Comparison of algorithm for dominant algae classification in water supply source site in Yeongsan, Seomjin river basin using Statistical Machine Learning Technique

Corresponding author:

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,

Seong J. Yang³\*, Kang-Young Jung²\*\*

*¹Yeongsan River Environment Research Center, National Institute of Environmental Research*

*²Education Planning Division, National Institute of Environmental Human Resources Development*

*³Department of Statistics (Institute of Applied Statistics), Jeonbuk National University*

**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 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.

**Key words**: Water Quality, Yeongsan·Seomjin River, Correlation Analysis, Self-Organizing Map, Statistical Machine Learning Algorithm, Classification

**1. Introduction**

In South Korea, sites that serve a role in providing living water to local residents are designated and managed as water protection zones. The extreme drought in the Honam region of South Korea in 2022 further highlighted the importance of managing water quality in water supply sources. To protect the water quality of water supply sources, which provide drinking water for residents, the Korean government established the algae alert system in 1998. To minimize toxic damage caused by harmful cyanobacteria occurring in water supply source sites, the system issues alerts based on the number of harmful cyanobacteria cells: Caution (at least 1,000 cells/mL two consecutive times), Warning (at least 10,000 cells/mL two consecutive times), Outbreak (at least 1,000,000 cells/mL two consecutive times), and Release (number of cyanobacteria cells below the alert threshold two consecutive times) (Kim, S. G. (2017)). Particularly, four representative harmful cyanobacteria species*, Aphanizomenon, Anabaena, Oscillatoria,* and *Microcystis,* release harmful toxins, causing acute liver disease in humans (Falconer et al. (2005)) and posing a serious threat to the stability of aquatic ecosystems (Fleming et al. (2002)). Researchers have studied various reduction methods to control the occurrence of these harmful cyanobacteria, including algal blocking mat (ABM), a physical method: plant-mineral composite (PMC), a chemical method; and Unio douglasiae, a biological method (Kim, Y. H. (2022) and Joo, J. H. (2017)). Using an artificial neural network technique, Kim, H, G. (2017) reviewed the method’s suitability for predicting the chlorophyll-a concentration at the Maegok site, a midstream site of the Nakdong River in South Korea. Moreover, Lee et al. (2020) investigated methods for predicting the chlorophyll-a concentration through 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. However, the study was limited as it could not compare the performance of various algorithms. The greatest difference between the two previous studies and this one is that the two studies focused on how accurately the techniques predicted the measured values of the water quality item chlorophyll-a, a continuous variable, whereas this study seeks to accurately classify dominant algae, a categorical variable. Nutrients such as nitrogen (N) and phosphorus (P), as well as water quality factors like water temperature, are the most representative factors that influence the occurrence of algae. However, hydraulic and hydrological factors such as water level and water storage capacity also have an impact, so all factors must be considered (Caissie et al. (2007)). Therefore, given diverse variables related to water quality and hydraulic and hydrological factors, if the dominant algae can be predicted in advance based on this data, then authorities can more rapidly prepare for water pollution accidents caused by algae. Based on water quality monitoring network data, algae alert system data, and hydraulic and hydrological data in Juam Lake and Tamjin Lake, representative water supply sources in the Yeongsan River and Seomjin River system, measured at 7-day intervals from January 2017 to December 2022 through the NIER Water Environment Information System, we compared and analyzed which of various statistical machine learning algorithms accurately classify the dominant algae.

**2. Materials and Methods**

**2.1 Study Area**

This study examined the Juam Lake and Tamjin Lake sites, representative water supply source sites in the Yeongsan River and Seomjin River system in South Korea. To respond to the algae alert system and monitor the water quality of water supply sources providing living water to local residents, the NIER Yeongsan River Environment Research Laboratory collects weekly samples from the dam front (J1) and Shinpyeong Bridge (J2) sites of Juam Lake and the dam front (T1) and Yuchi Stream Confluence (T2) sites of Tamjin Lake, and then measures various water quality items based on this data to survey the occurring algae. Other hydraulic and hydrological variables such as water storage capacity are measured daily by the Korea Water Resources Corporation. Juam Lake is a lake created by the freshwater of Juam Dam, which has a height of 58m and length of 330m. It is located in Daegwang-ri, Juam-myeon, Suncheon-si, Jeollanam-do and has a total basin area of 1,029.41km² and total water storage capacity of 457×10⁶ tons. Juam Dam supplies about 640×10³ tons of living water to the western part of Jeollanam-do, including Gwangju, Naju, Mokpo, and Hwasun (Choi et al. (2012)). Tamjin Lake is an artificial lake created by the construction of Jangheung Dam, which has a height of 53m and length of 403m. It has a total basin area of 193km² and total water storage capacity of 191×10⁶ tons, 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 (Park et al. (2017)). Fig. 1 shows the sampling sites in Juam Lake and Tamjin Lake.

|  |  |
| --- | --- |
|  | |
|  |  |
| (a) Juam lake (J1, J2) | (b) Tamjin lake (T1, T2) |

Fig. 1. Sampling site in the Juam lake and Tamjin lake

**2.2. Data Collection**

For analysis, we collected and organized data needed to comprehensively analyze hydraulic and hydrological data, algae alert system data, and water quality monitoring network data measured at 7-day intervals from January 2017 to December 2022 in the survey sites by the NIER Water Environment Information System. Tables 1 and 2 present the variables of the collected data and the number of observations per survey site, respectively.

Table 1. Variables in data

|  |  |  |  |
| --- | --- | --- | --- |
| **Response variable**  **(categorical)** | | **Explanatory variables**  **(continuous)** | |
| **water quality** | **hydraulics, hydrological** |
| **Dominant Algae**  **(based sum of cell count)** | | 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(10,000m³) |
| **category** | **Cyanophytes** |
| **Diatoms** |
| **Chlorophytes** |
| **Others** |

Table 2. Number of observations per sampling site

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

Of the variables listed above, Biochemical Oxygen Demand(BOD), Chemical Oxygen Demand(COD), Total Nitrogen(TN), Total Phosphorus(TP), Total Organic Carbon(TOC), Suspended Solids(SS), and Electrical Conductivity(EC) were collected from the water quality monitoring network data, and power of Hydrogen(pH), Dissolved Oxygen(DO), Temperature, Turbidity, Transparency, Chlorophyll-a(Chla), and Dominant Algae were collected from the algae alert system data. The remaining variables, Low water level (LowWaterLevel), Inflow, Discharge, and Reservoir, were collected from the National Water Resources Management Information System (http://www.wamis.go.kr/). The algae species that occurred at the survey sites when the data was measured are presented in Table 3. Fig. 2 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 Fig. 2, overall, 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 (Seo et al. (2013)). All of the data analyses in this study were performed through the program R version 4.2.1.

Table 3. Species of algae that appeared at the sampling site

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

|  |  |
| --- | --- |
| **J1** |  |
| **J2** |  |

|  |  |
| --- | --- |
| **T1** |  |
| **T2** |  |

Fig. 2. Line graph of cell count of algae

**2.3. Explanatory Data Analysis**

An explanatory data analysis is performed to investigate the overall data’s characteristics before analyzing the data (Cox, V. (2017)). As there is no set analysis method or process, different researchers may prefer different analysis methods. However, generally, one first determines whether the variables included in the data are continuous variables or categorical variables. If it is a continuous variable, then the variable’s mean, standard deviation, density, etc. are calculated to examine the distributional characteristics; if it is a categorical variable, then the number of categories in the variable and the number of observations for each category are examined.

**2.4. Correlation Analysis**

A correlation analysis is a representative statistical analysis method for investigating the correlation of continuous variables in data. For this purpose, the correlation coefficient is calculated using the Pearson correlation coefficient as in Equation (1), and a significance test is conducted on it. Typically, the validity of the analysis results can be proven only when normality is assumed to be satisfied through a normality test, such as the Shapiro-Wilk (SW) test (Das et al. (2016)). However, a limitation of this method is that it can only be applied when the variable has the properties of a random variable that satisfies independency. Given that all measurement variables in this study are time series variables measured over time rather than random variables that satisfy independency, the Jarque-Bera (JB) test method is more appropriate (Thadewald, T., & Büning, H. (2007)).

(1)

However, most environment-related measurement variables usually do not satisfy normality and have considerable fluctuations. As a result, the results lose reliability if the analysis is conducted using a Pearson correlation coefficient for data with such variables. Therefore, we conducted the correlation analysis using the Spearman correlation coefficient, a non-parametric method that can analyze correlation using rank, as expressed in Equation (2).

(2)

**2.5. Pattern Analysis based Self Organizing Map**

A self-organizing map is an artificial neural network technique that simultaneously performs dimension reduction and clustering (Kohonen, T. (1990)). When using this technique, many nodes contained in the high-dimensional data are clustered through competition, and based on the winning node that emerges through competition, the learning results that preserve similarity as much as possible in the reduced dimensions are obtained. This principle is illustrated in Fig. 3.

|  |
| --- |
|  |

Fig. 3. Self Organizing Map

Through this process, the algorithm shown in Equation (3) is repeated until convergence, and the th lattice vector at time is updated.

(3)

In Equation (3) above, is a learning rate parameter that reduces the learning rate to prevent overfitting, and 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 items measured at 28 sampling sites in the Nakdong River system in South Korea. Using a grading process through a cluster analysis according to the characteristics of each site, they evaluated which branches should be prioritized for management and proposed policies based on this. In this study, we performed a pattern analysis on 17 measurement variables through this method, and identified the correlations between the measurement variables.

**2.6. 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 James et al. (2013) and Sugiyama, M. (2015).

**2.6.1. Decision Tree(DT)**

A decision tree is a method for creating a decision model with a tree-like structure. It typically involves four steps: the growing step maximizes the size of the tree model, the pruning step removes unnecessary branches from the grown tree model, the optimal tree model selection step selects the optimal model from a set of tree models refined through pruning, and the testing step performs prediction or classification based on the finally selected tree model. When reviewing the impurity of nodes 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 the created model. However, this technique is not guaranteed to obtain the optimal tree model (Park et al. (2019)).

(4)

(5)

(6)

**2.6.2. Bagging(Bag)**

Bagging uses the 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 final conclusion. “Mode” refers to the value with the highest frequency. Using the bagging technique can greatly reduce the variance of the created model compared to creating a decision tree model once, through which more reliable prediction or classification results can be obtained (Han, S. W. (2016) and Hwang, S. Y. (2017)). Fig. 4 illustrates the principle of the bagging method.

|  |
| --- |
|  |

Fig. 4. Bagging

**2.6.3. AdaBoost(Ada)**

AdaBoost creates a strong learner with improved performance by appropriately combining weak learners with low prediction performance through weights. Through this process, it can yield more accurate results by correcting or supplementing incorrectly predicted or classified ones by the learner used in the previous step. As shown in Equation (7), AdaBoost uses the principle of creating a strong learner through a weighted linear combination of multiple weak learners.

(7)

where is the finally obtained strong learner, are the weak learners, and are the weights of the weak learners. Fig. 5 illustrates the principle of AdaBoost, and a more detailed explanation can be found in Schapire, R. E. (2013).

|  |
| --- |
|  |

Fig. 5. AdaBoost

**2.6.4. Gradient Boosting(GB)**

Gradient boosting repeats the process of using a gradient to create a model and then using the residual from this to create another model. Since it reduces the part that the model in the previous step could not explain, it can reduce bias; however, this also causes 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. Fig. 6 illustrates the principle of gradient boosting, and a more detailed explanation can be found in Natekin et al. (2013).

|  |
| --- |
|  |

Fig. 6. Gradient Boosting

**2.6.5. Random Forest(RF)**

Random forest was proposed to compensate for the disadvantages of bagging, and its overall principle is similar to bagging. The process of extracting multiple samples with replacement from the training data and fitting multiple decision tree models through them is identical between the two techniques. In random forest, however, only some explanatory variables are randomly selected and used for each sample, which improves the prediction or classification performance compared to bagging. Of course, the types of explanatory variables selected for each sample all differ. Through this process, the correlation between each sample that can occur when applying bagging is reduced, which can greatly enhance prediction or classification performance (Cutler et al. (2012)).

**2.6.6. Extreme Gradient Boosting(XGB)**

Extreme gradient boosting is an improved method that additionally supports parallel learning to compensate for the slow execution time and overfitting risk in gradient boosting. Regular gradient boosting does not have a function to regulate overfitting in the algorithm itself, whereas extreme gradient boosting has a function to regulate overfitting itself, making it more stable and durable. It can also independently conduct cross-validation tests and has an early stopping function that detects when overfitting occurs 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. This is expressed in Fig. 7. In this study, we applied the commonly used 5-fold cross-validation test. The point showing the lowest error value for the validation data was determined to be the best iteration. A detailed explanation can be found in Chen et al. (2015).

|  |
| --- |
|  |

Fig. 7. Example of -fold cross-validation test

**2.6.7. 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 are different from each other as much as possible, and that the variances within each category are small. Through this process, it is possible to find the linear decision boundary as shown in Fig. 8. A more detailed explanation can be found in Izenman, A. J. (2013).

|  |
| --- |
|  |

Fig. 8. Linear Discriminant Analysis

**2.6.8. Flexible Discriminant Analysis(FDA)**

Flexible discriminant analysis compensates for the shortcomings of linear discriminant analysis that uses the linear decision boundary. Using the spline method that creates a curve, it forms a non-linear decision boundary and then performs classification through this. Generally, the distribution of variables in data almost never shows a linear relationship. Hence, as most cases have complex non-linear distributions, it is difficult to apply linear discriminant analysis. This problem can be address through flexible discriminant analysis (Reynès et al. (2006)).

**2.6.9. 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 the value of is , then linear discriminant analysis is performed, and if the value of is , then quadratic discriminant analysis is performed. Here, , which serves as the weight for the linear decision boundary and quadratic curved decision boundary (Friedman, J. H. (1989)).

**2.6.10. Support Vector Machine(SVM)**

Support vector machine uses the kernel trick to maximize the margin, the distance between the decision boundary and support vector, to perform prediction or classification. The kernel trick is a technique that uses the mapping function without the need to set an appropriate transformation function beforehand to move the original data in an input space with a complex non-linear distribution to a high-dimensional feature space, thus converting it to a linear distribution and making it easier to find the decision boundary (Schölkopf, B. (2000)). This is expressed in Fig. 9. Of the many kernel types, this study used the radial basis kernel as shown in Equation (10), which is known as the most flexible for all data distributions.

(10)

Fig. 10 illustrates the principle of support vector machine, and a detailed explanation can be found in Pisner et al. (2020).

|  |
| --- |
|  |

Fig. 9. kernel trick

|  |
| --- |
|  |

Fig. 10. Support Vector Machine

**2.6.11. 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 layer and output layer. 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 such as Equation (12), and the cost function is set to a cross entropy function such as Equation (13). In this study, the array of hidden layers was set to for the analysis. A detailed explanation of deep neural networks can be found in Montavon et al. (2018).

(12)

(13)

**3. Result and Interpretation**

**3.1. Result of Data Analysis**

**3.1.1. Explanatory Data Analysis for Monitoring Data**

Based on the data collected for this study, the descriptive statistics of the explanatory variables for each survey site are organized in Table 4. This shows an overview of the distributions of measurement variables for each survey site (Jung et al. (2016)). 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 greater than 0 indicates that the center of the distribution is sharp, and a value less than 0 indicates that the center of the distribution is smooth (Blanca et al. 2013). According to Table 4, none of the measurement variables showed a value of 0 for skewness or kurtosis at any of the survey sites. Also, excluding pH at all survey sites, the JB test p-value was much smaller than the general significance level of 0.05. Hence, since normality is violated in most of the cases, the Spearman correlation coefficient must be used instead of the Pearson correlation coefficient in the correlation analysis (De Winter et al. (2016)). To better visualize the results, Fig. 11 presents boxplots of the measurement variable distributions at each survey site. As shown, the Tamjin Lake survey sites have mostly higher water quality item concentrations than Juam Lake, and most hydraulic and hydrological variables show higher values as well. In contrast, the values of turbidity and transparency are higher in the Juam Lake survey sites than Tamjin Lake, while DO and temperature show similar trends in the sites of the two lakes.

Table 4. Descriptive Statistics

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **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)** | **Tempe-rature**  **(℃)** | **Turbid-ity**  **(NTU)** | **Trans-parency**  **(m)** | **Chla**  **(mg/㎥)** | **Low Water Level**  **(cm)** | **Inflow**  **(cms)** | **Discha-rge**  **(cms)** | **Reserv-oir**  **(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 |

|  |  |
| --- | --- |
|  |  |
| **BOD** | **COD** |
|  |  |
| **TN** | **TP** |

|  |  |
| --- | --- |
|  |  |
| **TOC** | **SS** |
|  |  |
| **EC** | **pH** |

|  |  |
| --- | --- |
|  |  |
| **DO** | **Temperature** |
|  |  |
| **Turbidity** | **Transparency** |

|  |  |
| --- | --- |
|  |  |
| **Chla** | **LowWaterLevel** |
|  |  |
| **Inflow** | **Discharge** |

|  |
| --- |
|  |
| **Reservoir** |

Fig. 11. Boxplot

Table 5 is a contingency table of the variable “Dominant Algae”, a categorical variable. As shown, 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’

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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 Fig. 12. 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 correlation analysis results, though the results varied at each survey site, overall, the mutually related water quality items (BOD, COD, TN, TP, etc.) showed positive correlations, while the water quality item variables and hydraulic and hydrological variables showed negative correlations. An analysis that can support this is a self organizing map-based pattern analysis, the results of which are shown in Fig. 13. Through this, we identified the overall movements of the measurement variables at each survey site during the monitoring period. In the above figure, colors closer to red indicate higher measurement variable values, and colors closer to blue indicate lower measurement variable values. According to the results, water quality items that exhibited significant positive correlations in the correlation analysis showed similar patterns, while the water quality variables and hydraulic and hydrological variables, which exhibited significant negative correlations, showed opposite patterns. However, since this study used time series data measured according to time, they do not satisfy independency. Consequently, this process of calculating the normality test p-value for each time-dependent measurement variable and then performing a correlation analysis and interpretation based on this has clear limitations (Bai, J., & Ng, S. (2005)).

|  |  |  |
| --- | --- | --- |
| **J1** |  |  |
| **Spearman correlation coefficient** | **Testing significance** |
| **J2** |  |  |
| **Spearman correlation coefficient** | **Testing significance** |

|  |  |  |
| --- | --- | --- |
| **T1** |  |  |
| **Spearman correlation coefficient** | **Testing significance** |
| **T2** |  |  |
| **Spearman correlation coefficient** | **Testing significance** |

Fig. 12. Correlation Analysis

|  |  |  |  |
| --- | --- | --- | --- |
| **J1** |  |  |  |
| **BOD(mg/L)** | **COD(mg/L)** | **TN(mg/L)** |
| **J1** |  |  |  |
| **TP(mg/L)** | **TOC(mg/L)** | **SS(mg/L)** |

|  |  |  |  |
| --- | --- | --- | --- |
| **J1** |  |  |  |
| **EC(μS/cm)** | **pH** | **DO(mg/L)** |
| **J1** |  |  |  |
| **Temperature(℃)** | **Turbidity(NTU)** | **Transparency(m)** |

|  |  |  |  |
| --- | --- | --- | --- |
| **J1** |  |  |  |
| **Chla(mg/㎥)** | **LowWaterLevel(cm)** | **Inflow(cms)** |
| **J1** |  |  |  |
| **Discharge(cms)** | **Reservoir(10,000㎥)** |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **J2** |  |  |  |
| **BOD(mg/L)** | **COD(mg/L)** | **TN(mg/L)** |
| **J2** |  |  |  |
| **TP(mg/L)** | **TOC(mg/L)** | **SS(mg/L)** |

|  |  |  |  |
| --- | --- | --- | --- |
| **J2** |  |  |  |
| **EC(μS/cm)** | **pH** | **DO(mg/L)** |
| **J2** |  |  |  |
| **Temperature(℃)** | **Turbidity(NTU)** | **Transparency(m)** |

|  |  |  |  |
| --- | --- | --- | --- |
| **J2** |  |  |  |
| **Chla(mg/㎥)** | **LowWaterLevel(cm)** | **Inflow(cms)** |
| **J2** |  |  |  |
| **Discharge(cms)** | **Reservoir(10,000㎥)** |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **T1** |  |  |  |
| **BOD(mg/L)** | **COD(mg/L)** | **TN(mg/L)** |
| **T1** |  |  |  |
| **TP(mg/L)** | **TOC(mg/L)** | **SS(mg/L)** |

|  |  |  |  |
| --- | --- | --- | --- |
| **T1** |  |  |  |
| **EC(μS/cm)** | **pH** | **DO(mg/L)** |
| **T1** |  |  |  |
| **Temperature(℃)** | **Turbidity(NTU)** | **Transparency(m)** |

|  |  |  |  |
| --- | --- | --- | --- |
| **T1** |  |  |  |
| **Chla(mg/㎥)** | **LowWaterLevel(cm)** | **Inflow(cms)** |
| **T1** |  |  |  |
| **Discharge(cms)** | **Reservoir(10,000㎥)** |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **T2** |  |  |  |
| **BOD(mg/L)** | **COD(mg/L)** | **TN(mg/L)** |
| **T2** |  |  |  |
| **TP(mg/L)** | **TOC(mg/L)** | **SS(mg/L)** |

|  |  |  |  |
| --- | --- | --- | --- |
| **T2** |  |  |  |
| **EC(μS/cm)** | **pH** | **DO(mg/L)** |
| **T2** |  |  |  |
| **Temperature(℃)** | **Turbidity(NTU)** | **Transparency(m)** |

|  |  |  |  |
| --- | --- | --- | --- |
| **T2** |  |  |  |
| **Chla(mg/㎥)** | **LowWaterLevel(cm)** | **Inflow(cms)** |
| **T2** |  |  |  |
| **Discharge(cms)** | **Reservoir(10,000㎥)** |  |

Fig. 13. Pattern Analysis based Self Organizing Map

**3.2. Compare performance of Statistical Machine Learning Algorithm**

This chapter presents the results of an analysis performed to determine the dominant algae classification accuracy of 11 statistical machine learning algorithms. The data measured from 2017 to 2021 at each survey site was used as training data to train the algorithm, while the remaining data measured in 2022 was used as test data. As the main objective of this study is to determine the classification accuracy, evaluation criteria are required; three representative criteria for this are accuracy, sensitivity, and specificity (Parikh et al. (2008)). As shown in Table 6, these are calculated through a confusion matrix that organizes the actual correct answers and the answers predicted through classification.

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)

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 that are properly classified as positive out of those that are actually positive, and specificity is the ratio of observations that are properly classified as negative out of those that are actually negative (Xu et al. 2020). 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 (Li et al. (2013)). 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 (Trevethan, R. (2017)). Therefore, we additionally defined G mean, which can serve as a suitable supplementary point for these two metrics. This is obtained by taking the square root of the product of weighted sensitivity and weighted specificity as in Equation (17). We apply this form because the measurement data is imbalanced towards the diatoms category.

(17)

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

Of the algorithms whose classification performance is compared in this study, bagging, AdaBoost, gradient boosting, random forest, and extreme gradient boosting, which are made based on the tree model, are all referred to as tree-based algorithms. When classifying a specific response variable, these algorithms numerically calculate which explanatory variable has the most influence, which is called variable importance (Gregorutti et al. (2017)). Variable importance increases as the reduction of the Gini coefficient or the sum of squared errors, an impurity metric, increases. Particularly, in extreme gradient boosting, variable importance is calculated based on three measurement criteria: gain, cover, and frequency.

Let us examine several figures. Fig. 14 shows graphs of the error calculated when applying random forest based on the training data at each survey site. In the legend, OOB (out-of-bag) error refers to the error obtained by using the remaining data not included in sampling with replacement, which allows duplication, from the training data as validation data (Genuer, R. and Poggi, J. M. (2020)). The other items in the legend indicate the probability of an incorrect answer calculated as the error for each category when the dominant algae is arbitrarily classified as cyanophytes, diatoms, chlorophytes, and other algae for each specific individual. Fig. 14 shows that each error converges to a specific value as the number of tree models used in random forest increases. and the probability of error is the lowest when probabilistically judging that the dominant algae are diatoms. This proves the fact that during the survey period, time points when diatoms were the dominant algae were the most frequent. Extreme gradient boosting can conduct cross-validation tests on its own, the graphs for which are shown in Fig. 15. Through this, the point showing the smallest error value can be set as the best iteration. As shown in Fig. 15, for the training data, the error value continues to decrease with each iteration, whereas for the test data, the error value increases after a certain point, which indicates overfitting (Roelofs et al. (2019)). Hence, an advantage of extreme gradient boosting is that it can reduce the risk of overfitting through cross validation. Through this process, the analysis was performed using the point showing the smallest error value for the test data as the optimal condition.

|  |  |
| --- | --- |
|  |  |
| **J1** | **J2** |
|  |  |
| **T1** | **T2** |

Fig. 14. Graph of error in Random Forest

|  |  |
| --- | --- |
|  |  |
| **J1** | **J2** |
|  |  |
| **T1** | **T2** |

Fig. 15. Cross-validation test in Extreme Gradient Boosting

Based on this process, we calculated the variable importance of each algorithm for the training data by survey site, the results of which are shown in Tables 8 and 9. In the dominant algae classification, the top three measurement variable values based on variable importance for each survey site and algorithm are bolded, and in the case of identical values, both variables are bolded. According to the results, while the variable importance calculations varied according to the survey site and algorithm, overall, temperature and DO were more important than the other measurement variables in determining and classifying the dominant algae at a specific point for each survey site. This suggests a high correlation between water temperature and oxygen in terms of the possibility of algae occurrence. It is also consistent 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 the water temperature and dissolved oxygen. However, at the Tamjin Lake Yuchi Stream Confluence (T2) site, the variable importance of nutrient-related measurement variables such as BOD, TN, and Chla rather than DO was relatively high, and at the Tamjin Lake dam front (T1) site, the variable importance of EC was relatively high. This indicates that nutrients such as nitrogen and phosphorus have a greater influence on algae growth at the Tamjin Lake site than the Juam Lake site.

Table 8. Variable Importance of Explanatory Variables for dominant algae classification (Bagging, AdaBoost, Gradient Boosting, Random Forest)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **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** |
| **LowWaterLevel** | 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 9. Variable Importance of Explanatory Variables for dominant algae classification (Extreme Gradient Boosting)

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Algorithm** | **Extreme Gradient Boosting** | | | | | | | | | | | |
| **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 Compare algorithm based 4 criterion**

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. We then trained each algorithm using the training data, and based on the derived results, compared the classification performance based on accuracy, weighted sensitivity, weighted specificity, and G mean according to the test data. Table 10 shows the calculations of these four metrics for each algorithm based on the classification results by survey site, and Table 11 shows the best algorithms based on the four criteria. According to the results, the optimal algorithm varied according to the survey site and evaluation criteria. Furthermore, our findings indicated that algorithms with complex structures and training processes do not always yield optimal performance, and that even simple algorithms can sometimes sufficiently analyze the given data. The data used in this study is imbalanced, in which diatoms are the dominant algae in the majority of cases. As such, it is most desirable to select the optimal algorithm based not on accuracy but on G mean, which appropriately combines the harmonic average of weighted sensitivity and weighted specificity. 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 Stream Confluence (T2) site. That the best algorithm differs for each survey site suggests that the environmental characteristics of each survey site also differ.

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

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | **Algorithm** | | | | | | | | | | |
| **site** | **criterion** | **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. Best algorithm based 4 criterion

|  |  |  |
| --- | --- | --- |
| **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. Conclusion**

This study surveyed the dominant algae that occurred from 2017 to 2022 at the sites of Juam Lake and Tamjin Lake, representative water supply sources in the Yeongsan River and Seomjin River system in South Korea, and briefly examined their seasonal characteristics. Additionally, water quality items and hydraulic and hydrological variables considered related to algae occurrence were collected based on water quality monitoring network data, algae alert system data, and hydraulic and hydrological data to construct the data needed for analysis. Then, we performed an explanatory data analysis, correlation analysis, and self-organizing map-based pattern analysis for each measurement variable according to the four survey sites to investigate the overall relationships between the variables and their distributional characteristics. We then examined the dominant algae classification accuracy of 11 statistical machine learning algorithms for each survey site based on four algorithm evaluation criteria. The results are as follows.

1. 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.

2. Through an explanatory data analysis, correlation analysis, and self-organizing map-based pattern analysis of the monitoring data, we analyzed the water quality items and hydraulic and 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 items (BOD, COD, TN, TP, etc.) showed positive correlations, while the water quality item variables and hydraulic and hydrological variables showed negative correlations. However, the data used in this study are time series data measured according to time that do not satisfy independency. Consequently, the process of calculating the normality test p-value for each measurement variable in the data and then performing a correlation analysis and interpretation based on this has clear limitations.

3. 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 Stream Confluence (T2) site. That the best algorithm differs for each survey site indicates that the environmental characteristics of each survey site also differ. Moreover, through variable importance based on the tree-based algorithms, we confirmed that water temperature and oxygen are highly correlated with the possibility of algae occurrence.

However, the above results were derived only for the Juam Lake and Tamjin Lake sites; different results may be obtained if we add other measurement variables such as precipitation, collect more data by lengthening the survey period, or analyze data collected from water supply source sites in water systems other than the Yeongsan River and Seomjin River system. This is because as the amount of data increases, more prior information can be obtained and the algorithms can conduct more training, thus enhancing their performance. Moreover, since other water systems exhibit different water quality, hydraulic, and hydrological characteristics from the Yeongsan River and Seomjin River system, even the same algorithms will yield different results if applied to other water systems. We expect this research method to help manage the water quality of water supply sources more efficiently by accurately classifying the dominant algae occurring in water supply source sites.

**Acknowledgement**

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).

**References**

Kim, S. G. (2017). Green algae and algae warning system. *Water for future, 50*(7), 22-26.

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.

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.

Kim, Y. H. (2022). *Harmful Cyanobacterial Bloom and Application of Physical, Chemical and Biological Control Methods* (Doctoral dissertation, Hanyang University).

Joo, J. H. (2017). *Field application and development of biologically derived substances (BDSs) to mitigate freshwater harmful cyanobacterial blooms* (Doctoral dissertation, Hanyang University).

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.*

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.

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.

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.

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.

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.

Cox, V. (2017). Exploratory data analysis. In *Translating Statistics to Make Decisions* (pp. 47-74). Apress, Berkeley, CA.

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.

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.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE, 78*(9), 1464-1480.

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.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

Sugiyama, M. (2015). *Introduction to statistical machine learning.* Morgan Kaufmann.

Park, K. Y., & JW, K. (2019). A short guide to machine learning for economists. *Korean J. Econ, 26*(2), 367-408.

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.*

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.*

Schapire, R. E. (2013). Explaining adaboost. In *Empirical inference* (pp. 37-52). Springer, Berlin, Heidelberg.

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics, 7,* 21.

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random forests. In *Ensemble machine learning* (pp. 157-175). Springer, Boston, MA.

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.

Izenman, A. J. (2013). Linear discriminant analysis. In *Modern multivariate statistical techniques* (pp. 237-280). Springer, New York, NY.

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.

Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American statistical association, 84*(405), 165-175.

Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In *Machine learning* (pp. 101-121). Academic Press.

Montavon, G., Samek, W., & Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. *Digital signal processing, 73*, 1-15.

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.

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.

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.

Bai, J., & Ng, S. (2005). Tests for skewness, kurtosis, and normality for time series data. *Journal of Business & Economic Statistics, 23*(1), 49-60.

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.

Xu, J., Zhang, Y., & Miao, D. (2020). Three-way confusion matrix for classification: A measure driven view. *Information sciences, 507,* 772-794.

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.

Trevethan, R. (2017). Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice. *Frontiers in public health, 5,* 307.

Gregorutti, B., Michel, B., & Saint-Pierre, P. (2017). Correlation and variable importance in random forests. *Statistics and Computing, 27*(3), 659-678.

Genuer, R., & Poggi, J. M. (2020). Random forests. In *Random Forests with R* (pp. 33-55). Springer, Cham.

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.*

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.