Journal of information and communication convergence engineering
http://jicce.org

J. Inf. Commun. Converg. Eng. 22(2): 133-138, Jun. 2024 Regular paper

Prediction of Dissolved Oxygen at Anyang-stream using XG-Boost and Artificial Neural Networks

Keun Young Lee[1], Bomchul Kim[2], and Gwanghyun Jo[3]*

[1]Independent scholar, Republic of Korea
[2]Department of Environmental Science, Kangwon National University, Chuncheon Kangwon-do, Republic of Korea
[3]Department of Mathematical Data Analysis, Hanyang University ERICA, Ansan Gyeonggi-do, Republic of Korea

Abstract

Dissolved oxygen (DO) is an important factor in ecosystems. However, the analysis of DO is frequently rather complicated because of the nonlinear phenomenon of the river system. Therefore, a convenient model-free algorithm for DO variable is required. In this study, a data-driven algorithm for predicting DO was developed by combining XGBoost and an artificial neural network (ANN), called ANN-XGB. To train the model, two years of ecosystem data were collected in Anyang, Seoul using the Troll 9500 model. One advantage of the proposed algorithm is its ability to capture abrupt changes in climate-related features that arise from sudden events. Moreover, our algorithm can provide a feature importance analysis owing to the use of XGBoost. The results obtained using the ANN-XGB algorithm were compared with those obtained using the ANN algorithm in the Results Section. The predictions made by ANN-XGB were mostly in closer agreement with the measured DO values in the river than those made by the ANN.

Index Terms: XGBoost, artificial neural network, dissolved oxygen, feature importance

I. INTRODUCTION

Dissolved oxygen (DO) significantly influences biological and chemical processes in a river system [1-4]. Maintaining a certain level of DO is essential for the survival of aquatic organisms. In [5], it was discovered that fish mortality in an urban river was caused by the depletion of the DO concentration. Therefore, there is a need for real-time prediction of DO levels in advance of sudden decreases to maintain a healthy ecosystem. Some researchers have adopted ecosystem models [6-8] to predict water-quality variable. These approaches consider both dynamic relations and stochastic variances. However, reconstructing solutions are complicated because of the nonlinearity of the model. Moreover, it is difficult to include singular outer-events, such as sudden heavy rain or snow near the river.

Recently, empirical models have been used to analyze and simulate the complex nonlinear relationships between variables in water resources. In particular, artificial neural network (ANN) models have been employed efficiently to predict water quality-related features [9-15]. Some researchers have found that hybrid models that combine ANN with other machine learning algorithms are superior to simple ANN for water quality analysis. Kavousi-Fard developed an ANN-teacher learning algorithm [16], and Jha and Sahoo proposed an ANN-genetic algorithm to simulate groundwater levels [17]. In particular, Ravansalar et al. developed a hybrid Wavelet-ANN model to predict DO levels in the River Calder [18].

In this study, we developed a hybrid method that combines

Received 1 January 2024, **Revised** 20 February 2024, **Accepted** 4 March 2024
***Corresponding Author** Gwanghyun Jo (E-mail: gwanghyun@hanyang.ac.kr)
Department of Mathematical Data Science, Hanyang University ERICA

Open Access https://doi.org/10.56977/jicce.2024.22.2.133 print ISSN: 2234-8255 online ISSN: 2234-8883

an ANN and extreme gradient boosting (XG-boost) to predict DO in the Anyang Stream in Seoul, Republic of Korea. XG-boost [19] is a gradient boosting machine [20] type algorithm with a decision tree as the basis function. Because of its scalability and robustness, XGBoost has been successfully employed in various research areas. For example, forecasting the crude oil price [21], classifying rock facies [22], detecting DDOS [23], and forecasting the load in a power system [24]. The first step of our algorithm was to train the ANN using time-series DO data. The naïve version of the ANN-predicted DO was then substituted into the XGBoost machine together with meteorological data. Therefore, our method can capture the effects of sudden events in meteorological systems such as heavy rain. This algorithm was called ANN-XGB. One of the advantages of ANN-XGB is its capability of feature importance analysis. We analyzed the correlation of each feature with DO by counting the number of appearances of each feature in the tree branches. The performance of ANN-XGB is reported in the Results Section, where ANN-XGB outperformed simple ANN in most cases. This feature importance indicates that tidal level and water temperature play important roles in predicting DO variable.

The remainder of this paper is organized as follows. In Section 2, the entire procedure for the development of the ANN-XGB algorithms is presented, including data acquisition, pre-processing, and algorithm design. The results are reported in Section 3, and conclusions are presented in Section 4.

II. METHODS

In this section, we describe the ANN-XGB algorithm for predicting DO variables.

A. Data acquisition

Anyang Stream is a tributary of the largest river in Republic of Korea (the Han River) that flows through the capital city and converges with the downstream reaches of the Han River 50 km from the river mouth. The river finally reaches the Yellow Sea, which has a large tidal range of up to 10 m. Because of this large tidal range, the downstream reach of the Han River shows the effect of tides on the water level. The present study site is located 400 m from the main stream of the Han River within the reach of tidal influence; the water level increases and the tributary flow velocity decreases during high tide, and vice versa during low tide. Occasional short-term depletion of dissolved oxygen and high turbidity in rain events caused fish deaths and is regarded as a major stress factor to aquatic organisms. Located in a densely populated residential area, the clarity of the water and scenery of swimming fish is of concern to many local residents visiting the riverside trails of Anyang Stream.

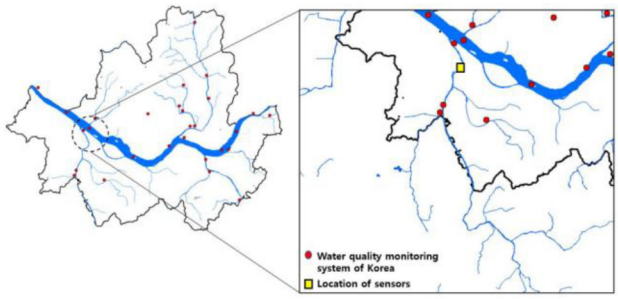

Fig. 1. Maps of the Han River and the Anyang Stream flowing through the Seoul city, showing the governmental water quality monitoring sites (red dots) and the study site where a water quality sensor is installed (rectangle).

Table 1. Types and ranges of water quality related data.

Parameters	Type	Range	Accuracy
Water temperature	Platinum resistance thermometer	-5~50 *C*	±0.1 *C*
Turbidity	Nephelonmeter, 90 light scattering 860 nm LED	0~2,000 NTU	2 NTU
Dissolved oxygen	Optical fluorescence quenching	0~ 20 mg/L	±0.1 mg/L
Conductivity	4-cell	5 to 2,000 µS/cm	2 µS/cm

A water quality sensor (Troll 9500 model, In-situ Inc. USA) was installed at the foot of a bridge (Yangpyung Bridge) in the Anyang Stream to monitor variations in water quality (Fig. 1). Using the Troll 9500 model, temperature, dissolved oxygen (DO), electric conductivity, and turbidity data were collected at 15-minute interval during the period 2010-2012 year. The sensors were cleaned and maintained every one or two weeks to remove biofouling and for calibration. Table 1 summarizes the types and ranges of water-quality-related data collated using the Troll 9500 model.

Together with the water quality data, we also considered the meteorological data provided by the Korean Meteorological Administration. Meteorological data included air temperature, wind speed, solar radiation, cloud, and precipitation at one-hour interval (Table 2). The water levels of the Han

Table 2. Types and ranges of weather-related data

Parameters	Range	Average	Standard deviation
Air Temperature	-17.7~33.7 *C*	10.86 *C*	±11.3 *C*
Wind speed	0~13 m/s	2.76 m/s	±1.48 m/s
Radiation	0~398 mj/m^2	94.23 mj/m^2	±87.9 mj/m^2
Precipitation	0.68~8.4 m	2.72 mm	±13.94 mm
Water level	0.68~8.4 m	2.23 m	±0.68 m
Cloud amount	0~10	5.52	±3.99

River were provided by the Han River Flood Control Office and monitored every 10 min.

B. Input variables and data preprocessing

Hydrological and meteorological data measured from 2010 to 2012 were selected as input variables. Hydrologic data included water temperature, turbidity, DO, conductivity, water height, and flow rate (Table 1), and meteorological data included air temperature, air speed, solar radiation, cloud, and precipitation (Table 2). Time series of DO of form $X_{t-\tau}$, $X_{t-\tau+1}$, ..., X_t are also adopted to train prediction model. While it is natural to assume that there is correlation between X_t and $X_{t-\tau}$, it is important to determine the length of such sequence. If the length, which is determined by delay parameter-τ, is too short, the prediction model will lose accuracy. Conversely, if τ is too large, we may include meaningless information. Mutual information between X_t and $X_{t-\tau}$ is used to determine the proper length of sequence. The mutual information between the two variables X and Y is calculated as follows:

$$MI(X,Y) = \sum_{y\in Y}\sum_{x\in X} P_{(X,Y)}(x,y)\log\left(\frac{P_{(X,Y)}(x,y)}{P_X(x)P_Y(y)}\right)$$

where $P_{(x,y)}$ is the joint probability mass function of X and Y. We measure $MI(X_t, Xt_{-\tau})$ with increasing τ (Fig. 2).

Since $MI(X_t, X_{t-\tau})$ function tends to be flat at seven days, we choose τ as seven days.

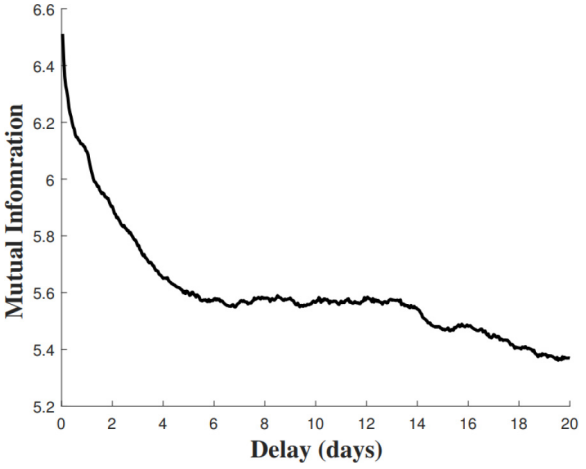

Fig. 2. Mutual information between X_t and $X_{t-\tau}$ with changing delay parameter τ.

For the ANN-XGB model, which is described in the next subsection, water quality data and a three-year meteorological dataset (7,368 samples × 11 variables) were divided into two sets. The first two years were selected as the training data and another one year as the test set. The training and test datasets have dimensions of 5,184 samples × 11 features and 2184 × 11 features, respectively. Because normalization is crucial, we normalize the DO variables between 0 and 1, that is,

$$X = \frac{X - X_{min}}{X_{max} - X_{min}}$$

C. Development of ANN-XGB for prediction of DO

In this subsection, we proposed ANN-XGB, which combines ANN and XG-Boosting to predict DO p-hour ahead. The XG-ANN model structure developed in this study is shown in Fig. 3. First, the ANN was trained using a time series of DO variables. Given a typical time t, the series of values $\{DO_{t-\tau\Delta T}, DO_{t-(\tau+1)\Delta T}, ..., DO_t\}$ is substituted into ANN to predict $\{DO_{t+\Delta T}, DO_{t+2\Delta T}, ..., DO_{t+p\Delta T}\}$. Here, time sampling rate and delay parameters are chosen as $\Delta T = 1$ h and τ =198. We report the performance of the ANN-XGB in the Results Section by varying p between 1 and 12 h. The sample data $(\boldsymbol{X}_j, \boldsymbol{Y}_j)$ are described as follows:

$$\boldsymbol{X}_j = \{DO_\ell \mid \ell = t^j - 198\Delta DeltaT, \ldots, t^j\},$$
$$\boldsymbol{Y}_j = \{DO_\ell \mid \ell = t^j + \Delta DeltaT, \ldots, t^j + p\Delta DeltaT\}.$$

The ANN is trained by standard supervised learning on dataset $(\boldsymbol{X}_j, \boldsymbol{Y}_j)$ which is generated by the measured DO from 2010 to 2012. We denote the ANN-predicted DO as $\{DO_{t^j+\Delta T}, \ldots, DO_{t^j+p\Delta T}\}$. However, this naive version of ANN cannot capture sudden changes in weather-related data. Therefore, we employed XG-Boost, which accompanies the hydrological and meteorological data at t^j, to enhance the predictions made by the ANN. We include hydrological and meteorological data together with $\{DO_{t^j+\Delta T}, \ldots, DO_{t^j+p\Delta T}\}$ when training XG-Boost model (see Fig. 3), which results in final prediction of DO. In the Results Section, we describe the XGB and ANN-related parameters used in ANN-XGB.

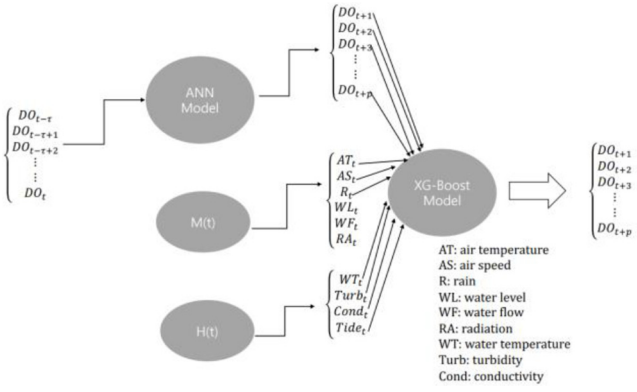

Fig. 3. Structure of XG-ANN model.

III. Results

In this section, the performance of ANN-XGB for predicting $DO_{t+p\Delta T}$ are presented.

We now describe the ANN- and XGB-related parameters. We experimentally determined the hyper-parameters in the ANN. Interestingly, with a small number of parameters, the ANN accurately predicted the DO variables. Therefore, instead of using complex models, such as LSTM or transformers, we chose a simple version of the ANN and used two hidden layers with a sigmoid function for activation, where the first hidden layer had 80 nodes and the second layer had 40 nodes. Next, for XG-boost, we chose a learning rate of 0.01. The number of estimators (trees) used in the model was. 400. To prevent overfitting, the subsample was set to 0.75 and the maximum depth was set to 12.

The model was trained on the period (2010. 3~2012. 1.2) and tested on the period (2012. 1.4~2012. 2.7) and (2012. 2.11~2013. 3.12), and (2012. 3. 14~2012. 4. 7).

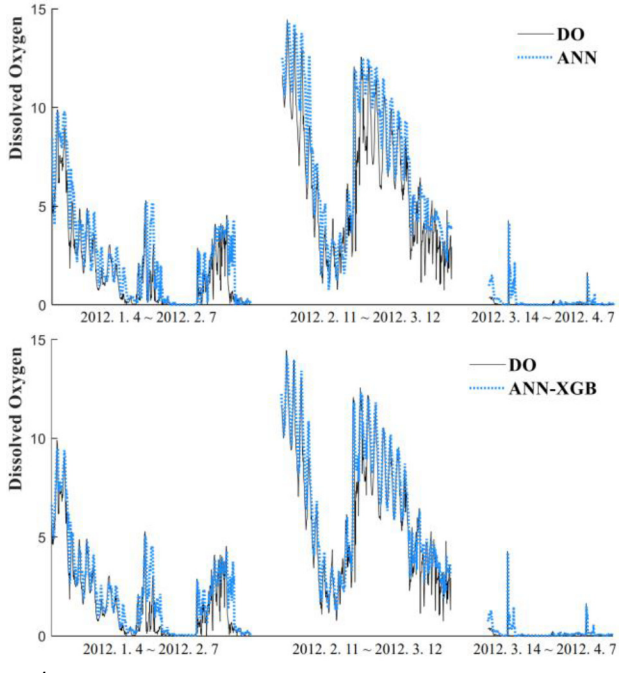

Fig. 4. Prediction of 1-hour ahead DO by ANN (top) and ANN-XGB (bottom).

We described several errors in the evaluation of the prediction model. First, the R^2 represents the percentage of variability that can be predicted by the model and is calculated as:

$$\mathrm{R}^2 = \frac{[N\sum_{i=1}^{N} y_i\tilde{y}_i - (\sum_{i=1}^{N} y_i)^2(\sum_{i=1}^{N} \tilde{y}_i)^2]^2}{[N\sum_{i=1}^{N} y_i^2 - (\sum_{i=1}^{N} y_i)^2][N\sum_{i=1}^{N} \tilde{y}_i^2 - (\sum_{i=1}^{N} \tilde{y}_i)^2]}$$

where $\tilde{y}_i$ and y_i are the predicted and target values, respectively, and N is the number of samples. The Nash-Sutcliffe efficiency coefficient (NSE) was used:

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N}(y_i - \tilde{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2}$$

where, $\bar{y}_i$ is the mean value of $\{y_i\}$. The root mean square error (RMSE) and mean absolute error (MAE) were also used.

Let us first compare the prediction results of 1-hour ahead DO by ANN and ANN-XGB (see Fig. 4). Overall, both the predictions by the ANN and ANN-XGB represent the actual shape of the DO during the test period. However, the predictions made by the ANN-XGB were closer to the actual DO than those made by the ANN.

Now, we compare the ANN-XGB with the ANN by varying p from 1 to 12 has shown in Table 3. ANN-XGB tended to outperform ANN when $p < 3$. However, the performance of the two models was similar. This suggests that ANN-XGB is preferable for short-time prediction.

Table 3. Accuracy of prediction of $DO_{t+p\Delta T}$ in terms of R^2, NSE, RMSE and MAE

p	R^2		NSE		RMSE		MAE	
	ANN	ANN-XGB	ANN	ANN-XGB	ANN	ANN-XGB	ANN	ANN-XGB
1	0.88	0.94	0.84	0.94	1.35	0.89	0.88	0.55
2	0.87	0.90	0.83	0.89	1.38	1.12	0.90	0.72
3	0.88	0.87	0.87	0.85	1.24	1.32	0.80	0.86
4	0.87	0.86	0.85	0.84	1.31	1.37	0.84	0.89
5	0.85	0.85	0.85	0.82	1.32	1.41	0.85	0.93
6	0.84	0.85	0.84	0.82	1.35	1.41	0.86	0.92
7	0.82	0.83	0.83	0.81	1.39	1.46	0.89	0.96
8	0.80	0.82	0.82	0.81	1.41	1.47	0.91	0.96
9	0.79	0.82	0.82	0.80	1.43	1.50	0.92	0.97
10	0.78	0.82	0.82	0.80	1.44	1.51	0.91	0.97
11	0.78	0.82	0.81	0.80	1.45	1.50	0.92	0.96
12	0.79	0.81	0.81	0.80	1.45	1.49	0.93	0.96

One of the advantages of the proposed scheme is its capability for correlation analysis of the target and feature variables via the feature importance score given by the XGB algorithm. For example, *weight* type feature importance was obtained by counting the number of features appearing in tree branches. Thus, a high feature importance score for a certain variable implies a strong influence on the target variable. Fig. 5 shows the *weight* type importance scores for the ANN-XGB model. The tidal level, water temperature, and turbidity have high scores, implying that they are correlated with changes in DO variable.

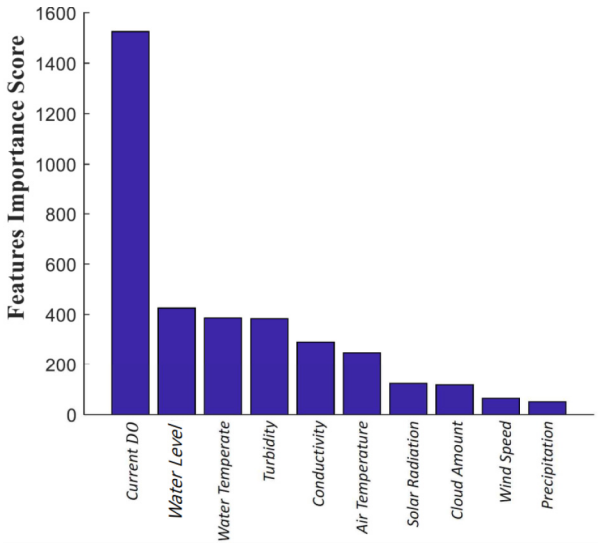

Fig. 5. Feature importance scores obtained by XGB.

IV. CONCLUSION

We proposed hybrid models combining ANN and XGboost to predict DO in the Anyang Stream. First, an ANN model was developed to predict the DO, where the hyperparameters were tuned optimally. Next, XGB was adopted in the second stage, where the predictions of the first stage were used as input features together with hydrologic and meteorological data. The R^2 and NSE coefficients are approximately 0.94 for the XGB-ANN, outperforming the ANN with R^2 and NSE coefficients of 0.88 and 0.84. We also document the feature importance scores obtained using XGB. The results indicate that tidal level and water temperature are correlated with changes in DO variable.

A limitation of the proposed algorithm is that our methods are model-free. Therefore, the accuracy of the proposed algorithm relies heavily on the range of the training period, because it is more likely to occur during various types of meteorological events during the training period. Combining ecosystem models [6-8] might enhance the generalization capability of our algorithm.

In future work, we will develop prediction models for other water quality related parameters, such as water temperature and turbidity. Additionally, to validate the proposed algorithms, downstream data collected at other sites should be considered in future studies.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) (No. 2020R1C1C1A01005396).

REFERENCES

[1] K. P. Singh, A. Basant, A. Malik, and G. Jain, "Artificial neural network modeling of the river water quality—a case study," *Ecological Modelling*, vol. 220, no. 6, pp. 888-895, Mar. 2009. DOI: 10.1016/j.ecolmodel.2009.01.004.

[2] H. J. Wu, Z. Y. Lin, and S. L. Gao, "The application of artificial neural networks in the resources and environment," *Resources and Environment in the Yangtze Basin*, vol. 9, no. 2, pp. 241-246, 2000. DOI: 10.1007/s11434-010-4183-3.

[3] S. Xiang, Z. Liu, and L. Ma, "Study of multivariate linear regression analysis model for ground water quality prediction," *Environmental Science*, vol. 24, no. 1, pp. 60-62, 2006.

[4] M. I. Hejazi, X. Cai, and B. L. Ruddell, "The role of hydrologic information in reservoir operation--learning from historical releases," *Advances in water resources*, vol. 31, no. 12, pp. 1636-1650, Dec. 2008. DOI: 10.1016/j.advwatres.2008.07.013.

[5] J. Y. Lee, K. Y. Lee, S. Lee, J. Choi, S. J. Lee, S. Jung, M. S. Jung, and B. Kim, "Recovery of fish community and water quality in streams where fish kills have occurred," *Korean Journal of Ecology and Environment*, vol. 46 no. 2, pp. 154-165, Jun. 2013. DOI: 10.11614/KSL.2013.46.2.154.

[6] J. E. Nash, and J. V. Sutcliffe, "River flow forecasting through conceptual models part i—a discussion of principles," *Journal of hydrology*, vol. 10, pp. 282-290, Apr. 1970. DOI: 10.1016/0022-1694(70)90255-6.

[7] M. A. Pena, S. Katsev, T. Oguz, and D. Gilbert, "Modeling dissolved oxygen dynamics and hypoxia," *Biogeosciences*, vol. 7, no. 3, pp. 933-957. Mar. 2010. DOI: 10.5194/bg-7-933-2010.

[8] C.-Y. Liaw, N. Islam, K. K. Phoon, S. Y. Liong, "Comment on does the river run wild? assessing chaos in hydrological systems," *Advances in water resources*, vol. 24, no. 5, pp. 575-580, 2001. DOI: 10.1016/S0309-1708(00)00053-1.

[9] V. Z. Antonopoulos and S. K. Gianniou, "Simulation of water temperature and dissolved oxygen distribution in lake vegoritis, Greece," *Ecological Modelling*, vol. 160, no. 1-2, pp. 39-53, Feb. 2003. DOI: 10.1016/S0304-3800(02)00286-7.

[10] S. Lek, and J. F. Guegan, "Artificial neural networks as a tool in ecological modelling, an introduction," *Ecological modelling*, vol. 120, no. 2-3, pp. 65-73, Aug. 1999. DOI: 10.1016/S0304-3800(99)00092-7.

[11] J. Bowers, and C. Shedrow, "Predicting stream water quality using artificial neural networks (ANN)," *WIT Transactions on Ecology and the Environment*, vol. 41, 2000. DOI: 10.2495/ENV000081.

[12] Y. M. Kuo, C. W. Liu, and K. H. Lin, "Evaluation of the ability of an artificial neural network model to assess the variation of groundwater quality in an area of blackfoot disease in Taiwan," *Water research*, vol. 38, no. 1, pp. 148-158, Jan. 2004. DOI: 10.1016/j.watres.2003.09.026.

[13] G. Sahoo, S. Schladow, and J. Reuter, "Forecasting stream water temperature using regression analysis, artificial neural network, and chaotic non-linear dynamic models," *Journal of hydrology*, vol. 378, no. 3-4, pp. 25-342, Nov. 2009. DOI: j.jhydrol.2009.09.037.

[14] S. Palani, S. Y. Liong, and P. Tkalich, "An ANN application for water quality forecasting," *Marine pollution bulletin*, vol. 56, no. 9, pp. 15861597, Sep. 2008. DOI: 10.1016/j.marpolbul.2008.05.021.

[15] V. Nourani, M. Komasi, and A. Mano, "A multivariate ANN-wavelet approach for rainfall--runoff modeling," *Water resources management*, vol. 23, pp. 2877-2894, Feb. 2009. DOI: 10.1007/s11269-009-9414-5.

[16] A. Kavousi-Fard, "A new fuzzy-based feature selection and hybrid TLA-ANN modelling for short-term load forecasting," *Journal of*

Experimental & Theoretical Artificial Intelligence, vol. 25, no. 4, pp. 543-557. May 2013. DOI: 10.1080/0952813X.2013.782350.

[17] M. K. Jha, and S. Sahoo, "Efficacy of neural network and genetic algorithm techniques in simulating spatio-temporal fluctuations of groundwater," *Hydrological processes*, vol. 29, no. 5, pp. 671-691, Feb. 2015. DOI: 10.1002/hyp.10166.

[18] M. Ravansalar, T. Rajaee, and M. Ergil, "Prediction of dissolved oxygen in river calder by noise elimination time series using wavelet transform," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 28, no. 4, pp. 689-706, May 2015.DOI: 10.1080/0952813X.2015.1042531.

[19] T. Chen, and C. Guestrin, "Xgboost: A scalable tree boosting system," *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, ACM*, pp. 785-794, Aug. 2016. DOI: 10.1145/2939672.2939785.

[20] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *The Annals of statistics*, vol. 29, no. 5, pp. 1189-1232, Oct. 2001.

[21] M. Gumus, and M. S. Kiran, "Crude oil price forecasting using XGboost," in *2017 International Conference on Computer Science and Engineering (UBMK),* Antalya, TR, pp. 1100-1103, 2017. DOI: 10.1109/UBMK.2017.8093500.

[22] L. Zhang, and C. Zhan, "Machine learning in rock facies classification: an application of XGboost," in *International Geophysical Conference,* Qingdao, CN, pp. 1371-1374, 2017. DOI: 10.1190/IGC2017-351.

[23] Z. Chen, F. Jiang, Y. Cheng, X. Gu, W. Liu, and J. Peng, "Xgboost classifier for DDoS attack detection and analysis in SDN-based cloud," in *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)*, Shanghai, CN, pp. 251-256, 2018. DOI: 10.1109/BigComp.2018.00044.

[24] H. Zheng, J. Yuan, and L. Chen, "Short-term load forecasting using EMD-LSTM neural networks with a XGboost algorithm for feature importance evaluation," *Energies*, vol. 10, no. 18, Aug. 2017. DOI: 10.3390/en10081168.

Keun Young Lee

received his Ph. D. degree from the Department of Mathematical Science, KAIST, in 2009. From 2017 to 2020, he was a faculty member of the Department of Mathematics at Sejong University, Republic of Korea. From 2020 to the present, he has been an independent scholar in Republic of Korea. His research interests include Banach space theory, machine learning, and fuzzy theory.

Bomchul Kim

received his B.S. degree in oceanography from Seoul University in 1977, received his M. S. in bio-engineering from KAIST in 1980, and received a Ph. D. in oceanography from Seoul University in 1987. He has been a faculty member of the Department of Environmental Science at Kangwon National University since 1981. Currently, he is a professor emeritus at Kangwon National University.

Gwanghyun Jo

received his M.S. and Ph. D. degree in the Department of Mathematical Science, KAIST, in 2013 and 2018, respectively. From 2019 to 2023, he was a faculty member of the Department of Mathematics at Kunsan University, Republic of Korea. From 2023 to the present, he has been a faculty member of the Department of Mathematical Data Science, Hanyang University, ERICA. His research interests include numerical analysis, computational fluid dynamics, and machine learning.