Article

Comparison of algorithms for dominant algae classification in water supply source site in Yeongsan and Seomjin river basins using Statistical Machine Learning Technique

Seong-Yun Hwang1+, Byung-Woong Choi1, Jong-Hwan Park1, Dong-Seok Shin1, Hyeon-Su Chung1, Mi-Sun Son1, Chae-Hong Lim1, Hyeon-Mi Choi1, Don-Woo Ha1, Kang-Young Jung2\*

|  |
| --- |
| **Citation:** To be added by editorial staff during production.  Academic Editor: Firstname Lastname  Received: date  Revised: date  Accepted: date  Published: date    **Copyright:** © 2023 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/). |

1 Yeongsan River Environment Research Center, National Institute of Environmental Research; e-mail@e-mail.com

2 Education Planning Division, National Institute of Environmental Human Resources Development; e-mail@e-mail.com

**\*** Correspondence: e-mail@e-mail.com; Tel.: (optional; include country code; if there are multiple corresponding authors, add author initials)

**Abstract:** ~~South Korea operates an algae alert system through the Water Environment Information System of the National Institute of Environmental Research (NIER). The system discloses data organized from the analysis results of samples collected weekly from water supply source sites, which provide living water to the public. Of the cyanobacteria that occur in water supply source sites, four harmful cyanobacteria species (Aphanizomenon, Anabaena, Oscillatoria, and Microcystis) pose the risk of toxic damage, so it is important to accurately predict the dominant algae that occur in water supply source sites. In this study, using information collected and organized based on water quality monitoring network data, algae alert system data, and hydraulic and hydrological data measured at 7-day intervals from January 2017 to December 2022 through the NIER Water Environment Information System in sites of Juam Lake and Tamjin Lake, representative water supply source sites in the Yeongsan River and Seomjin River system in South Korea, we performed an explanatory data analysis of each measurement variable to examine the overall fluctuations. Additionally, using data from 2017 to 2021 as training data and data from 2022 as test data, we compared and evaluated the dominant algae classification accuracy of 11 statistical machine learning algorithms. The results showed that the optimal algorithm varied with the survey site and algorithm evaluation criteria. This means that the environmental characteristics are different for each survey site. By predicting the dominant algae in advance through this process, it will be possible to more effectively prepare for water source contamination accidents. This study’s findings demonstrate the applicability of machine learning algorithms as tools to efficiently manage the water quality of water supply source systems using monitoring data.~~

Proposed edited abstract within 200 word-limit:

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 cyanobacteria genera, such as Aphanizomenon, Anabaena, Oscillatoria, and Microcystis, is crucial for managing water source contamination risks. This study utilized data from Juam Lake and Tamjin Lake, representative water supply source sites in the Yeongsan River and Seomjin River basins, collected between January 2017 and December 2022. We performed an exploratory data analysis on the water quality parameters monitored to understand overall fluctuations. Using data from 2017-2021 as training data and 2022 data as test data, we compared the dominant algae 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, we 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

1. Introduction

In South Korea, sites crucial for providing potable water to local residents are designated and managed as water protection zones. The extreme drought in the Honam region of South Korea in 2022 underscored the importance of managing water quality in water supply sources. To safeguard the water quality of these sources, the Korean government established the algae alert system in 1998. To minimize toxic effects caused by harmful cyanobacteria, the system issues alerts based on harmful cyanobacteria cell counts: Caution (at least 1,000 cells/mL for two consecutive times), Warning (at least 10,000 cells/mL for two consecutive times), Outbreak (at least 1,000,000 cells/mL for two consecutive times), and Release (number of cyanobacteria cells below the alert threshold for two consecutive times) [1-3]. Specifically, four representative harmful cyanobacteria genera, Aphanizomenon, Anabaena, Oscillatoria, and Microcystis, release harmful toxins causing acute liver disease in humans [4] and threatening aquatic ecosystem stability [5]. Researchers have explored various reduction methods to control the occurrence of these harmful cyanobacteria, such as algal blocking mat (ABM), a physical method; plant-mineral composite (PMC), a chemical method; and Unio douglasiae, a biological method [6-8].

Additionally, recent studies have actively explored the evaluation of water quality-related data and prediction of changes in specific water quality parameters using various statistical machine learning techniques have recently been actively conducted. Kim, H. G. (2017) assessed the suitability of the artificial neural network technique for predicting the chlorophyll-a concentration at the Maegok site, a midstream location in South Korea’s Nakdong River [9]. Moreover, Lee et al. (2020) investigated methods for predicting the chlorophyll-a concentrations using four statistical machine learning algorithms at the Kangjeong-Goryeong Weir and Dalseong Weir sites, which are representative weir sites in the Nakdong River system in South Korea [10]. Bui et al. (2020) also conducted a study to predict the Water Quality Index (WQI) based on water quality parameters using 16 novel hybrid machine learning algorithms [11]. However, the study was limited in its ability to compare the performance of various algorithms. The primary difference between these previous and current studies is that the former focused on accurately predicting the measured values of the water quality parameter chlorophyll-a (a continuous variable). In contrast, this study aims to accurately classify dominant algae (a categorical variable). Nutrients such as nitrogen (N) and phosphorus (P), as well as water quality parameters like water temperature, are the most influential factors in algal growth. 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 can enable authorities to prepare rapidly for water pollution accidents caused by algae. We utilized the water quality monitoring network data, algae alert system data, and hydraulic/hydrological data in 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 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 algae occurrences, we aim to provide a more efficient approach to water quality management.

2. Materials and Methods

2.1. Study Area

This study focused on the 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 from the dam front (J1, coordinates) and Shinpyeong Bridge (J2 coordinates) sites of Juam Lake and the dam front (T1, coordinates) and Yuchi stream confluence (T2, coordinates) sites of Tamjin Lake to monitor water quality and respond to the algae alert system. Additionally, the Korea Water Resources Corporation conducts daily measurements of hydraulic/hydrological variables, such as water storage capacity.

Juam Lake is formed by the freshwater of 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 1,029.41 km² and a total water storage capacity of 457×10⁶ tons. Juam Dam supplies about 640×10³ 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 km² and a total water storage capacity of 191×10⁶ tons. It is located in Yuchi-myeon, Jangheung-gun, Jeollanam-do, and supplies 73×10⁶ tons of living water to nine cities and counties in Jeollanam-do [14]. Figure 1 shows the survey sites in Juam Lake and Tamjin Lake.

|  |  |
| --- | --- |
|  |  |



**Figure 1.** Survey sites in the 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 monitoring site is as follows: In Juam Lake, there are 307 observations at both the dam front (J1) and Shinpyeong Bridge (J2) sites. In Tamjin Lake, the dam front (T1) and Yuchi Stream Confluence (T2) sites each have 304 observations. Overall, this study includes a total of 1,222 observations. Tables 1 ~~and 2~~ display the variables of the collected data ~~and the number of observations per monitoring site, respectively~~. Figure 2 presents a methodology flowchart illustrating the structure and process of this study.

**Table 1.** Data variables used in this study.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Response variable (categorical)** | | **Explanatory variables (continuous)** | | |
| (based on total cell count) | | Water Quality | | Hydraulic/Hydrological |
|  | | Biological Oxygen Demand (BOD), mg/L  Chemical Oxygen Demand (COD), mg/L  Total Nitrogen (TN), mg/L  Total Phosphorus (TP), mg/L  Total Organic Carbon (TOC), mg/L  Suspended Solids (SS), mg/L  Electrical Conductivity (EC), μS/cm  pH  Dissolved Oxygen (DO), mg/L  Temperature, ℃  Turbidity, NTU  Transparency, m  Chlorophyll a (Chla), mg/㎥) | | LowWaterLevel, cm  Inflow Rate (Inflow), cms  Discharge Rate (Discharge), cms  Water Storage Capacity  (Reservoir), 10,000 m³ |

**~~Table 2.~~** ~~Number of observations per survey site.~~

|  |  |  |  |
| --- | --- | --- | --- |
| ~~Juam Lake~~ | | ~~Tamjin Lake~~ | |
| ~~J1~~  ~~(dam front)~~ | ~~J2~~  ~~(Shinpyeong Bridge)~~ | ~~T1~~  ~~(dam front)~~ | ~~T2~~  ~~(Yuchi Stream Confluence)~~ |
| ~~307~~ | ~~307~~ | ~~304~~ | ~~304~~ |
| ~~total : 1222~~ | | | |

Diagram, timeline

Description automatically generated

**Figure 2.** Methodology flowchart used in this study.

Of the variables listed in Table 1, BOD, COD, TN, TP, TOC, SS, and EC were collected from the water quality monitoring network data, while pH, DO, Temperature, Turbidity, Transparency, Chla, and Dominant Algae were collected from the algae alert system data. The remaining variables, LowWaterLevel, Inflow, Discharge, and Reservoir, were collected from the National Water Resources Management Information System (http://www.wamis.go.kr/). The genera of algae that occurred at the monitoring sites when the data was measured are presented in Table 3. Figure 3 shows line graphs of the monthly mean number of algae cells occurring during the survey period, divided according to the survey site and type of algae. According to the results in Table 3 and Figure 3, during the survey period, chlorophytes or diatoms tended to dominate in spring, cyanophytes in early summer and summer, and chlorophytes and diatoms in autumn and early winter [15]. All of the data analyses in this study were performed through the program R version 4.2.1.

Table 3. Genera of algae that were identified in water samples collected from survey sites.

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

|  |  |
| --- | --- |
| (a) |  |
| (b) |  |
| (c) |  |
| (d) |  |

**Figure 3.** Line graphs of average algal cell count at survey sites (a) J1, (b) J2, (c) T1, and (d) T2 from January 2017 to November 2022.

2.3. Methods

This section describes the data analysis methods employed in this study. We start with exploratory data analysis, which 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 analysis data. We also briefly explained the principles of 11 statistical machine learning algorithms used to compare predictive power for dominant algae classification.

2.3.1. Exploratory Data Analysis

An exploratory data analysis is performed to investigate the overall data’s characteristics before analyzing the data [16] . While no specific analysis 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 pattern analysis using a self-organizing map to visually confirm the results.

1. Correlation Analysis

Correlation is a widely used statistical analysis method for investigating the relationships between continuous variables in data. For this purpose, the Pearson correlation coefficient was calculated as shown in Equation (1), and a significance test was conducted on the resulting coefficient. Generally, the validity of the analysis 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 the variables have the properties of random variables that satisfy independency. Since all measurement variables in this study are time series data measured over a period rather than random variables that satisfy independence, the Jarque-Bera (JB) test method is more appropriate [18].

|  |  |
| --- | --- |
|  | (1) |

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

|  |  |
| --- | --- |
|  | (2) |

1. Pattern Analysis using Self-Organizing Map

A self-organizing map 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 th lattice vector at time is updated.

|  |  |
| --- | --- |
|  | (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 self-organizing maps, 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]. They used a grading process through cluster analysis according to the characteristics of each site to determine which branches should be prioritized for management and proposed policies based on these findings. In this study, we performed a pattern analysis on 17 measurement variables using this method and identified correlations between the measurement variables.

2.3.2. Statistical Machine Learning algorithm for dominant algae 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. Decision Tree(DT)

A decision tree is a method for creating a decision model with a tree-like structure. It typically involves the growing step, the pruning step, the optimal tree model selection step, and the testing step. The impurity of nodes is reviewed to select the optimal separation criteria for pruning. Mean squared error, such as Equation (4), is used for regression, and the Gini coefficient, such as Equation (5), or entropy coefficient, such as Equation (6), is used for classification. Compared to other algorithms, decision trees are visually simple and it is relatively easy to interpret [23]. This makes it easier to identify which variables were used as the main criteria when predicting the dominant algae.

|  |  |
| --- | --- |
|  | (4) |
|  | (5) |
|  | (6) |

1. Bagging Method

Bagging involves sampling with replacement method, which allows observations extracted from the analysis data to be re-extracted to extract multiple samples () for the analysis, creates multiple decision tree models () based on these, and then averages () the prediction results obtained through this or performs multiple voting () based on the classification results to derive the conclusion. “Mode” refers to the value with the highest frequency. The bagging technique greatly reduces the variance of the created model compared to creating a decision tree model once, as it uses the survey with replacement method [24-25]. This makes it likely to perform better than simple decision trees in predicting dominant algae. Figure 5 illustrates the principle of the bagging method.

|  |
| --- |
| Diagram  Description automatically generated |

**Figure 5.** Schematic diagram of the bagging method.

1. AdaBoost (Ada)

AdaBoost is a boosting algorithm that combines multiple weak learners with low prediction performance through weights to create a strong learner with improved performance. By correcting or supplementing incorrectly predicted or classified instances from previous steps, it can yield more accurate results. The algorithm creates a strong learner by taking a weighted linear combination of multiple weak learners, as shown in Equation (7).

|  |  |
| --- | --- |
|  | (7) |

where is the finally obtained strong learner, are the weak learners, and are the weights of the weak learners. Figure 6 illustrates the principle of AdaBoost, and a more detailed explanation can be found in the literature [26].

|  |
| --- |
| Diagram  Description automatically generated |

**Figure 6.** Schematic diagram of the AdaBoost method.

1. Gradient Boosting (GB)

Gradient boosting involves repeatedly using the gradient to create a model, and then using the residual from this to create another model. This process reduces the part that the previous model could not explain, thereby reducing bias. However, there is a risk of overfitting, which occurs when the model overlearns the training data. If the given training data is and the previously created model is , then gradient boosting undergoes through the process of finding function that models the residual, which is the difference between the actual value and the predicted value, as shown in Equation (8).

|  |  |
| --- | --- |
|  | (8) |

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

|  |  |
| --- | --- |
|  | (9) |

where parameter is the learning rate, which reduces the risk of overfitting. Figure 7 illustrates the principle of gradient boosting, and a more detailed explanation can be found in Natekin et al. (2013) [27].

|  |
| --- |
| Diagram  Description automatically generated |

**Figure 7.** Schematic diagram of the gradient boosting method.

1. Random Forest (RF)

Random forest was proposed to address the shortcomings of bagging, and it operates on a similar principle. Like bagging, it involves extracting multiple samples with replacements 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 improves the prediction or classification performance compared to bagging. The types of variables selected for each sample differ, reducing the correlation between each sample that can occur with bagging. This can greatly enhance prediction or classification performance [28]. As a result, random forest is likely to have better predictive power for dominant algae than bagging.

1. Extreme Gradient Boosting (XGB)

Extreme gradient boosting is an improved method that addresses the slow execution time and overfitting risks of gradient boosting by supporting parallel learning. Regular gradient boosting does not have an in-built mechanism to regulate overfitting, but extreme gradient boosting has a self-regulating function that makes it more stable and durable. It can also independently conduct cross-validation tests and has an early stopping function that detects overfitting in advance. Traditionally, after randomly dividing the training data into parts, data are used as new training data and the remaining data are used as new test data to evaluate the performance of the algorithm. The cross-validation test performs this process on all parts of the data, as shown in Figure 8. A detailed explanation can be found in Chen et al. (2015) [29].

|  |
| --- |
| Chart, bar chart  Description automatically generated |

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

1. Linear Discriminant Analysis (LDA)

Linear discriminant analysis is a classification method using R. A. Fisher’s linear decision boundary. The given data is projected on a specific one-dimensional axis, followed by a process that finds the optimal straight line that properly distinguishes the categories. To find this straight line, it is necessary to ensure that the means of the categories differ as much as possible and that the variances within each category are small. This process makes it possible to find the linear decision boundary, as shown in Figure 9. A more detailed explanation can be found in Izenman, A. J. (2013) [30].

|  |
| --- |
| Diagram  Description automatically generated |

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

1. Flexible Discriminant Analysis (FDA)

Flexible discriminant analysis is a method that addresses the limitations of linear discriminant analysis, which relies on linear decision boundaries. Instead, it uses splines to create a non-linear decision boundary for classification. In reality, the distribution of variables in data rarely exhibits linear relationships, and most cases have complex non-linear distributions, making it difficult to apply linear discriminant analysis. This is where flexible discriminant analysis comes in, as it allows for non-linear relationships to be captured and better classification accuracy to be achieved. This problem can be addressed through flexible discriminant analysis [31]

1. Regularized Discriminant Analysis (RDA)

When the data contains many explanatory variables, regularized discriminant analysis improves the estimation of the covariance matrix through regularization (e.g., shrinkage) to create a decision boundary with better classification performance. It can be viewed as a generalized method of linear discriminant analysis, which uses a linear decision boundary, and quadratic discriminant analysis, which uses a quadratic curved decision boundary. For this, the optimal parameter is estimated based on the training data; if = , then the linear discriminant analysis is performed, and if = , then the quadratic discriminant analysis is performed. Here, , which serves as the weight for the linear decision boundary and quadratic curved decision boundary [32].

1. Support Vector Machine (SVM)

Support vector machine (SVM) is a classification algorithm that maximizes the margin, i.e., the distance between the decision boundary and the support vectors, to perform prediction or classification. 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 trick. This technique applies a mapping function without the need to set an appropriate transformation function beforehand. This converts the data into a linear distribution and makes it easier to find the decision boundary (Schölkopf, B. (2000)) [33]. Figure 10 illustrates this concept. This study uses the radial basis kernel, shown in Equation (10), which is known to be the most flexible kernel type for all data distributions. This kernel is widely applicable to water quality-related data, such as the data used in this study. Figure 11 illustrates the principle of the support vector machine, and a detailed explanation can be found in Pisner et al. (2020) [34].

|  |  |
| --- | --- |
|  | (10) |

|  |
| --- |
| Chart  Description automatically generated |

**Figure 10.** Schematic diagram of the kernel trick.

|  |
| --- |
| Chart, scatter chart  Description automatically generated |

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

1. 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. Through this, it is possible to model the complex non-linear relationship between the response variable and explanatory variable in the data. The model is trained through a backpropagation algorithm that updates the weights through stochastic gradient descent, as shown in Equation (11).

|  |  |
| --- | --- |
|  | (11) |

where is the parameter that controls the learning rate, and 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 Montavon et al. (2018) [35].

|  |  |
| --- | --- |
|  | (12) |
|  | (13) |

3. Results

3.1. Data Analysis

3.1.1. Exploratory Data Analysis for Monitoring Data

Based on the data collected for this study, the descriptive statistics of the variables for each survey site are organized in Table 4. This shows an overview of the distributions of measurement variables for each survey site [36]. We also calculated the Jarque-Bera (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 to the right and a negative value when the tail is long to 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 [37]. According to Table 4, none of the measurement variables showed a value of 0 for skewness or kurtosis at any of the survey sites.

Furthermore, except for pH at all survey sites, the JB test p-value was significantly lower than the significance level of 0.05. Hence, since normality is often violated, the Spearman correlation coefficient must be used instead of the Pearson correlation coefficient in the correlation analysis [38].

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

Table 4. Descriptive statistics of the water quality and hydraulic/hydrological parameters at each survey site.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Survey Site | Statistics | BOD  (mg/L) | COD  (mg/L) | TN  (mg/L) | TP  (mg/L) | TOC  (mg/L) | SS  (mg/L) | EC  (μS/cm) | pH | DO  (mg/L) | Temperature  (℃) | Turbidity  (NTU) | Transparency  (m) | Chla  (mg/㎥) | LowWaterLevel  (cm) | Inflow  (cms) | Discharge  (cms) | Reservoir  (10,000㎥) |
| J1 | mean | 0.95 | 3.3 | 0.68 | 0.01 | 2.34 | 1.8 | 81.8 | 7.26 | 8.62 | 14.16 | 2.32 | 3.17 | 3.33 | 7260.49 | 3.61 | 4.12 | 9945.48 |
| sd | 0.38 | 0.43 | 0.13 | 0.01 | 0.41 | 1.04 | 8.82 | 0.43 | 2.22 | 5.63 | 2.01 | 1 | 2.69 | 620.75 | 11.33 | 7.7 | 3298.51 |
| median | 0.9 | 3.3 | 0.66 | 0.01 | 2.3 | 1.5 | 80 | 7.2 | 8.5 | 14.7 | 1.7 | 3.2 | 2.7 | 7238 | 0.8 | 2.7 | 9968 |
| min | 0.4 | 2.4 | 0.44 | 0 | 1.4 | 0.5 | 62 | 6.1 | 4.6 | 2.1 | 0.1 | 0.7 | 0.3 | 6167 | 0 | 1.7 | 3552 |
| max | 2.6 | 4.7 | 1.08 | 0.05 | 3.5 | 10.6 | 101 | 8.8 | 12.9 | 24.6 | 15.4 | 7.2 | 25.2 | 9638.62 | 162.63 | 93.21 | 16947 |
| skewness | 1.31 | 0.59 | 0.38 | 1.13 | 0.86 | 3.1 | 0.45 | 0.12 | 0.07 | -0.09 | 3.08 | 0.27 | 3.6 | 1.43 | 9.92 | 8.26 | 0.05 |
| kurtosis | 2.44 | 0.19 | -0.45 | 2.49 | 0.47 | 17.92 | -0.76 | 0.03 | -1.33 | -1.3 | 14.73 | 0.4 | 19.16 | 4.05 | 127.3 | 78.22 | -0.68 |
| J-B 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 | 0.95 | 3.3 | 0.68 | 0.01 | 2.34 | 1.8 | 81.8 | 7.31 | 8.8 | 14.83 | 2.43 | 3.12 | 3.72 | 7260.49 | 3.61 | 4.12 | 9945.48 |
| sd | 0.38 | 0.43 | 0.13 | 0.01 | 0.41 | 1.04 | 8.82 | 0.45 | 2.17 | 6 | 3.58 | 0.93 | 2.16 | 620.75 | 11.33 | 7.7 | 3298.51 |
| median | 0.9 | 3.3 | 0.66 | 0.01 | 2.3 | 1.5 | 80 | 7.3 | 8.8 | 15.2 | 1.6 | 3 | 3.4 | 7238 | 0.8 | 2.7 | 9968 |
| min | 0.4 | 2.4 | 0.44 | 0 | 1.4 | 0.5 | 62 | 5.8 | 4.1 | 2.1 | 0.1 | 0.7 | 0.2 | 6167 | 0 | 1.7 | 3552 |
| max | 2.6 | 4.7 | 1.08 | 0.05 | 3.5 | 10.6 | 101 | 8.6 | 12.9 | 25.9 | 34.6 | 6 | 22.2 | 9638.62 | 162.63 | 93.21 | 16947 |
| skewness | 1.31 | 0.59 | 0.38 | 1.13 | 0.86 | 3.1 | 0.45 | -0.16 | -0.07 | -0.08 | 5.6 | 0.34 | 2.97 | 1.43 | 9.92 | 8.26 | 0.05 |
| kurtosis | 2.44 | 0.19 | -0.45 | 2.49 | 0.47 | 17.92 | -0.76 | 0.42 | -1.11 | -1.31 | 38 | 0.43 | 18.68 | 4.05 | 127.3 | 78.22 | -0.68 |
| J-B 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 | 2.27 | 5.41 | 1.63 | 0.1 | 3.87 | 10.63 | 185.99 | 7.27 | 9.17 | 13.81 | 2.42 | 2.63 | 5.16 | 9908.66 | 15.01 | 17.03 | 23247.54 |
| sd | 1.36 | 1.57 | 0.66 | 0.04 | 1 | 11.02 | 65.12 | 0.42 | 2.19 | 5.47 | 2.84 | 0.73 | 2.93 | 724.55 | 39.11 | 37.24 | 8015.62 |
| median | 2 | 5 | 1.47 | 0.09 | 3.7 | 8.4 | 178 | 7.3 | 9.1 | 14.5 | 1.7 | 2.6 | 4.4 | 10063 | 3.7 | 11.38 | 23560 |
| min | 0.5 | 2.6 | 0.64 | 0.03 | 1.9 | 1.3 | 68 | 6.1 | 4.8 | 1.5 | 0.1 | 1 | 0.4 | 6805 | 0 | 1.94 | 7105 |
| max | 8.8 | 13 | 4.94 | 0.23 | 6.7 | 93.2 | 600 | 8.6 | 13.6 | 24.8 | 36 | 7 | 19 | 10704 | 310.63 | 464.6 | 37807 |
| skewness | 2.01 | 1.32 | 1.75 | 0.91 | 0.49 | 4.08 | 1.53 | 0.17 | 0.03 | -0.19 | 6.65 | 1.01 | 1.41 | -2.52 | 4.58 | 8.54 | -0.08 |
| kurtosis | 5.42 | 2.53 | 4.61 | 0.14 | -0.25 | 20.93 | 5.79 | -0.05 | -1.13 | -1.08 | 66.11 | 3.87 | 2.85 | 8.1 | 23.37 | 83.5 | -0.96 |
| J-B 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 | 2.27 | 5.41 | 1.63 | 0.1 | 3.87 | 10.63 | 185.99 | 7.39 | 8.84 | 13.72 | 2.41 | 2.56 | 4.87 | 9908.66 | 15.01 | 17.03 | 23247.54 |
| sd | 1.36 | 1.57 | 0.66 | 0.04 | 1 | 11.02 | 65.12 | 0.59 | 2.34 | 5.51 | 2.24 | 0.73 | 3.04 | 724.55 | 39.11 | 37.24 | 8015.62 |
| median | 2 | 5 | 1.47 | 0.09 | 3.7 | 8.4 | 178 | 7.3 | 8.7 | 14.15 | 1.8 | 2.5 | 4.1 | 10063 | 3.7 | 11.38 | 23560 |
| min | 0.5 | 2.6 | 0.64 | 0.03 | 1.9 | 1.3 | 68 | 5.6 | 4.1 | 1.7 | 0.1 | 1 | 0 | 6805 | 0 | 1.94 | 7105 |
| max | 8.8 | 13 | 4.94 | 0.23 | 6.7 | 93.2 | 600 | 9 | 13.2 | 28 | 20.7 | 7 | 17.6 | 10704 | 310.63 | 464.6 | 37807 |
| skewness | 2.01 | 1.32 | 1.75 | 0.91 | 0.49 | 4.08 | 1.53 | -0.01 | 0.03 | -0.04 | 4.05 | 1.09 | 1.26 | -2.52 | 4.58 | 8.54 | -0.08 |
| kurtosis | 5.42 | 2.53 | 4.61 | 0.14 | -0.25 | 20.93 | 5.79 | 0.03 | -1.17 | -0.96 | 23.38 | 4.22 | 1.76 | 8.1 | 23.37 | 83.5 | -0.96 |
| J-B 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 |

|  |  |  |
| --- | --- | --- |
| Chart, box and whisker chart  Description automatically generated | Chart, box and whisker chart  Description automatically generated | Chart, box and whisker chart  Description automatically generated |
| Chart, box and whisker chart  Description automatically generated | Chart, box and whisker chart  Description automatically generated | Chart, box and whisker chart  Description automatically generated |
| Chart, box and whisker chart  Description automatically generated | Chart, box and whisker chart  Description automatically generated | Chart, box and whisker chart  Description automatically generated |
| Chart, box and whisker chart  Description automatically generated | Chart, box and whisker chart  Description automatically generated | Chart, box and whisker chart  Description automatically generated |
| Chart, box and whisker chart  Description automatically generated | Chart  Description automatically generated | Chart  Description automatically generated |
| Chart, box and whisker chart  Description automatically generated |  |  |

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

Table 5 presents a contingency table of the variable ‘Dominant Algae’, a categorical variable. The table indicates that diatoms were dominant at all survey sites during the monitoring period, followed by chlorophytes, cyanophytes, and other algae.

Table 5. Contingency table of 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 method using rank, must be applied for the correlation analysis. On this basis, we performed a correlation analysis for each survey site, the results of which are shown in Figure 13. The figures on the left for each survey site show the calculated Spearman correlation coefficients, and those on the right show the corresponding significance test results in which non-significant correlations are marked with an “X.” According to the results of the correlation analysis, there were variations in the results at each survey site; however, in general, the water quality parameters that were mutually related (BOD, COD, TN, TP, etc.) showed positive correlations, while the water quality variables and hydraulic/hydrological variables showed negative correlations. Pattern analysis of the self-organizing maps supports these results, as shown in Figure 14-17. The analysis helped identify the overall movement of the measurement variables at each survey site during the survey period. The water quality parameters that exhibited significant positive correlations in the correlation analysis showed similar patterns, while the water quality and hydraulic/hydrological variables that exhibited significant negative correlations showed opposite patterns. However, it should be noted that since this study used time series data, which are measured over a certain period and not independent, calculating the normality test p-value for each time-dependent measurement variable and performing a correlation analysis and interpretation based on this has limitations [39].

|  |  |  |
| --- | --- | --- |
|  | **(a) Spearman correlation coefficient** | **(b) Significance test results** |
| J1 |  |  |
| J2 |  |  |
| T1 |  |  |
| T2 |  |  |

**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

|  |  |  |  |
| --- | --- | --- | --- |
| A picture containing chart  Description automatically generated | A picture containing chart  Description automatically generated | Chart  Description automatically generated with low confidence | Chart  Description automatically generated |
| A picture containing chart  Description automatically generated | A picture containing chart  Description automatically generated | A picture containing chart  Description automatically generated | A picture containing chart  Description automatically generated |
| A picture containing chart  Description automatically generated | A picture containing chart  Description automatically generated | A picture containing background pattern  Description automatically generated | A picture containing chart  Description automatically generated |
| A picture containing chart  Description automatically generated | A picture containing chart  Description automatically generated | A picture containing background pattern  Description automatically generated | A picture containing qr code  Description automatically generated |
| A picture containing diagram  Description automatically generated |  |  |  |

**Figure 14.** Self-organizing map for survey site J1.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Figure 15.** Self-organizing map for survey site J2.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Figure 16.** Self-organizing map for survey site T1.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Figure 17.** Self-organizing map for survey site T1.

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

This section presents the results of analyzing the dominant algae classification accuracy using 11 statistical machine learning algorithms. The training data consisted of the measurements from 2017 to 2021 at each survey site, while the test data was the remaining measurements from 2022. To evaluate the classification accuracy, three representative criteria were used: accuracy, sensitivity, and specificity [40]. These criteria were calculated using a confusion matrix that organized the actual correct answers and the predicted answers from the classification, as shown in Table 6.

**Table 6.** Confusion matrix of dominant algae classification

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Predicted | | | |
| Cyanophytes | Diatoms | Chlorophytes | Others |
| Actual | Cyanophytes |  |  |  |  |
| Diatom |  |  |  |  |
| Chlorophytes |  |  |  |  |
| Others |  |  |  |  |

“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.

|  |  |
| --- | --- |
|  | (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 calculated sensitivity and specificity for the four algae categories (cyanophytes, diatoms, chlorophytes, and others), calculated weighted sensitivity and weighted specificity by taking the weighted average according to the number of data, and 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 7.

**Table 7.** 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 out of those that are actually positive, and specificity is the ratio of observations properly classified as negative out of those that are actually negative [41]. Both ratios range from 0 to 1, with values closer to 1 indicating better algorithm performance. This is expressed in Equation (15).

|  |  |
| --- | --- |
|  | (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, and weighted sensitivity and weighted specificity, the weighted average of each, are used [42]. Hence, to create a binary confusion matrix for the diatoms 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). , , , and is the serial number for each category, the probability of being included in each category, and the sensitivity and specificity for each category, respectively.

|  |  |
| --- | --- |
|  | (16) |

Moreover, there is a trade-off relationship between sensitivity and specificity, where one decreases if the other increases [43]. 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 is imbalanced towards the diatoms category.

|  |  |
| --- | --- |
|  | (17) |

**3.2.1 Tree-based algorithm for Assessing Variable Importance**

In this study, the classification performance of five tree-based algorithms, namely bagging, AdaBoost, gradient boosting, random forest, and extreme gradient boosting, was compared. Each algorithm computes variable importance to determine which explanatory variable has the most influence on the response variable [44]. Variable importance increases as the reduction of 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 graphs of the error calculated when applying the random forest algorithm based on the training data at each survey site. The OOB (out-of-bag) error in the legend refers to the error obtained by using the remaining data not included in survey 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 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 16 presents the cross-validation tests conducted by extreme gradient boosting, wherein the point indicating the smallest mlogloss error value was considered as the best iteration. As illustrated in Fig. 16, the mlogloss error value progressively decreased with each iteration for the training data, but increased 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.

|  |  |
| --- | --- |
| Chart  Description automatically generated | Chart, histogram  Description automatically generated |
| **(a)** | **(b)** |
| Chart, histogram  Description automatically generated | Chart, text  Description automatically generated |
| **(c)** | **(d)** |

**Figure 18.** Graph of error in Random Forest for survey site (a) J1, (b) J2, (c) T1, and (d) T2.

|  |  |
| --- | --- |
| Chart, line chart  Description automatically generated | Chart, line chart  Description automatically generated |
| **(a)** | **(b)** |
| Chart, line chart  Description automatically generated | Chart, line chart  Description automatically generated |
| **(c)** | **(d)** |

**Figure 19.** Cross-validation test in Extreme Gradient Boosting for survey site (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, and the results are shown in Tables 8 and 9. Based on the results, the variable importance calculations varied depending on the survey site and algorithm. Overall, temperature and DO (dissolved oxygen) were 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 algae 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 was relatively high, surpassing that of DO. At the Tamjin Lake dam front (T1) site, the variable importance of EC was relatively high. This indicates that nutrients like nitrogen and phosphorus have a more significant influence on algae growth at the Tamjin Lake site compared to the Juam Lake site.

**Table 9.** Variable importance of explanatory variables for dominant algae 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 |
| LowWaterLevel | **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 algae classification performance of the 11 statistical machine learning algorithms explained in Section 2.6, 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 10 presents the calculations of these four criteria for each algorithm based on the classification results by survey site, and Table 11 displays the best algorithms based on the four criteria.

The results show that the optimal algorithm varied depending on the survey site and evaluation criteria. Moreover, our findings indicated 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 is 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.

**Table 10.** Result of dominant algae classification by 11 statistical machine learning algorithms.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **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 |

**Table 11.** The best algorithm based on four criteria.

|  |  |  |
| --- | --- | --- |
| **Site** | **Criterion** | **Best Algorithm** |
| J1 | Accuracy | Decision Tree |
| Weighted Sensitivity | Decision Tree |
| Weighted Specificity | Regularized Discriminant Analysis |
| G mean | Decision Tree |
| J2 | Accuracy | Random Forest |
| Weighted Sensitivity | Random Forest |
| Weighted Specificity | Extreme Gradient Boosting |
| G mean | Random Forest |
| T1 | Accuracy | Random Forest |
| Weighted Sensitivity | Random Forest |
| Weighted Specificity | Linear Discriminant Analysis, Flexible Discriminant Analysis |
| G mean | Support Vector Machine |
| T2 | Accuracy | Gradient Boosting |
| Weighted Sensitivity | Gradient Boosting |
| Weighted Specificity | Extreme Gradient Boosting |
| G mean | Gradient Boosting |

4. Discussion

This study analyzed the dominant algae from 2017 to 2022 at the sites of Juam Lake and Tamjin Lake, representative water supply sources in the Yeongsan River and Seomjin River systems in South Korea, and briefly examined their seasonal characteristics. Additionally, water quality parameters and hydraulic/hydrological parameters related to algae 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 self-organizing map 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 algae classification accuracy of 11 statistical machine learning algorithms for each survey site.

According to the algorithm examination, we found that the best algorithm differs for each survey site, indicating that the environmental characteristics of each survey site also differ. Unlike previous studies [48-49], which mainly used traditional multivariate statistical analysis techniques like principal component analysis (PCA) or clustering analysis (CA) to evaluate the environmental characteristics of the survey site, our study is meaningful as it attempted to evaluate the environmental characteristics of the survey site using the latest version of statistical machine learning algorithms. The main results of this study are as follows. According to the line graphs showing the monthly average number of cells for each algae type measured during the survey period from 2017 to 2022 at the Juam Lake and Tamjin Lake sites, chlorophytes or diatoms tended to dominate in spring, cyanophytes in early summer and summer, and chlorophytes and diatoms in autumn and early winter.

Through an exploratory data analysis, using correlation analysis and pattern analysis of the self-organizing map 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. According to the results, overall, the mutually related water quality parameters (BOD, COD, TN, TP, etc.) showed positive correlations, while the water quality item variables and hydraulic/hydrological variables showed negative correlations.

Using the data measured 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 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 may limit the generalizability of the findings to other water sources worldwide. Future research should explore the prediction of dominant algae in a larger number of investigation sites to obtain more universal results. This would facilitate the development of a way to evaluate the 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. Conclusion

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, extending the survey period, or analyzing data from water supply sources outside of the Yeongsan River and Seomjin River system, may result in different outcomes. As the amount of data increases, more prior knowledge can be obtained, and algorithms can be further trained, potentially improving their 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, continuous research on various water source points and result comparisons are necessary. This research approach can aid in the efficient management of water supply source quality by accurately classifying dominant algae occurrences.

**Supplementary Materials:** The following supporting information can be downloaded at: www.mdpi.com/xxx/s1, Figure S1: title; Table S1: title; Video S1: title.

**Author Contributions:** For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used “Conceptualization, X.X. and Y.Y.; methodology, X.X.; software, X.X.; validation, X.X., Y.Y. and Z.Z.; formal analysis, X.X.; investigation, X.X.; resources, X.X.; data curation, X.X.; writing—original draft preparation, X.X.; writing—review and editing, X.X.; visualization, X.X.; supervision, X.X.; project administration, X.X.; funding acquisition, Y.Y. All authors have read and agreed to the published version of the manuscript.” Please turn to the [CRediT taxonomy](https://img.mdpi.org/data/contributor-role-instruction.pdf) for the term explanation. Authorship must be limited to those who have contributed substantially to the work reported.

**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:** We encourage all authors of articles published in MDPI journals to share their research data. In this section, please provide details regarding where data supporting reported results can be found, including links to publicly archived datasets analyzed or generated during the study. Where no new data were created, or where data is unavailable due to privacy or ethical restrictions, a statement is still required. Suggested Data Availability Statements are available in section “MDPI Research Data Policies” at https://www.mdpi.com/ethics.

**Acknowledgments:** In this section, you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments).

**Conflicts of Interest:** Declare conflicts of interest or state “The authors declare no conflict of interest.” Authors must identify and declare any personal circumstances or interest that may be perceived as inappropriately influencing the representation or interpretation of reported research results. Any role of the funders 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 must be declared in this section. If there is no role, please state “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. (2017). Green algae and algae warning system. *Water for future, 50*(7), 22-26.
2. Kim, K. B., Jung, M. K., Tsang, Y. F., & Kwon, H. H. (2020). Stochastic modeling of chlorophyll-a for probabilistic assessment and monitoring of algae blooms in the Lower Nakdong River, South Korea. *Journal of Hazardous Materials, 400,* 123066.
3. Srivastava, A., Ahn, C. Y., Asthana, R. K., Lee, H. G., & Oh, H. M. (2015). Status, alert system, and prediction of cyanobacterial bloom in South Korea. *BioMed Research International, 2015.*
4. Falconer, I. R., & Humpage, A. R. (2005). Health risk assessment of cyanobacterial (blue-green algal) toxins in drinking water. *International Journal of Environmental Research and Public health, 2*(1), 43-50.
5. Fleming, L. E., Rivero, C., Burns, J., Williams, C., Bean, J. A., Shea, K. A., & Stinn, J. (2002). Blue green algal (cyanobacterial) toxins, surface drinking water, and liver cancer in Florida. *Harmful Algae, 1*(2), 157-168.
6. Kim, Y. H. (2022). *Harmful Cyanobacterial Bloom and Application of Physical, Chemical and Biological Control Methods* (Doctoral dissertation, Hanyang University).
7. Joo, J. H. (2017). *Field application and development of biologically derived substances (BDSs) to mitigate freshwater harmful cyanobacterial blooms* (Doctoral dissertation, Hanyang University).
8. Guillaume, M. C., & Dos Santos, F. B. (2023). Assessing and reducing phenotypic instability in cyanobacteria. *Current Opinion in Biotechnology, 80,* 102899.
9. Kim, H. G. (2017). Prediction of chlorophyll-a in the middle reach of the Nakdong River at Maegok using artificial neural networks. *Department of Integrated Biological Science, Master’s Thesis, The Graduate School of Busan National University, Busan, Korea.*
10. Lee, S. M., Park, K. D., & Kim, I. K. (2020). Comparison of machine learning algorithms for Chl-a prediction in the middle of Nakdong river (focusing on water quality and quantity factors). *Journal of Korean Society of Water and Wastewater, 34*(4), 277-288.
11. Bui, D. T., Khosravi, K., Tiefenbacher, J., Nguyen, H., & Kazakis, N. (2020). Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Science of the Total Environment*, *721*, 137612.
12. Caissie, D., Satish, M. G., & El-Jabi, N. (2007). Predicting water temperatures using a deterministic model: Application on Miramichi River catchments (New Brunswick, Canada). *Journal of Hydrology, 336*(3-4), 303-315.
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. (2012). Estimation of pollutant load delivery ratio for flow duration using LQ equation from the Oenam-cheon watershed in Juam Lake. *Journal of Environmental Science International, 21*(1), 31-39.
14. Park, H. G., Kang, D. W., Shin, K. H., & Ock, G. Y. (2017). Tracing source and concentration of riverine organic carbon transporting from Tamjin River to Gangjin Bay, Korea. *Korean Journal of Ecology and Environment, 50*(4), 422-431.
15. Seo, K. A., Jung, S. J., Park, J. H., Hwang, K. S., & Lim, B. J. (2013). Relationships Between the Characteristics of Algae Occurrence and Environmental Factors in Lake Juam, Korea. *Journal of Korean Society on Water Environment, 29*(3), 317-328.
16. Cox, V. (2017). Exploratory data analysis. In *Translating Statistics to Make Decisions* (pp. 47-74). Apress, Berkeley, CA.
17. Das, K. R., & Imon, A. H. M. R. (2016). A brief review of tests for normality. *American Journal of Theoretical and Applied Statistics, 5*(1), 5-12.
18. Thadewald, T., & Büning, H. (2007). Jarque–Bera test and its competitors for testing normality–a power comparison. *Journal of Applied Statistics, 34*(1), 87-105.
19. Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE, 78*(9), 1464-1480.
20. Jung, K. Y., Cho, S. H., Hwang, S. Y., Lee, Y. J., Kim, K. H., & Na, E. H. (2020). Identification of High-Priority Tributaries for Water Quality Management in Nakdong River Using Neural Networks and Grade Classification. *Sustainability, 12*(21), 9149.
21. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: Springer.
22. Sugiyama, M. (2015). *Introduction to statistical machine learning.* Morgan Kaufmann.
23. Park, K. Y., & JW, K. (2019). A short guide to machine learning for economists. *Korean Journal of Economics, 26*(2), 367-408.
24. Han, S. W. (2016). A study on kernel ridge regression using ensemble method. *Department of Statistics, Master’s Thesis, The Graduate School of Hankuk University of Foreign Studies, Seoul, Korea.*
25. Hwang, S. Y. (2017). A study on efficiency of kernel ridge logistic regression classification using ensemble method. *Department of Statistics, Master’s Thesis, The Graduate School of Hankuk University of Foreign Studies, Seoul, Korea.*
26. Schapire, R. E. (2013). Explaining adaboost. In *Empirical Inference* (pp. 37-52). Springer, Berlin, Heidelberg.
27. Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics, 7,* 21.
28. Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random forests. In *Ensemble Machine Learning* (pp. 157-175). Springer, Boston, MA.
29. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., & Chen, K. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2, 1*(4), 1-4.
30. Izenman, A. J. (2013). Linear discriminant analysis. In *Modern Multivariate Statistical Techniques* (pp. 237-280). Springer, New York, NY.
31. Reynès, C., Sabatier, R., & Molinari, N. (2006). Choice of B-splines with free parameters in the flexible discriminant analysis context. *Computational Statistics & Data Analysis, 51*(3), 1765-1778.
32. Schölkopf, B. (2000)).
33. Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association, 84*(405), 165-175.
34. Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In *Machine Learning* (pp. 101-121). Academic Press.
35. Montavon, G., Samek, W., & Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing, 73*, 1-15.
36. Jung, K. Y., Lee, I. J., Lee, K. L., Cheon, S. U., Hong, J. Y., & Ahn, J. M. (2016). Long-term trend analysis and exploratory data analysis of Geumho River based on seasonal Mann-Kendall test. *Journal of Environmental Science International, 25*(2), 217-229.
37. Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 9*(2), 78.
38. De Winter, J. C., Gosling, S. D., & Potter, J. (2016). Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychological Methods, 21*(3), 273.
39. Bai, J., & Ng, S. (2005). Tests for skewness, kurtosis, and normality for time series data. *Journal of Business & Economic Statistics, 23*(1), 49-60.
40. Parikh, R., Mathai, A., Parikh, S., Sekhar, G. C., & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian Journal of Ophthalmology, 56*(1), 45.
41. Xu, J., Zhang, Y., & Miao, D. (2020). Three-way confusion matrix for classification: A measure driven view. *Information Sciences, 507,* 772-794.
42. Li, D. L., Shen, F., Yin, Y., Peng, J. X., & Chen, P. Y. (2013). Weighted Youden index and its two-independent-sample comparison based on weighted sensitivity and specificity. *Chinese Medical Journal, 126*(06), 1150-1154.
43. Trevethan, R. (2017). Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice. *Frontiers in Public Health, 5,* 307.
44. Gregorutti, B., Michel, B., & Saint-Pierre, P. (2017). Correlation and variable importance in random forests. *Statistics and Computing, 27*(3), 659-678.
45. Genuer, R., & Poggi, J. M. (2020). Random forests. In *Random Forests with R* (pp. 33-55). Springer, Cham.
46. Roelofs, R., Shankar, V., Recht, B., Fridovich-Keil, S., Hardt, M., Miller, J., & Schmidt, L. (2019). A meta-analysis of overfitting in machine learning. *Advances in Neural Information Processing Systems, 32.*
47. Woo, C. Y., Yun, S. L., Kim, S. G., & Lee, W. T. (2020). 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. *Korean Society of Environmental Engineers, 42*(12), 664-673.
48. Jung, K. Y., Ahn, J. M., Kim, K., Lee, I. J., & Yang, D. S. (2016). Evaluation of water quality characteristics and water quality improvement grade classification of Geumho River tributaries. *Journal of Environmental Science International, 25*(6), 767-787.
49. Sun, X., Zhang, H., Zhong, M., Wang, Z., Liang, X., Huang, T., & Huang, H. (2019). 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. *International Journal of Environmental Research and Public Health, 16*(6), 1020.