









Development of a short-term water quality prediction model for urban rivers using real-time water quality data

J. H. Lee ^a, J. Y. Lee^a, M. H. Lee ^a, M. Y. Lee ^a, Y. W. Kim ^b, J. S. Hyung ^c, K. B. Kim ^c, Y. K. Cha ^b and J. Y. Koo ^{b,*}

^a Department of Water Environment Research, Seoul Metropolitan Government Research Institute of Public Health and Environment, Seoul 13818, Korea

^b School of Environmental Engineering, University of Seoul, Seoul 02504, Korea

^c Construction Engineering and Management, Purdue University, Indiana, USA

*Corresponding author. E-mail: jykoo@uos.ac.kr

 JHL, 0000-0002-2989-1682; MHL, 0000-0001-8894-9605; MYL, 0000-0001-7227-5415; YWK, 0000-0001-7123-6190; JSH, 0000-0002-3957-9385; KBK, 0000-0002-3957-9385; YKC, 0000-0001-9638-9476; JYK, 0000-0001-8313-3033

ABSTRACT

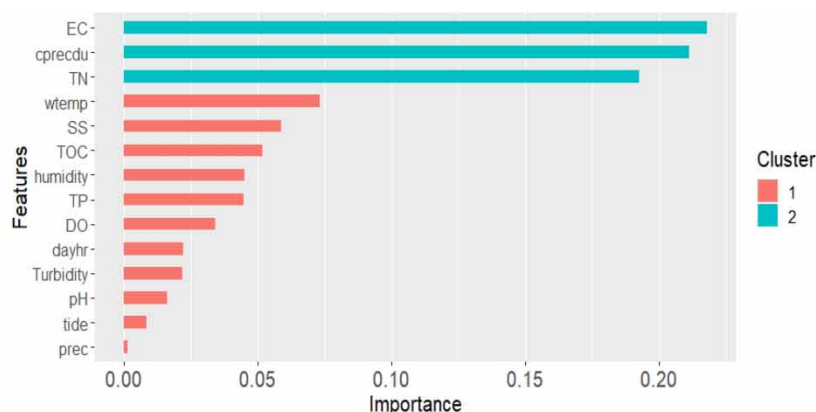
We developed a classification model and a real-time prediction model for short-term dissolved oxygen (DO) at the junction of the Han River in Anyangcheon, where water quality accidents occur frequently. The classification model is an analysis model that derives the main factors affecting DO changes in the Anyangcheon mobile water quality monitoring network using decision tree, random forest, and XGBoost. The model identified the key factors affecting DO changes to be electrical conductivity, cumulative precipitation, total nitrogen, and water temperature. Random forest (sensitivity, 0.9962; accuracy, 0.9981) and XGBoost (sensitivity, 1.0000; accuracy, 0.9822) showed excellent classification performance. The real-time prediction model for short-term DO that we developed adopted artificial neural network (ANN), long short-term memory (LSTM), and gated recurrent unit (GRU) algorithms. LSTM ($R^2 = 0.93 - 0.97$, first half; $R^2 = 0.95 - 0.96$, second half) and GRU ($R^2 = 0.94 - 0.98$, first half; $R^2 = 0.96 - 0.98$, second half) significantly outperformed ANN ($R^2 = 0.64 - 0.86$). The LSTM and GRU models we developed used real-time automatic measurement data, targeting urban rivers that are sensitive to water quality changes and are waterfront areas for citizens. They can quickly reflect and simulate short-term, real-time changes in water quality compared with existing static models.

Key words: classification model, dissolved oxygen prediction model, real-time automatic measurement data, urban river, water quality accident

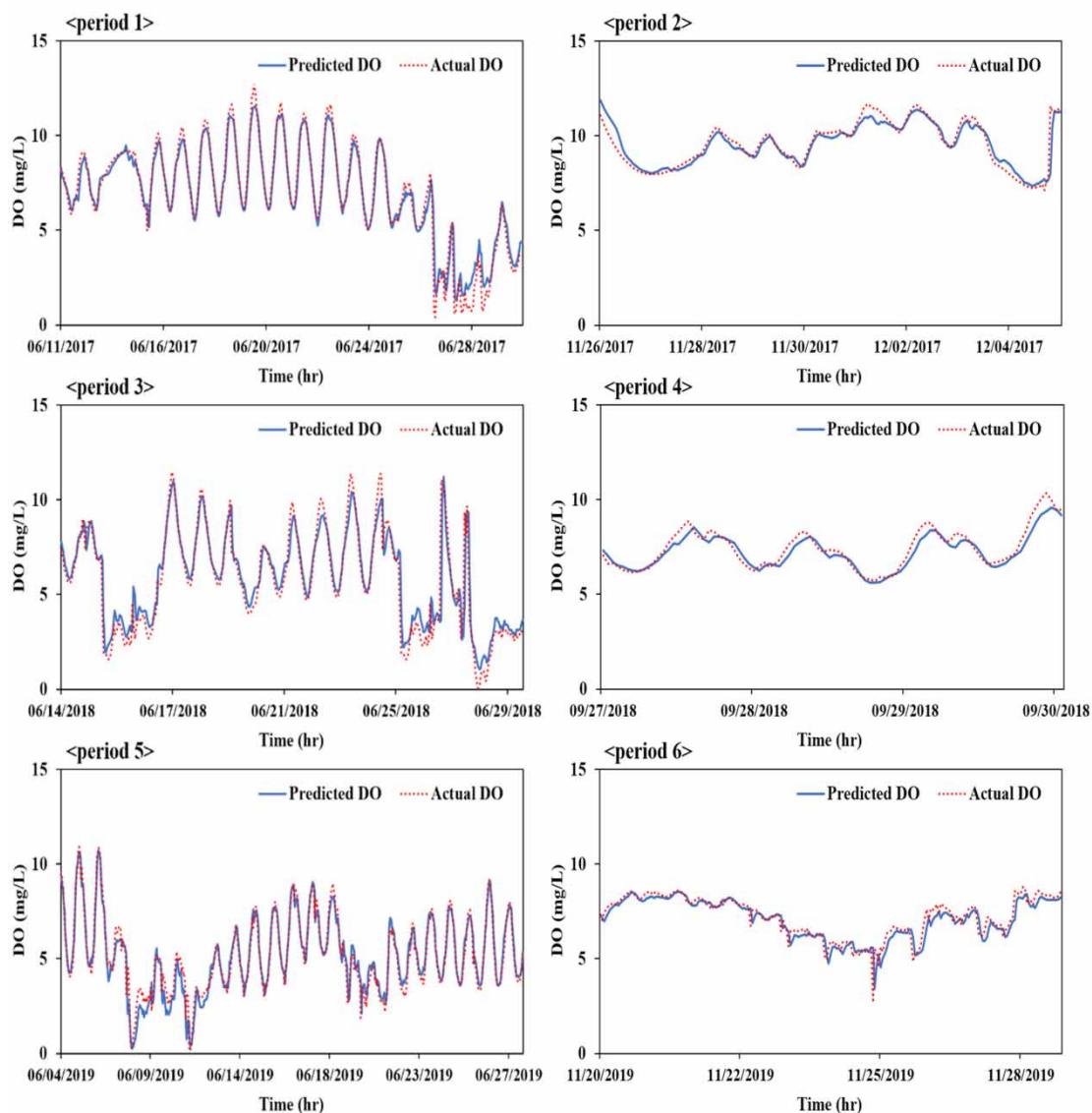
HIGHLIGHTS

- We developed a classification model and a real-time prediction model on urban rivers.
- The classification model identified the key factors affecting DO changes.
- The LSTM and GRU models can quickly reflect and simulate short-term, real-time changes in water quality.
- The dissolved oxygen water quality prediction model developed in this study is an ensemble model grafted with a classification model.

GRAPHICAL ABSTRACT



Main features affecting DO changes of Anyangcheon mobile WQMN



Predictive performance results of LSTM model for test data.

1. INTRODUCTION

Urban rivers are streams that run through the center of the city and serve as a passageway for river water, forming an element of the urban landscape and serving as an ecological space in which various organisms live. Since urban rivers flow into the main stream, which is an important water resource management target, efficient water quality management of urban rivers is of utmost importance. Recently, water-friendly activities in major urban rivers have increased rapidly, and accordingly, citizens' expectations for the water quality of urban rivers have risen. However, owing to global climate change, the impacts of rapid urbanization and industrialization have progressed in severity, especially in the case of Korea, which is seeing the drying of the urban environment and rivers. Other impacts include urban flooding and the increase in non-point pollution flowing into urban rivers, and lack of dissolved oxygen (DO) in urban rivers. Consequently, the water quality, health, and stability of aquatic ecosystems are gradually deteriorating (Cho & Lee 2018). To secure their ecological environment and provide pleasant hydrophilic activities, cities should prioritize the advanced management of urban rivers, such as via real-time water quality monitoring and prediction of water quality changes.

In 1975, the Noryangjin automatic water quality monitoring network (WQMN) was first installed in the main stream of the Han River in Seoul. In the intervening decades, the Seonyu automatic WQMN was installed in the main Han River, and three automatic WQMN were installed in Anyangcheon, Jungnangcheon, and Tancheon, which are major urban rivers. Through five automatic WQMN, water quality of the main stream of Han River and major urban rivers is constantly measured to support the efforts to prevent water pollution in public waters. In addition, many damages caused by water pollution accidents can be minimized by early detection of water pollution accidents by using WQMN. At the same time, they are actively used as a tool for Seoul's water quality conservation policy by grasping the status of water pollution through real-time water quality data from five automatic WQMN. As measured parameters, 17 items (water temperature, pH, DO, conductivity, turbidity, suspended solids, total nitrogen, total phosphorus, total organic carbon, CN, Alkyl Benzene Sulfonate (ABS), Phenols, chlorophyll-a, *Daphnia magna*, fish, algae, microbes) for the main Han River and 14 items (water temperature, pH, DO, conductivity, turbidity, suspended solids, total nitrogen, total phosphorus, total organic carbon, CN, Alkyl Benzene Sulfonate (ABS), Phenols, Hg, *Daphnia magna*) for urban rivers, are being measured in real time (Lee *et al.* 2018). Since 2017, the mobile WQMN has been operating in urban rivers where water quality accidents frequently occur. This network allows real-time water quality measurement using ICT and IoT to monitor sudden changes in water quality and inflow of harmful substances, helping the authorities cope with water pollution accidents. However, despite these efforts, dozens of large and small unexpected water quality accidents continue to occur every year, such as green algal bloom and fish kill in the Han River and major urban rivers in Seoul. In the case of accidents according to natural types, such as initial rainfall causing a lack of DO that results in fish kill, the initial response tends to be delayed, and consequently, the evidence disappears. Most of the response has been in the form of follow-up investigation. In addition, complex factors limit the capability to identify the exact cause and prepare the appropriate countermeasures. Related studies and application cases are also insufficient for urban rivers in domestic cases.

Therefore, in this study, we used decision tree (Kass 1980; Fournier & Cremilleux 2002), random forest (Breiman 2001; Krzysztof *et al.* 2019), and eXtreme Gradient Boosting (XGBoost) (Friedman 2001) classification models (Guidotti *et al.* 2018) using ICT and IoT-based water quality big data. This is because the decision tree model is intuitive and it is easy to understand analysis process, random forest model with multiple decision trees is possible to solve the overfitting problem and produce stable prediction results and XGBoost model has high accuracy and at the same time, analysis speed is very fast, and artificial neural network (ANN) (Fitore *et al.* 2019), long short-term memory (LSTM) (Hochreiter & Schmidhuber 1997), and gated recurrent unit (GRU) (Giuseppe & Balaji 2017) algorithms to develop a real-time short-term DO prediction model (Avi & Shani 2005; Durdu 2009; Ruiz-Aguilar *et al.* 2014) that can predict even non-linear changes in DO, a representative water quality factor closely related to water quality accidents using time series data (Guzman *et al.* 2017; Chen *et al.* 2018) for urban rivers. In the case of a real-time DO prediction model, it is difficult to model short-period data and nonlinear relationships with only the past traditional time series models. LSTM and GRU recurrent neural network models developed to solve the vanishing gradient problem were used.

Recently, classification models have been applied in the environmental field such as machine learning-based algae forecasting system (Muttill & Chau 2006), and prediction of air pollutant concentrations using time series data (Snezhana *et al.* 2019). At the same time, they are being used in various fields such as emotion classification using bio-signals and pattern information (Murugappan *et al.* 2013) and vehicle classification based on image features such as setting a region of interest (Kim & Kang

2017). Similarly, prediction models based on machine learning and deep learning also have high accuracy and can predict even non-linear changes in the target variable, for example, monthly NASDAQ index prediction (Sima & Akbar 2018) and rice cultivation prediction (Kiran & Arjun 2019) using LSTM, voice signal data analysis using GRU (Chung *et al.* 2014), DO concentration prediction in the downstream area of Yamuna river through various input data such as flow rate, pH, BOD, etc., (Sarkar & Pandeym 2015) and DO and TP concentration prediction using a modified LSTM model for Lake Taihu (Wang *et al.* 2017). They are widely used in various fields along with classification models.

2. METHODS

2.1. Materials

In this study, to develop a classification model that could derive the main factors affecting DO changes in the Anyangcheon mobile WQMN and a real-time DO prediction model, as shown in Figure 1, we selected as the research target area the area with frequent water quality accidents within the Anyangcheon water system. This area is characterized by the frequent occurrence of fish kill and availability of data related to water quality.

We used the following as key input data: water quality data from the Anyangcheon automatic and mobile WQMN; hydrological data from the Han River Flood Control Office; weather data from the Korea Meteorological Administration; tide data from the Korea Hydrographic and Oceanographic Agency for 2017–2019, consisting of 16 items, namely, water temperature (wtemp), pH, electrical conductivity (EC), DO, turbidity, suspended solids (SS), total nitrogen (TN), total phosphorus (TP), total organic carbon (TOC), water level, flow rate, tide level (tide), precipitation (prec), cumulative precipitation (cprecdu), humidity, and sunlight (dayhr). All were time unit data. Tide level variable among input variables is data measured at Ganghwa bridge, a point near the Anyangcheon downstream area because it is expected to affect DO changes in the downstream area of Anyangcheon. In addition, in order to improve prediction power of the model by reflecting weather conditions of study area into the model, weather data such as cumulative precipitation from the Korea Meteorological Administration

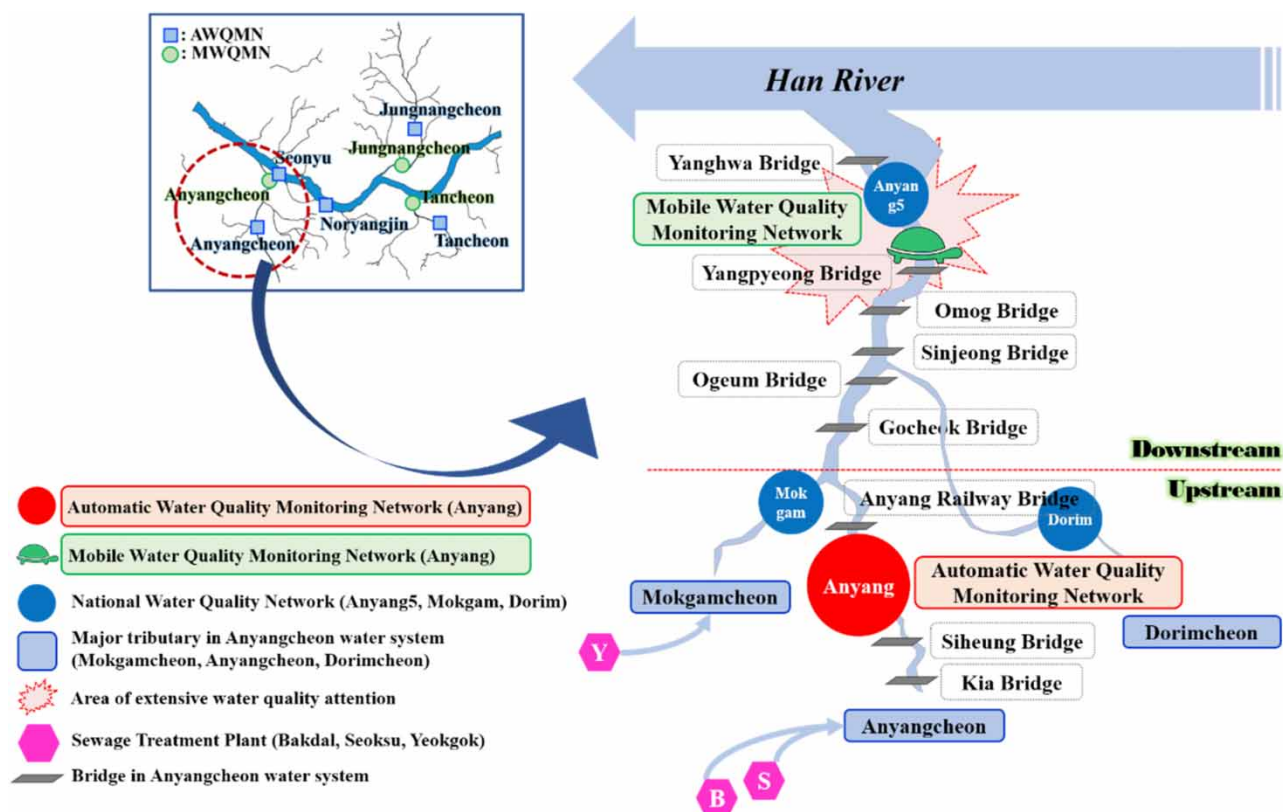


Figure 1 | Study area for classification and DO prediction model derivation.

(Seoul branch) were added as input variables. The type of dependent variable mobile WQMN DO varied according to the model. In the classification model, it was a categorical variable divided into category 0 (DO of 3 mg/L or higher, criterion fit) and category 1 (DO of less than 3 mg/L, criterion nonconforming). In the real-time DO prediction model, it was applied as a numeric variable. In general, a categorical variable refers to data classified by a category like a specific class, and a numeric variable refers to data that has a meaning by itself and is used as a statistical indicator.

As shown in Table 1, the data used as input and dependent variables for model development are only those within the period of March to June (first half, Periods 1, 3, 5) and September to December (second half, Periods 2, 4, 6). When missing values occurred owing to equipment calibration, data error, repairs, power cuts, and other reasons, we used the Kalman filter method for the input and dependent variable values.

2.2. Classification model development

Starting with statistical classification theory and image analysis provided by Duda & Hart (1973), experts have developed classification models of various algorithms. The main algorithms include a rule-based decision tree that can understand classification and prediction results, as well as the variables that affect these results. Other examples are logistic regression, which interprets the result value between 0 and 1 as a probability value based on the linear relation between the input variable and the dependent variable; support vector machine, which classifies between individual groups through simple and efficient calculation using a maximum margin classifier; naive Bayes statistics; and generative models. One of the important features of the classification model is that the variable importance can be checked through each classification algorithm. Variable importance in the classification model refers to the relative importance of the variables that affect the classification result.

We used R-3.6.1 and Python 3.7 to develop the classification model. Data pre-processing involved determining the optimal time lag using CART algorithm (algorithm that recursively divides data in a direction that minimizes data impurity) (Breiman *et al.* 1984), categorizing the dependent variable (0, 1), improving data imbalance using Upsampling, and running 10-fold cross validation (80% training data and 20% test data). Next, we derived the optimal classification model by evaluating classification model performance using the decision tree, random forest, and XGBoost classification model algorithm and confusion matrix. Accuracy, sensitivity, and specificity were used as performance evaluation indices, shown in Equations (1)–(3). Larger index values indicated better predictive performance:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (2)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (3)$$

where TP (true positive) is the frequency at which both the actual value and predicted value are true, TN (true negative) is the frequency at which both the actual value and the predicted value are false, FP (false positive) is the frequency at which the actual value is false but predicted as true, and FN (false negative) is the frequency at which the actual value is true but predicted as false.

Table 1 | Data period of Anyangcheon mobile water quality monitoring network (WQMN) for model development

		Data period					
		2017		2018		2019	
		Period 1 Mar 27–Jun 30 (n ^a = 2,282)	Period 2 Oct 18–Dec 5 (n = 1,156)	Period 3 Apr 11–Jun 29 (n = 1,897)	Period 4 Sep 10–30 (n = 487)	Period 5 Mar 1–Jun 28 (n = 2,869)	Period 6 Oct 11–Nov 29 (n = 1,179)
Anyangcheon Mobile WQMN DO + Other input variables							
Data partitioning	Train (80%)	Mar 27–Jun 11	Oct 18–Nov 25	Apr 11–Jun 13	Sep 10–26	Mar 1–Jun 4	Oct 11–Nov 19
	Test (20%)	Jun 12–30	Nov 26–Dec 5	Jun 14–29	Sep 27–30	Jun 5–28	Nov 20–29

^an = number of data points for each period.

2.3. Real-time dissolved oxygen prediction model development

Recent efforts to develop water quality prediction models have used time series analysis of large sized data and non-normal data distribution. Existing time series modeling methods are limited in terms of the consideration of accurate prediction, trend reflection, and interaction between variables. These limits have given rise to the development of models based on machine learning, such as ANN, which is a non-parametric methodology, and recurrent neural networks (RNNs), in which the concept of time series is substituted for ANN. Machine learning and deep learning-based models, such as LSTM and GRU, which are recurrent neural network variants, have also emerged to compensate for the shortcomings of ANN and RNN models and to achieve more accurate prediction.

RNN is particularly limited by the vanishing gradient problem: as learning time increases, it becomes difficult to reflect information from the distant past in the present. In LSTM, a new node called a memory cell can be added to solve optimization errors, such as the vanishing gradient appearing in RNN when past information affects the present. LSTM is effective in capturing long-term temporal dependencies. Therefore, LSTM is an effective and scalable model for learning problems related to various time series data. It is widely used in various fields in combination with other neural networks, as well as in language modeling, image analysis, and speech recognition. GRU is one of the LSTM transformation models that more simply processes the structure of LSTM for overcoming the vanishing gradient problem.

In this study, we used R-3.6.1 (ANN) and Python 3.7 (LSTM, GRU) programs to develop a model for real-time DO prediction of the Anyangcheon mobile WQMN. First, we developed an ANN model, in which the main influencing factors of DO derived from the classification model and DO of the Anyangcheon automatic WQMN were selected as input variables. Hyperparameters, such as hidden layer, hidden node, activation function, and learning algorithm, were determined through a trial and error method along with a 24-hour time difference between input and dependent variables (Stephan & Lucila 2002). We applied root mean square error (RMSE), mean absolute error (MAE), and R^2 (coefficient of determination) as evaluation indices. We also considered volatility (%) (volatility relative to average value), considering that a low RMSE does not necessarily give a high accuracy model – average values must be considered along with the RMSE. Second, we created the LSTM and GRU models using all variables as input variables. The optimization technique was applied to each half year (the first and second halves). Hyperparameters were divided into nine detailed items, including network structure, loss function, activation function, learning rate, optimization algorithm, and dropout probability, and these values were selected based on the literature (Chen *et al.* 2018). In addition, we considered a time delay (1–6 h) for the LSTM and GRU input variables to reflect the optimal flow time derived from the classification model results as well as the 24-hour time difference between the input and dependent variables. The evaluation indices were RMSE, R^2 , and volatility (%).

3. RESULTS AND DISCUSSION

3.1. Classification model development results

3.1.1. Decision tree model

The time lag between the input and dependent variables was changed from 2 to 8 h based on the physical range of flow time between Anyangcheon's automatic and mobile WQMN through the decision tree model. As a result, as shown in Table 2, the optimal time lag was determined to be 6 h, which showed the best accuracy and sensitivity. We implemented pruning using

Table 2 | Matrix according to time lag using decision tree model

Time lag	Sensitivity	Accuracy
2	0.7643	0.8588
3	0.8154	0.8458
4	0.7831	0.8501
5	0.7820	0.8548
6	0.8144	0.8599
7	0.7540	0.8409
8	0.7593	0.8381

the complexity parameter (cp) when the error rate was the lowest; the following results were obtained: sensitivity was 0.8144, accuracy was 0.8599, and specificity was 0.9049. Overall, categories 0 and 1 were classified properly.

In general, in the decision tree model, separation occurs first in the variable with the most explanatory power. Therefore, as shown in Figure 2, the main factors affecting DO changes in the Anyangcheon mobile WQMN derived from the decision tree model included cumulative precipitation, electrical conductivity, and TN. In addition, looking at Figure 2, the result of the decision tree model, such as numbers, colors, and ratios (%) in the figures, is visually expressed, where 'Yes' is category 0 (DO of 3 mg/L or higher, criterion fit), 'No' is category 1 (DO of less than 3 mg/L, criterion nonconforming), and the middle number means the proportion of 'No' in the corresponding node, based on 0.5, if it is greater than 0.5, the figures are classified as 'No', and if it is less than 0.5, the figures are classified as 'Yes'. The last number indicates the percentage (%) of the node data out of the total data. By synthesizing these results, we could easily check under what conditions the factors affecting DO changes were classified, where and in what proportion each data belonged. In addition, the contents of Figure 2 are summarized and presented in Table 3, and through this we confirmed that DO at the junction (near the end of Anyangcheon) of the Han River in Anyangcheon did not meet Seoul's automatic WQMN self-monitoring standards for tributaries (unsuitable) under five conditions. Therefore, to manage the water quality of the Anyangcheon downstream area, the authorities need to check the various environmental factors, such as water quality of the effluent from nearby sewage treatment plants, influence of tides, and status of river maintenance water. In particular, various treatment facilities need to be installed and operated to reduce the outflow of non-point pollutants brought by the initial rain during rainfall.

3.1.2. Random forest model

When setting each node, we set the number of variables to be included in the tree (mtry) and minimum number of nodes (node size) to 2 and 5, respectively, whereas the number of trees was set to 100. Thus, we constructed a random forest model with the highest accuracy and lowest OOB (out of bags) error rate. As a result of random forest model execution, classification of category 0 and category 1 was performed accurately, with a sensitivity of 0.9962, accuracy of 0.9981, and specificity of 1.000.

Through the random forest model (Figure 3), we derived the main factors affecting DO changes in the Anyangcheon mobile WQMN. Increasing mean decrease accuracy (MDA) and mean decrease Gini (MDG) values, two indicators measuring the importance of variables in the random forest model, indicated the increasing importance of variables.

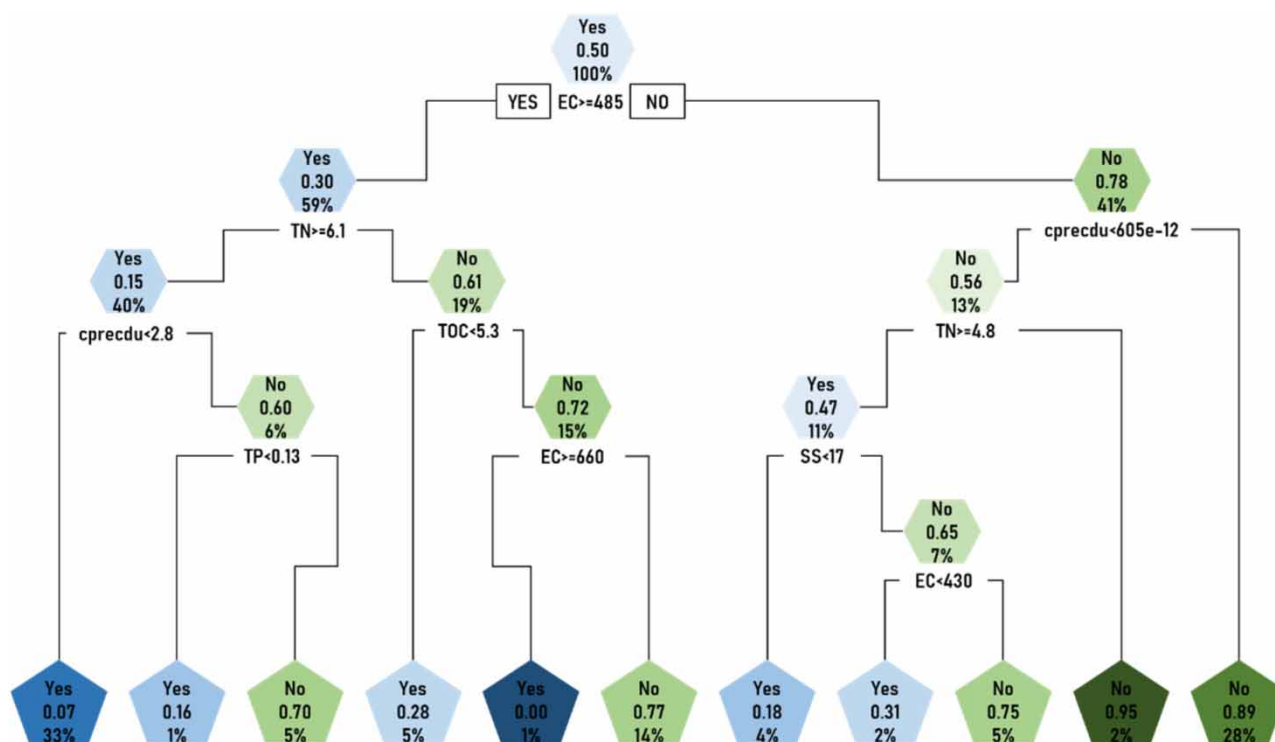
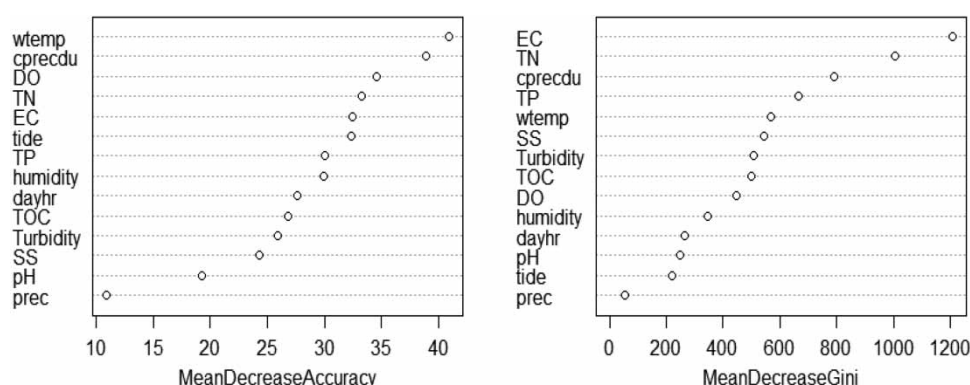


Figure 2 | Decision tree plot for DO classification.

Table 3 | Cases affecting the DO of Anyangcheon mobile WQMN

Case 1	EC ^a	Cumulative precipitation	TN	TOC	TP	SS
1	≥ 485	≥ 2.8	≥ 6.1	–	≥ 0.13	–
2	$485 \leq \text{EC} < 660$	–	< 6.1	≥ 5.3	–	–
3	$430 \leq \text{EC} < 485$	–	≥ 4.8	–	–	≥ 17
4	< 485	–	< 4.8	–	–	–
5	< 485	over 0	–	–	–	–

^aEC = electrical conductivity.**Figure 3** | Rank of factors affecting changes in DO of Anyangcheon mobile WQMN (wtemp: water temperature, cprecdu: cumulative precipitation, EC: electrical conductivity, tide: tide level, dayhr: sunlight, prec: precipitation).

Therefore, we confirmed that water temperature, EC, cumulative precipitation, and TN, which had high MDA and MDG values, had a major influence on DO changes in the downstream area of Anyangcheon. This result is consistent with those of previous studies (Senlin & Salim 2020; Yi *et al.* 2020) and the variable importance results in the aforementioned decision tree model.

3.1.3. XGBoost model

We built the XGBoost model by determining hyperparameters with 30,000 iterations, 10-fold cross validation, eval_metric mlogloss, early_stopping_rounds 10, max depth 5, gamma 3, and other functions. The XGBoost model showed good classification performance: the random forest model had a sensitivity of 1.0000, accuracy of 0.9822, specificity of 0.9650. Moreover, categories 0 and 1 were classified better compared with the decision tree model. As shown in Figure 4, the main factors affecting DO changes in the Anyangcheon mobile WQMN derived from the XGBoost model were EC, cumulative precipitation, TN, and water temperature. The results showed a marked similarity to other models.

As shown in Table 4, per the classification models, the main factors influencing the DO changes of the Anyangcheon mobile WQMN through the three types of classification models (i.e. decision tree, random forest, and XGBoost) were EC, cumulative precipitation, TN, and water temperature. All three models showed good performance, but random forest and XGBoost showed excellent classification performance.

Therefore, the authorities need to focus on water quality management based on DO at the junction of the Han River in Anyangcheon, paying particular attention to EC, water temperature, and TN, as well as pollutant management, given that the inflow of non-point pollutants into urban rivers increases during rainfall. In addition, if the main factors derived through these classification models are actively used as input variables in real-time DO prediction models for urban rivers, water quality abnormalities, such as water quality accidents, can be prevented. These models may enable cause analyses of water quality accidents.

3.2. Real-time dissolved oxygen prediction model development results

3.2.1. ANN model

After building the ANN model, we determined that the model with the best prediction performance was a model with two hidden layers and 12×12 hidden nodes. Table 5 shows the hyperparameters constituting the optimized ANN model. In

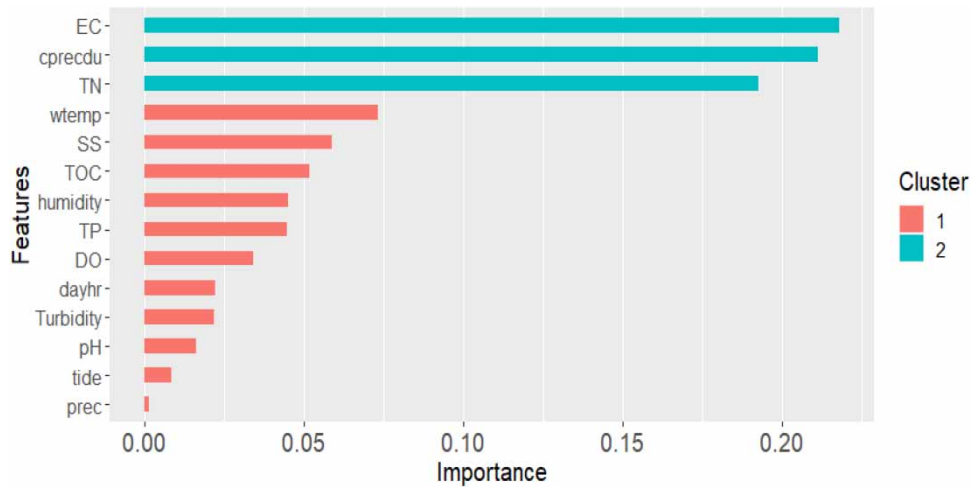


Figure 4 | Main features affecting DO changes of Anyangcheon mobile WQMN (wtemp: water temperature, cprecdu: cumulative precipitation, EC: electrical conductivity, tide: tide level, dayhr: sunlight, prec: precipitation).

Table 4 | Variable importance in classification model

Classification model	Major variable importance
Decision tree	Cumulative precipitation, EC, TN, TP
Random forest	Cumulative precipitation, EC, TN, water temperature
XGBoost	Cumulative precipitation, EC, TN, water temperature

Table 5 | Hyperparameters of the 12×12 ANN model

Normalization method	Min-Max [0, 1]
Training set/test set	8:2
Hidden layer	2
Hidden node	12×12
Activation function	Sigmoid
Error function	SSE
Learning rate	0.01
Learning algorithm	Resilient back propagation
Threshold	0.01
Evaluation (test data)	RMSE
	MSE
	MAE
	MASE
	R^2
	1.423
	2.025
	1.055
	1.132
	0.60

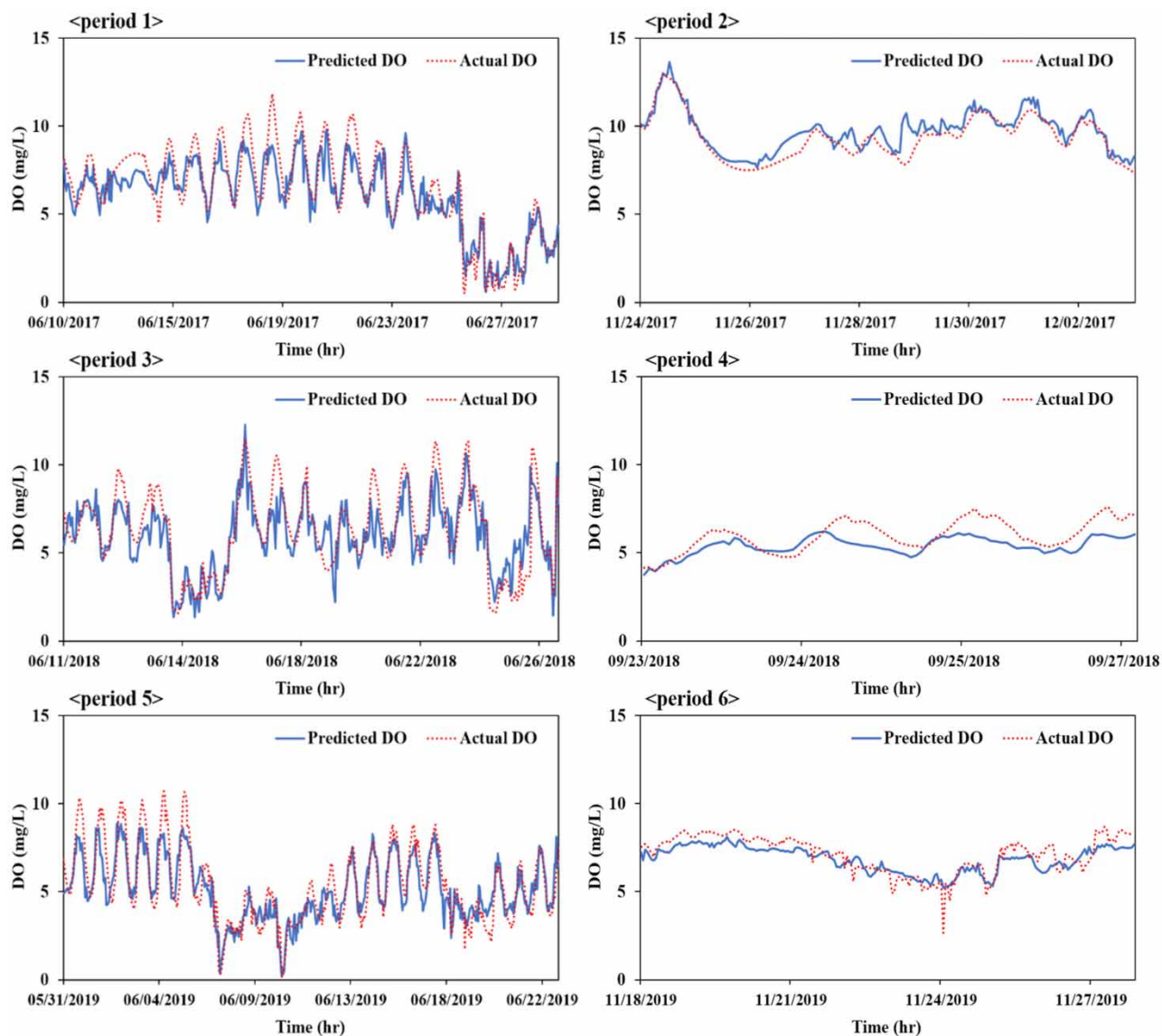
addition, we compared the predicted and actual values for the entire test data using the optimized ANN model, subsequently confirming good prediction performance (RMSE of 1.423 and R^2 of 0.60).

Based on the selected 12×12 ANN model, we verified the predictive performance of the model for each period of test data and confirmed good predictive performance (Table 6 and Figure 5).

As shown in Table 6, the ANN model for each period demonstrated good predictive performance. However, in period 4 ($n = 487$), the DO prediction performance decreased owing to the overall poor amount of data and, looking at volatility (%)

Table 6 | Predictive performance results of the ANN model for test data by period

Period	RMSE	MSE	MAE	MASE	R ²	Volatility (%)
1	1.046	1.095	0.835	2.354	0.86	16.17
2	0.611	0.373	0.453	3.422	0.86	6.53
3	1.316	1.731	0.985	2.108	0.71	20.95
4	1.679	2.819	1.583	8.603	0.64	28.11
5	1.022	1.046	0.790	1.921	0.78	19.09
6	1.433	2.052	1.346	6.223	0.79	20.12

**Figure 5** | Predictive performance results of the ANN model for test data.

for each period, all but period 2 showed somewhat unstable fluctuation characteristics of 15% or higher. Therefore, securing sufficient data is crucial when applying the ANN model. A better model can be developed if the training process is preceded by data securing, such as cross validation.

3.2.2. LSTM model

Table 7 shows the structure of the optimal LSTM model derived using training data for the first and second halves. In the first half, the average RMSE was 0.050 (0.037–0.057) and the average R^2 was 0.91 (0.88–0.95); in the second half, the average RMSE was 0.029 (0.021–0.034) and the average R^2 was 0.96 (0.94–0.98), showing higher predictive performance compared with the ANN model.

The prediction performance of the second-half model was slightly higher than that of the first-half model. In the case of the first-half model, the autocorrelation between data by time may be high because of the nature of time series data. When data from a relatively long period are used, the autocorrelation between input variables may increase, and the predictive performance of the model may decrease somewhat. Therefore, when constructing a real-time short-term prediction model using time series data, it is important to secure sufficient data to increase the predictive performance of the model. It is also important to select an appropriate level of data section considering the autocorrelation between time series data.

Table 8 and Figure 6 show the results of the DO prediction performance based on the LSTM model finally selected for each semi-annual period.

According to the results of the LSTM model given in Table 8, we confirmed that the DO prediction performance of the LSTM model was very good and stable compared with the ANN model (R^2 0.75–0.85). In addition, the values derived by the LSTM model for each semi-annual period in this study can be used to predict DO in the 24 h future at the junction of the Han River in Anyangcheon.

3.2.3. GRU model

Table 9 shows the structure of the optimal GRU model derived using the first and second halves of the training data. In the first half, the average RMSE was 0.029 (0.026–0.031) and the average R^2 was 0.97 (0.96–0.98). In the second half, the average RMSE was 0.028 (0.022–0.036) and the average R^2 was 0.96 (0.94–0.98), similar to the LSTM model. The GRU model showed higher predictive performance than the ANN model.

The GRU model showed higher accuracy than the ANN model and had the advantages of a simpler model structure and shorter training time compared with the LSTM model. In the future, the GRU model may be used in real-time to predict major water quality items, such as DO in the field. Table 10 and Figure 7 show the results of the DO prediction performance based on the GRU model finally selected for each semi-annual period.

Table 7 | Optimal model of LSTM for predicting DO for the first and second half

Hyper parameter	First half	Second half
Network structure	Layer 5 (256, 128, 64, 32, 16)	Layer 4 (256, 128, 64, 32)
Activation function	tanh	tanh
Loss function	Mean squared error	Mean squared error
Learning rate	10^{-2}	10^{-2}
Optimization algorithm	Adam	Adam
Dropout	0.1	0.1
Time delay	6 h	2 h

Table 8 | Predictive performance results of LSTM model for test data by period

Period	Date	RMSE	R^2	Volatility (%)
1	Jun 11–30, 2017	0.457	0.97	6.53
2	Nov 26–Dec 5, 2017	0.279	0.95	2.92
3	Jun 14–29, 2018	0.552	0.96	9.60
4	Sep 27–30, 2018	0.238	0.95	3.21
5	Jun 4–28, 2019	0.506	0.93	9.78
6	Nov 19–29, 2019	0.223	0.96	3.09

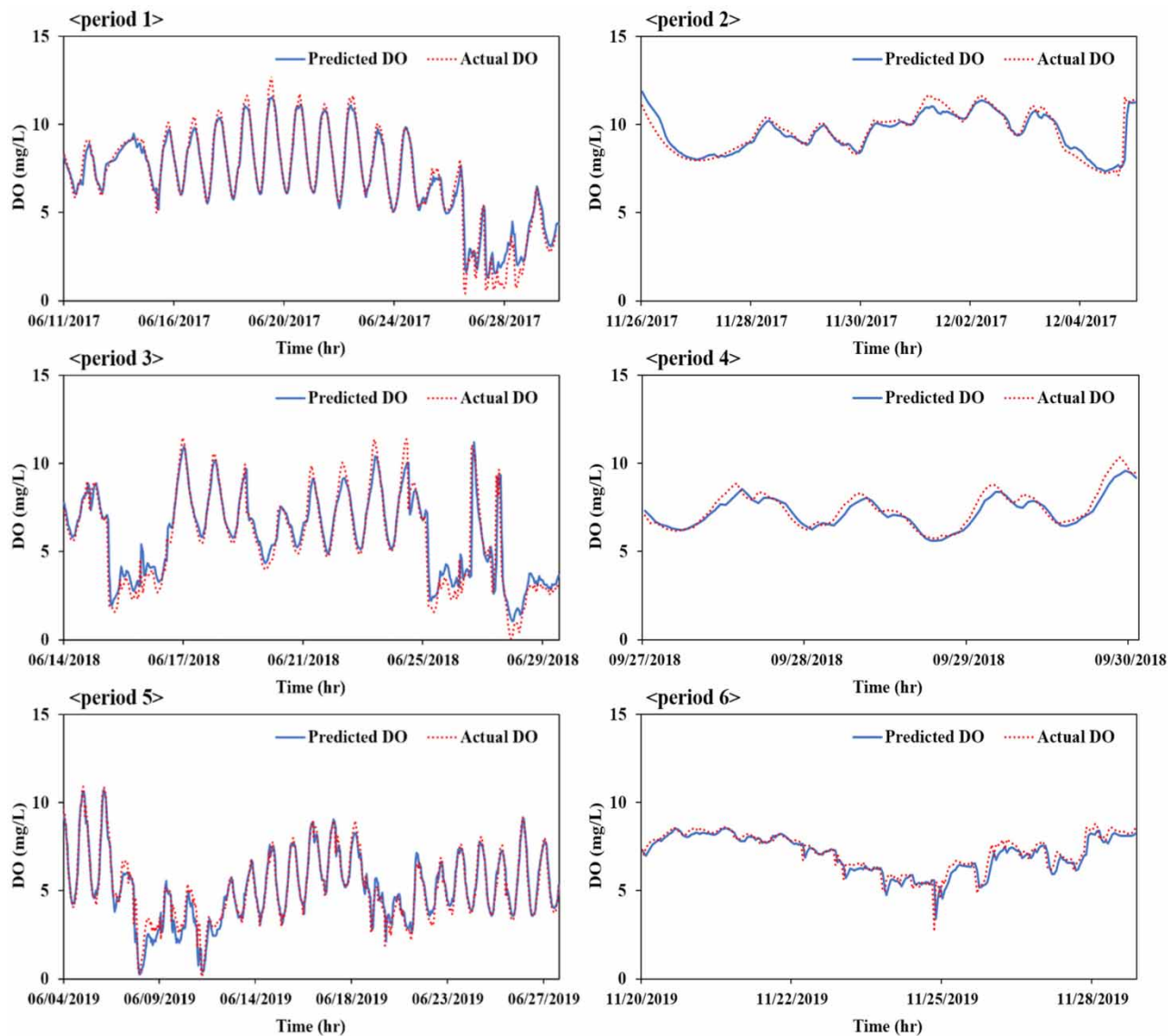


Figure 6 | Predictive performance results of LSTM model for test data.

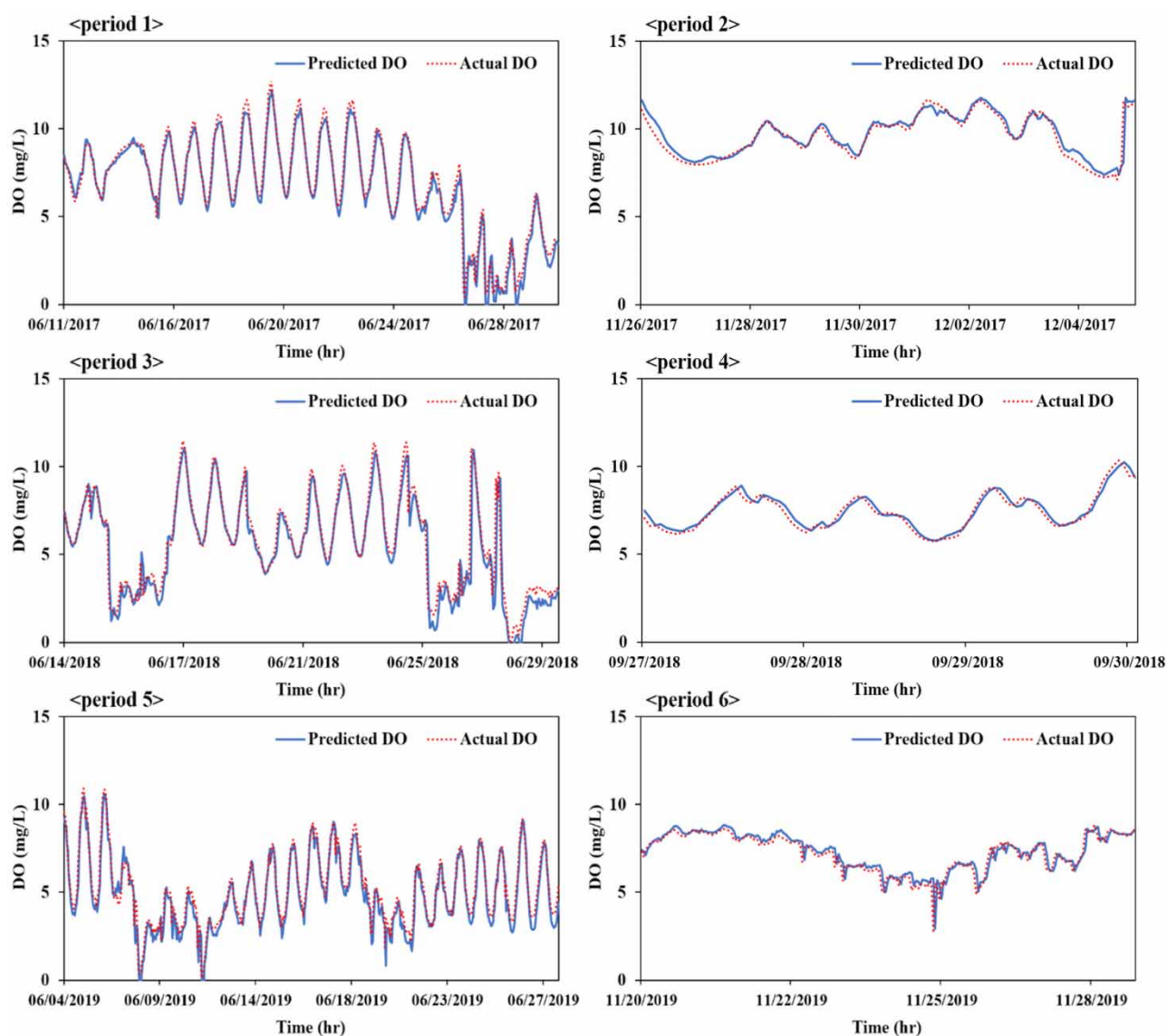
Table 9 | Optimal model of GRU for predicting DO for the first and second half data

Hyper parameter	First half	Second half
Network structure	Layer 4 (256, 128, 64, 32)	Layer 3 (128, 64, 32)
Activation function	tanh	tanh
Loss function	Mean squared error	Mean squared error
Learning rate	10^{-2}	10^{-2}
Optimization algorithm	Adam	Adam
Dropout	0.1	0.1
Time delay	6 hour	1 hour

As shown in Figure 8, The DO prediction performance of the GRU model was similar to that of the LSTM model and was very good and stable compared with the ANN model. In addition, similar to the LSTM model, the GRU model derived for each semi-annual in this study can be used to predict DO in the 24 h future at the junction of the Han River in Anyangcheon.

Table 10 | Predictive performance results of GRU model for test data by period

Period	Date	RMSE	R ²	Volatility (%)
1	Jun 11–30, 2017	0.432	0.98	6.17
2	Nov 26–Dec 5, 2017	0.243	0.96	2.54
3	Jun 14–29, 2018	0.505	0.97	8.79
4	Sep 27–30, 2018	0.166	0.98	2.24
5	Jun 4–28, 2019	0.501	0.94	9.68
6	Nov 19–29, 2019	0.144	0.98	2.00

**Figure 7** | Predictive performance results of GRU model for test data.

Also, as shown in Figure 9, it was possible to confirm how much the error rate of the LSTM and GRU models was reduced compared to the ANN model. For the LSTM model, prediction error was significantly reduced to 50.5–85.8% and the GRU model to 51.0–90.1%.

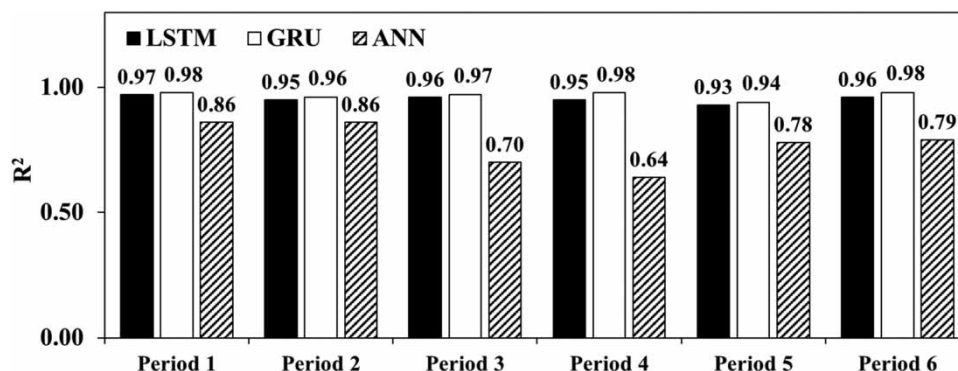


Figure 8 | Predictive performance results (R^2) for test data.

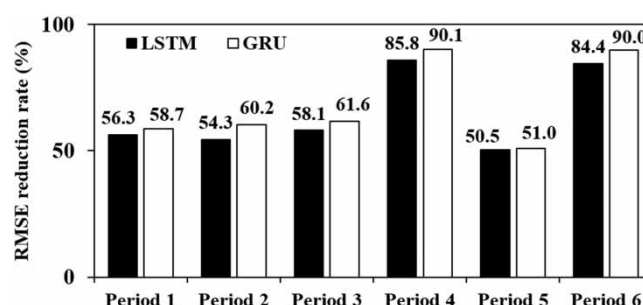


Figure 9 | RMSE reduction rate (%) of LSTM, GRU compared to ANN.

4. CONCLUSIONS

We developed a classification model and a real-time prediction model for short-term DO based on machine learning using water quality big data from the Anyangcheon water system. This model can be used as a method for advanced water quality management of urban rivers. The main results are as follows.

First, the classification model, applying decision tree, random forest, and XGBoost algorithms, identified EC, cumulative precipitation, TN, and water temperature as the main factors affecting DO changes in the Anyangcheon mobile WQMN. Based on the results of this classification model, water quality management may be carried out focusing on the water quality parameters of EC, cumulative precipitation, TN, and water temperature, which are closely related to DO changes in normal times. Doing so may prevent water quality abnormalities, such as water quality accidents, and improve water quality. In the event of a water quality abnormality, the cause analysis can be performed by considering the derived factors first.

In terms of classification performance, the random forest model had a sensitivity of 0.9962 and accuracy of 0.9981, whereas the XGBoost model had a sensitivity of 1.0000 and accuracy of 0.9822, confirming excellent classification performance. Therefore, in the case of insufficient data or an overfitting model, a random forest model with a bagging algorithm is advantageous. When using large amounts of data or when analysis time is important, the XGBoost model, which is advantageous for its fast calculation and hyperparameter adjustment, is advantageous through the use of GPUs. The variable importance ranking may yield slight differences among the classification models. When applying the classification model for more precise cause analysis and deriving influencing factors, it is necessary to derive common important variables by executing various classification models.

Second, we developed three models, namely, ANN, LSTM, and GRU, to predict real-time DO concentrations in the 24 h future at the junction of the Han River in Anyangcheon. A comparison of the prediction performance of the models for each period confirmed the excellent predictive performance of LSTM and GRU. These two models showed high predictive performance and model consistency regardless of the period, with R^2 of 0.95 and 0.98, respectively, even in the period with the smallest number of data points ($n = 487$). Therefore, we confirmed that DO could be predicted relatively accurately.

The LSTM and GRU models, which are machine learning-based RNN models, are considered to be suitable for predicting DO with strong nonlinearity in urban rivers, which have the characteristics of rapidly changing DO concentration according to various environmental factors. The LSTM and GRU models could account for all factors that can affect DO changes, such as cumulative precipitation, water temperature, EC, and TN, as input variables.

Notably, the LSTM and GRU models developed in this study do not suggest medium and long-term water quality management measures for large rivers, lakes, and watersheds. However, as waterfront areas for citizens, urban rivers are ideal targets for real-time automatic measurement. Their sensitivity to water quality changes necessitates the use of data that consider water quality environment in case of an emergency, such as after heavy rainfall. Short-term real-time changes in water quality could be quickly detected by our models, compared with existing static models, such as QUAL2 and HSPE.

In the future, it is judged that it is necessary to further improve the accuracy by securing sufficient data as the data-based model. In addition, in order to increase the applicability and scalability of a real-time short-term DO prediction model, various methods of calculating the importance of input variables and related researches such as transfer learning should be conducted.

ACKNOWLEDGEMENTS

The authors would like to thank the editors and reviews for useful comments which were helpful in improving the quality of the manuscript.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

REFERENCES

- Avi, O. H. & Shani, S. A. 2005 Hybrid genetic-instance based learning algorithm for CE-QUAL-W2 calibration. *Journal of Hydrology* **310**, 122–142.
- Breiman, L. 2001 Random forests. *Machine Learning* **45**, 5–32.
- Breiman, L., Friedman, J., Charles, J. S. & Olshen, R. A. 1984 *Classification and Regression Trees (The Wadsworth and Brooks-Cole Statistics Probability Series)*. Chapman & Hall, UK, pp. 27–35.
- Chen, L., Hongqing, L., Mingjun, L. & Qingyun, D. 2018 Dongting lake water level forecast and its relationship with the Three Gorges Dam based on a long short-term memory network. *Water* **10** (10), 1–20.
- Cho, Y. M. & Lee, J. H. 2018 *The Core Direction of Seoul's integrated Water Management Policy is set as Regional Focus, Watershed Management, and Expansion of Governance*. Report The Seoul Institute, Seoul, Korea.
- Chung, J. Y., Gulcehre, C., Cho, K. H. & Bengio, Y. 2014 *Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling*. arXiv preprint arXiv:1412.3555, Cornell University, New York, USA.
- Duda, R. O. & Hart, P. E. 1973 *Pattern Classification and Scene Analysis*. Report Stanford Research Institute, California, USA.
- Durdu, O. F. 2009 A hybrid neural network and ARIMA model for water quality time series prediction. *Engineering Applications of Artificial Intelligence* **23**, 586–594.
- Fitore, M., Doina, L. & Florin, L. 2019 Machine learning approaches for anomaly detection of water quality on a real-world data set. Fourier analysis for demand forecasting in a fashion company. *Journal of Information and Telecommunication* **3** (3), 294–307.
- Fournier, D. & Cremilleux, B. 2002 A quality index for decision tree pruning. *Knowledge Based Systems* **15**, 37–43.
- Friedman, J. H. 2001 Greedy function approximation; a gradient boosting machine. *The Annals of Statistics* **29** (5), 1189–1232.
- Giuseppe, C. & Balaji, V. 2017 *Neural Networks with R*. Packt Publishing Ltd, Birmingham, UK, pp. 259–261.
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F. & Pedreschi, D. 2018 A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)* **51** (5), 1–45.
- Guzman, S. M., Paz, J. O. & Tagert, M. M. L. 2017 The use of NARX neural networks to forecast daily groundwater levels. *Water Resources Management* **31**, 1591–1603.
- Hochreiter, S. & Schmidhuber, J. 1997 Long short-term memory. *Neural Computation* **9** (8), 1735–1780.
- Kass, G. V. 1980 An exploratory technique for investigating large quantities of categorical data. *Applied Statistics* **29** (2), 119–127.
- Kim, S. H. & Kang, S. J. 2017 Image feature-based electric vehicle detection and classification system using machine learning. *The Transactions of The Korean Institute of Electrical Engineers* **66** (7), 1092–1099.
- Kiran, K. P. & Arjun, B. 2019 Forecasting of rice cultivation in india-A comparative analysis with ARIMA and LSTM-NN models. *EAI Endorsed Transactions on Scalable Information System* **24** (7), 1–11.
- Krzysztof, G., Iga, G. & Tomasz, Z. 2019 Reducing false arrhythmia alarms using different methods of probability and class assignment in random forest learning methods. *Sensors* **19** (7), 1–23.

- Lee, J. H., Yoon, H. K., Ha, H. J., Cho, S. S., Yang, I. H., Lee, S. K., Chun, C. W., Lee, T. H., Yang, J. Y., Kim, J. H., Kil, H. K., Lee, M. Y., Jung, K. & Koo, J. Y. 2018 *Utilization Plan of Automatic Water Quality Monitoring Networks Data with Statistical Models (Focusing on Setting the Monitoring Standard)*. Report Seoul Metropolitan Government Research Institute of Public Health and Environment, Seoul, Korea.
- Murugappan, M., Subbulakshmi, M. & Bong, S. Z. 2013 Frequency band analysis of electrocardiogram (ECG) signal for human emotional state classification using discrete wavelet transform (DWT). *Journal of Physical Therapy Science* **25** (7), 753–759.
- Muttil, N. & Chau, K. W. 2006 Neural network and genetic programming for modelling coastal algal blooms. *International Journal of Environment and Pollution* **28** (3), 223–238.
- Ruiz-Aguilar, J. J., Turias, J. J. & Jiménez-Come, M. J. 2014 Hybrid approaches based on SARIMA and artificial neural networks for inspection time series forecasting. *Transportation Research Part E* **67**, 1–13.
- Sarkar, A. & Pandey, P. 2015 River water quality modelling using artificial neural network technique. *Aquatic Procedia* **4**, 1070–1077.
- Senlin, Z. & Salim, H. 2020 Prediction of dissolved oxygen in urban rivers at the Three Gorges Reservoir, China: extreme learning machines (ELM) versus artificial neural network (ANN). *Water Quality Research Journal* **55** (1), 106–118.
- Sima, S. N. & Akbar, S. N. 2018 *Forecasting Economics and Financial Time Series: ARIMA vs. LSTM*. arXiv:1803.06386, Cornell University, New York, USA.
- Snezhana, G. G., Desislava, S. V., Maya, P. S., Atanas, V. I. & Ilycho, P. I. 2019 Regression trees modeling of time series for air pollution analysis and forecasting. *Neural Computing and Applications* **31**, 311–317.
- Stephan, D. & Lucila, O. M. 2002 Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics* **35** (5-6), 352–359.
- Wang, Y., Zhou, J., Chen, K., Wang, Y. & Liu, L. 2017 Water quality prediction method based on LSTM neural network. In: *IEEE*. pp. 1–5.
- Yi, F. Z., Peter, F. & Peter, J. T. 2020 Prediction the trend of dissolved oxygen based on the kPCA-RNN model. *Water* **12**, 1–15.

First received 16 August 2021; accepted in revised form 19 January 2022. Available online 2 February 2022