Research Article

The Prediction of Diatom Abundance by Comparison of Various Machine Learning Methods

Yuna Shin,¹ Heesuk Lee,² Young-Joo Lee,² Dae Keun Seo,² Bomi Jeong,³ Seoksu Hong ,³ Jaehoon Kim,³ Taekgeun Kim,³ Jae-Kyeong Lee,⁴ and Tae-Young Heo ³

¹Department of Water Environment Research, National Institute of Environmental Research, Incheon 22689, Republic of Korea

²K-water, Daejeon 34045, Republic of Korea

³Department of Information & Statistics, Chungbuk National University, Chungbuk 28644, Republic of Korea

⁴Idea Commercialization Center, Korea Institute of Science Technology Information, Seoul 02456, Republic of Korea

Correspondence should be addressed to Tae-Young Heo; theo@cbnu.ac.kr

Received 24 July 2018; Revised 24 November 2018; Accepted 20 December 2018; Published 27 May 2019

Academic Editor: Mohammad A. Hariri-Ardebili

Copyright © 2019 Yuna Shin et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This study adopts two approaches to analyze the occurrence of algae at Haman Weir for Nakdong River; one is the traditional statistical method, such as logistic regression, while the other is machine learning technique, such as kNN, ANN, RF, Bagging, Boosting, and SVM. In order to compare the performance of the models, this study measured the accuracy, specificity, sensitivity, and AUC, which are representative model evaluation tools. The ROC curve is created by plotting association of sensitivity and (1-specificity). The AUC that is area of ROC curve represents sensitivity and specificity. This measure has two competitive advantages compared to other evaluation tools. One is that it is scale-invariant. It means that purpose of AUC is how well the model predicts. The other is that the AUC is classification-threshold-invariant. It shows that the AUC is independent of threshold because it is plotted association of sensitivity and (1-specificity) obtained by threshold. We chose AUC as a final model evaluation tool with two advantages. Also, variable selection was conducted using the Boruta algorithm. In addition, we tried to distinguish the better model by comparing the model with the variable selection method and the model without the variable selection method. As a result of the analysis, Boruta algorithm as a variable selection method suggested $\text{PO}_4\text{-P}$, DO, BOD, $\text{NH}_3\text{-N}$, Susp, pH, TOC, Temp, TN, and TP as significant explanatory variables. A comparison was made between the model with and without these selected variables. Among the models without variable selection method, the accuracy of RF analysis was highest, and ANN analysis showed the highest AUC. In conclusion, ANN analysis using the variable selection method showed the best performance among the models with and without variable selection method.

1. Introduction

Recently, the decrease of river flow due to the diverse effects of climate change and the increase of residence time of the river due to the influence of the installation of structures installed on the river, such as beam, weir, and drainage gate, have caused frequent occurrence of algae proliferation. This phenomenon adversely affects the utilization of rivers and the health of the aquatic ecosystem. In particular, these problems are emerging in recent years as social and environmental issues, and efforts to analyze the causes and to present solutions to these issues through academic approaches in various fields are urgently needed. Various conditions, such

as sufficient nutrient supply, retention time, adequate water temperature, and light quantity condition, are essential for the growth of algae. However, under the current water quality monitoring system, it is difficult for researchers to identify and analyze the limiting factors and various control variables for algae growth by each aquatic ecosystem and region. Therefore, it is necessary to study the physical, chemical, and biological environmental factors and optimal control variables that are critical for the occurrence of algae, in order to establish an effective algal monitoring system for each country and region.

Algae are small microorganisms that live in the water of river, sea, lake, or ponds; they contain chlorophyll, which

lets them use energy from light and carbon dioxide for photosynthesis, creating organic materials and oxygen. Algae, found in fresh water, mostly consist of very small microalga containing chlorophyll, the green pigment required for photosynthesis; they are classified into diatoms, chlorophyceae, cyanophyceae, and others. Diatoms, chlorophyceae, and cyanophyceae have respective pigments of brown, green, and blue colors.

The optimal conditions of the amount of nutrients, radiation of light, water temperature, etc. required for the growth of each type of algae differ from each other. Depending on conditions, such as the amount of nutrients like phosphorus, sunlight, and water temperature, the types of algae that appear and proliferate vary. In general, diatoms are frequently found during the interval from winter to spring, when water temperature remains below 10°C, while chlorophyceae are frequently found during the interval from spring to early summer, when water temperature remains between (10 and 20)°C. The optimal water temperature for the fast proliferation of cyanophyceae falls in the range (20–30)°C, and accordingly, algal bloom is a seasonal phenomenon that is frequently observable during summertime.

In addition, the occurrence of algae varies significantly, depending on the growth conditions and locations. In an identical aquatic ecosystem, algae appear intensively at points at which the inflow of nutrients, including phosphorus and nitrogen, is concentrated. Points of stagnant flow of water where the flowing water meanders, or where tributaries concentrate, often render proliferating algae.

Algae are one of the natural members essential for the sustainment of an aquatic ecosystem. However, excessive proliferation affects the aquatic ecosystem for the worse. The algae covering the water surface block sunlight and disturb the photosynthesis of aquatic plants in water. In an ecosystem, in which diverse kinds of biological species and abiotic environment share integrated interactions, the prediction of an occurrence of algal proliferation in nature, or clarification of the algal generation process, would be quite difficult.

Despite such difficulties, several techniques for machine learning and deep learning have been developed recently to more accurately predict algal occurrence. Artificial neural network (ANN) was reported useful for the nonlinear ordination of ecological data, and for the visualization, multiple regression, and time series modeling [1]. The approach has been mainly employed for the prediction of phytoplankton species abundance time series [2–4]; ANN was also employed for the selection of input variables for the prediction of phytoplankton proliferation [5, 6].

Cho et al. [7] predicted the TOC, pH, and atmospheric and water temperature as factors affecting algal bloom in Lake Juam, while Jeong et al. [8] carried out the time series modeling of phytoplankton dynamics in the Nakdong-gang river. Park et al. [9] used ANN and three-dimensional hydrodynamic models to clarify the relationship between cyanobacterial blooms and environmental factors in the Baekje Reservoir (inland water). In addition, an early prediction of the concentration of chlorophyll-a was examined in the Juam Reservoir and Yeongsan Reservoir with the use of ANN and support vector machine (SVM) [10].

Wang et al. [11] employed the backpropagation neural network for the prediction of cyanobacterial bloom in Dianchi Lake, whereas Segura et al. [12] used the Random Forest (RF) approach to predict *Microcystis aeruginosa* complex (MAC) colonies. The four machine learning models of the approaches of regression trees (RT), RF, SVM, and ANN were used to predict changes in phytoplankton community composition in the Miyun reservoir [13], whereas Hollister et al. [14] used the RF technique for the modeling of chlorophyll-a and trophic state of lake. In addition, Bourel et al. [15] predicted the presence/absence of marine phytoplankton species based on the four machine learning techniques.

To date, the concentration of chlorophyll-a, which is taken as an indirect indicator of algae, has been examined by many studies. However, studies that delve into the correlation between the number of cells of each taxon of algae and environmental factors are rare. Each species of algae reacts sensitively to environmental changes and is affected by an integration of diverse factors. This is why the employment of simple correlation between algae and factors affecting water quality for the prediction of behaviors of algae is limited. Especially, diatoms are sensitive to nutrient and organic matter concentrations [16] and integrate environmental conditions with long-term water quality [17]. Also, diatoms create substances that cause fishy smell and deteriorate the efficiency of coagulation and sedimentation in the process of water purification. In particular, some of them cause clogging of filter paper. This necessitates an integrated and multidimensional approach for the preparation of scientific and fundamental measures necessary for the control of diatom proliferation.

The prediction of the behavior of algae has been carried out many times in previous studies with the use of ANN and SVM. However, the cases of the employment of other machine learning approaches, such as kNN, Bagging, and Boosting, for the prediction of algal behaviors are rare. In this study, various machine learning techniques were applied, and the performances were compared with each other, to determine an optimal model for the prediction of behaviors of diatoms, and to derive factors affecting the proliferation of diatoms.

2. Materials and Methods

2.1. Data. In this study, the variables determining water quality, such as pH, DO, BOD, COD, Susp, TN, TP, TOC, water temperature, NH₃-N, NO₃-N, and PO₄-P, were taken as explanatory variables, whereas the Dia (Diatom Abundance) was taken as a dependent variable. The data collected for this study were measurements obtained from Haman Weir for Nakdong River. The decision to choose the site where the data is collected has been made thoroughly. The Four Major Rivers Restoration Project had been conducted as a green growth purpose on the Han River, Nakdong River, Geum River, and Yeongsan River in South Korea from January 2009 to October 2011. With only environmental aspect, one of initial purposes of improving water quality by restoring river ecosystem seems to be failed because many rivers suffers from the algal blooms

FIGURE 1: Location of Haman Weir for Nakdong River.

right after the completion of the project. Haman weir was built at the southern tip of the Nakdong River in Figure 1. The algal bloom issue recently emerges around Haman weir due to the proliferation of diatom. This is why the study uses the water quality data from Haman weir to identify which independent variables help diatoms grow.

Figure 2 is missing map showing the pattern of missing values. It represents that a total of 181 cases (69.9%) selected among an entire 259 cases excluding missing data were used for the analyses conducted in this study.

Table 1 presents descriptions of the variables of water quality; we used diatom abundance as a response variable. And pH, DO, BOD, COD, Susp, TN, TP, TOC, Temp, $\text{NH}_3\text{-N}$, $\text{NO}_3\text{-N}$, and $\text{PO}_4\text{-P}$ were used as explanatory variables. The data that support the findings of this study are available from the corresponding author upon reasonable request.

Sunil et al. [18] converted abundance estimates of diatom cells to presence-absence for the prediction of diatom in the continental US to apply the selected machine learning techniques and to compare the results of application and each other. Similarly, data were divided into presence and absence groups from the value of the median of measurements (3,519) of diatoms to apply the selected machine learning techniques and to compare the results of application and each other. “Presence” was defined as greater than median of measurements. Likewise, we defined “absence” as smaller than median of measurements because distribution of diatom abundance is right-skewed.

Table 2 shows the results obtained from the “Independent Samples T-test” conducted to determine the presence of differences between the two groups divided by the degree of occurrence of diatoms. As a result of analysis for differences between the two groups, pH ($P=0.0050$), DO ($P<.0001$),

FIGURE 2: Missing map of variables.

BOD ($P=0.0100$), TN ($P=0.0074$), Temp ($P<.0001$), $\text{NH}_3\text{-N}$ ($P<.0001$), $\text{NO}_3\text{-N}$ ($P=0.0100$), and $\text{PO}_4\text{-P}$ ($P<.0001$) are significant.

2.2. Machine Learning Method

2.2.1. Logistic Regression. Logistic regression is a model of regression analysis for categorical dependent variables. The

TABLE 1: Characteristics of variables.

Characteristic	Variable Type	Descriptions	Mean	SD	Minimum	Maximum
Dia	Response	Diatom Abundance	4735.453	5194.028	0.000	42453.000
pH	Explanatory Variable	pH value	8.183	0.493)	7.200	9.600
DO	Explanatory Variable	Dissolved Oxygen	11.070	2.265	6.500	15.500
BOD	Explanatory Variable	Biochemical Oxygen Demand	2.317	0.882	0.900	4.800
COD	Explanatory Variable	Chemical Oxygen Demand	6.283	1.077	4.100	10.700
Susp	Explanatory Variable	Suspended solid	9.494	5.586	2.200	44.300
TN	Explanatory Variable	Total Nitrogen	2.826	0.796	1.361	5.483
TP	Explanatory Variable	Total Phosphorus	0.053	0.026	0.015	0.174
TOC	Explanatory Variable	Total Organic Carbon	3.925	0.797)	2.600	7.600
Temp	Explanatory Variable	Water Temperature	16.382	8.657	2.600	34.300
NH ₃ -N	Explanatory Variable	Ammonia Nitrogen	0.080	0.062	0.008	0.378
NO ₃ -N	Explanatory Variable	Nitric Nitrogen	2.021	0.690	0.650	4.003
PO ₄ -P	Explanatory Variable	Phosphate Phosphorus	0.009	0.015	0.000	0.092

TABLE 2: Characteristics of various water quality variables.

algal bloom variables		Presence(n=90) Mean(S.E)	Absence(n=91) Mean(S.E)	P-value
Diatom Criteria: 3519	pH	8.291(0.0491)	8.076(0.0523)	0.005
	DO	11.738(0.2160)	10.410(0.2395)	<0.001
	BOD	2.501(0.0889)	2.135(0.0928)	0.01
	COD	6.276(0.1123)	6.291(0.1148)	0.8055
	Susp	9.457(0.4236)	9.532(0.7144)	0.8388
	TN	2.987(0.0931)	2.667(0.0700)	0.0074
	TP	0.050(0.0019)	0.056(0.0032)	0.0857
	TOC	3.839(0.0797)	4.011(0.0871)	0.1406
	Temp	15.604(0.8659)	17.152(0.9490)	<0.001
	NH ₃ -N	0.065(0.0056)	0.096(0.0070)	<0.001
	NO ₃ -N	2.153(0.0765)	1.891(0.0661)	0.01
	PO ₄ -P	0.005(0.0009)	0.013(0.0020)	<0.001

model of logistic regression is constructed of random component, systematic component, and link function construct. The random component signifies the dependent variable Y that includes individual error and is assumed to be the probability distribution of y_i . Systematic components signify independent variables $X_1, X_2, \dots, X_n$; the expression of logistic regression is a linear combination of independent variables:

$$g(p) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n \quad (1)$$

where $p = E(Y | X)$ denotes the probability of Y_i corresponding to the observations of X . The link function $g(p) = \log(p/(1-p))$ connects p to the systematic component in logistic regression.

2.2.2. Support Vector Machine. SVM model is an algorithm that is designed to search for nonlinear formula to maximize the distance between each category of binary data. The hyper-plane is a nonlinear formula for conducting classification

FIGURE 3: A schematic of support vector machine.

for machine learning. The SVM model uses support vector to find grounds for classification. Support vectors nearest to binary classifications are searched for to obtain (non)linear formulae that maximize the margins between the hyperplane and each support vector. In the case where linear classification is disabled, then a kernel function is employed for the conversion of linear function into a higher dimensional nonlinear one, to conduct the classification. Figure 3 shows a hyperplane separating two classes with the maximum margin.

2.2.3. Artificial Neural Network. ANN model is a statistical learning algorithm that emulates the human brain. It is a nonlinear model that is designed to use data for making predictions. The ANN model is constructed of an input layer, hidden layer, and output layer. Each layer is constructed of input node, hidden node, and output node. The input layer corresponds to independent variable X , while the output layer corresponds to the dependent variable Y . Values of hidden nodes are obtained through the nonlinear combination of input nodes. The values of output nodes are available through the nonlinear combination of hidden nodes. Weights are assigned to each nonlinear combination. ANN then performs a training to adjust each weight. The training mainly focuses on the reduction of errors of output values to adjust weights. The adjusted weights are transferred from output nodes to hidden nodes and from hidden nodes to input nodes. This is referred to as a backpropagation algorithm. Figure 4 shows the example of simple artificial neural network.

2.2.4. K-Nearest Neighbor. kNN model is an instance-based learning model to classify the characteristics of data newly found through an identification of existing k training data adjacent to new data. Here, the Euclidean distance, as expressed below, is used to determine the similarity of training data to new data.

$$\text{dist}(X_1, X_2) = \sqrt{\sum_{i=1}^n (x_{1i} - x_{2i})^2} \quad (2)$$

In the kNN model, the value of training data (X_1) and the value of new data (X_2) are used to obtain Euclidean distance to classify the characteristics of new data according to the characteristics of k adjacent data.

2.2.5. Bagging. Bagging is an abbreviation of “Bootstrapping Aggregating”. It is an algorithm deriving new results from the results of several prediction models to be created after sampling with the replacement of given training data. Average values of the predicted results of prediction models are used for the values of continuous dependent variables, and the categories of higher portions among results obtained from prediction models are used as predicted results for the categorical dependent variables. The results of prediction models are variable in terms of statistical learning; the combination of increased prediction models can reduce the variability of the prediction results.

2.2.6. Boosting. Boosting model is an algorithm that uses the technique of sampling with replacement from given training data to create several prediction models using the samples. Respective weights are then assigned to each prediction model to convert the given prediction results, to derive the final prediction results. Types of Boosting comprise Adaboost, Gamboost, and Glmboost, etc. Glmboost method is used as a prediction model in this study. Glmboost is a model derived from the generalized linear model. Glmboost is different from the generalized linear model that creates a basic model after selecting variables. Basic models of respective iterations are obtained to calculate the negative gradient of the loss function for each basic model. The sum of the negative gradient vector of the corresponding iteration and the coefficient of regression is assigned to the model of the next iteration. The final prediction model is then created after completion of the predetermined number of iterations.

2.2.7. CART. CART is one of the decision tree models. It is an algorithm that repeats the process of bifurcation of variables to construct tree models to classify categorical dependent variables, or to conduct regression analysis for continuous dependent variables. CART model proceeds in a way to increase purity, which is a concept opposite to uncertainty; the decrease in uncertainty attained in accordance with increasing purity will ensure more accurate information. Thus the CART algorithm takes criteria that are capable of maximizing purity when it performs bifurcation of the branches in the tree.

2.2.8. Random Forest. Random Forest model exploits the scheme of sampling with replacement from given training data to create several decision tree models. This is an algorithm that creates new results from the results derived previously. The Bagging algorithm, which uses the decision tree model as a prediction model, can be regarded as a sub-model of RF model that adjusts the number of independent variables, which is different from the Bagging algorithm. The independence of each model can be improved statistically by adjusting the number of independent variables to be used for each decision tree model. This can significantly reduce the variability of results.

2.2.9. Model Evaluation Index. As a means to test the results obtained from this study, the “K-fold cross validation” was

FIGURE 4: A schematic description of artificial neural network.

used. The “K-fold cross validation” scheme tests and evaluates the data collected in this study for which the collected data are divided into k-subsets, in which the k-1 subsets, except for just one subset, are used for training of the model, to evaluate its performance with the use of the one subset. Thereafter, the evaluation of the model will be repeated k times, for which all subsets are used as test data. Here, the average value of k evaluative measurements obtained from each stage of test is calculated to make it one evaluative measurement. The formula rendering the average value is

$$T = \frac{1}{k} \sum_{i=1}^k T_i \quad (3)$$

where T_i denotes the i-th measurement among k measurements.

Accuracy, specificity, sensitivity, and area under the curve (AUC) were taken as evaluative indicators; the specificity corresponds to the ratio of accurate prediction in the case of the prediction of diatom occurrence higher than the reference level, whereas the sensitivity corresponds to the ratio of accurate prediction in the case of the prediction of diatom occurrence lower than the reference level. AUC signifies an area resulting from the ROC curve showing the probabilities of an occurrence of data higher and lower than the reference level from the model prediction with given data; the respective probabilities of the results of the corresponding data are predicted from a specified cut-off. The ROC curve results from the sensitivity (=1-specificity) obtained by varying the specified cut-off; the wider area corresponds to better prediction. The formulae of each evaluative indicator are expressed as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\begin{aligned} \text{AUC} &= \sum_{i=1}^k \{1 - \text{Spec}_i - (1 - \text{Spec}_{i-1})\} * \text{Sens}_i \\ &= \sum_{i=1}^k \{\text{Spec}_i + \text{Spec}_{i-1}\} * \text{Sens}_i \end{aligned}$$

(4)

where TP is an acronym for True Positive that corresponds to the ratio of data predicted by the prediction model as higher than the reference level and was also concluded as higher than the reference level among actual data; TN is an acronym for True Negative that corresponds to the ratio concluded by the prediction model as higher than the reference level and was also concluded as higher than the reference level among actual data; FP denotes False Positive that also corresponds to the ratio of data predicted by the prediction model as higher than the reference level and was also concluded as higher than the reference level among actual data. In addition, FN is an acronym for False Negative that corresponds to the ratio of data predicted by the prediction model as higher than the reference level and was also concluded as higher than the reference level among actual data; and k denotes the number of cut-off.

The ROC curve is created by plotting association of sensitivity and (1-specificity). The AUC that is area of ROC curve represents sensitivity and specificity. This measure has two competitive advantages compared to other evaluation tools. One is that it is scale-invariant. It means that purpose of AUC is how well the model predicts. The other is that the AUC is classification-threshold-invariant. It shows that the AUC is independent of threshold because it is plotted association of sensitivity and (1-specificity) obtained by threshold. We chose AUC as a final model evaluation tool with two advantages.

2.2.10. Variable Selection. Variable selection is a method of creating a model by selecting variables that increase the performance of each model. There are traditional statistical

TABLE 3: Results of logistic regression.

	Estimate	Std. Error	z value	p-value	Std. Estimate
Intercept	-8.9789	4.8464	-1.8530	0.0639	-0.0436
pH	0.2108	0.6188	0.3410	0.7334	0.1039
DO	0.3519	0.1939	1.8150	0.0695	0.7971
BOD	0.07631	0.3113	0.2450	0.8064	0.0673
COD	0.1162	0.4113	0.2830	0.7775	0.1252
Susp	0.0971	0.0524	1.8520	0.0639	0.5423
TN	1.2665	0.7994	1.5840	0.1131	1.0086
TP	-26.4483	17.4433	-1.5160	0.1295	-0.6749
TOC	-0.7778	0.4644	-1.6750	0.0939	-0.6198
Temp	0.1763	0.0679	2.5940	0.0095	1.5261
NH ₃ -N	-9.5813	4.1612	-2.3030	0.0213	-0.5936
NO ₃ -N	0.1732	0.9129	0.1900	0.8495	0.1195
PO ₄ -P	-10.9478	27.6156	-0.3960	0.6918	-0.1678

techniques using “Correlation” and “Regression”. The “Correlation” technique identifies multicollinearity through correlation between the paired variables. The “Regression” technique identifies significant variables through regression analysis and multicollinearity variables through “VIF” statistic. There are machine learning techniques using “Variable Important Plot”. The RF model and Bagging model were used to evaluate the importance of each variable and to select variables for modeling. The importance of the variable signifies the degree of contribution of the selected variable corresponding to the improvement of ROC of each model that we chose as final model evaluation tool. In addition, the Boruta algorithm selects variables by the use of the RF model. In order to illustrate the importance of each variable, the colors of green, yellow, red, and blue correspond to the variables of high importance, potential importance, low importance, and those randomly created, respectively.

3. Results and Discussion

3.1. Variable Selection. In this study, the “Correlation Plot”, “Logistic Regression”, and “Variable Importance Plot” were used to select variables. The “Correlation Plot” identifies the presence of multicollinearity through correlation between each variable. The “Logistic Regression” determines variables of statistical significance. In addition, the “Variable Importance Plot” shows the importance of independent variables corresponding to dependent variables. “Logistic Regression” and “Correlation Plot” are based on traditional statistical techniques, and the method of selecting variables using the “Variable Importance Plot” can be understood as a machine learning method. Thus, it is important to compare the traditional method with the machine learning method. We described the three methods and compared between the machine learning method and the traditional method through VIF statistic.

First, the “Correlation Plot” was used to see the correlation between variables of data. Figure 5 shows that the coefficient of correlation between NO₃-N and TN is 0.95, the biggest coefficient of correlation; the coefficient of correlation between TOC and COD is 0.84, while the coefficient of correlation between PO₄-P and TP is 0.74. The coefficient of correlation between Temp and DO is the negative value of -0.80, the biggest negative correlation, whereas the coefficient of correlation between NO₃-N and Temp is -0.77, while the coefficient of correlation between Temp and TN is -0.71.

Second, the standardized estimates of corresponding variables were compared to each other, and the logistic regression was used to find variables affecting diatom presence. Table 3 shows the results of logistic regression analysis in which the variables of Temp and NH₃-N were found to be statistically significant at the level of significance of 5%; the variables elsewhere appeared as statistically insignificant at the level of significance of 5%. Here, the unit increment of Temp that appeared corresponded to an increase of 1.2 times ($= \exp(0.1763)$) of the odds of presence of diatom, while the unit increment of NH₃-N corresponded to the decrease of <0.001 times ($= \exp(-9.5813)$) of the odds of presence of diatom. In addition, the influence of Temp over diatom presence was found to be greater than that of NH₃-N from the “Standardized Estimate”.

Finally, Figure 6 illustrates the RF of algae, the importance of the variables of the RF and Bagging models, and the importance of the variables of Boruta algorithm. In terms of the importance of variables, the variables of the following order of PO₄-P > DO > NH₃-N > BOD > Temp > TOC > TN > Susp > NO₃-N > pH > TP > COD appeared in the RF model, whereas they appeared in the following order: PO₄-P > NH₃-N > DO > BOD > pH > NO₃-N > TN > TOC > Susp > Temp > COD > TP in the Bagging model. They appeared in the following order: PO₄-P > DO > BOD > NH₃-N > Susp > pH > TOC > Temp > TN > TP in the Boruta; COD and NO₃-N can be regarded as variables of potential importance.

FIGURE 5: Correlation Plot.

As a result of variable selection, the variables of $\text{PO}_4\text{-P}$, $\text{NH}_3\text{-N}$, and DO were found to be of higher common influence from the importance of the variables. The variables of Temp and $\text{NH}_3\text{-N}$ were concluded as significant variables from the Logistic Regression. In particular, the variable Temp appeared as highly correlated mostly with other variables, suggesting relatively the low importance for diatom presence, owing to its multicollinearity.

The Boruta method was used as the final variable selection method. The selected variables are pH, DO, BOD, Susp, TN, TP, TOC, Temp, $\text{NH}_3\text{-N}$, and $\text{PO}_4\text{-P}$ because of green that means high importance variables. As a result of VIF, the Temp variable has multicollinearity problem ($\text{VIF} = 11.72$). The multicollinearity is defined if the VIF value is greater than 10. Although Temp variable has multicollinearity problem, we

choose the Temp variable in the model because of association between Temp variable and response variable.

3.2. Classification. In this study, the variables $\text{PO}_4\text{-P}$, DO, BOD, $\text{NH}_3\text{-N}$, Susp, pH, TOC, Temp, TN, and TP were selected with the use of Boruta algorithm described in Section 3.1. The difference in results between the variables selected model and other models were compared with each other.

Each scheme of analysis may have respective advantages and disadvantages, among which the best scheme of analysis fit for respective purposes can be found. Thus, the scheme of analysis fit for characteristics of data should be selected. In addition, the most appropriate scheme of analysis can be

FIGURE 6: Plot of variable importance for diatom using (a) Random Forest, (b) Bagging, and (c) Boruta algorithm results.

derived from comparison of the results obtained from the case of models of the application of selected variables that reflect the characteristics of data, with the results of the case of models done without the application of variables selected through Boruta algorithm.

Table 4 presents the summarized results of “Accuracy”, “Sensitivity”, “Specificity”, and “AUC” resulting from each model of LOGIT, KNN, CART, ANN, RF, Bagging, Boosting, and SVM that predicted the occurrence of diatoms inhabiting the points of investigation.

The results obtained from models that employed variables taken without variable selection showed the highest values of “Accuracy” (71.27%) of RF model and “AUC” (76.65%) of ANN model. Among the results obtained from models

that employed variables taken through variable selection, the highest values of “Accuracy” (71.27%) of RF model and “AUC” (77.93%) of ANN model were obtained. Therefore, the variable selection was found to have influenced models to render mostly higher values of AUC; the RF and ANN models were also found as the most appropriate ones for the prediction of algal proliferation.

3.3. Selected Model Output. CART model was used to identify the relationship between independent variables and dependent variables. CART model repeats the process of bifurcation of each variable to construct a decision tree that enables an interpretation of results that is different from other models. Figure 7 illustrates the relationship between variables

Rattle 2019-2-26 12:56:58 kjh

FIGURE 7: Description of CART.

used for the CART model. When $\text{PO}_4\text{-P}$ and $\text{NH}_3\text{-N}$ exceed respective levels of 0.0035 and 0.047, the diatom is classified as “Absence”, whereas the combination of conditions of $\text{PO}_4\text{-P}$ over 0.0035, $\text{NH}_3\text{-N}$ less than 0.047, and TN less than 2.3 corresponds also to the “Absence” of diatom. The combination of conditions of $\text{PO}_4\text{-P}$ over 0.0035, $\text{NH}_3\text{-N}$ less than 0.047, and TN over 2.3 corresponds to the “Presence” of diatom, whereas the combination of $\text{PO}_4\text{-P}$ less than 0.0035 and TP less than 0.03 corresponds to the “Absence” of diatom.

The combination of $\text{PO}_4\text{-P}$ less than 0.0035, TP over 0.03, and Temp over 23 corresponds to the “Absence”, whereas the combination of $\text{PO}_4\text{-P}$ less than 0.0035, TP over 0.03, Temp less than 23, and $\text{NH}_3\text{-N}$ over 0.13 corresponds to the “Absence” of diatom. Finally, the combination of $\text{PO}_4\text{-P}$ less than 0.0035, TP over 0.03, Temp less than 23, and $\text{NH}_3\text{-N}$ less than 0.13 corresponds to the “Presence” of diatom.

Figure 8 illustrates the structure of independent variables and dependent variables obtained from ANN, in which the thicker lines connecting each node signify greater influence

of the nodes to the next nodes. The diatoms, having hidden nodes of 3, 4, and 5, exhibited the highest accuracy.

4. Conclusion

The measurements of water quality obtained from the Haman Weir for Nakdong River were analyzed with the use of various machine learning techniques, such as kNN, CART, ANN, RF, BAG, BOOST, and SVM, to determine factors of water quality that affect the proliferation of diatoms (factors that have influence on the proliferation of freshwater diatoms). Among the machine learning techniques, the ANN and RF rendered the highest accuracy in the prediction of diatom proliferation. As a result of the variable selection, the variables of $\text{PO}_4\text{-P}$, BOD, $\text{NH}_3\text{-N}$, and DO were found as common influential variables in the models of Bagging and RF because these variables also occupied higher position in the models of RF and Bagging. Also, the major variables that were selected

FIGURE 8: Structure of artificial neural network.

by Boruta algorithm indicated significant improvement of AUC in the analytic schemes of kNN, ANN, BAG, and SVM.

This study was carried out to find the appropriate techniques that were suitable for the prediction of diatom proliferation among various machine learning techniques, from which the ANN and RF were found as appropriate models for the prediction of diatom proliferation. Since diatoms react sensitively to spatiotemporal changes, further studies that focus on the comparison and analysis of diverse data seem necessary to determine the optimal machine learning schemes that are fit for the prediction of the proliferation of each taxon of algae.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Heesuk Lee and Young-Joo Lee investigated the data. Dae Keun Seo validated the data. Bomi Jeong, Seoksu Hong, Jaehoon Kim, Taekgeun Kim, and Jae-Kyeong Lee analyzed data. Yuna Shin and Tae-Young Heo wrote the original paper. All authors read and approved the final manuscript.

Acknowledgments

This study was performed under the project "A Study on the Characteristics of Algae by the Types of Rivers in Korea (NIER-2018-01-01-034)" by the National Institute of Environmental Research.

TABLE 4: Performance comparison for various models using 10-fold cross validation.

			True Value		Accuracy	Sensitivity	Specificity	AUC
			Absence	Presence				
Without Boruta algorithm	LOGISTIC	Absence	57	30	64.64%	65.52%	63.83%	72.19%
		Presence	34	60				
	kNN	Absence	61	33	65.19%	67.00%	63.33%	74.66%
		Presence	30	57				
	CART	Absence	58	33	63.54%	63.89%	63.33%	68.84%
		Presence	33	57				
	ANN	Absence	65	29	69.61%	71.33%	67.78%	76.65%
		Presence	26	61				
	RF	Absence	67	28	71.27%	74.00%	68.89%	76.28%
		Presence	24	62				
	BAG	Absence	58	26	67.40%	64.11%	71.11%	74.44%
		Presence	33	64				
	BOOST	Absence	54	32	63.54%	60.67%	66.67%	73.40%
		Presence	37	58				
	SVM	Absence	61	36	63.54%	67.22%	60.00%	75.41%
		Presence	30	54				
With Boruta algorithm	LOGISTIC	Absence	59	29	66.29%	67.05%	65.59%	73.61%
		Presence	32	61				
	kNN	Absence	56	32	62.98%	61.56%	64.44%	76.34%
		Presence	35	58				
	CART	Absence	60	37	62.43%	66.11%	58.89%	68.68%
		Presence	31	53				
	ANN	Absence	66	32	68.51%	72.78%	64.44%	77.93%
		Presence	25	58				
	RF	Absence	67	28	71.27%	73.89%	68.89%	76.01%
		Presence	24	62				
	BAG	Absence	60	26	68.51%	66.22%	71.11%	75.77%
		Presence	31	64				
	BOOST	Absence	56	28	65.19%	61.78%	68.89%	72.95%
		Presence	35	62				
	SVM	Absence	60	36	62.98%	66.22%	60.00%	76.22%
		Presence	31	54				

References

- [1] F. Recknagel, "Applications of machine learning to ecological modelling," *Ecological Modelling*, vol. 146, no. 1-3, pp. 303–310, 2001.
- [2] H. R. Maier and G. C. Dandy, "Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications," *Environmental Modelling & Software*, vol. 15, no. 1, pp. 101–124, 2000.
- [3] D. F. Millie, G. R. Weckman, G. L. Fahnenstiel et al., "Using artificial intelligence for cyanoHAB niche modeling: discovery and visualization of microcystis–environmental associations within western Lake Erie," *Canadian Journal of Fisheries & Aquatic Sciences*, vol. 71, no. 11, pp. 1642–1654, 2014.
- [4] F. Recknagel, M. French, P. Harkonen, and K.-I. Yabunaka, "Artificial neural network approach for modelling and prediction of algal blooms," *Ecological Modelling*, vol. 96, no. 1–3, pp. 11–28, 1997.
- [5] K.-S. Jeong, D.-K. Kim, and G.-J. Joo, "River phytoplankton prediction model by artificial neural network: model performance and selection of input variables to predict time-series phytoplankton proliferations in a regulated river system," *Ecological Informatics*, vol. 1, no. 3, pp. 235–245, 2006.
- [6] N. Muttill and K.-W. Chau, "Machine-learning paradigms for selecting ecologically significant input variables," *Engineering Applications of Artificial Intelligence*, vol. 20, no. 6, pp. 735–744, 2007.
- [7] S. Cho, B. Lim, J. Jung et al., "Factors affecting algal blooms in a man-made lake and prediction using an artificial neural network," *Measurement*, vol. 53, pp. 224–233, 2014.
- [8] K.-S. Jeong, G.-J. Joo, H.-W. Kim, K. Ha, and F. Recknagel, "Prediction and elucidation of phytoplankton dynamics in the

- Nakdong River (Korea) by means of a recurrent artificial neural network,” *Ecological Modelling*, vol. 146, no. 1-3, pp. 115–129, 2001.
- [9] Y. Park, J. Pyo, Y. S. Kwon et al., “Evaluating physico-chemical influences on cyanobacterial blooms using hyperspectral images in inland water, Korea,” *Water Research*, vol. 126, pp. 319–328, 2017.
 - [10] Y. Park, K. H. Cho, J. Park, S. M. Cha, and J. H. Kim, “Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea,” *Science of the Total Environment*, vol. 502, pp. 31–41, 2015.
 - [11] Z. Wang, K. Huang, P. Zhou, and H. Guo, “A hybrid neural network model for cyanobacteria bloom in Dianchi Lake,” *Procedia Environmental Sciences*, vol. 2, pp. 67–75, 2010.
 - [12] A. M. Segura, C. Piccini, L. Nogueira, I. Alcántara, D. Calliari, and C. Kruk, “Increased sampled volume improves *Microcystis aeruginosa* complex (MAC) colonies detection and prediction using random forests,” *Ecological Indicators*, vol. 79, pp. 347–354, 2017.
 - [13] Q. Zeng, Y. Liu, H. Zhao, M. Sun, and X. Li, “Comparison of models for predicting the changes in phytoplankton community composition in the receiving water system of an inter-basin water transfer project,” *Environmental Pollution*, vol. 223, pp. 676–684, 2017.
 - [14] J. W. Hollister, W. R. Milstead, and B. J. Kreakie, “Modelling lake trophic state: a random forest approach,” Tech. Rep. 2167-9843, Peer J Preprints, 2015.
 - [15] M. Bourel, C. Crisci, and A. Martínez, “Consensus methods based on machine learning techniques for marine phytoplankton presence–absence prediction,” *Ecological Informatics*, vol. 42, pp. 46–54, 2017.
 - [16] D. Hering, R. K. Johnson, S. Kramm, S. Schmutz, K. Szoszkiewicz, and P. F. M. Verdonchot, “Assessment of European streams with diatoms, macrophytes, macroinvertebrates and fish: A comparative metric-based analysis of organism response to stress,” *Freshwater Biology*, vol. 51, no. 9, pp. 1757–1785, 2006.
 - [17] E. D. Reavie, T. M. Jicha, T. R. Angradi, D. W. Bolgrien, and B. H. Hill, “Algal assemblages for large river monitoring: comparison among biovolume, absolute and relative abundance metrics,” *Ecological Indicators*, vol. 10, no. 2, pp. 167–177, 2010.
 - [18] S. Kumar, S. A. Spaulding, T. J. Stohlgren, K. A. Hermann, T. S. Schmidt, and L. L. Bahls, “Potential habitat distribution for the freshwater diatom *didymosphenia geminata* in the continental US,” *Frontiers in Ecology and the Environment*, vol. 7, no. 8, pp. 415–420, 2009.

Submit your manuscripts at
www.hindawi.com