water

MDPI

Article

Evaluating Statistical Machine Learning Algorithms for Classifying Dominant Algae in Juam Lake and Tamjin Lake, Republic of Korea

Seong-Yun Hwang [1,†], Byung-Woong Choi [1], Jong-Hwan Park [1], Dong-Seok Shin [2], Hyeon-Su Chung [1], Mi-Sun Son [1], Chae-Hong Lim [1], Hyeon-Mi Chae [1], Don-Woo Ha [1] and Kang-Young Jung [3,*]

[1] Yeongsan River Environment Research Center, National Institute of Environmental Research, 5, Cheomdangwagi-ro 208beon-gil, Buk-gu, Gwangju 61011, Republic of Korea; hsyliark@korea.kr (S.-Y.H.); bchoi628@korea.kr (B.-W.C.); thanks@korea.kr (J.-H.P.); jys7246@korea.kr (H.-S.C.); miza03@korea.kr (M.-S.S.); chaehong@korea.kr (C.-H.L.); chm2022@korea.kr (H.-M.C.); hahaha9909@korea.kr (D.-W.H.)

[2] Freshwater Bioresources Culture Research Division, Nakdonggang National Institute of Biological Resources, 137, Donam 2-gil, Sangju-si 37242, Republic of Korea; sds8488@korea.kr

[3] Education Planning Division, National Institute of Environmental Human Resources Development, 42, Hwangyeong-ro, Seo-gu, Incheon 22689, Republic of Korea

* Correspondence: happy3313@korea.kr; Tel.: +82-32-560-7795

† This author is the primary author of this study.

Abstract: South Korea's National Institute of Environmental Research (NIER) operates an algae alert system to monitor water quality at public water supply source sites. Accurate prediction of dominant harmful cyanobacterial genera, such as *Aphanizomenon*, *Anabaena*, *Oscillatoria*, and *Microcystis*, is crucial for managing water source contamination risks. This study utilized data collected between January 2017 and December 2022 from Juam Lake and Tamjin Lake, which are representative water supply source sites at the Yeongsan River and Seomjin River basins. We performed an exploratory data analysis on the monitored water quality parameters to understand overall fluctuations. Using data from 2017 to 2021 as training data and 2022 data as test data, we compared the dominant algal classification accuracy of 11 statistical machine learning algorithms. The results indicated that the optimal algorithm varied depending on the survey site and evaluation criteria, highlighting the unique environmental characteristics of each site. By predicting dominant algae in advance, stakeholders can better prepare for water source contamination accidents. Our findings demonstrate the applicability of machine learning algorithms as efficient tools for managing water quality in water supply source systems using monitoring data.

Keywords: water quality; Yeongsan River; Seomjin River; correlation analysis; self-organizing map; statistical machine learning algorithm; classification

Citation: Hwang, S-Y.; Choi, B-W.; Park, J-H.; Shin, D-S.; Chung, H-S.; Son, M-S.; Lim, C-H.; Chae, H-M.; Ha, D-W.; Jung, K-Y. Evaluating Statistical Machine Learning Algorithms for Classifying Dominant Algae in Juam Lake and Tamjin Lake, Republic of Korea. *Water* **2023**, *15*, 1738. https://doi.org/10.3390/w15091738

Academic Editor: Guangyi Wang

Received: 4 April 2023
Revised: 23 April 2023
Accepted: 27 April 2023
Published: 30 April 2023

1. Introduction

In South Korea, sites crucial for providing potable water to local residents are designated and managed as water protection zones. The importance of properly managing these water sources was underscored by the extreme drought in the Honam region of South Korea in 2022. To safeguard the water quality of these sources, the Korean government established an algae alert system in 1998. This system minimizes toxic effects caused by large numbers of harmful cyanobacteria by issuing alerts based on harmful cyanobacterial cell counts: Caution (at least 1000 cells mL^{-1} for 2 consecutive counts), Warning (at least 10,000 cells mL^{-1} for 2 consecutive counts), Outbreak (at least 1,000,000 cells mL^{-1} for 2 consecutive counts), and Release (number of cyanobacterial cells below the alert threshold for 2 consecutive counts) [1–3]. Specifically, four representative genera of

cyanobacteria, *Aphanizomenon*, *Anabaena*, *Oscillatoria*, and *Microcystis*, release harmful toxins causing acute liver disease in humans [4] and threatening the stability of aquatic ecosystems [5]. Researchers have explored various methods to reduce the abundance of these harmful cyanobacteria, including physical methods, such as algal blocking mats (ABM); chemical methods, such as plant–mineral composite (PMC) coagulants; and biological methods, such as using *Unio douglasiae* [6–8]. However, such methods are predominantly used reactively rather than proactively, i.e., they are used when water quality is declining or has already declined.

To predict future changes in water quality and, thus, enable more proactive management of water sources, recent studies have explored how to predict changes in specific water quality parameters, with particular focus on statistical machine learning techniques. Such techniques are being investigated because they are capable of processing large amounts of water quality-related data and can be used to compare the usefulness of different water quality parameters. In particular, multiple studies have focused on predicting values of a water quality parameter, chlorophyll-a (Chla). For instance, Kim, H. G. (2017) assessed the suitability of an artificial neural network technique for predicting Chla concentration at a midstream location in South Korea's Nakdong River [9]. Moreover, Lee et al. (2020) investigated the ability of four statistical machine learning algorithms to predict Chla concentrations [10]. Similarly, Bui et al. (2020) used 16 novel hybrid machine learning algorithms and various water quality parameters to predict changes in the Water Quality Index (WQI) [11]. However, this study was limited in its ability to thoroughly compare the performance of the 16 algorithms. In particular, this study did not apply the latest algorithms, such as AdaBoost or Gradient boosting. The primary difference between these previous studies and the current study is that the former studies focused on accurately predicting the measured values of the water quality parameter Chla (a continuous variable). In contrast, this study aims to accurately classify dominant algae (a categorical variable). Algal growth is influenced by many factors; the most important of which is the availability of nutrients, such as nitrogen (N) and phosphorus (P), and quality parameters, such as water temperature. However, hydraulic/hydrological factors, such as water level and water storage capacity, also play a role, necessitating the consideration of all factors [12].

Therefore, considering the diverse variables related to water quality and hydraulic/hydrological factors, accurately predicting the dominant algae could enable authorities to better prepare for and respond to algal water pollution incidents. For this study, we utilized the water quality monitoring network data, algae alert system data, and hydraulic/hydrological data collected from Juam Lake and Tamjin Lake. These representative water supply sources in the Yeongsan River and Seomjin River systems had measurements taken at seven-day intervals from January 2017 to December 2022 through the National Institute Environmental Research (NIER) Water Environment Information System. We compared and analyzed various statistical machine learning algorithms to determine their accuracy in classifying the dominant algae. By developing and implementing a predictive method for dominant algal occurrences, we aim to provide a more efficient approach to water quality management.

2. Materials and Methods

The methods for this study consisted of three main stages, namely data collection, exploratory data analysis, and a comparison of the classification performance of 11 selected algorithms. A flowchart of these key steps for the methodology is shown in Figure 1.

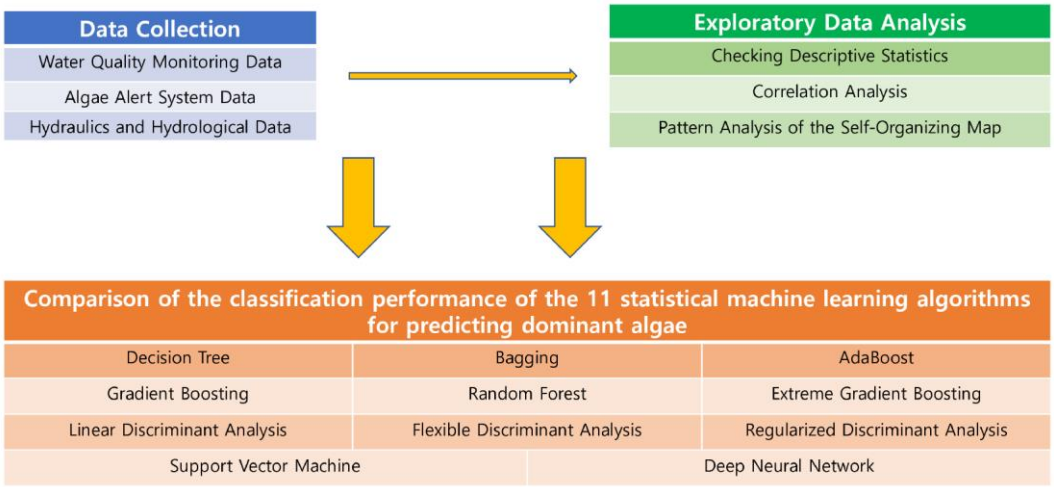

Figure 1. Methodological flowchart used in this study.

2.1. Study Area

This study focused on Juam Lake and Tamjin Lake, two representative water supply sources in the Yeongsan River and Seomjin River systems in South Korea. The NIER Yeongsan River Environment Research Laboratory collects weekly samples to monitor water quality and respond to the algae alert system from the dam front (J1, 127°14′26.74″E/35°03′23.78″N) and Shinpyeong Bridge (J2, 127°13′59.11″E/35°00′50.37″N) at Juam Lake, and the dam front (T1, 126°52′52.01″E/34°45′07.09″N) and Yuchi stream confluence (T2, 126°52′11.82″E/34°46′02.99″N) at Tamjin Lake. Additionally, the Korea Water Resources Corporation conducts daily measurements of hydraulic/hydrological variables, such as water storage capacity.

Juam Lake is an artificial lake formed by the freshwater held back by Juam Dam, which has a height of 58 m and a length of 330 m. It is located in Daegwang-ri, Juam-myeon, Suncheon-si, Jeollanam-do, and has a total basin area of 1010 km^2 and a total water storage capacity of 457 × 10^6 tons. Juam Dam supplies about 640 × 10^3 tons of potable water to the western part of Jeollanam-do, including Gwangju, Naju, Mokpo, and Hwasun [13]. Tamjin Lake is an artificial lake created by the construction of Jangheung Dam, which has a height of 53 m and a length of 403 m. It has a total basin area of 193 km^2 and a total water storage capacity of 191 × 10^6 tons. It is located in Yuchi-myeon, Jangheung-gun, Jeollanam-do, and supplies 73 × 10^6 tons of potable water to 9 cities in Jeollanam-do [14]. Figure 2 shows the sampling sites at Juam Lake and Tamjin Lake.

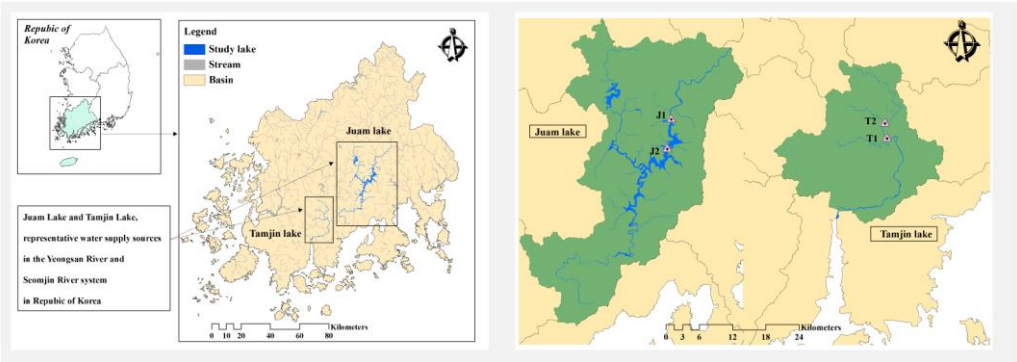

Figure 2. Sampling sites at Juam Lake and Tamjin Lake.

2.2. Data Collection

To conduct a comprehensive analysis, we collected and organized hydraulic/hydrological data, algae alert system data, and water quality monitoring network data from the survey sites. These data were measured at seven-day intervals from January 2017 to December 2022 and were obtained through the NIER Water Environment Information System. The number of observations for each sampling site was as follows: in Juam Lake, 307 observations at both the dam front (J1) and Shinpyeong Bridge (J2) sites, and in Tamjin Lake, 304 observations at both the dam front (T1) and Yuchi Stream Confluence (T2) sites. Overall, this study comprised a total of 1222 observations. For comparison of the performance of the statistical machine learning algorithms, the training data consisted of the measurements from 2017 to 2021 at each survey site, while the test data consisted of the remaining measurements from 2022. For the J1 site and J2 site for Juam Lake, the number of observations included in the training data and test data was 257 and 50, respectively. For the T1 site and T2 site for Tamjin Lake, the number of observations included in the training data and test data was 255 and 49, respectively. Table 1 shows the data variables used in this study.

Table 1. Data variables used in this study.

Response Variable (Categorical)	**Explanatory Variables (Continuous)**	
Dominant Algae (Based on Total Cell Count)	**Water Quality**	**Hydraulic/Hydrological**
Cyanophytes Diatoms Chlorophytes Others	Biological Oxygen Demand (BOD), mg L^{-1} Chemical Oxygen Demand (COD), mg L^{-1} Total Nitrogen (TN), mg L^{-1} Total Phosphorus (TP), mg L^{-1} Total Organic Carbon (TOC), mg L^{-1} Suspended Solids (SS), mg L^{-1} Electrical Conductivity (EC), μS cm^{-1} pH Dissolved Oxygen (DO), mg L^{-1} Temperature, °C Turbidity, NTU Transparency, m Chlorophyll a (Chla), mg m^{-3}	Low Water Level, cm Inflow Rate (Inflow), cms Discharge Rate (Discharge), cms Water Storage Capacity (Reservoir), 10,000 m^3

Of the variables listed in Table 1, biological oxygen demand (BOD), chemical oxygen demand (COD), total nitrogen (TN), total phosphorous (TP), total organic carbon (TOC), suspended solids (SS), and electrical conductivity (EC) were collected from the water quality monitoring network data, while pH, dissolved oxygen (DO), temperature, turbidity, transparency, Chla, and dominant algae were obtained from the algae alert system data. The remaining variables, including low water level, inflow, discharge, and reservoir, were collected from the National Water Resources Management Information System (http://www.wamis.go.kr/ (accessed on 5 February 2023)). The genera of algae that were found from the data collection at the sampling sites are presented in Table 2. Figure 3 shows line graphs of the monthly mean number of algal cells sampled during the survey period, categorized according to the survey site and algal genus. Based on the results in Table 2 and Figure 3, during the survey period, chlorophytes or diatoms tended to dominate in spring, cyanophytes in early summer and summer, and chlorophytes along with diatoms in autumn and early winter [15]. For clarity, in South Korea, the period from 25

June to 19 July is considered early summer, and the period from 20 July to 7 September is considered summer. All data analyses in this study were performed using the statistical software R, version 4.2.1.

Table 2. Genera of algae that were identified in the water samples collected from the sampling sites.

Cyanophytes		Diatoms	Chlorophytes	Others
Normal	**Harmful**			
Aphanocapsa	*Anabaena*	*Acanthoceras*	*Actinastrum*	*Ceratium*
Chroococcus	*Aphanizomenon*	*Achnanthes*	*Ankistrodesmus*	*Cryptomonas*
Merismopedia	*Microcystis*	*Asterionella*	*Ankyra*	*Dinobryon*
Phormidium	*Oscillatoria*	*Attheya*	*Chlamydomonas*	*Euglena*
Pseudanabaena		*Aulacoseira*	*Chlorella*	*Kephyrion*
Worinochinia		*Coccoineis*	*Chodatella*	*Mallomonas*
		Cyclotella	*Closteriopsis*	*Peridinium*
		Cymbella	*Closterium*	*Phacus*
		Fragilaria	*Coelastrum*	*Strombomonas*
		Gomphonema	*Coenochloris*	*Trachelomonas*
		Melosira	*Cosmarium*	
		Navicula	*Crucigenia*	
		Nitzschia	*Dictyosphaerium*	
		Rhizosolenia	*Dimorphococcus*	
		Stephanodiscus	*Elakatothrix*	
		Surirella	*Euastrum*	
		Synedra	*Eudorina*	
			Eunotia	
			Gloeocystis	
			Golenkinia	
			Gonium	
			Kirchnerionella	
			Micractinium	
			Monoraphidium	
			Mougeotia	
			Nephrocystium	
			Oocystis	
			Pandorina	
			Pectodictyon	
			Pediastrum	
			Scenedesmus	
			Schroederia	
			Selenastrum	
			Sphaerocystis	
			Spondylosium	
			Staurastrum	
			Tetraedron	
			Tetrastrum	
			Treubaria	

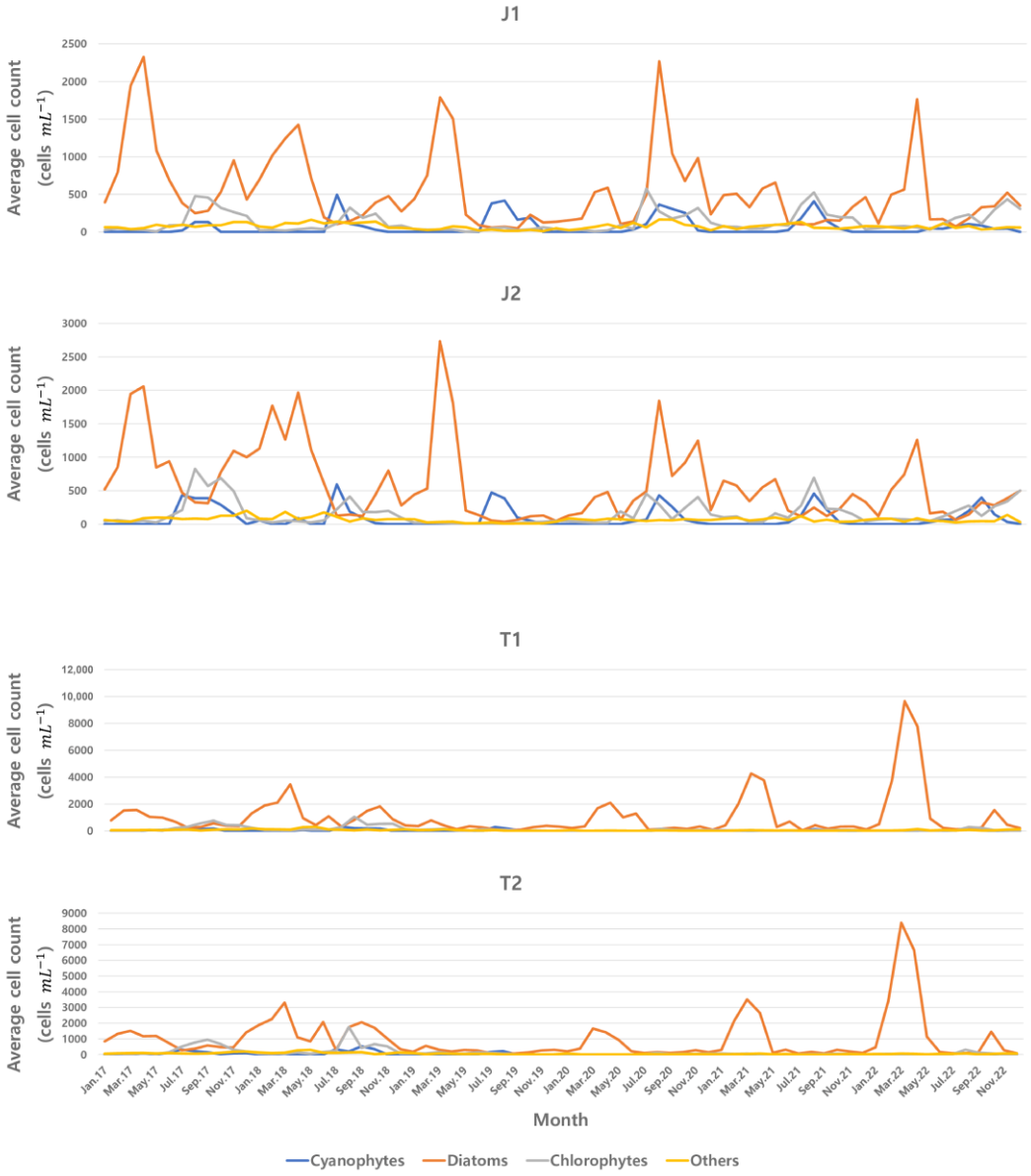

Figure 3. Line graphs of average algal cell count at the sampling sites J1, J2, T1, and T2 from January 2017 to November 2022.

2.3. Data Analysis Methods

This section describes the data analysis methods employed in this study, starting with exploratory data analysis. This includes correlation analysis and pattern analysis using a self-organizing map (SOM) to examine the overall distribution of water quality parameters and the hydraulic/hydrological variables included in the data. We also briefly explain the principles of the 11 statistical machine learning algorithms, which we compared against each other to assess their relative predictive power in classifying the dominant algae.

2.3.1. Exploratory Data Analysis

Before analyzing the data, an exploratory data analysis was performed to investigate the overall characteristics of the variables in the data, including descriptive statistics, such as mean or variance, and distribution [16]. While no specific analytical method or process exists, researchers may prefer different methods depending on their objectives. Generally, the first step is to determine whether the variables included in the data are continuous or

categorical. The mean, standard deviation, density, and other distributional characteristics were calculated for continuous variables. For categorical variables, the number of categories and the number of observations for each category were examined. In this study, we employed correlation analysis to investigate the relationship between water quality parameters and hydraulic/hydrological parameters, and we applied pattern analysis using an SOM to visually confirm the results.

1. Correlation analysis

Correlation is a widely used statistical analysis method for investigating the relationships between continuous variables in a dataset. For this purpose, the Pearson correlation coefficient was calculated as shown in Equation (1), and a significance test was conducted on the resultant coefficient. Generally, the validity of the analytical results can be confirmed only when normality is assumed to be satisfied through a normality test, such as the Shapiro–Wilk (SW) test [17]. However, this method is limited because it can only be applied when variables have the properties of random variables that satisfy independency. Since all measurement variables in this study are time series data measured over a given time period rather than random variables that satisfy independence, the Jarque–Bera (JB) test method was deemed more appropriate [18].

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}} \quad (1)$$

However, environment-related measurement variables typically do not satisfy normality and instead fluctuate considerably. Consequently, the analytical results lose reliability if conducted using a Pearson correlation coefficient for data with such variables. Therefore, we performed correlation analysis using the Spearman correlation coefficient, a non-parametric method that analyzes correlation based on ranks, as expressed in Equation (2):

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^{\,2}}{n(n^2-1)},\ d_i = \mathrm{rank}(x_i) - \mathrm{rank}(y_i) \quad (2)$$

2. Pattern analysis using SOM

An SOM is an artificial neural network technique that simultaneously performs dimension reduction and clustering [19]. With this technique, numerous nodes in high-dimensional data are clustered through competition. Based on the winning node that emerges from this competition, the learning results that preserve similarity as much as possible in the reduced dimensions are obtained. This principle is illustrated in Figure 4.

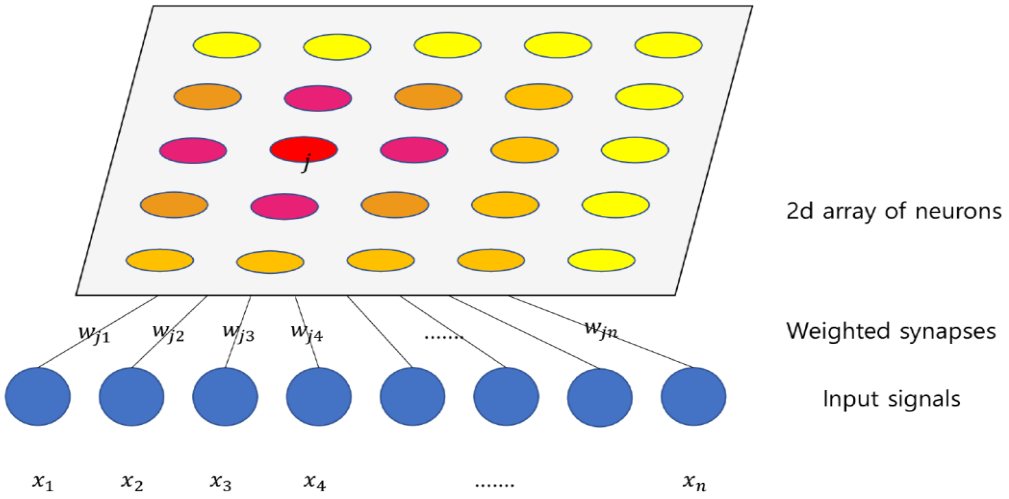

Figure 4. Schematic diagram of a self-organizing map.

This process repeats the algorithm shown in Equation (3) until convergence, and the jth lattice vector at time t is updated:

$$w_j^{t+1} = w_j^t + \eta_t \lambda_x^{j,t}(x - w_j^t) \quad (3)$$

In the above Equation (3), η is a learning rate parameter that reduces the learning rate to prevent overfitting, and λ is a parameter that makes the neighborhood size larger for the winning node and smaller for the distant nodes.

Through SOMs, Jung et al. (2020) performed a pattern analysis based on the water quality parameters measured at 28 sampling sites in the Nakdong River system in South Korea [20]. To determine which branches should be prioritized for management, they used a grading process through cluster analysis based on the characteristics of each site. From these findings, they were able to propose policy recommendations. In this study, we performed a pattern analysis on 17 measurement variables using this method and identified the correlations between them.

2.3.2. Statistical Machine Learning Algorithms for Dominant Algal Classification

We compared the performance of 11 statistical machine learning algorithms for classifying the dominant algae at each survey site. Detailed explanations of the principles of the applied algorithms can be found in the literature [21,22].

1. Three tree-based methods

A decision tree (DT) is a method for creating a decision model with a tree-like structure. The impurity of nodes is reviewed to select the optimal separation criteria for pruning. Mean squared error, calculated using Equation (4), is used for regression, and the Gini coefficient, calculated using Equation (5), or the entropy coefficient, calculated using Equation (6), is used for classification. Compared to other algorithms, decision trees are visually simple and relatively easy to interpret [23].

$$\mathrm{MSE}(t) = \frac{1}{n_t}\sum_{i=1}^{n_t}[y_i(t) - \overline{y_t}]^2 \quad (4)$$

$$\mathrm{Gini}(t) = 1 - \sum_{j=1}^{J} p_j{}^2\,(t) \quad (5)$$

$$\mathrm{Entropy}(t) = -\sum_{j=1}^{J} p_j\,(t)\log_2 p_j(t) \quad (6)$$

In contrast to the decision tree method, the bagging (Bag) method involves sampling with replacement. This allows observations extracted from the analysis data to be re-extracted in multiple samples (L_b, $b = 1,2,\ldots,B$) for analysis. This analysis first creates multiple decision tree models ($\varphi(x, L_b)$, $b = 1,2,\ldots,B$) and then averages ($\varphi_B(x) = \frac{1}{B}\sum_{b=1}^{B}\varphi\,(x, L_b)$) the prediction results obtained through this, or performs multiple voting ($\varphi_B(x) = \mathrm{Mode}\,\varphi(x, L_b)$) based on the classification results. "Mode" refers to the value with the highest frequency. Since the bagging technique uses survey data with the replacement method, it greatly reduces the variance of the created model compared to that of the decision tree model which is created once [24,25]. Figure 5 illustrates the principle of the bagging method.

Thirdly, the random forest (RF) method was proposed to address the shortcomings of bagging, such as a correlation between multiple decision tree models made by multiple samples. Similar to bagging, RF involves extracting multiple samples with replacement from the training data and fitting multiple decision tree models through them. However, in random forest, only a subset of the variables is randomly selected and used for each sample. This results in a better prediction or classification performance compared to

bagging. Moreover, the types of variables selected for each sample differ, reducing the erroneous correlation between each sample that can occur with bagging [26].

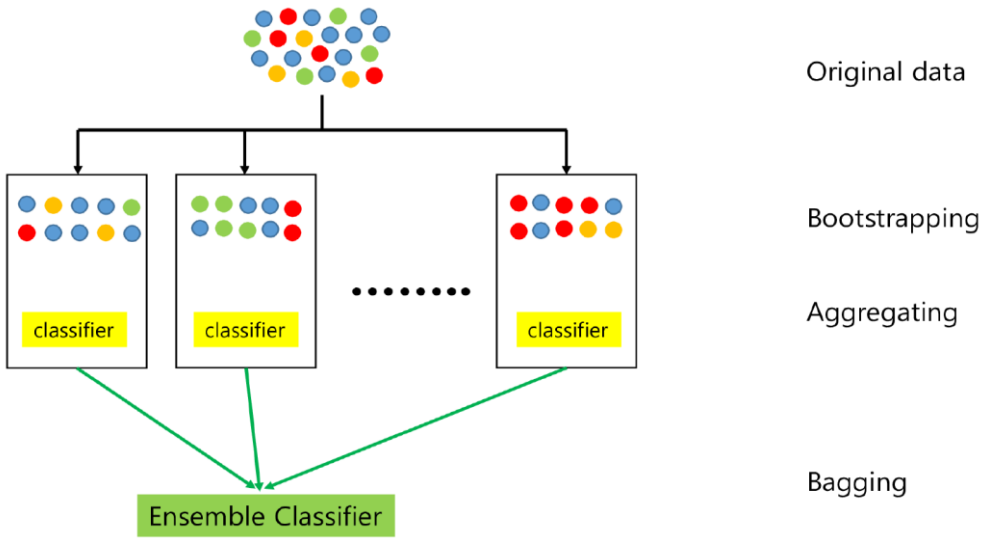

Figure 5. Schematic diagram of the bagging method.

2. AdaBoost (Ada)

AdaBoost is a boosting algorithm that creates a strong learner by taking a weighted linear combination of multiple weak learners, as shown in Equation (7). By correcting or supplementing incorrectly predicted or classified instances from previous steps, it can yield more accurate results than possible through the three tree-based methods:

$$H(x) = w_1 h_1(x) + w_2 h_2(x) + \cdots + w_T h_T(x) = \sum_{t=1}^{T} w_t\, h_t(x) \quad (7)$$

where $H(x)$ is the final strong learner obtained, $h_t(x)$, $t = 1,2,\ldots,T$ are the T weak learners, and w_t, $t = 1,2,\ldots,T$ are the weights of the weak learners. Figure 6 illustrates the principle of AdaBoost. A more detailed explanation can be found in the literature [27].

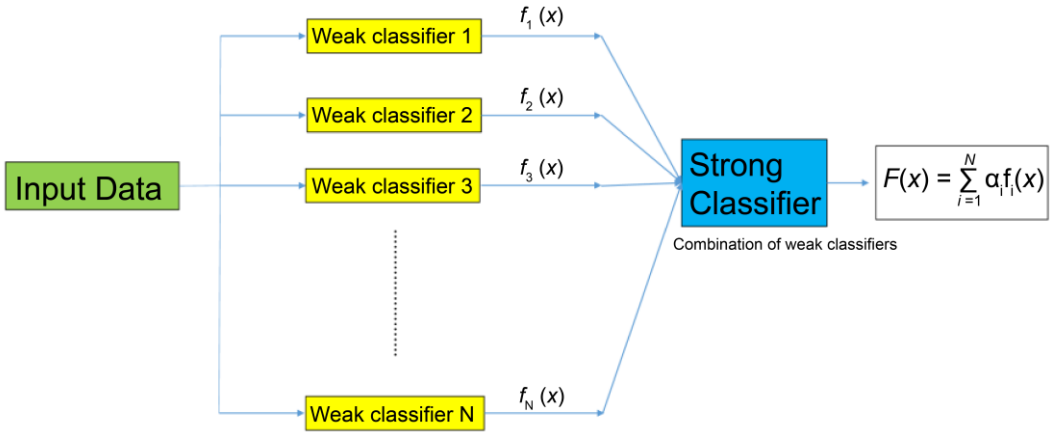

Figure 6. Schematic diagram of the AdaBoost method.

3. Two gradient boosting methods

Gradient boosting (GB) involves iteratively using a gradient to create a model, and then using the residual from this to create another model. This process reduces the portion of variation that the previous model could not explain, thereby reducing bias. If a given training dataset is (x_i, y_i), $i = 1,2,\ldots,n$ and the previously created model is G_{old}, then

gradient boosting gradually finds the function r that models the residual, which is the difference between the actual value and the predicted value, as shown in Equation (8):

$$y_i = G_{old}(x_i) + r(x_i),\ i = 1,2,\ldots,n \tag{8}$$

After the function r is found in this process, the new model is updated as shown in Equation (9):

$$G_{update}(x) = G_{old}(x) + \lambda r(x),\ 0 \langle \lambda \langle 1 \tag{9}$$

where parameter λ is the learning rate. This process reduces the risk of overfitting. Figure 7 illustrates the principle of gradient boosting. A more detailed explanation can be found in the study by Natekin et al. [28].

Extreme gradient boosting (XGB) is an improved method that addresses the slow execution time and overfitting risks of gradient boosting by supporting parallel learning. It has a self-regulating function that makes it more stable and durable. Traditionally, after randomly dividing the training data into n parts, $n-1$ data parts are used as new training data and the remaining 1 data part are used as new test data to evaluate the performance of the algorithm. The cross-validation test performs this process on all n parts of the data, as shown in Figure 8. A detailed explanation can be found in a paper by Chen et al. [29].

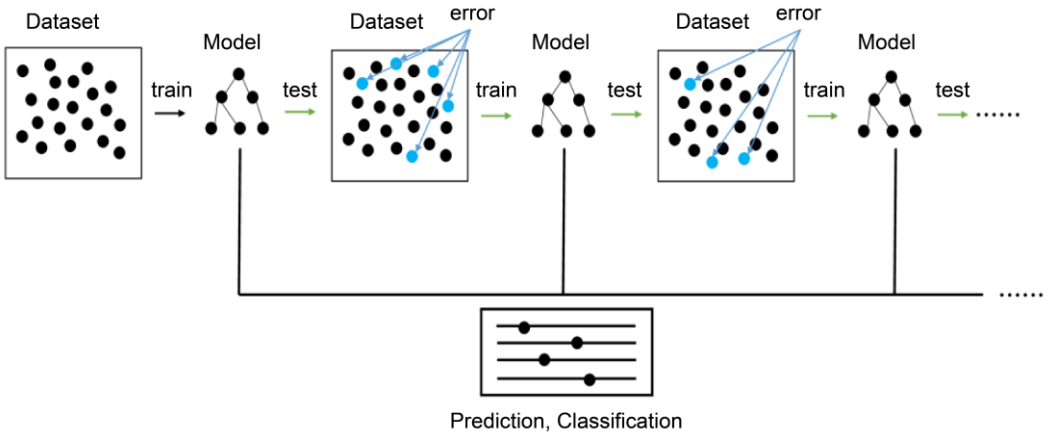

Figure 7. Schematic diagram of the gradient boosting method.

Figure 8. Example of k-fold cross-validation test.

4. Three discriminant analysis methods

Linear discriminant analysis (LDA) is a classification method using R. A. Fisher's linear decision boundary. The given data are projected onto a specific one-dimensional axis, followed by a process that finds the optimal straight line that properly distinguishes the categories. This process makes it possible to find the linear decision boundary, as shown in Figure 9. A more detailed explanation can be found in an article by Izenman, A. J. [30].

Flexible discriminant analysis (FDA) is a method that addresses the limitations of linear discriminant analysis. Instead of relying on linear decision boundaries, FDA uses splines to create a non-linear decision boundary for classification. This allows non-linear relationships to be captured and improves the overall classification accuracy [31].

Finally, when the data contain many explanatory variables, regularized discriminant analysis (RDA) improves the estimation of the covariance matrix through regularization (e.g., shrinkage) to create a decision boundary with better classification performance. For this, the optimal parameter α is estimated based on the training data; if $\alpha = 1$, then linear discriminant analysis is performed, and if $\alpha = 0$, then quadratic discriminant analysis is performed. Here, $0 \leq \alpha \leq 1$, which serves as the weight for the linear decision boundary and quadratic curved decision boundary [32].

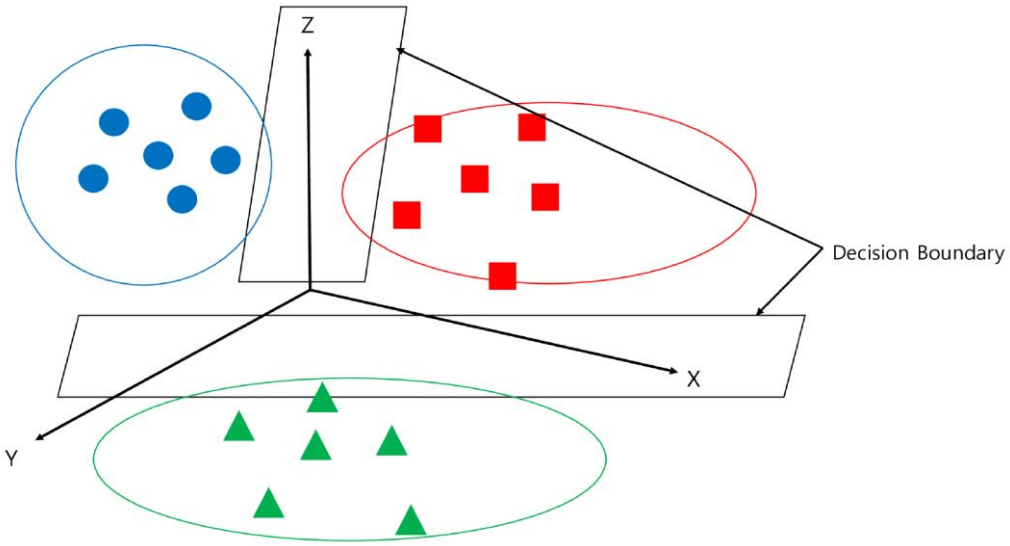

Figure 9. Schematic diagram of the linear discriminant analysis method.

5. Support Vector Machine (SVM)

SVM is a classification algorithm that maximizes the margin, i.e., the distance between the decision boundary and the support vectors. To move the original data in an input space with a complex non-linear distribution to a high-dimensional feature space, SVM uses the kernel method. This technique applies a mapping function without setting a transformation function beforehand. The kernel method converts the data into a linear distribution and makes it easier to find the decision boundary [33]. Figure 10 illustrates this concept. This study used the radial basis kernel, shown in Equation (10), which is known to be the most flexible kernel type for all data distributions. Figure 11 illustrates the principle of the support vector machine, and a detailed explanation can be found in the study by Pisner et al. [34].

$$k(u, v) = \langle \varphi(u), \varphi(v) \rangle = \exp[-\gamma \, || \, u - v \, ||^2] \quad (10)$$

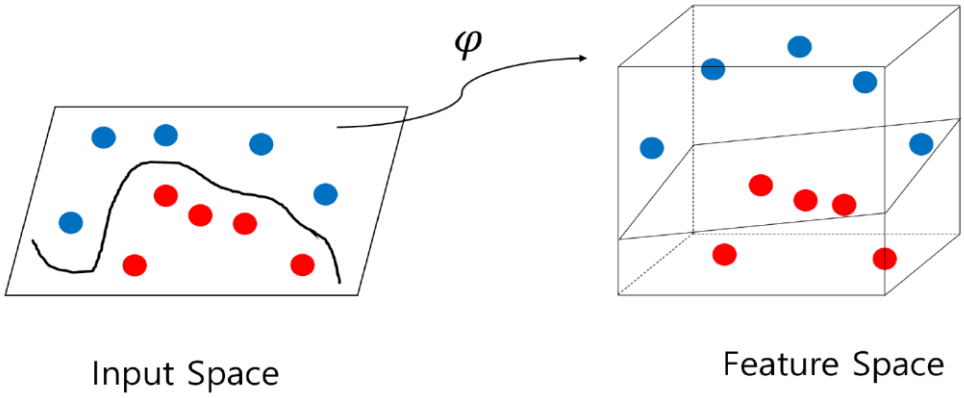

Figure 10. Schematic diagram of the kernel trick.

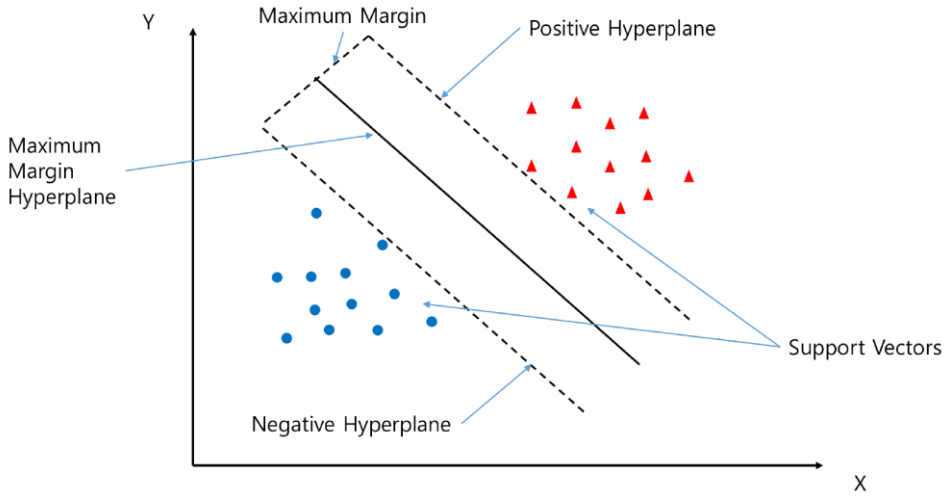

Figure 11. Schematic diagram that illustrates the concept of support vector machine.

6. Deep Neural Network (DNN)

A deep neural network is a model in the form of a neural network created by constructing multiple hidden layers between the input and output layers. The model is trained through a backpropagation algorithm that updates the weights through stochastic gradient descent, as shown in Equation (11).

$$w_{ij}(t+1) = w_{ij}(t) + \eta \frac{\partial C}{\partial w_{ij}(t)} \quad (11)$$

where η is the parameter that controls the learning rate, and C is the cost function. Typically, before executing a deep neural network, the appropriate activation function and cost function are determined according to the analysis conditions. In multiclass classification, the activation function is set to a softmax function, as shown in Equation (12), and the cost function is set to a cross-entropy function, as shown in Equation (13). A detailed explanation of deep neural networks can be found in a paper by Montavon et al. [35].

$$p_j = \frac{\exp(x_j)}{\sum_{k=1}^{K} \exp(x_k)} \quad (12)$$

$$C = -\sum_{j} p_j \log(p_j) \quad (13)$$

2.3.3. Evaluation Indexes

To evaluate the classification accuracy of the statistical machine learning algorithms, three representative criteria were used: accuracy, sensitivity, and specificity [36]. These criteria were calculated using a confusion matrix that organized the actual correct answers and those answers predicted from the classification, as shown in Table 3.

Table 3. Confusion matrix of dominant algal classification.

		Predicted			
		Cyanophytes	**Diatoms**	**Chlorophytes**	**Others**
Actual	Cyanophytes	n_1	n_2	n_3	n_4
	Diatom	n_5	n_6	n_7	n_8
	Chlorophytes	n_9	n_{10}	n_{11}	n_{12}
	Others	n_{13}	n_{14}	n_{15}	n_{16}

"Accuracy" simply refers to the ratio of observations that match the correct answer through classification among all observations and can be calculated as shown in Equation (14) using the table above.

$$acc = \frac{n_1+n_6+n_{11}+n_{16}}{\sum_{i=1}^{16} n_i} \quad (14)$$

The advantages of accuracy are that it is easy to calculate and can be understood intuitively. However, as it simply takes the arithmetic average, the imbalance between each class can be severe when using imbalanced data. To compensate for this shortcoming, we also calculated sensitivity and specificity for the four algae categories (cyanophytes, diatoms, chlorophytes, and others). Specifically, we calculated weighted sensitivity and weighted specificity by taking the weighted average of the data, and we used these two metrics as additional criteria to evaluate the algorithms. Sensitivity and specificity can be understood through the binary confusion matrix shown in Table 4.

Table 4. Binary confusion matrix.

		Predicted	
		Positive	**Negative**
Actual	**Positive**	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Sensitivity is the ratio of observations properly classified as positive compared to those that are actually positive, whilst specificity is the ratio of observations properly classified as negative compared to those that are actually negative [37]. Both ratios range from 0 to 1, with values closer to 1 indicating better algorithm performance. This is expressed in Equation (15):

$$sen = \frac{TP}{TP + FN}, spe = \frac{TN}{TN + FP} \quad (15)$$

For multiclass classification with at least three classes of categorical variables, as in this study, sensitivity and specificity are calculated using the binary confusion matrix for each class. For weighted sensitivity and weighted specificity, the weighted average of each class is used [38]. Hence, to create a binary confusion matrix for the diatom category, we can set diatoms to "positive" and the remaining categories (cyanophytes, chlorophytes, and others) to "negative". The weighted sensitivity and weighted specificity are expressed in Equation (16). i, p_i, sen_i, and spe_i are the serial number for each category, the probability of being included in each category, and the sensitivity and specificity for each category, respectively.

$$\mathrm{sen_w} = \sum_{i=1}^{4} p_i \times \mathrm{sen_i}, \quad \mathrm{spe_w} = \sum_{i=1}^{4} p_i \times \mathrm{spe_i} \tag{16}$$

Moreover, there is a trade-off relationship between sensitivity and specificity, where one decreases if the other increases [39]. Therefore, we additionally defined G mean, which can serve as a suitable supplementary point for these two metrics. This was obtained by taking the square root of the product of weighted sensitivity and weighted specificity as in Equation (17). We applied this form because the measurement data are imbalanced toward the diatom category.

$$G_m = \sqrt{\mathrm{sen_w} \times \mathrm{spe_w}} \tag{17}$$

3. Results

3.1. Data Analysis

3.1.1. Exploratory Data Analysis for Monitoring Data

The descriptive statistics of the variables for each sampling site are presented in Table 5. This shows an overview of the distributions of measurement variables for each sampling site [40]. We also calculated the JB test *p*-value for each variable to determine the normality test results. To identify the overall distribution of each explanatory variable, seven descriptive statistics were calculated: mean, standard deviation, median, minimum, maximum, skewness, and kurtosis. Skewness has a positive value when the tail is long toward the right and a negative value when the tail is long toward the left. A kurtosis value > 0 indicates that the center of the distribution is sharp, and a value < 0 suggests that the center of the distribution is smooth [41]. According to Table 5, none of the measurement variables show a value of zero for skewness or kurtosis at any of the sampling sites.

Furthermore, except for pH at all sampling sites, the JB test *p*-value is significantly lower than the significance level of 0.05. Hence, since normality is often violated, the Spearman correlation coefficient needed to be used instead of the Pearson correlation coefficient in the correlation analysis [42].

To better visualize the results, Figure 12 presents the boxplots of the parameters at each survey site. In general, the Tamjin Lake sampling sites have higher water quality parameter values and hydraulic/hydrological data values than those in Tamjin Lake. However, the turbidity and transparency values are higher at the Juam Lake sampling sites than those at Tamjin Lake, while DO and temperature show similar trends for the sampling sites of both lakes.

Table 5. Descriptive statistics of the water quality and hydraulic/hydrological parameters at each survey site.

Survey Site	Statistics	BOD (mg L^{-1})	COD (mg L^{-1})	TN (mg L^{-1})	TP (mg L^{-1})	TOC (mg L^{-1})	SS (mg L^{-1})	EC (μS cm^{-1})	pH	DO (mg L^{-1})	Temperature (°C)	Turbidity (NTU)	Transparency (m)	Chla (mg m^{-3})	Low Water Level (cm)	Inflow (cms)	Discharge (cms)	Reservoir (10,000 m^3)
J1	mean	0.9500	3.3000	0.6800	0.0100	2.3400	1.8000	81.8000	7.2600	8.6200	14.1600	2.3200	3.1700	3.3300	7260.4900	3.6100	4.1200	9945.4800
	sd	0.3800	0.4300	0.1300	0.0100	0.4100	1.0400	8.8200	0.4300	2.2200	5.6300	2.0100	1.0000	2.6900	620.7500	11.3300	7.7000	3298.5100
	median	0.9000	3.3000	0.6600	0.0100	2.3000	1.5000	80.0000	7.2000	8.5000	14.7000	1.7000	3.2000	2.7000	7238.0000	0.8000	2.7000	9968.0000
	min	0.4000	2.4000	0.4400	0.0000	1.4000	0.5000	62.0000	6.1000	4.6000	2.1000	0.1000	0.7000	0.3000	6167.0000	0.0000	1.7000	3552.0000
	max	2.6000	4.7000	1.0800	0.0500	3.5000	10.6000	101.0000	8.8000	12.9000	24.6000	15.4000	7.2000	25.2000	9638.6200	162.6300	93.2100	16947.0000
	skewness	1.3100	0.5900	0.3800	1.1300	0.8600	3.1000	0.4500	0.1200	0.0700	−0.0900	3.0800	0.2700	3.6000	1.4300	9.9200	8.2600	0.0500
	kurtosis	2.4400	0.1900	−0.4500	2.4900	0.4700	17.9200	−0.7600	0.0300	−1.3300	−1.3000	14.7300	0.4000	19.1600	4.0500	127.3000	78.2200	−0.6800
	JB test *p*-value	0.0000	0.0001	0.0071	0.0000	0.0000	0.0000	0.0002	0.6839	0.0000	0.0000	0.0000	0.0517	0.0000	0.0000	0.0000	0.0000	0.0536
J2	mean	0.9500	3.3000	0.6800	0.0100	2.3400	1.8000	81.8000	7.3100	8.8000	14.8300	2.4300	3.1200	3.7200	7260.4900	3.6100	4.1200	9945.4800
	sd	0.3800	0.4300	0.1300	0.0100	0.4100	1.0400	8.8200	0.4500	2.1700	6.0000	3.5800	0.9300	2.1600	620.7500	11.3300	7.7000	3298.5100
	median	0.9000	3.3000	0.6600	0.0100	2.3000	1.5000	80.0000	7.3000	8.8000	15.2000	1.6000	3.0000	3.4000	7238.0000	0.8000	2.7000	9968.0000
	min	0.4000	2.4000	0.4400	0.0000	1.4000	0.5000	62.0000	5.8000	4.1000	2.1000	0.1000	0.7000	0.2000	6167.0000	0.0000	1.7000	3552.0000
	max	2.6000	4.7000	1.0800	0.0500	3.5000	10.6000	101.0000	8.6000	12.9000	25.9000	34.6000	6.0000	22.2000	9638.6200	162.6300	93.2100	16947.0000
	skewness	1.3100	0.5900	0.3800	1.1300	0.8600	3.1000	0.4500	−0.1600	−0.0700	−0.0800	5.6000	0.3400	2.9700	1.4300	9.9200	8.2600	0.0500
	kurtosis	2.4400	0.1900	−0.4500	2.4900	0.4700	17.9200	−0.7600	0.4200	−1.1100	−1.3100	38.0000	0.4300	18.6800	4.0500	127.3000	78.2200	−0.6800
	JB test *p*-value	0.0000	0.0001	0.0071	0.0000	0.0000	0.0000	0.0002	0.1423	0.0004	0.0000	0.0000	0.0137	0.0000	0.0000	0.0000	0.0000	0.0536
T1	mean	2.2700	5.4100	1.6300	0.1000	3.8700	10.6300	185.9900	7.2700	9.1700	13.8100	2.4200	2.6300	5.1600	9908.6600	15.0100	17.0300	23247.5400
	sd	1.3600	1.5700	0.6600	0.0400	1.0000	11.0200	65.1200	0.4200	2.1900	5.4700	2.8400	0.7300	2.9300	724.5500	39.1100	37.2400	8015.6200
	median	2.0000	5.0000	1.4700	0.0900	3.7000	8.4000	178.0000	7.3000	9.1000	14.5000	1.7000	2.6000	4.4000	10063.0000	3.7000	11.3800	23560.0000
	min	0.5000	2.6000	0.6400	0.0300	1.9000	1.3000	68.0000	6.1000	4.8000	1.5000	0.1000	1.0000	0.4000	6805.0000	0.0000	1.9400	7105.0000
	max	8.8000	13.0000	4.9400	0.2300	6.7000	93.2000	600.0000	8.6000	13.6000	24.8000	36.0000	7.0000	19.0000	10704.0000	310.6300	464.6000	37807.0000
	skewness	2.0100	1.3200	1.7500	0.9100	0.4900	4.0800	1.5300	0.1700	0.0300	−0.1900	6.6500	1.0100	1.4100	−2.5200	4.5800	8.5400	−0.0800
	kurtosis	5.4200	2.5300	4.6100	0.1400	−0.2500	20.9300	5.7900	−0.0500	−1.1300	−1.0800	66.1100	3.8700	2.8500	8.1000	23.3700	83.5000	−0.9600
	JB test *p*-value	0.0000	0.0000	0.0000	0.0000	0.0016	0.0000	0.0000	0.4793	0.0003	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0029
T2	mean	2.2700	5.4100	1.6300	0.1000	3.8700	10.6300	185.9900	7.3900	8.8400	13.7200	2.4100	2.5600	4.8700	9908.6600	15.0100	17.0300	23247.5400
	sd	1.3600	1.5700	0.6600	0.0400	1.0000	11.0200	65.1200	0.5900	2.3400	5.5100	2.2400	0.7300	3.0400	724.5500	39.1100	37.2400	8015.6200
	median	2.0000	5.0000	1.4700	0.0900	3.7000	8.4000	178.0000	7.3000	8.7000	14.1500	1.8000	2.5000	4.1000	10063.0000	3.7000	11.3800	23560.0000

min	0.5000	2.6000	0.6400	0.0300	1.9000	1.3000	68.0000	5.6000	4.1000	1.7000	0.1000	1.0000	0.0000	6805.0000	0.0000	1.9400	7105.0000
max	8.8000	13.0000	4.9400	0.2300	6.7000	93.2000	600.0000	9.0000	13.2000	28.0000	20.7000	7.0000	17.6000	10704.0000	310.6300	464.6000	37807.0000
skewness	2.0100	1.3200	1.7500	0.9100	0.4900	4.0800	1.5300	−0.0100	0.0300	−0.0400	4.0500	1.0900	1.2600	−2.5200	4.5800	8.5400	−0.0800
kurtosis	5.4200	2.5300	4.6100	0.1400	−0.2500	20.9300	5.7900	0.0300	−1.1700	−0.9600	23.3800	4.2200	1.7600	8.1000	23.3700	83.5000	−0.9600
JB test *p*-value	0.0000	0.0000	0.0000	0.0000	0.0016	0.0000	0.0000	0.9832	0.0002	0.0033	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0029

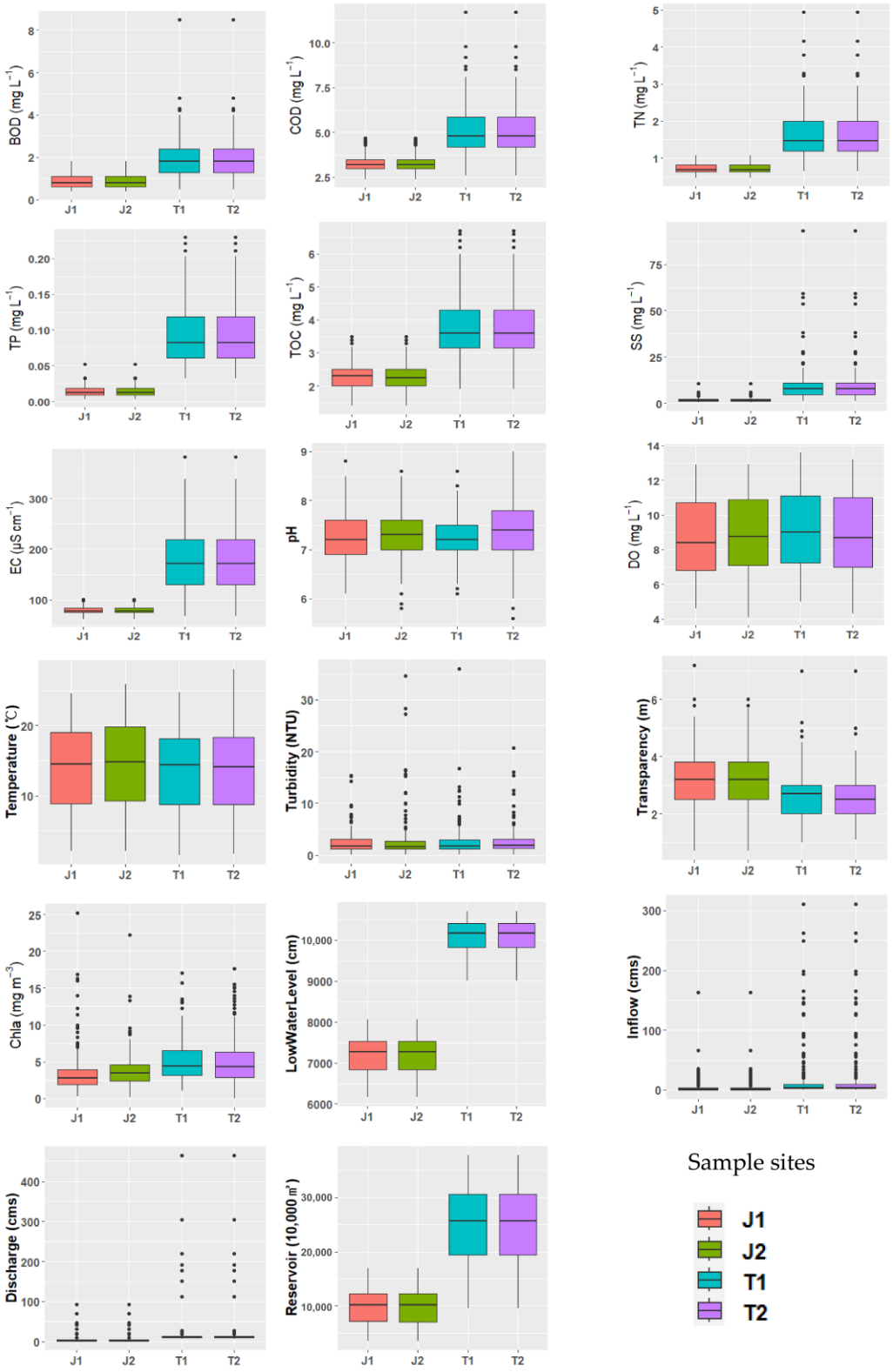

Figure 12. Boxplot of the data obtained at the four sampling sites J1, J2, T1, and T2.

Table 6 presents a contingency table of the variable "Dominant Algae", a categorical variable. The table indicates that diatoms are dominant at all sampling sites during the monitoring period, followed by chlorophytes, cyanophytes, and other algae.

Table 6. Contingency table of the variable "Dominant Algae".

Survey Site	Cyanophytes	Diatoms	Chlorophytes	Others
J1	23	215	52	17
J2	31	218	49	9
T1	14	250	36	4
T2	12	250	33	9

3.1.2. Correlation Analysis and SOM Pattern Analysis

In Section 3.1.1., we confirmed that the Spearman correlation coefficient, a non-parametric measure of rank correlation, must be applied for the correlation analysis. Using this, we performed a correlation analysis for each sampling site; the results of which are shown in Figure 13. The figures for each sampling site show the calculated Spearman correlation coefficients. According to the results of the correlation analysis, there are variations in the results at each survey site; however, in general, the water quality parameters that are mutually related (BOD, COD, TN, TP, etc.) show positive correlations, while the water quality and hydraulic/hydrological variables show negative correlations. The pattern analysis of the SOMs supports these results, as shown in Figures 14–17. This analysis helped identify the overall movement of the measurement variables at each survey site during the survey period. The water quality parameters that exhibit significant positive correlations in the correlation analysis show similar patterns, while the water quality and hydraulic/hydrological variables that exhibit significant negative correlations show opposite patterns. However, it should be noted that this study used time series data, which are measured over a certain period and are not independent. As such, calculating the normality test *p*-value for each time-dependent measurement variable and performing a correlation analysis and interpretation based on this have limitations [43].

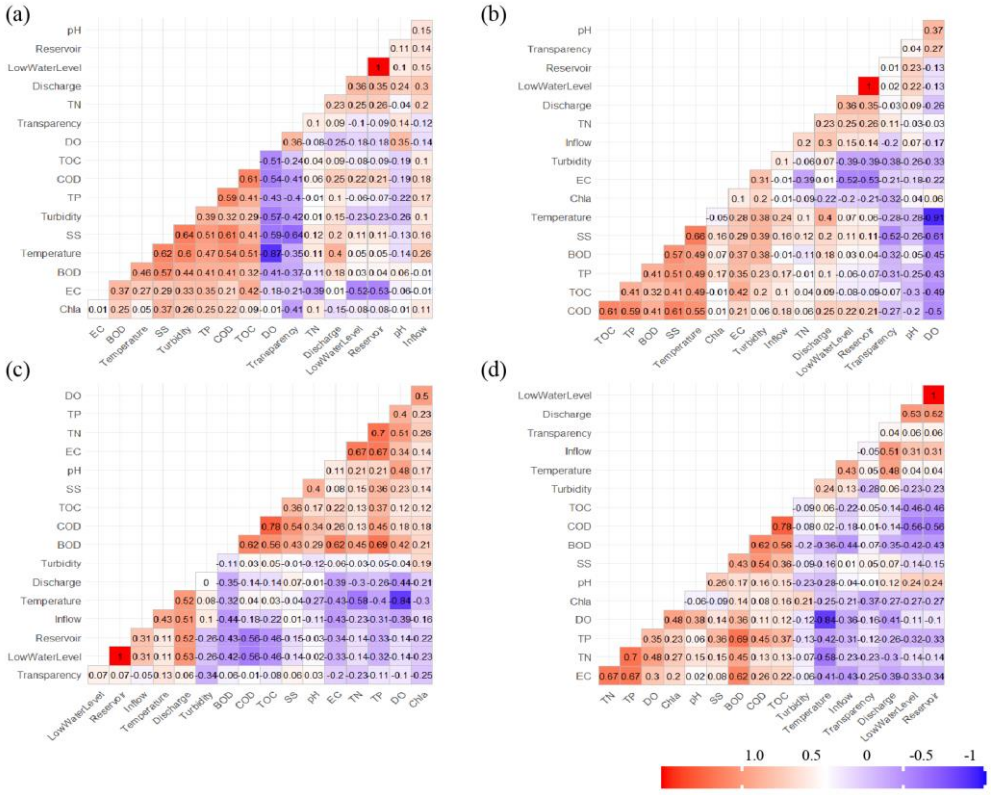

Figure 13. Correlation matrix showing Spearman's correlation analysis of water quality parameters at survey sites (**a**) J1, (**b**) J2, (**c**) T1, and (**d**) T2. The numbers inside the boxes represent the Spearman correlation coefficients.

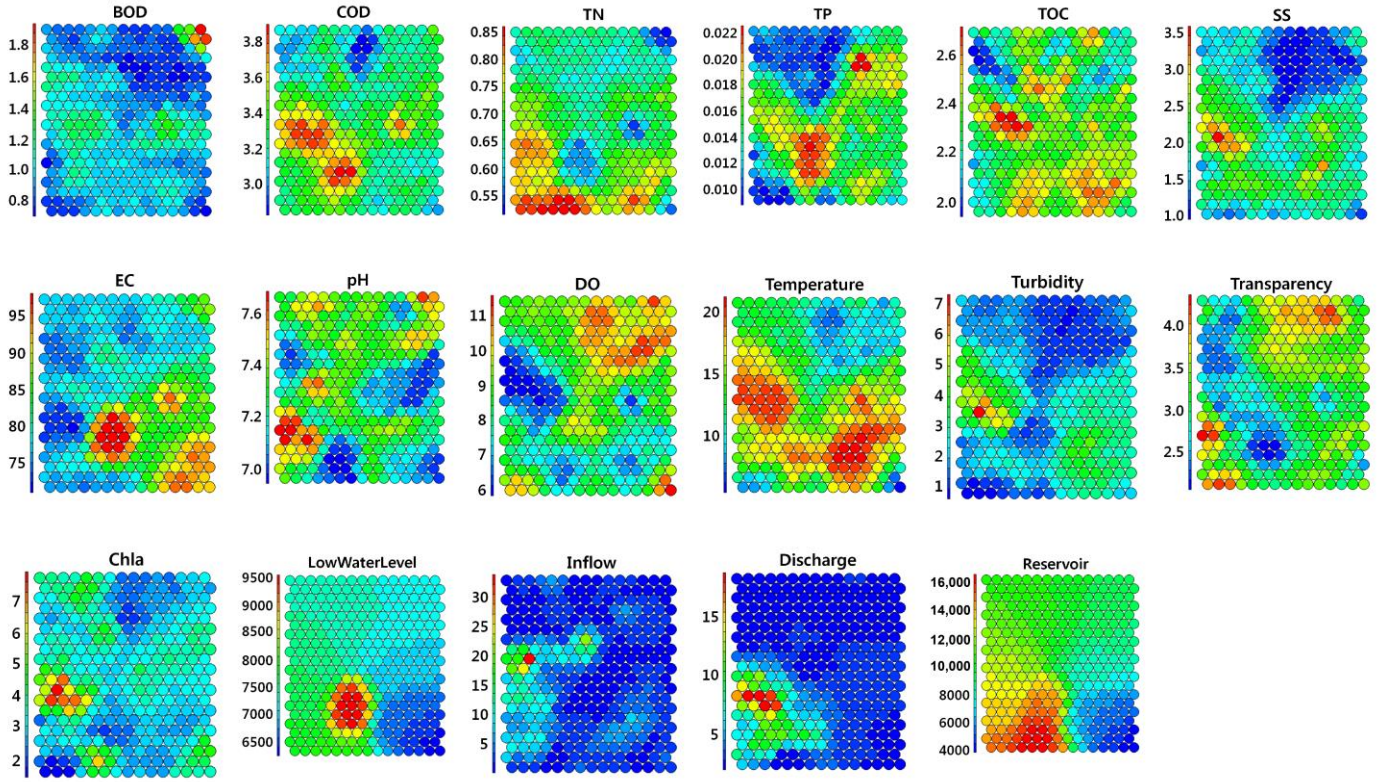

Figure 14. Self-organizing map for sampling site J1.

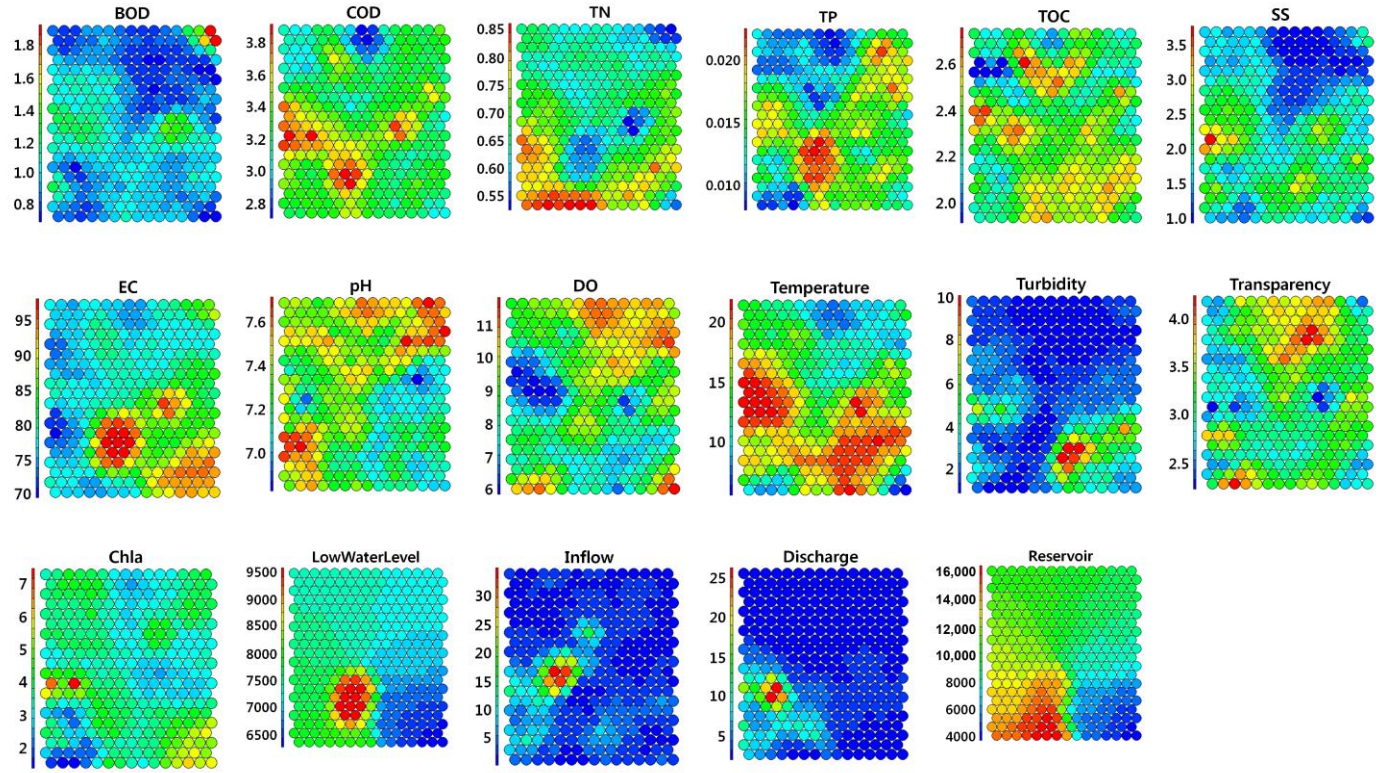

Figure 15. Self-organizing map for sampling site J2.

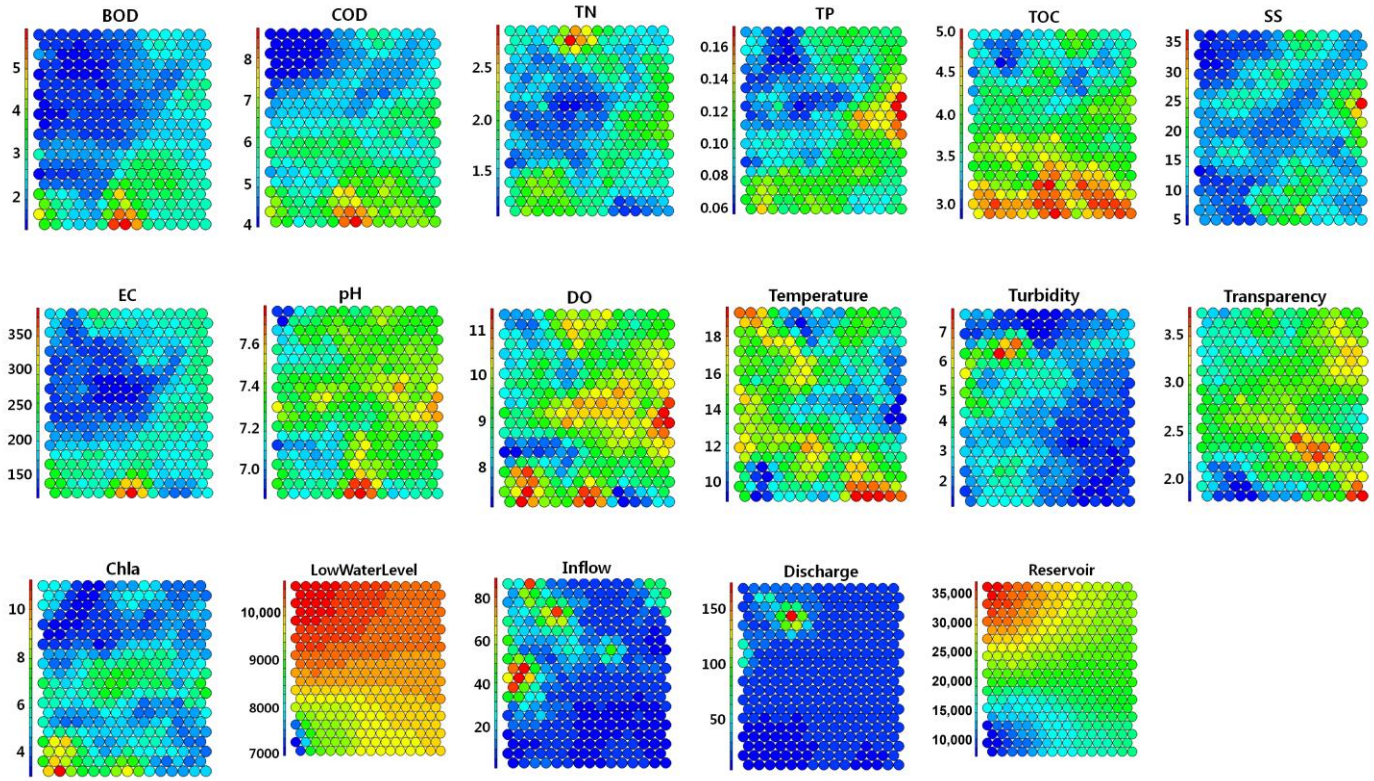

Figure 16. Self-organizing map for sampling site T1.

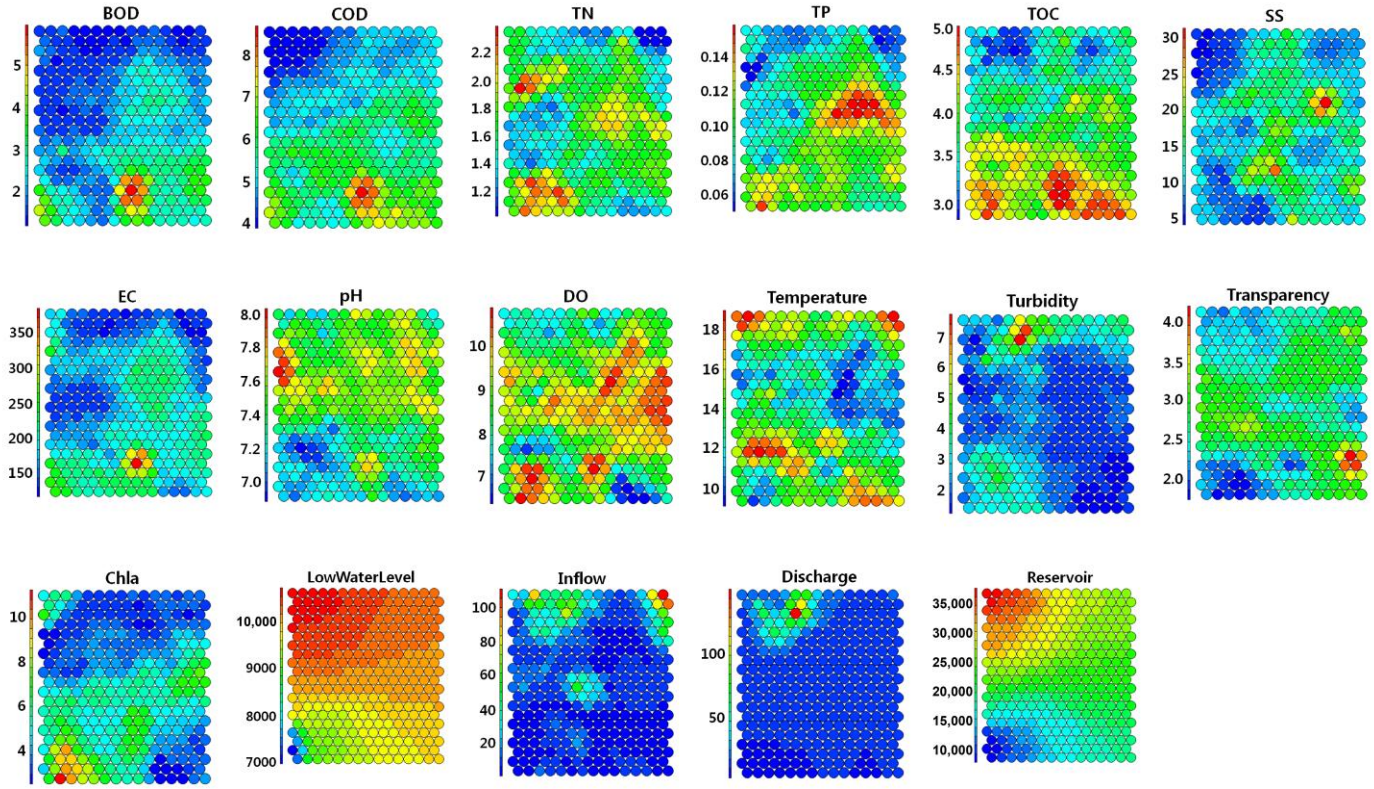

Figure 17. Self-organizing map for sampling site T2.

3.2. Comparison of the Performance of the Statistical Machine Learning Algorithms

This section presents the results of the analysis of the dominant algal classification accuracy using 11 statistical machine learning algorithms.

3.2.1. Tree-Based Algorithm for Assessing Variable Importance

In this study, the classification performance of five tree-based algorithms, namely random forest, bagging, AdaBoost, gradient boosting, and extreme gradient boosting, was compared. Each algorithm computes the importance of each variable to determine which explanatory variable has the most influence on the response variable [44]. Variable importance increases as the reduction in the Gini coefficient or the sum of squared errors increases. In extreme gradient boosting, variable importance is calculated using three measurement criteria: gain, cover, and frequency.

Figure 18 presents the graphs of the error calculated when applying the random forest algorithm based on the training data at each sampling site. The OOB (out-of-bag) error in the legend refers to the error obtained by using the remaining data not included in the sampling with replacement, which allows duplication, from the training data as validation data [45]. The other items in the legend indicate the probability of an incorrect answer calculated as the error for each category when the dominant algae are classified as either cyanophytes, diatoms, chlorophytes, or other algae. Figure 18 demonstrates that each error converges to a specific value as the number of tree models used in random forest increases. The probability of error is the lowest when probabilistically judging that the dominant algae are diatoms. This confirms that the most frequent time points during the survey period were those when diatoms dominated. Figure 19 presents the cross-validation tests conducted by extreme gradient boosting, where the point indicating the smallest mlogloss error value is deemed the best iteration. As illustrated in Figure 19, the mlogloss error value progressively decreases with each iteration for the training data, but it increases after a certain point for the test data, indicating overfitting [46]. Therefore, one of the advantages of extreme gradient boosting is that it reduces the risk of overfitting through cross-validation.

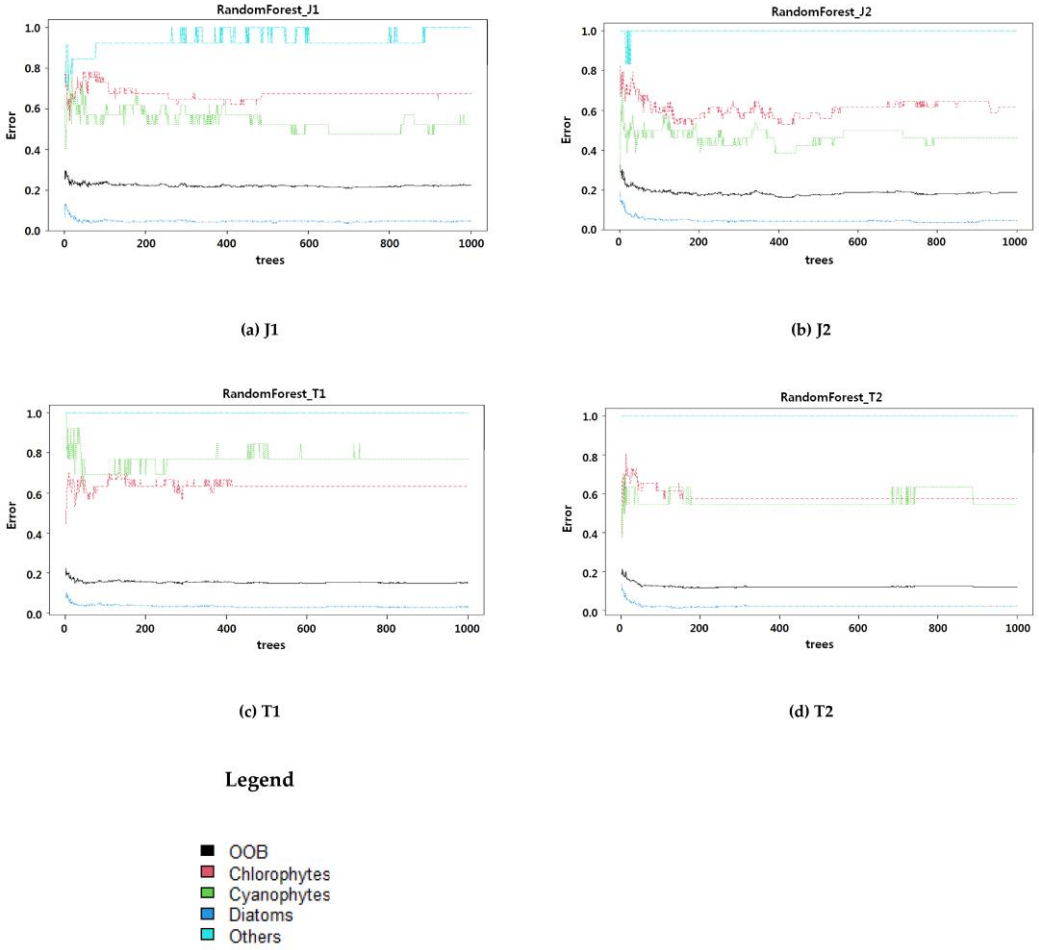

Figure 18. Graphs representing error when using random forest for the sampling sites (**a**) J1, (**b**) J2, (**c**) T1, and (**d**) T2.

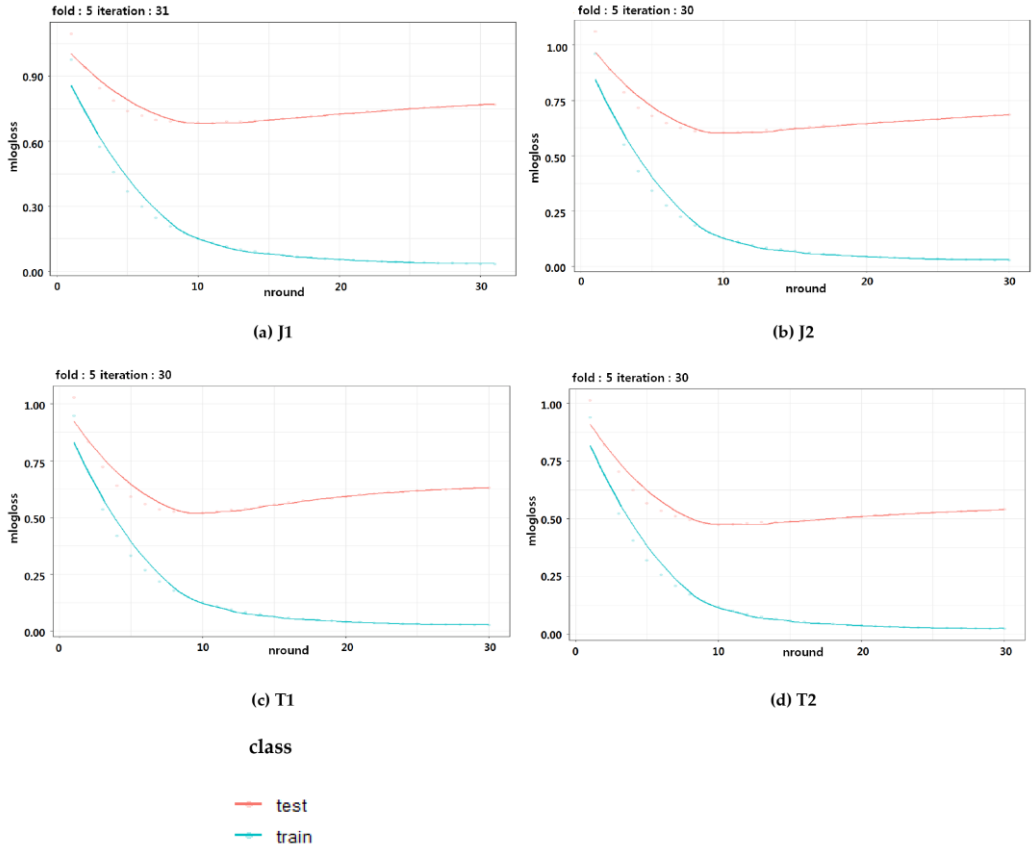

Figure 19. Cross-validation test when extreme gradient boosting for the sampling sites (**a**) J1, (**b**) J2, (**c**) T1, and (**d**) T2.

Using this process, the variable importance of each algorithm for the training data by survey site was calculated, with the results shown in Tables 7 and 8. Based on the results, the variable importance calculations vary depending on the survey site and algorithm. Overall, temperature and DO are more important than other measurement variables in determining and classifying the dominant algae at a specific point for each survey site. This observation suggests a high correlation between water temperature and oxygen in terms of the possibility of algal occurrence.

These results align with the findings of Woo et al. (2020), who reported that the amount of harmful cyanobacteria occurring at nine water supply source sites in the main stream of the Nakdong River in South Korea from 2012 to 2019 was highly correlated with water temperature and dissolved oxygen [47]. However, at the Tamjin Lake–Yuchi River confluence (T2) site, the variable importance of nutrient-related measurement variables, such as BOD, TN, and Chla, is relatively high, surpassing that of DO. In turn, the variable importance of EC is relatively high at the Tamjin Lake dam front (T1) site. This indicates that nutrients, such as nitrogen and phosphorus, have a more significant influence on algal growth at the Tamjin Lake site compared to the Juam Lake site.

Table 7. Variable importance of explanatory variables for dominant algal classification (bagging, AdaBoost, gradient boosting, and random forest). The top three measurement variable values, based on variable importance for each survey site and algorithm, are bolded. In instances where identical values are present, both variables are bolded.

Algorithm	Bagging				AdaBoost				Gradient Boosting				Random Forest			
Site	**J1**	**J2**	**T1**	**T2**	**J1**	**J2**	**T1**	**T2**	**J1**	**J2**	**T1**	**T2**	**J1**	**J2**	**T1**	**T2**
BOD	1.5151	1.2076	1.4199	**10.9415**	5.3688	2.3051	3.2752	6.1182	4.4015	3.5559	3.6883	**8.5063**	5.9707	4.3431	3.6987	5.0601
COD	3.9329	0.6218	3.3253	0.9804	3.8681	4.3541	5.9501	6.1057	3.4951	2.4247	2.3878	3.2978	4.6970	3.5746	3.0373	2.7818
TN	**7.6162**	3.3504	4.7121	6.3081	7.1730	7.1124	4.8650	**12.1125**	**6.8504**	4.5037	3.9922	7.0974	6.9187	4.7123	5.0819	**7.0660**
TP	1.4593	0.9727	1.4096	1.7724	5.2335	5.1272	4.5811	5.5618	3.4535	2.7191	2.9967	3.0749	5.2055	4.1090	3.5175	3.0822
TOC	1.5957	1.0353	3.9916	2.2245	2.9938	3.6203	3.6343	5.8841	2.1803	2.9575	3.4477	4.5815	3.9918	4.1497	3.5106	3.6138
SS	1.7521	1.6342	2.6280	3.3955	6.5863	5.8286	7.2618	5.5634	4.9319	3.1961	5.3275	**8.4605**	6.3064	3.9993	3.9027	4.3123
EC	4.7823	3.7292	**4.7572**	2.3607	5.4583	4.9772	**8.4684**	6.4223	4.1840	3.5315	5.1276	4.0014	6.4564	4.5444	4.7248	3.4153
pH	4.7826	4.8354	0.6660	0.5151	4.7689	**9.6268**	3.5256	5.3030	5.3392	5.3432	1.4144	1.7538	5.1586	5.6326	2.6790	3.1063
DO	**28.3646**	3.9285	**37.4578**	1.2421	**10.3830**	7.9277	**9.3204**	3.2858	**18.9131**	**11.9789**	**22.1672**	3.7293	**14.9578**	**12.9253**	**10.5199**	4.7113
Temperature	**26.6748**	**57.0166**	4.5654	**40.1099**	**8.7429**	**15.3805**	**11.5727**	**8.2058**	**15.4682**	**29.3555**	**8.6348**	**24.7510**	**14.2645**	**20.5988**	**6.3262**	**9.8519**
Turbidity	1.3844	**5.6681**	1.8321	1.0291	5.9557	4.3917	7.1061	7.4450	3.7155	5.8834	3.8942	2.7514	5.3849	6.3450	3.1732	3.5382
Transparency	0.9296	0.8488	0.6438	6.6063	4.8024	3.5518	2.8340	3.3015	2.6789	1.4298	2.2568	3.0288	3.8831	3.3025	2.1322	2.5311
Chla	2.9024	2.9814	2.5859	**10.6339**	6.1437	6.5267	7.3452	**11.5679**	4.5390	5.3894	8.4450	8.2374	5.5247	4.8334	4.2982	**5.5031**
Low Water Level	5.4858	**8.8656**	**24.4891**	6.2134	**8.3114**	**8.8755**	6.2390	4.1642	6.4153	**7.5803**	**11.3711**	5.7199	6.4287	**7.0199**	**6.3847**	4.3711
Inflow	2.6177	1.5479	1.4467	1.2481	6.6966	5.5075	8.0738	3.6056	6.0363	5.1891	5.8343	3.6946	5.3530	4.3713	4.1482	2.3601
Discharge	3.9374	1.7066	2.8512	4.3660	7.0177	4.7818	5.0124	5.2433	6.4306	2.5860	7.2105	6.5851	**7.6860**	5.9045	5.4156	4.0187
Reservoir	0.2672	0.0500	1.2183	0.0531	0.4960	0.1051	0.9349	0.1097	0.9672	2.3757	1.8039	0.7290	6.6168	6.8222	6.0186	4.5432

Table 8. Variable importance of explanatory variables for dominant algal classification using extreme gradient boosting. The top three measurement variable values, based on variable importance for each survey site and algorithm, are bolded. In instances where identical values are present, both variables are bolded.

Method	**Gain**				**Cover**				**Frequency**			
Site	**J1**	**J2**	**T1**	**T2**	**J1**	**J2**	**T1**	**T2**	**J1**	**J2**	**T1**	**T2**
BOD	0.0437	0.0191	0.0168	0.0817	0.0409	0.0394	0.0100	0.0332	0.0501	0.0487	0.0365	0.0601
COD	0.0447	0.0168	0.0361	0.0549	0.0324	0.0096	0.0621	**0.1455**	0.0537	0.0254	0.0547	**0.1148**
TN	0.0623	0.0515	0.0527	**0.0951**	0.0569	0.0350	0.0502	**0.1691**	0.0590	0.0742	**0.0833**	**0.1257**
TP	0.0284	0.0217	0.0264	0.0380	0.0747	0.0108	0.0259	0.0222	0.0555	0.0318	0.0547	0.0437
TOC	0.0201	0.0199	0.0669	0.0731	0.0240	0.0188	**0.0936**	0.0384	0.0358	0.0424	0.0781	0.0738
SS	0.0460	0.0276	0.0190	0.0604	0.0625	0.0251	0.0198	0.1181	0.0537	0.0403	0.0469	0.0902
EC	0.0546	0.0592	0.0774	0.0411	0.0445	0.0328	**0.1580**	0.0173	0.0644	0.0657	**0.1016**	0.0574
pH	0.0613	0.0626	0.0333	0.0130	**0.0857**	**0.1221**	0.0202	0.0580	0.0698	0.0869	0.0443	0.0410
DO	**0.2037**	**0.1009**	**0.2566**	0.0059	**0.1660**	0.1083	**0.2462**	0.0034	**0.1002**	**0.1102**	**0.1224**	0.0164
Temperature	**0.1870**	**0.3645**	**0.0880**	**0.2813**	**0.1396**	**0.1797**	0.0584	**0.2365**	**0.1091**	**0.0890**	0.0599	**0.1175**
Turbidity	0.0416	0.0593	0.0236	0.0304	0.0210	0.0682	0.0328	0.0185	0.0519	0.0678	0.0469	0.0492
Transparency	0.0196	0.0125	0.0097	0.0341	0.0193	0.0465	0.0061	0.0168	0.0358	0.0318	0.0234	0.0301
Chla	0.0247	0.0586	0.0516	**0.1026**	0.0462	0.0809	0.0441	0.0550	0.0465	**0.1017**	0.0677	0.0984
Low Water Level	**0.0650**	**0.0693**	**0.1378**	0.0090	0.0500	**0.1382**	0.0480	0.0093	**0.0751**	0.0742	0.0443	0.0164
Inflow	0.0304	0.0306	0.0602	0.0148	0.0460	0.0305	0.0569	0.0110	0.0608	0.0508	**0.0833**	0.0219
Discharge	0.0581	0.0258	0.0203	0.0644	0.0809	0.0541	0.0354	0.0478	0.0680	0.0593	0.0339	0.0437
Reservoir	0.0089	0.0000	0.0234	0.0000	0.0092	0.0000	0.0323	0.0000	0.0107	0.0000	0.0182	0.0000

3.2.2. Comparison of Algorithms Based on Four Criteria

To compare the dominant algal classification performance of the 11 statistical machine learning algorithms described in Section 2.3.2, we used the measurements at each survey site from 2017 to 2021 as the training data and the remaining measurements from 2022 as the test data. Each algorithm was trained using the training data, and the classification performance was compared based on accuracy, weighted sensitivity, weighted specificity, and G mean according to the test data. Table 9 presents the calculations of these four criteria for each algorithm based on the classification results by survey site. In this table, for each survey site, the criterion value for the algorithm that shows the best performance based on each of the four criteria is highlighted in bold.

The results show that the optimal algorithm varies depending on the survey site and evaluation criteria. Moreover, our findings indicate that algorithms with complex structures and training processes do not always yield optimal performance, and even simple algorithms can sometimes sufficiently analyze the given data. The data used in this study are imbalanced, with diatoms being the dominant algae in most cases. As such, it is most desirable to select the optimal algorithm based on the G mean, which appropriately combines the harmonic average of weighted sensitivity and weighted specificity rather than accuracy.

Accordingly, the best algorithms for classifying the dominant algae are as follows: decision tree for the Juam Lake dam front (J1) site, random forest for the Juam Lake Shinpyeong Bridge (J2) site, support vector machine for the Tamjin Lake dam front (T1) site, and gradient boosting for the Tamjin Lake–Yuchi River confluence (T2) site. The fact that the best algorithm differs for each survey site suggests that the environmental characteristics of each survey site also vary. This is because the statistical and distributional characteristics of the measured variables investigated for each survey site affect the operation of the algorithm, such as the optimal parameter estimation. As a result, the algorithm that shows the best performance for each survey site is different.

Table 9. Result of dominant algal classification using 11 statistical machine learning algorithms (values in bold represent the criterion for which each algorithm shows the best performance, at each of the four sites).

Site	Criterion	Algorithm										
		DT	**Bag**	**Ada**	**GB**	**RF**	**XGB**	**LDA**	**FDA**	**RDA**	**SVM**	**DNN**
J1	Accuracy	**0.7000**	0.6200	0.6000	0.5400	0.6200	0.6200	0.4000	0.4000	0.4200	0.6600	0.5800
	Weighted Sensitivity	**0.7000**	0.6200	0.6000	0.5400	0.6200	0.6200	0.4000	0.4000	0.4200	0.6600	0.5800
	Weighted Specificity	0.6239	0.6431	0.6949	0.7010	0.6699	0.6948	0.8791	0.8791	**0.9046**	0.6257	0.4200
	G mean	**0.6609**	0.6314	0.6462	0.6153	0.6445	0.6563	0.5930	0.5930	0.6164	0.6426	0.4936
J2	Accuracy	0.5800	0.5400	0.5400	0.5200	**0.6600**	0.5600	0.5800	0.5800	0.5400	0.6200	0.5400
	Weighted Sensitivity	0.5800	0.5400	0.5400	0.5200	**0.6600**	0.5600	0.5800	0.5800	0.5400	0.6200	0.5400
	Weighted Specificity	0.7620	0.7385	0.7046	0.7087	0.7179	**0.8067**	0.7131	0.7131	0.4600	0.6583	0.4600
	G mean	0.6648	0.6315	0.6168	0.6071	**0.6883**	0.6721	0.6431	0.6431	0.4984	0.6389	0.4984
T1	Accuracy	0.7551	0.8163	0.8367	0.8776	**0.9184**	0.7959	0.5918	0.5918	0.8367	0.8980	0.8367
	Weighted Sensitivity	0.7551	0.8164	0.8368	0.8775	**0.9184**	0.7960	0.5919	0.5919	0.8367	0.8980	0.8367
	Weighted Specificity	0.8641	0.7709	0.7762	0.7843	0.6834	0.8698	**0.8801**	**0.8801**	0.1633	0.7823	0.1633
	G mean	0.8078	0.7933	0.8059	0.8296	0.7922	0.8321	0.7218	0.7218	0.3696	**0.8382**	0.3696
T2	Accuracy	0.7551	0.7551	0.7551	**0.7755**	0.7551	0.7551	0.7143	0.7143	0.7551	0.7551	0.7551
	Weighted Sensitivity	0.7552	0.7552	0.7552	**0.7756**	0.7552	0.7552	0.7143	0.7143	0.7552	0.7552	0.7552
	Weighted Specificity	0.2448	0.2448	0.3043	0.3673	0.2448	**0.3698**	0.2439	0.2439	0.2448	0.2448	0.2448
	G mean	0.4300	0.4300	0.4794	**0.5337**	0.4300	0.5285	0.4174	0.4174	0.4300	0.4300	0.4300

4. Discussion

In this study, we analyzed the dominant algae from 2017 to 2022 at various sites in Juam Lake and Tamjin Lake, which are representative water supply sources in the Yeongsan River and Seomjin River systems in South Korea. We also briefly examined the seasonal characteristics of the dominant algae. Additionally, water quality and hydraulic/hydrological parameters related to algal occurrence were collected based on water quality monitoring network data, algae alert system data, and hydraulic/hydrological data to construct the data needed for analysis. We then performed an exploratory data analysis, including correlation analysis and pattern analysis of the SOM for each measurement variable according to the four survey sites, to investigate the overall relationships between the variables and their distributional characteristics. Based on four algorithm evaluation criteria, we also examined the dominant algal classification accuracy of 11 statistical machine learning algorithms for each survey site.

Through evaluating the algorithms, we found that the best one differs for each survey site, indicating that the environmental characteristics of each survey site also differ. In contrast to previous studies [48,49], which mainly used traditional multivariate statistical analysis techniques, such as principal component analysis (PCA) or clustering analysis (CA), to evaluate the environmental characteristics of a survey site, our study attempted to evaluate the environmental characteristics of each survey site using the latest versions of statistical machine learning algorithms. The main results of this study are as follows: chlorophytes or diatoms tended to dominate in spring, cyanophytes in early summer and summer, and chlorophytes and diatoms in autumn and early winter. These results are based on the monthly average number of cells for each algal type measured during the survey period from 2017 to 2022 at the Juam Lake and Tamjin Lake sites.

Through an exploratory data analysis using correlation analysis and pattern analysis of the SOM of the monitoring data, we analyzed the water quality parameters and hydraulic/hydrological variables measured at the Juam Lake and Tamjin Lake sites from 2017 to 2022. This revealed that, overall, mutually related water quality parameters (BOD, COD, TN, TP, etc.) showed positive correlations, while the water quality variables and hydraulic/hydrological variables showed negative correlations.

Using the data from 2017 to 2022 at the Juam Lake and Tamjin Lake monitoring sites of this study, we identified the best algorithms for classifying dominant algae. Based on the G mean, the following algorithms yielded the best performance and were selected: decision tree for the Juam Lake dam front (J1) site, random forest for the Juam Lake Shinpyeong Bridge (J2) site, support vector machine for the Tamjin Lake dam front (T1) site, and gradient boosting for the Tamjin Lake–Yuchi River confluence (T2) site.

This study presents rigorous analyses of water quality data from four survey sites to predict the dominant algae using machine learning algorithms. However, the limited number of survey sites in our study may limit the generalizability of these findings to other water sources, especially those in very different environments. Future research should, therefore, explore the prediction of dominant algae across a larger number of investigation sites to obtain more universal results. This would facilitate the development of a way to evaluate generalized environmental characteristics of water quality. Overall, this study provides valuable insights into the use of statistical machine learning algorithms for water quality management, highlighting the need for further research in this area.

5. Conclusions

The results presented in Section 4 were based solely on data collected from the Juam Lake and Tamjin Lake sites. It is important to note that incorporating additional measurement variables, such as precipitation, and extending the survey period, or analyzing data from water supply sources outside of the Yeongsan River and Seomjin River system, may give different results. As the amount of data increases, so does the prior knowledge

obtained, which can then be used to train the algorithms further. This iterative process can potentially improve algorithm performance. Additionally, different water systems have unique water quality and hydraulic/hydrological characteristics, meaning that even the same algorithms may produce varying results when applied to different water systems. Therefore, more research investigating and comparing a wide range of water source points is necessary. This research approach can support stakeholders and authorities to more accurately classify dominant algal occurrences and, thus, more efficiently manage the quality of important water sources.

Author Contributions: Conceptualization, S.-Y.H. and K.-Y.J.; methodology, S.-Y.H. and B.-W.C.; software, S.-Y.H. and K.-Y.J.; validation, S.-Y.H., B.-W.C., and K.-Y.J.; formal analysis, S.-Y.H.; investigation, H.-S.C., M.-S.S., C.-H.L., H.-M.C., and D.-W.H.; resources, H.-S.C. and B.-W.C.; data curation, S.-Y.H. and K.-Y.J.; writing—original draft preparation, S.-Y.H.; writing—review and editing, S.-Y.H. and K.-Y.J.; visualization, S.-Y.H. and K.-Y.J.; supervision, J.-H.P. and D.-S.S.; project administration, S.-Y.H. and D.-S.S.; funding acquisition, J.-H.P. and D.-S.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by a grant from the National Institute of Environmental Research (NIER), funded by the Ministry of Environment (ME) of the Republic of Korea (NIER-2023-01-01-043).

Data Availability Statement: The datasets used and analyzed during the current study are available from the corresponding author upon request.

Acknowledgments: We would like to thank the reviewers for their comments.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Kim, S.G. Green algae and algae warning system. *Water Future* **2017**, *50*, 22–26.
2. Kim, K.B.; Jung, M.K.; Tsang, Y.F.; Kwon, H.H. Stochastic modeling of chlorophyll-a for probabilistic assessment and monitoring of algae blooms in the Lower Nakdong River, South Korea. *J. Hazard. Mater.* **2020**, *400,* 123066. https://doi.org/10.1016/j.jhazmat.2020.123066.
3. Srivastava, A.; Ahn, C.Y.; Asthana, R.K.; Lee, H.G.; Oh, H.M. Status, alert system, and prediction of cyanobacterial bloom in South Korea. *Biomed. Res. Int.* **2015**, *2015*, 584696. https://doi.org/10.1155/2015/584696.
4. Falconer, I.R.; Humpage, A.R. Health risk assessment of cyanobacterial (blue-green algal) toxins in drinking water. *Int. J. Environ. Res. Public Health* **2005**, *2*, 43–50. https://doi.org/10.3390/ijerph2005010043.
5. Fleming, L.E.; Rivero, C.; Burns, J.; Williams, C.; Bean, J.A.; Shea, K.A.; Stinn, J. Blue green algal (cyanobacterial) toxins, surface drinking water, and liver cancer in Florida. *Harmful Algae* **2002**, *1*, 157–168. https://doi.org/10.1016/S1568-9883(02)00026-4.
6. Kim, Y.H. Harmful Cyanobacterial Bloom and Application of Physical, Chemical and Biological Control Methods. Doctoral Dissertation, Hanyang University, Seoul, Republic of Korea, 2022.
7. Joo, J.H. Field Application and Development of Biologically Derived Substances (BDSs) to Mitigate Freshwater Harmful Cyanobacterial Blooms. Doctoral Dissertation, Hanyang University, Seoul, Republic of Korea, 2017.
8. Guillaume, M.C.; Dos Santos, F.B. Assessing and reducing phenotypic instability in cyanobacteria. *Curr. Opin. Biotechnol.* **2023**, *80,* 102899. https://doi.org/10.1016/j.copbio.2023.102899.
9. Kim, H.G. Prediction of Chlorophyll-A in the Middle Reach of the Nakdong River at Maegok Using Artificial Neural Networks. Master's Thesis, Department of Integrated Biological Science, The Graduate School of Busan National University, Busan, Republic of Korea, 2017.
10. Lee, S.M.; Park, K.D.; Kim, I.K. Comparison of machine learning algorithms for Chl-a prediction in the middle of Nakdong river (focusing on water quality and quantity factors). *J. Korean Soc. Water Wastewater* **2020**, *34*, 277–288. https://doi.org/10.11001/jksww.2020.34.4.277.
11. Bui, D.T.; Khosravi, K.; Tiefenbacher, J.; Nguyen, H.; Kazakis, N. Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Sci. Total Environ.* **2020**, *721*, 137612. https://doi.org/10.1016/j.scitotenv.2020.137612.
12. Caissie, D.; Satish, M.G.; El-Jabi, N. Predicting water temperatures using a deterministic model: Application on Miramichi River catchments (New Brunswick, Canada). *J. Hydrol.* **2007**, *336*, 303–315. https://doi.org/10.1016/j.jhydrol.2007.01.008.

13. Choi, D.H.; Jung, J.W.; Lee, K.S.; Choi, Y.J.; Yoon, K.S.; Cho, S.H.; Park, H.N.; Lim, B.J.; Chang, N.I. Estimation of pollutant load delivery ratio for flow duration using LQ equation from the Oenam-cheon watershed in Juam Lake. *J. Environ. Sci. Int.* **2012**, *21*, 31–39. https://doi.org/10.5322/JES.2012.21.1.31.
14. Park, H.G.; Kang, D.W.; Shin, K.H.; Ock, G.Y. Tracing source and concentration of riverine organic carbon transporting from Tamjin River to Gangjin Bay, Korea. *KJEE* **2017**, *50*, 422–431. https://doi.org/10.11614/KSL.2017.50.4.422.
15. Seo, K.A.; Jung, S.J.; Park, J.H.; Hwang, K.S.; Lim, B.J. Relationships between the Characteristics of Algae Occurrence and Environmental Factors in Lake Juam, Korea. *J. Korean Soc. Water Environ.* **2013**, *29*, 317–328.
16. Cox, V. Exploratory data analysis. In *Translating Statistics to Make Decisions*; Apress: Berkeley, CA, USA, 2017; pp. 47–74.
17. Das, K.R.; Imon, A.H.M.R. A brief review of tests for normality. *Am. J. Ther. Appl. Stat.* **2016**, *5*, 5–12. https://doi.org/10.11648/J.AJTAS.20160501.12.
18. Thadewald, T.; Büning, H. Jarque–Bera test and its competitors for testing normality—A power comparison. *J. Appl. Stat.* **2007**, *34*, 87–105. https://doi.org/10.1080/02664760600994539.
19. Kohonen, T. The self-organizing map. *Proc. IEEE* **1990**, *78*, 1464–1480. https://doi.org/10.1109/5.58325.
20. Jung, K.Y.; Cho, S.H.; Hwang, S.Y.; Lee, Y.J.; Kim, K.H.; Na, E.H. Identification of High-Priority Tributaries for Water Quality Management in Nakdong River Using Neural Networks and Grade Classification. *Sustainability* **2020**, *12*, 9149. https://doi.org/10.3390/su12219149.
21. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*; Springer: New York, NY, USA, 2013; Volume 112, p. 18.
22. Sugiyama, M. *Introduction to Statistical Machine Learning*; Morgan Kaufmann: Burlington, MA, USA, 2015.
23. Park, K.Y.; JW, K. A short guide to machine learning for economists. *Korean J. Econ.* **2019**, *26*, 367–408.
24. Han, S.W. A Study on Kernel Ridge Regression Using Ensemble Method. Master's Thesis, Department of Statistics, The Graduate School of Hankuk University of Foreign Studies, Seoul, Republic of Korea, 2016.
25. Hwang, S.Y. A Study on Efficiency of Kernel Ridge Logistic Regression Classification Using Ensemble Method. Master's Thesis, Department of Statistics, The Graduate School of Hankuk University of Foreign Studies, Seoul, Republic of Korea, 2017.
26. Cutler, A.; Cutler, D.R.; Stevens, J.R. Random forests. In *Ensemble Machine Learning*; Springer: Boston, MA, USA, 2012; pp. 157–175.
27. Schapire, R.E. Explaining adaboost. In *Empirical Inference*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 37–52.
28. Natekin, A.; Knoll, A. Gradient boosting machines, a tutorial. *Front. Neurorobot.* **2013**, *7*, 21. https://doi.org/10.3389/fnbot.2013.00021.
29. Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H.; Chen, K. Xgboost: Extreme Gradient Boosting, R Package Version 0.4-2. 2015; pp. 1–4. Available online: https://cran.microsoft.com/snapshot/2017-12-11/web/packages/xgboost/vignettes/xgboost.pdf (accessed on 10 January 2023).
30. Izenman, A.J. Linear discriminant analysis. In *Modern Multivariate Statistical Techniques*; Springer: New York, NY, USA, 2013; pp. 237–280.
31. Reynès, C.; Sabatier, R.; Molinari, N. Choice of B-splines with free parameters in the flexible discriminant analysis context. *Comput. Stat. Data Anal.* **2006**, *51*, 1765–1778. https://doi.org/10.1016/j.csda.2005.11.018.
32. Schölkopf, B.; Smola, A.J.; Williamson, R.C.; Bartlett, P.L. New support vector algorithms. *Neural Comput.* **2000**, *12*, 1207–1245. https://doi.org/10.1162/089976600300015565.
33. Friedman, J.H. Regularized discriminant analysis. *J. Am. Stat. Assoc.* **1989**, *84*, 165–175. https://doi.org/10.1080/01621459.1989.10478752.
34. Pisner, D.A.; Schnyer, D.M. Support vector machine. In *Machine Learning*; Academic Press: Cambridge, MA, USA, 2020; pp. 101–121.
35. Montavon, G.; Samek, W.; Müller, K.R. Methods for interpreting and understanding deep neural networks. *Digit Signal Process.* **2018**, *73*, 1–15. https://doi.org/10.1016/j.dsp.2017.10.011.
36. Parikh, R.; Mathai, A.; Parikh, S.; Sekhar, G.C.; Thomas, R. Understanding and using sensitivity, specificity and predictive values. *Indian J. Ophthalmol.* **2008**, *56*, 45. https://doi.org/10.4103/0301-4738.37595.
37. Xu, J.; Zhang, Y.; Miao, D. Three-way confusion matrix for classification: A measure driven view. *Inf. Sci.* **2020**, *507*, 772–794. https://doi.org/10.1016/j.ins.2019.06.064.
38. Li, D.L.; Shen, F.; Yin, Y.; Peng, J.X.; Chen, P.Y. Weighted Youden index and its two-independent-sample comparison based on weighted sensitivity and specificity. *Chin. Med. J.* **2013**, *126*, 1150–1154. https://doi.org/10.3760/cma.j.issn.0366-6999.20123102.
39. Trevethan, R. Sensitivity, specificity, and predictive values: Foundations, pliabilities, and pitfalls in research and practice. *Front. Public Health* **2017**, *5*, 307. https://doi.org/10.3389/fpubh.2017.00307.
40. Jung, K.Y.; Lee, I.J.; Lee, K.L.; Cheon, S.U.; Hong, J.Y.; Ahn, J.M. Long-term trend analysis and exploratory data analysis of Geumho River based on seasonal Mann-Kendall test. *J. Environ. Sci. Int.* **2016**, *25*, 217–229.
41. Blanca, M.J.; Arnau, J.; López-Montiel, D.; Bono, R.; Bendayan, R. Skewness and kurtosis in real data samples. *Methodology* **2013**, *9*, 78–84. https://doi.org/10.1027/1614-2241/a000057.

42. De Winter, J.C.; Gosling, S.D.; Potter, J. Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychol. Methods* **2016**, *21*, 273. https://doi.org/10.1037/met0000079.
43. Bai, J.; Ng, S. Tests for skewness, kurtosis, and normality for time series data. *J. Bus. Econ. Stat.* **2005**, *23*, 49–60. https://doi.org/10.1198/073500104000000271.
44. Gregorutti, B.; Michel, B.; Saint-Pierre, P. Correlation and variable importance in random forests. *Stat. Comput.* **2017**, *27*, 659–678. https://doi.org/10.1007/s11222-016-9646-1.
45. Genuer, R.; Poggi, J.M. Random forests. In *Random Forests with R*; Springer: Cham, Switzerland, 2020; pp. 33–55.
46. Roelofs, R.; Shankar, V.; Recht, B.; Fridovich-Keil, S.; Hardt, M.; Miller, J.; Schmidt, L. A meta-analysis of overfitting in machine learning. In Proceedings of the Advances in Neural Information Processing Systems 32 (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019.
47. Woo, C.Y.; Yun, S.L.; Kim, S.G.; Lee, W.T. Occurrence of Harmful Blue-green Algae at Algae Alert System and Water Quality Forecast System Sites in Daegu and Gyeongsangbuk-do between 2012 and 2019. *J. Korean Soc. Environ. Eng.* **2020**, *42*, 664–673. https://doi.org/10.4491/KSEE.2020.42.12.664.
48. Jung, K.Y.; Ahn, J.M.; Kim, K.; Lee, I.J.; Yang, D.S. Evaluation of water quality characteristics and water quality improvement grade classification of Geumho River tributaries. *J. Environ. Sci. Int.* **2016**, *25*, 767–787. https://doi.org/10.5322/JESI.2016.25.6.767.
49. Sun, X.; Zhang, H.; Zhong, M.; Wang, Z.; Liang, X.; Huang, T.; Huang, H. Analyses on the temporal and spatial characteristics of water quality in a seagoing river using multivariate statistical techniques: A case study in the Duliujian River, China. *Int. J. Environ. Res. Public Health* **2019**, *16*, 1020. https://doi.org/10.3390/ijerph16061020.