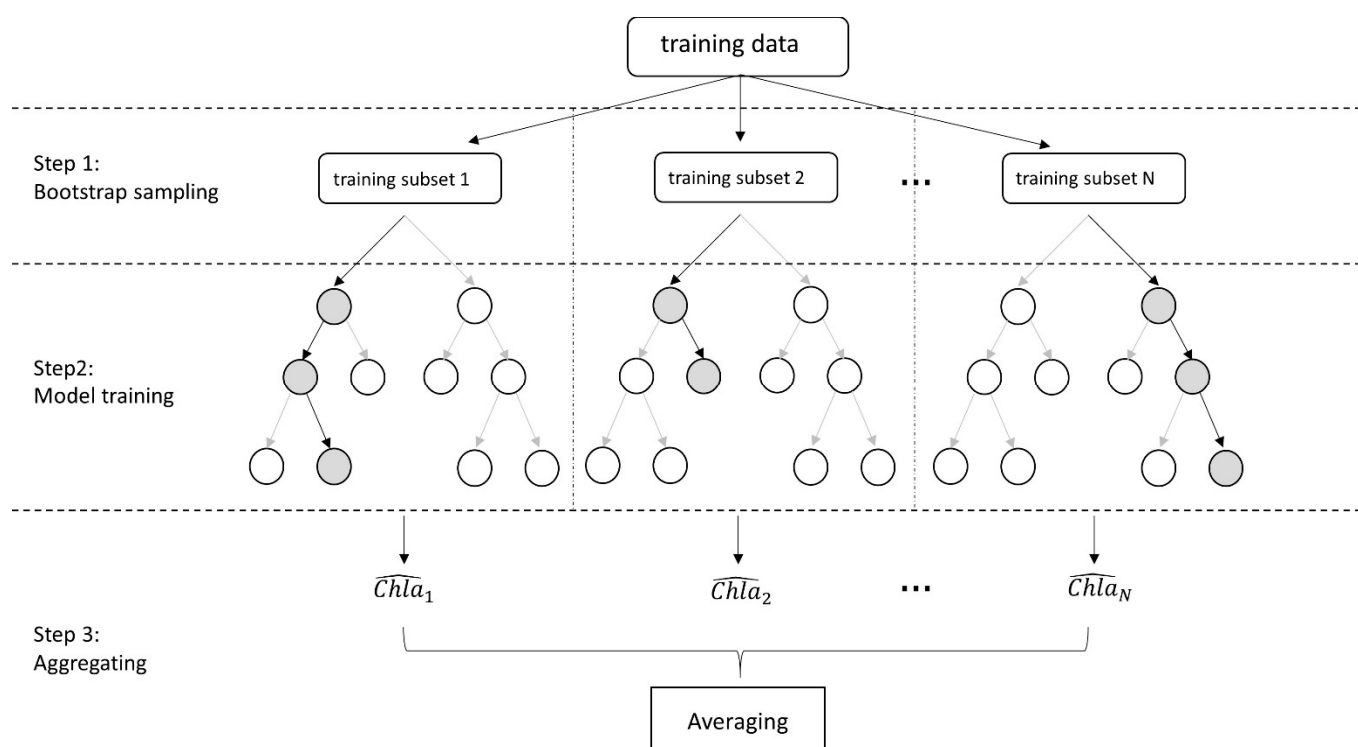


**Figure 6.** An example diagram of a univariate time series k-neighbor with  $w = 3$  and  $k = 3$ . The black dot is the observed value at each time.

### 2.4.3. Bagging

“Bagging” is shorthand for bootstrap aggregating. It is a technique for producing a large number of samples by bootstrapping from training data, training a model with each sample, and aggregating the results into ensembles.

Bagging is usually used for classification and regression, and in this study, the average value of the predicted results for each sample is used as the predicted value for fitting chlorophyll-*a* concentration, which is a dependent variable, as shown in Figure 7.



**Figure 7.** Bagging structure.

Random forest (RF) [28] is a method for reducing variance by ensuring variety and unpredictability in data. When the bagging model is trained, the RF method randomly selects fewer variables than the number of original variables, and it chooses the best variable among the variables to branch out when it searches for the decision tree’s branching point. This procedure continues until a fully developed tree is formed. The method uses the average of the estimates from each bootstrap model sampled.

Extremely randomized trees (ET) [19] is similar to RF, but it achieves greater randomness than RF when searching for branches of a decision tree, and is much faster because it uses a random variable rather than the best variable.

### 2.4.4. Boosting

As one of the ensemble strategies of machine learning, boosting is an algorithm that improves the performance of classification or regression models by producing multiple weak learners and by successively generating and merging models.

After all data are randomly sampled, the weights of all samples are initialized. We create a weak learner, and after training, we output the model weights of the weak learner according to the results. The data are updated using the weights of the output model. After repeating this process *N* times, *N* model weights are ensembled to create a new learner. In this study, adaptive boosting (AdaBoost) [29], gradient boosting (GB) [30], and extreme gradient boosting (XGBoost) [31], which are based on the idea of the boosting algorithm, were used. Figure 8 summarizes the boosting method described above.

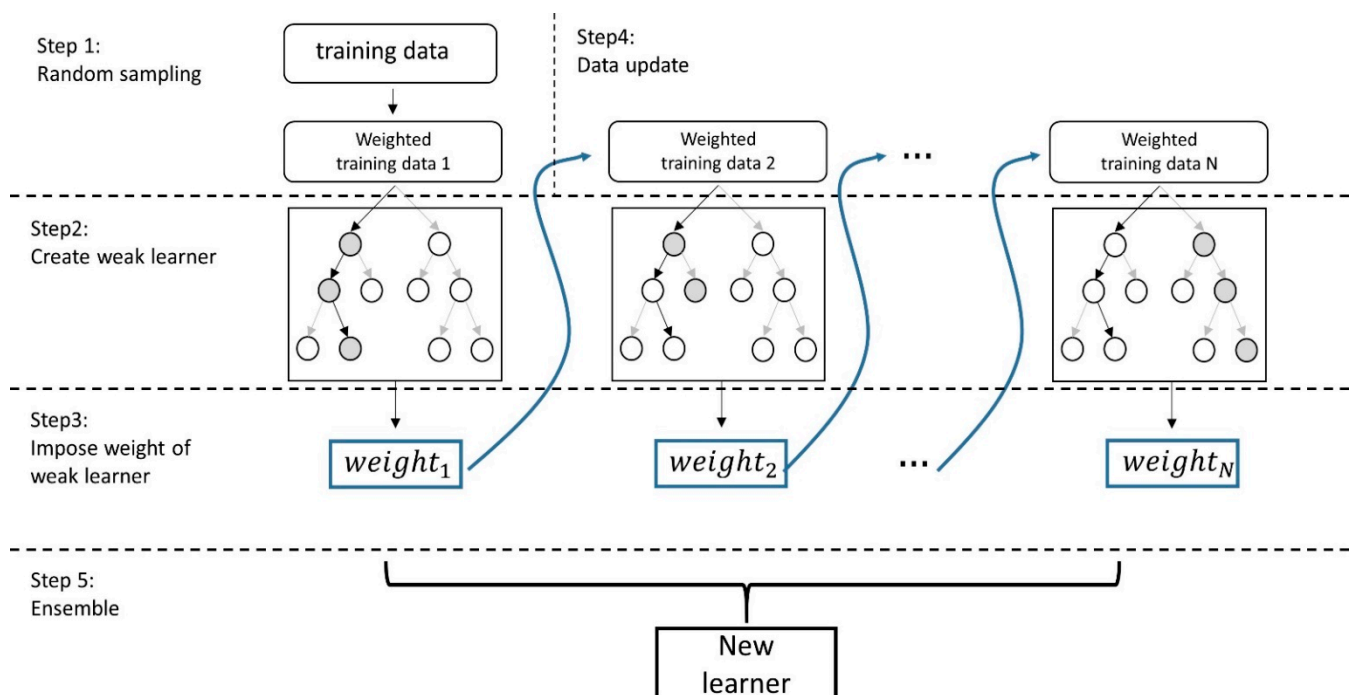


Figure 8. Boosting structure.

AdaBoost, which was developed from boosting techniques, is a method for increasing the weight of observations with large training errors and lowering the weight of small observations. After random sampling of the training data, the weights of the data are initialized.

$$W_i^t = \frac{1}{m}, i = 1, 2, \dots, m, t = 1, 2, \dots, T \quad (1)$$

In Equation (1),  $W$  is the data weight,  $i$  is the number of observations selected through sampling, and  $t$  is the number of weak learners. To find the best learner in the ensemble, we need to find the weak learners and extract the weights.

$$\varepsilon_t = \frac{\sum_{i=1}^n W_i^t |h_t(x_i) - y_i|}{\sum_{i=1}^n W_i^t} \quad (2)$$

$$\alpha_t = \log\left(\frac{\varepsilon_t}{1 - \varepsilon_t}\right) \quad (3)$$

$$W_i^{t+1} = W_i^t \alpha_t^{1 - |h_t(x_i) - y_i|} \quad (4)$$

In Equation (2),  $\varepsilon$  is the error rate loss function, and  $h$  is the weak learner. In Equation (3),  $\alpha$  is the weight of the weak learner. A new weight vector is obtained by finding  $h_t$  when  $\varepsilon_t$  is at a minimum, and then finding the corresponding weight  $\alpha$ . Equations (2)–(4) are repeated to ensemble the newly created weight  $W_i$  and the corresponding  $h_t$  to finally obtain the boosted model. Equation (5) defines the final model.

$$h(x) = \text{sign}\left[\sum_{i=1}^m \alpha_i h_t(x)\right] \quad (5)$$

GB is a regression approach that uses boosting and a strong ensemble model. It employs a gradient in the loss function, which assesses the ability of the model coefficient to match the data. By forming an ensemble of many basic decision tree models, GB creates a predictive model. The GB model inherits all of the benefits of decision tree models while further improving resilience and accuracy. The GB model has additional advantages, such

as the capacity to manage huge datasets without preprocessing, resilience to outliers, the ability to manage missing data, robustness to complicated data, and resistance to overfitting. The GB model, in general, begins by fitting the data using a basic decision tree model that has a certain level of accuracy.

XGBoost is a distributed processing algorithm that uses the greedy approach to quickly discover classifiers and acceptable parameters. All leaves are tied to the final score of the model when a tree, which is a classifier, is generated using an ensemble model called classification and regression tree (CART); therefore, it is feasible to compare the scores of different models having the same classification result. This technique can be used to identify the best performing model.

### 3. Results

This section focuses on the performance of the model, the influence of the four pre-processing methods, and presents a comparison between the estimated and observed data values. Table 1 lists the formulas for the evaluation metrics used in this study. We applied six evaluation metrics to select the optimal model from among various models. A detailed evaluation of the performance of the eight machine learning models and the four pre-processing methods are summarized in Table 2. The optimal models were those achieving good results in at least four of the six metrics.

**Table 1.** Equations for evaluation metrics used in machine learning analysis.

| Metric | Equation   | Range (Optimal Value)       |
|--------|--|-----------------------------|
| $R^2$  | $\left\{ \frac{\sum_{i=1}^n (y_i^{obs} - \bar{y}^{obs})(y_i^{pred} - \bar{y}^{pred})}{\sqrt{\sum_{i=1}^n (y_i^{obs} - \bar{y}^{obs})^2} \sqrt{\sum_{i=1}^n (y_i^{pred} - \bar{y}^{pred})^2}} \right\}^2$ | 0.0~1.0 (1.0)               |
| NSE    | $1 - \frac{\sum_{i=1}^n (y_i^{obs} - y_i^{pred})^2}{\sum_{i=1}^n (y_i^{obs} - \bar{y}^{obs})^2}$   | $-\infty \sim 1.0$ (1.0)    |
| d      | $1 - \frac{\sum_{i=1}^n (y_i^{obs} - y_i^{pred})^2}{\sum_{i=1}^n ( y_i^{pred} - \bar{y}^{obs}  +  y_i^{obs} - \bar{y}^{obs} )^2}$  | 0.0~1.0 (1.0)               |
| RMSE   | $\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i^{obs} - y_i^{pred})^2}$   | 0.0~ $\infty$ (0.0)         |
| RSR    | $\frac{\sqrt{\sum_{i=1}^n (y_i^{obs} - y_i^{pred})^2}}{\sqrt{\sum_{i=1}^n (y_i^{obs} - \bar{y}^{pred})^2}}$  | 0.0~ $\infty$ (0.0)         |
| PBIAS  | $\frac{\sum_{i=1}^n y_i^{obs} - y_i^{pred}}{\sum_{i=1}^n y_i^{obs}} \times 100$  | $-\infty \sim \infty$ (0.0) |

This table summarizes the equations, ranges, and optimal values of the evaluation metrics used in this study, where  $y_i^{obs}$  is the  $i$ -th observed value,  $y_i^{pred}$  is the  $i$ -th predicted value,  $\bar{y}^{obs}$  is the mean of the observations, and  $\bar{y}^{pred}$  is the mean of the predicted values.

**Table 2.** Machine learning results for the estimation of chlorophyll-*a*.

| Method          |          | $R^2$        | MSE          | MAPE          | NSE          | d            | PSR          |
|-----------------|----------|--------------|--------------|---------------|--------------|--------------|--------------|
| Baseline        | OLS      | 0.380        | 126.665      | 77.598        | −7.592       | 0.402        | 2.121        |
|                 | RF       | 0.919        | 20.070       | 44.153        | −0.361       | 0.557        | 0.922        |
|                 | ET       | 0.986        | 18.878       | 42.474        | −0.281       | 0.569        | 0.912        |
|                 | GB       | 0.941        | 12.765       | 33.110        | 0.134        | 0.701        | 0.809        |
|                 | AdaBoost | 0.908        | 22.93        | 47.754        | −0.555       | 0.502        | 0.963        |
|                 | KNN      | 0.925        | 17.284       | 39.489        | −0.172       | 0.617        | 0.898        |
|                 | SVR      | <b>0.991</b> | <b>1.299</b> | <b>8.365</b>  | <b>0.912</b> | <b>0.977</b> | <b>0.297</b> |
|                 | XGBoost  | 0.948        | 10.908       | 30.194        | 0.26         | 0.742        | 0.765        |
| Standard Scaler | OLS      | 0.122        | 337.646      | 178.614       | −21.904      | 0.305        | 1.04         |
|                 | RF       | 0.919        | 19.967       | 43.996        | −0.354       | 0.558        | 0.921        |
|                 | ET       | 0.986        | 18.918       | 42.529        | −0.283       | 0.568        | 0.913        |
|                 | GB       | 0.947        | 11.282       | 31.21         | 0.235        | 0.732        | 0.772        |
|                 | AdaBoost | 0.908        | 23.348       | 48.034        | −0.584       | 0.507        | 0.963        |
|                 | KNN      | 0.929        | 16.617       | 39.044        | −0.127       | 0.632        | 0.882        |
|                 | SVR      | <b>0.986</b> | <b>2.132</b> | <b>10.473</b> | <b>0.855</b> | <b>0.961</b> | <b>0.379</b> |
|                 | XGBoost  | 0.948        | 10.908       | 30.194        | 0.260        | 0.742        | 0.765        |

Table 2. Cont.

| Method         |          | R <sup>2</sup> | MSE          | MAPE          | NSE          | d            | PSR          |
|----------------|----------|----------------|--------------|---------------|--------------|--------------|--------------|
| Min-Max scaler | OLS      | 0.737          | 45.016       | 48.533        | −2.054       | 0.685        | 1.604        |
|                | RF       | 0.919          | 19.967       | 43.996        | −0.354       | 0.558        | 0.921        |
|                | ET       | 0.986          | 18.878       | 42.474        | −0.281       | 0.569        | 0.912        |
|                | GB       | 0.947          | 11.282       | 31.21         | 0.235        | 0.732        | 0.772        |
|                | AdaBoost | 0.908          | 23.348       | 48.034        | −0.584       | 0.507        | 0.963        |
|                | KNN      | 0.930          | 16.797       | 39.028        | −0.139       | 0.632        | 0.879        |
|                | SVR      | <b>0.987</b>   | <b>1.950</b> | <b>10.113</b> | <b>0.868</b> | <b>0.965</b> | <b>0.363</b> |
|                | XGBoost  | 0.948          | 10.908       | 30.194        | 0.260        | 0.742        | 0.765        |
| PCA            | OLS      | 0.170          | 330.055      | 184.408       | −21.389      | 0.314        | 1.003        |
|                | RF       | 0.964          | 6.675        | 23.326        | 0.547        | 0.853        | 0.636        |
|                | ET       | <b>0.986</b>   | 5.820        | 22.345        | 0.605        | 0.861        | 0.589        |
|                | GB       | 0.980          | 3.428        | 16.316        | 0.767        | 0.935        | 0.472        |
|                | AdaBoost | 0.962          | 8.479        | 27.148        | 0.425        | 0.82         | 0.68         |
|                | KNN      | 0.928          | 16.752       | 39.179        | −0.136       | 0.63         | 0.884        |
|                | SVR      | 0.982          | <b>2.602</b> | <b>11.776</b> | <b>0.824</b> | <b>0.953</b> | <b>0.419</b> |
|                | XGBoost  | 0.981          | 3.229        | 15.499        | 0.781        | 0.935        | 0.459        |
| PLS            | OLS      | 0.171          | 330.243      | 184.598       | −21.402      | 0.314        | 1.002        |
|                | RF       | 0.983          | 4.291        | 17.468        | 0.709        | 0.928        | 0.51         |
|                | ET       | 0.986          | 4.475        | 20.171        | 0.696        | 0.905        | 0.514        |
|                | GB       | 0.988          | 1.875        | 10.351        | 0.873        | 0.969        | 0.356        |
|                | AdaBoost | 0.977          | 7.481        | 25.6          | 0.493        | 0.868        | 0.625        |
|                | KNN      | 0.932          | 14.624       | 36.223        | 0.008        | 0.663        | 0.861        |
|                | SVR      | 0.981          | 2.828        | 12.363        | 0.808        | 0.948        | 0.436        |
|                | XGBoost  | <b>0.990</b>   | <b>1.595</b> | <b>10.416</b> | <b>0.892</b> | <b>0.972</b> | <b>0.327</b> |

### 3.1. Performance Measures

We used different performance measures to show the significant differences between the fit models. The performance metrics used in this study are given below [32–37].

1. R<sup>2</sup>: This is an applied evaluation metric for fit regression models that is used mainly in hydrological studies [32]. However, there is a disadvantage that R<sup>2</sup> increases unconditionally when the number of variables increases.
2. Nash-Sutcliffe efficiency (NSE): This metric reflects the desirable and undesirable features of a model of interest, and it increases as the quality of the model increases [33]. However, it is sensitive to extreme values because it uses squared differences, and it cannot identify model bias [32].
3. d: This consists of MSE and potential error (PE) [34]. It offers the advantage that errors and differences are given an appropriate weightage that is not inflated by squared values. However, it is sensitive to extreme values because it uses squared differences [32].
4. Root mean square error (RMSE): This metric is obtained by applying the root to the mean of the total squared error (the sum of the individual squared errors). Therefore, it increases when the variance associated with the frequency distribution of error magnitudes increase [37].
5. RSR (RMSE-observations standard deviation ratio): This metric standardizes RMSE using the standard deviation of the observations. Therefore, a lower RSR means better model performance and a lower RMSE [35].
6. Percent bias (PBIAS): This measures the average tendency of the simulated data to be larger or smaller than their observed counterparts. That is, positive values indicate a model underestimation bias, and negative values indicate a model overestimation bias. It is useful for continuous long-term simulations and can help identify the average model simulation bias [36].

### 3.2. Estimation

In this section, we focus on the performance of the machine learning models in estimating chlorophyll-*a*, the impact of the four preprocessing methods, and the comparison between the estimated and measured chlorophyll-*a* values. Table 2 summarizes the machine learning results estimated in combination with the preprocessing methods. Each machine learning model's hyperparameters were optimized using GridSearch. Table 3 summarizes

the GridSearch results of the best-performing model for each of the machine learning techniques summarized in Table 2.

**Table 3.** GridSearch results for each machine learning technique.

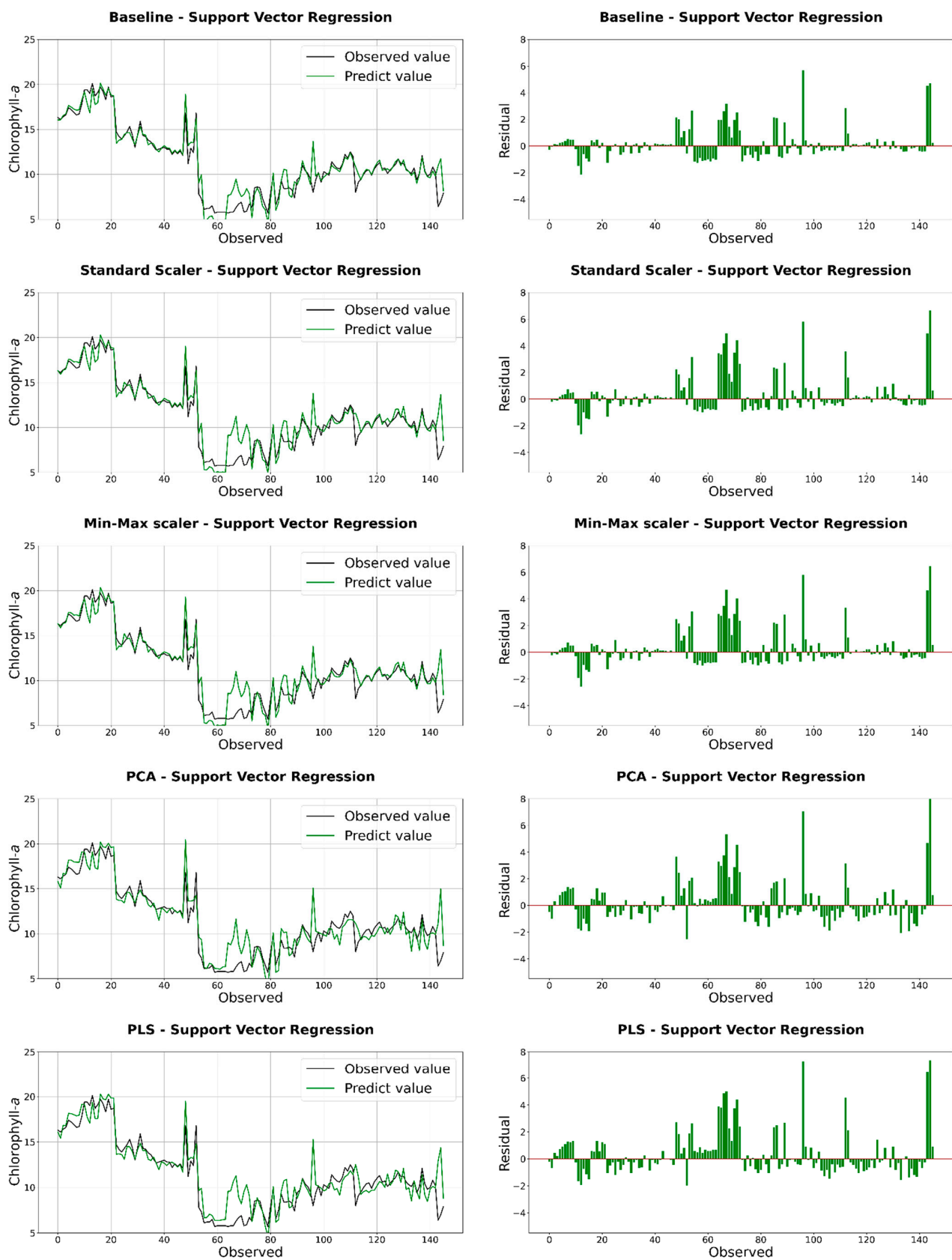
| ML Model (with Preprocessing) | Hyperparameters                     | Type                                   | Search Space  | Optimal Parameters         |
|-------------------------------|-------------------------------------|--|---|----------------------------|
| OLS                           | -                                   | -                                      | -   | -                          |
| Random Forest (PLS)           | min_samples_leaf<br>max_depth       | Discrete<br>Discrete                   | 3, 5, 7, 10<br>3, 4, 5, 6   | 3<br>6                     |
| Extreme Tree (PLS)            | min_samples_leaf<br>max_depth       | Discrete<br>Discrete                   | 3, 5, 7, 10<br>3, 4, 5, 6   | 5<br>6                     |
| Gradient Boost (PLS)          | min_samples_leaf<br>n_estimators    | Discrete<br>Discrete                   | 3, 5, 7, 10<br>100, 200, 300  | 10<br>300                  |
| AdaBoost (PLS)                | n_estimators<br>learning_rate       | Discrete<br>Discrete                   | 100, 200, 300<br>0.1, 0.05, 0.02, 0.01                                  | 300<br>0.1                 |
| KNN (PLS)                     | n_neighbors<br>weights<br>algorithm | Discrete<br>Categorical<br>Categorical | 3,5,7,9,11,<br>“uniform”, “distance”<br>“ball_tree”, “kd_tree”, “brute” | 5<br>distance<br>ball_tree |
| SVR (Baseline)                | Kernel<br>C                         | Categorical<br>Discrete                | “rbf”, “sigmoid”<br>10,30,100,300,1000                                  | rbf<br>1000                |
| XGBoost (PLS)                 | max_depth<br>learning_rate          | Discrete<br>Discrete                   | 5, 6, 7<br>0.03, 0.05, 0.07   | 5<br>0.07                  |

Among all the models, SVR obtained the best results, and XGBoost and ET showed good results. OLS showed very poor performance for all preprocessing methods aside from the OLS method;  $R^2$  was generally greater than 90%. Among the methods, PCA and PLS generally achieved an  $R^2$  greater than 95%. SVR showed the best performance for all preprocessing combinations; SVR combined with raw data showed the best result with  $R^2 = 99.1\%$ .

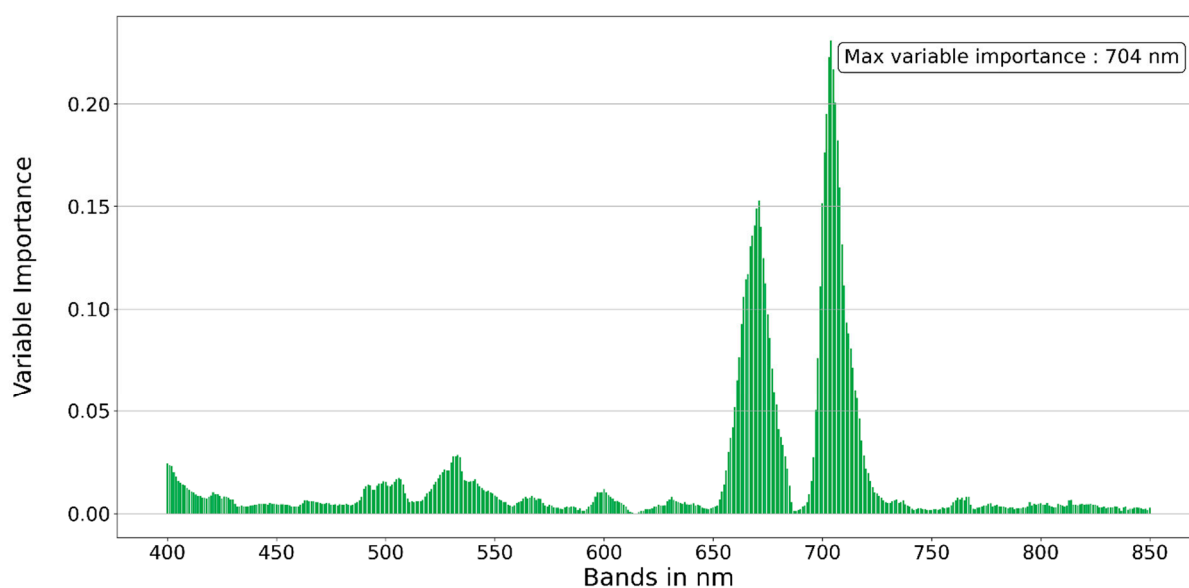
The values presented in bold text represent the best values obtained for the evaluating metrics for each model. PCA and PLS provided better estimates than the other preprocessing methods, and SVR was the best model for each preprocessing approach. The left side of Figure 9 shows chlorophyll-*a* values estimated using SVR with each pretreatment method compared with the observed values, and the right side shows the corresponding residual values.

### 3.3. Variable Importance

In Figure 10, we show the variable importance distribution for the chlorophyll-*a* hyperspectral input data generated using SVR without preprocessing. As can be seen in Table 2, the feature importance of SVR was checked because SVR showed the best performance baseline. Feature importance had the highest peak at 704 nm, and it was confirmed that wavelengths between 650 nm and 730 nm are important variables for estimating chlorophyll-*a*.



**Figure 9.** Left-side figure shows the difference between the observed and fitted values using support vector regression for each preprocessing scheme, while the right-side figure shows the residual value.



**Figure 10.** Feature importance for support vector regression without preprocessing (baseline).

#### 4. Discussion

A variety of solutions have been used to prevent HABs. We used hyperspectral data to estimate chlorophyll-*a* concentration in an attempt to avert dangerous algal blooms in advance. Methods for predicting, preventing, and eliminating algae have been proposed in various fields [38,39]. In this study, we presented a machine learning algorithm to quickly observe algae. One such initiative was to use the chlorophyll-*a* estimation algorithm presented in this paper to readily identify areas of dangerous algal blooms using a hyperspectral image recording device for data that are collected using a drone or an unmanned aerial vehicle. Previously, an analysis was carried out to estimate chlorophyll-*a* using a water quality parameter and a specific wavelength band of the hyperspectral sensor. Shafique et al. [13] and Murugan et al. [15] estimated chlorophyll-*a* by computing specific wavelength band values. However, in the present study, we used a range of machine learning algorithms to estimate and evaluate chlorophyll-*a* concentration using just hyperspectral sensor data. As a result, high  $R^2$  values were obtained for seven machine learning methods, except for OLS, with the PLS preprocessing method. Using this method, chlorophyll-*a* can be estimated immediately without measuring water quality data using hyperspectral data acquired by drones and unmanned aerial vehicles.

Our study obtained reasonable results for several models. In particular, Keller [18] was able to obtain an  $R^2$  of 0.914 using PCA with ET. Our data obtained generally good values from all models using PLS. Moreover, the results were considered appropriate because the values obtained using the evaluation metrics were not significantly different compared with the chlorophyll-*a* estimation results of previous studies [18].

In this study, we used six evaluation metrics to explain the machine learning results. Various evaluation indicators were applied to increase the objectivity of the analysis results. In addition, the dimensions were reduced using PLS and PCA during pretreatment, which led to better results.

However, we found several outcomes of interest in this study. Considering the fact that algal blooms show significant differences from season to season, and that HABs occur in the summer, the inability to use data over a long period of time may have an impact on the application of the developed model. Moreover, because the data are not time-dependent, we cannot apply deep learning models for estimation. Considering that the ultimate solution for algal blooms is to predict the occurrence of algal blooms in advance, the inability to use machine learning and deep learning for chlorophyll-*a* prediction could be a limitation. This study presents the applicability and characteristics of machine learning

models to predict algal bloom in a field reservoir. The collection of field data including four-seasons observation in the future would improve the precision and applicability of the model developed in this study.

## 5. Conclusions

In this study, we developed a method for estimating chlorophyll-*a* concentration with high accuracy by applying various machine learning models to hyperspectral data and comparing the results using various performance measures. The feasibility of estimating chlorophyll-*a* using hyperspectral data was investigated, and the future application of hyperspectral data was discussed [40,41]. Dimensionality reduction approaches, such as PCA and PLS, as well as min-max scaling and standard scaling, were used in the chlorophyll-*a* estimation preprocessing procedure. A framework for analysis was presented, which included eight machine learning methods and data preprocessing methods. The estimation performance of each framework was compared, and it was finally determined that SVR was the best model among all the analysis frameworks.

**Author Contributions:** Software, G.I. and D.L.; Formal analysis, S.L. (Sanghun Lee); Investigation, S.L. (Sungjong Lee); Data curation, G.I.; Writing—original draft, G.I. and J.L.; Writing—review & editing, J.P.; Supervision, T.-Y.H.; Funding acquisition, T.-Y.H. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported by Chungbuk National University Korea National University Development Project (2022).

**Data Availability Statement:** The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Steidinger, K.A. Historical perspective on *Karenia brevis* red tide research in the Gulf of Mexico. *Harmful Algae* **2009**, *8*, 549–561. [\[CrossRef\]](#)
2. Gobler, C.J. Climate change and harmful algal blooms: Insights and perspective. *Harmful Algae* **2020**, *91*, 101731. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Van Apeldoorn, M.E.; van Egmond, H.P.; Speijers, G.J.; Bakker, G.J. Toxins of cyanobacteria. *Mol. Nutr. Food Res.* **2007**, *51*, 7–60. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Paerl, H.W.; Otten, T.G. Harmful cyanobacterial blooms: Causes, consequences, and controls. *Microb. Ecol.* **2013**, *65*, 995–1010. [\[CrossRef\]](#)
5. Min, S.K.; Son, S.W.; Seo, K.H.; Kug, J.S.; An, S.I.; Choi, Y.S.; Lee, M.I. Changes in weather and climate extremes over Korea and possible causes: A review. *Asia-Pac. J. Atmos. Sci.* **2015**, *51*, 103–121. [\[CrossRef\]](#)
6. Hallegraeff, G.M.; Anderson, D.M.; Belin, C.; Bottein, M.-Y.D.; Bresnan, E.; Chinain, M.; Enevoldsen, H.; Iwataki, M.; Karlson, B.; McKenzie, C.H.; et al. Perceived global increase in algal blooms is attributable to intensified monitoring and emerging bloom impacts. *Commun. Earth Environ.* **2021**, *2*, 117. [\[CrossRef\]](#)
7. Karlson, B.; Andersen, P.; Arneborg, L.; Cembella, A.; Eikrem, W.; John, U.; West, J.J.; Klemm, K.; Kobos, J.; Lehtinen, S.; et al. Harmful algal blooms and their effects in coastal seas of Northern Europe. *Harmful Algae* **2021**, *102*, 101989. [\[CrossRef\]](#)
8. Maniyar, C.B.; Kumar, A.; Mishra, D.R. Continuous and Synoptic Assessment of Indian Inland Waters for Harmful Algae Blooms. *Harmful Algae* **2022**, *111*, 102160. [\[CrossRef\]](#)
9. Filstrup, C.T.; Downing, J.A. Relationship of chlorophyll to phosphorus and nitrogen in nutrient-rich lakes. *Inland Waters* **2017**, *7*, 385–400. [\[CrossRef\]](#)
10. Sellner, K.G.; Doucette, G.J.; Kirkpatrick, G.J. Harmful algal blooms: Causes, impacts and detection. *J. Ind. Microbiol. Biotechnol.* **2003**, *30*, 383–406. [\[CrossRef\]](#)
11. Xing, Q.; Chen, C.; Shi, H.; Shi, P.; Zhang, Y. Estimation of chlorophyll-*a* concentrations in the Pearl River Estuary using in situ hyperspectral data: A case study. *Mar. Technol. Soc. J.* **2008**, *42*, 22–27. [\[CrossRef\]](#)
12. Shafique, N.A.; Fulk, F.; Autrey, B.C.; Flotemersch, J. Hyperspectral remote sensing of water quality parameters for large rivers in the Ohio River basin. In Proceedings of the First Interagency Conference on Research in the Watershed, Benson, AZ, USA, 27–30 October 2003; pp. 216–221.
13. Shin, Y.; Kim, T.; Hong, S.; Lee, S.; Lee, E.; Hong, S.; Lee, C.; Kim, T.; Park, M.S.; Park, J.; et al. Prediction of chlorophyll-*a* concentrations in the Nakdong River using machine learning methods. *Water* **2020**, *12*, 1822. [\[CrossRef\]](#)

14. Murugan, P.; Sivakumar, R.; Pandiyanc, R. Chlorophyll-A estimation in case-II water bodies using satellite hyperspectral data. In Proceedings of the ISPRS TC VIII International Symposium on Operational Remote Sensing Applications: Opportunities, Progress and Challenges, Hyderabad, India, 9–12 December 2014; p. 536.
15. Glukhovets, D.I.; Goldin, Y.A. Express method for chlorophyll concentration assessment. *J. Photochem. Photobiol.* **2021**, *8*, 100083. [\[CrossRef\]](#)
16. Chi, M.; Plaza, A.; Benediktsson, J.A.; Sun, Z.; Shen, J.; Zhu, Y. Big Data for Remote Sensing: Challenges and Opportunities. *Proc. IEEE* **2016**, *104*, 2207–2219. [\[CrossRef\]](#)
17. Keller, S.; Maier, P.M.; Riese, F.M.; Norra, S.; Holbach, A.; Börsig, N.; Wilhelms, A.; Moldaenke, C.; Zaaake, A.; Hinz, S. Hyperspectral data and machine learning for estimating CDOM, chlorophyll *a*, diatoms, green algae and turbidity. *Int. J. Environ. Res. Public Health* **2018**, *15*, 1881. [\[CrossRef\]](#)
18. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, *63*, 3–42. [\[CrossRef\]](#)
19. Levy, J.; Cary, S.; Joy, K.; Lee, C. Detection and community-level identification of microbial mats in the McMurdo Dry Valleys using drone-based hyperspectral reflectance imaging. *Antarct. Sci.* **2020**, *32*, 367–381. [\[CrossRef\]](#)
20. Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52. [\[CrossRef\]](#)
21. Kim, D.W.; Min, J.H.; Yoo, M.; Kang, M.; Kim, K. Long-term effects of hydrometeorological and water quality conditions on algal dynamics in the Paldang dam watershed, Korea. *Water Sci. Technol. Water Supply* **2014**, *14*, 601–608. [\[CrossRef\]](#)
22. Li, Z.; Shin, H.H.; Lee, T.; Han, M.S. Resting stages of freshwater algae from surface sediments in Paldang Dam Lake, Korea. *Nova Hedwig.* **2015**, *101*, 475–500. [\[CrossRef\]](#)
23. Peters, S.; Laanen, M.; Groetsch, P.; Ghezehegn, S.; Poser, K.; Hommersom, A.; de Reus, E.; Spaia, L. WISPstation: A new autonomous above water radiometer system. In Proceedings of the Ocean Optics XXIV Conference, Dubrovnik, Croatia, 7–12 October 2018; pp. 7–12.
24. Lee, D.H.; Woo, S.E.; Jung, M.W.; Heo, T.Y. Evaluation of Odor Prediction Model Performance and Variable Importance according to Various Missing Imputation Methods. *Appl. Sci.* **2022**, *12*, 2826. [\[CrossRef\]](#)
25. Smola, A.J.; Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199–222. [\[CrossRef\]](#)
26. Altman, N.S. An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Stat.* **1992**, *46*, 175–185.
27. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [\[CrossRef\]](#)
28. Freund, Y.; Schapire, R.E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139. [\[CrossRef\]](#)
29. Friedman, J.H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **2002**, *38*, 367–378. [\[CrossRef\]](#)
30. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
31. Moriasi, D.N.; Gitau, M.W.; Pai, N.; Daggupati, P. Hydrologic and water quality models: Performance measures and evaluation criteria. *Trans. ASABE* **2015**, *58*, 1763–1785.
32. Zeybek, M. Nash-sutcliffe efficiency approach for quality improvement. *J. Appl. Math. Comput.* **2018**, *2*, 496–503. [\[CrossRef\]](#)
33. Legates, D.R.; McCabe, G.J., Jr. Evaluating the use of “goodness-of-fit” measures in hydrologic and hydroclimatic model validation. *Water Resour. Res.* **1999**, *35*, 233–241. [\[CrossRef\]](#)
34. Moriasi, D.N.; Arnold, J.G.; van Liew, M.W.; Bingner, R.L.; Harmel, R.D.; Veith, T.L. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Trans. ASABE* **2007**, *50*, 885–900. [\[CrossRef\]](#)
35. Gupta, H.V.; Sorooshian, S.; Yapo, P.O. Status of automatic calibration for hydrologic models: Comparison with multilevel expert calibration. *J. Hydrol. Eng.* **1999**, *4*, 135–143. [\[CrossRef\]](#)
36. Willmott, C.J.; Matsuura, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim. Res.* **2005**, *30*, 79–82. [\[CrossRef\]](#)
37. Pyo, J.; Hong, S.M.; Jang, J.; Park, S.; Park, J.; Noh, J.H.; Cho, K.H. Drone-borne sensing of major and accessory pigments in algae using deep learning modeling. *GIScience Remote Sens.* **2022**, *59*, 310–332. [\[CrossRef\]](#)
38. Wong, K.T.; Lee, J.H.; Hodgkiss, I.J. A simple model for forecast of coastal algal blooms. *Estuar. Coast. Shelf Sci.* **2007**, *74*, 175–196. [\[CrossRef\]](#)
39. Huh, J.H.; Choi, Y.H.; Lee, H.J.; Choi, W.J.; Ramakrishna, C.; Lee, H.W.; Lee, S.-H.; Ahn, J.W. The use of oyster shell powders for water quality improvement of lakes by algal blooms removal. *J. Korean Ceram. Soc.* **2016**, *53*, 1–6. [\[CrossRef\]](#)
40. Zhu, W.; Sun, Z.; Yang, T.; Li, J.; Peng, J.; Zhu, K.; Li, S.; Gong, H.; Lyu, Y.; Li, B.; et al. Estimating leaf chlorophyll content of crops via optimal unmanned aerial vehicle hyperspectral data at multi-scales. *Comput. Electron. Agric.* **2020**, *178*, 105786. [\[CrossRef\]](#)
41. Gai, Y.; Yu, D.; Zhou, Y.; Yang, L.; Chen, C.; Chen, J. An improved model for chlorophyll-*a* concentration retrieval in coastal waters based on UAV-Borne hyperspectral imagery: A case study in Qingdao, China. *Water* **2020**, *12*, 2769. [\[CrossRef\]](#)