Water Supply

IWA PUBLISHING

Water Supply Vol 22 No 1, 1168 doi: 10.2166/ws.2021.202

BOD5 prediction using machine learning methods

Kai Sheng Ooi[a], ZhiYuan Chen[a,*], Phaik Eong Poh[b] and Jian Cui[c]
[a] School of Computer Science, University of Nottingham Malaysia, Jln Broga, 43500 Semenyih, Selangor, Malaysia
[b] School of Engineering, Monash University Malaysia, Jalan Lagoon Selatan, 47500 Subang Jaya, Selangor, Malaysia
[c] Jiangsu Province and Chinese Academy of Sciences, Institute of Botany, Nanjing, China
*Corresponding author. E-mail: zhiyuan.chen@nottingham.edu.my

ZYC, 0000-0002-4915-1593; PEP, 0000-0002-4215-5284

ABSTRACT

Biological oxygen demand (BOD5) is an indicator used to monitor water quality. However, the standard process of measuring BOD5 is time consuming and could delay crucial mitigation works in the event of pollution. To solve this problem, this study employed multiple machine learning (ML) methods such as random forest (RF), support vector regression (SVR) and multilayer perceptron (MLP) to train a best model that can accurately predict the BOD5 values in water samples based on other physical and chemical properties of the water. The training parameters were optimized using genetic algorithm (GA) and feature selection was made using the sequential feature selection (SFS) method. The proposed machine learning framework was first tested on a public dataset (Waterbase). The MLP method produced the best model, with an R^2 score of 0.7672791942775417, relative mean squared error (MSE) and relative mean absolute error (MAE) of approximately 15%. Feature importance calculations indicated that chemical oxygen demand (CODCr), ammonium and nitrate are features that highly correlate to BOD5. In the field study with a small private dataset consisting of water samples collected from two different lakes in Jiangsu Province of China, the trained model was found to have a similar range of prediction error (around 15%), a similar relative MAE (around 14%) and achieved about 6% better relative RMSE.

Key words: biological oxygen demand, multilayer perceptron, random forest, supervised regression, support vector regression, water quality

HIGHLIGHTS

- Multiple machine learning (ML) methods used to train a best model to predict the BOD5 values in water samples.
- Permutation feature importance (PFI) values calculated indicate that CODCr, ammonium and nitrate are features with the highest correlation with BOD5.
- The best model trained using MLP yielded the best performance with an R^2 score of 0.7672791942775417 together with relative RMSE and relative MAE of approximately 15%.

1. INTRODUCTION

Water is essential to sustain lives on Earth. Approximately 70% of the surface of the Earth is covered by water, but 97% of it is seawater which cannot be used directly for drinking. Out of the remaining water source, only 1.2% is in the form of surface freshwater which is used for the vast majority of human activities (Shiklomanov 1993). According to the United Nations, the world's population has grown tremendously from 2.54 billion (2.54×10^9) in 1950 to 7.79 billion in 2020 and is projected to reach 10.88 billion by the end of the century (United Nations n.d.). The increase in human population will contribute to the escalating demand for safe and clean water. Recently researchers have been working on new technologies such as nanotechnologies to improve water quality and water purification, and on environmental remediation (Zinatloo-Ajabshir *et al.* 2020). Nevertheless, anthropogenic activities contributed to water pollution, making safe and clean water sources scarce. Therefore, management of water resources is imperative to sustain development of civilization.

Two major categories of pollutants contribute to water pollution: (i) organic and (ii) inorganic contaminants. Organic contaminants can be naturally decomposed by aquatic microorganisms through oxidation and this process eventually depletes the dissolved oxygen in water. Meanwhile, the widespread usage in the agricultural sector of artificial fertilizers which contain nitrates and phosphates also causes algae blooms and eutrophication, covering the water surface, emitting a foul smell and preventing sunlight from reaching plants in littoral zones. The plants die and contribute to the increased amounts of dead

organic matter that is also then oxidized by the aquatic microorganisms, further depleting the oxygen concentration in water. These factors contribute to creation of hypoxic regions where the dissolved oxygen level is too low to sustain lives of most organisms, thus destroying the marine ecosystems (Chislock *et al.* 2013).

As such, quantifying the extent of organic pollution in water is essential for water resources management. Biochemical oxygen demand – a measure of the amount of oxygen needed by aerobic biological organisms to break down organic material in a water sample over time, is widely used as an indicator for organic pollution in water (USGS n.d.). The standard method of measuring biochemical oxygen demand (BOD), known as BOD5, was proposed by the UK's Royal Commission on Sewage Disposal in 1912 and measures the oxygen consumed per liter of water over 5 days of incubation at 20 °C.

The process of measuring BOD5 in a large number of water samples requires substantial work and time which consumes the limited resources of water management agencies (Delzer & McKenzie 2003). The delay between start of measurement and availability of results can also cost water management agencies precious time for mitigation works in case of pollution events. Alternative methods for BOD5 measurement using biosensors have been tested but the measurement results often deviate significantly from the measured BOD5 values (Reshetilov *et al.* 2013). On the other hand, machine learning models can predict the BOD5 values more accurately by using data from previous BOD5 measurements. Therefore, this study aims to develop an effective system to predict BOD5 values from water samples using machine learning supervised regression methods, which would help water management agencies to focus their resources in other crucial areas as well as carry out pollution mitigation efforts in a timely manner.

The traditional method used to predict BOD5 values is to use statistical methods; however, this can suffer from lower performance compared to machine learning methods. Alvarez-Guerra *et al.* (Chan *et al.* 2021) studied the use of traditional statistical methods in prediction of amphipod toxicity in contaminated sediments and reported machine learning methods achieving significantly higher prediction accuracy compared to any of the statistical methods. Li *et al.* 2020 studied the use of a nonlinear autoregressive exogenous model (NARX) in predicting concentrations of three toxic metals in the Elbe river in Europe and demonstrated it was inferior to other statistical methods. Cipullo et al. (2019) presented the research work of predicting the bioavailable concentration of chemicals in soil samples with a better prediction performance using the random forest (RF) method. Other machine learning methods such as boosted regression trees (BRT), M5 trees and random forest regression (RFR) have also been presented in various studies (Ransom *et al.* 2017; Chou *et al.* 2018; Huang *et al.* 2018; Nieto *et al.* 2019; Yuchi *et al.* 2019).

Among machine learning methods, artificial neural networks (ANNs) and support vector machines (SVMs) have been widely used in prediction of chemical components. Counter-propagation artificial neural networks (CP-ANN) (Chan *et al.* 2021), back-propagation neural networks (BPNN) (Li *et al.* 2020) and deep neural network (DNN) (Chan *et al.* 2021) were found to have excellent performance and were superior to from traditional statistical methods. Even traditional neural networks (NNs) (Park *et al.* 2014; Cipullo *et al.* 2019; Chou *et al.* 2018) have also shown good performance in prediction. Particularly Dogan *et al.* (2009) used ANN for predicting BOD5 in water samples with data obtained from the Melen river in Turkey, and found it yielded good results ($R^2 = 0.875$). Many researchers (Wang *et al.* 2008; Sapankevych & Sankar 2009; Chou *et al.* 2018; Nieto *et al.* 2019) have surveyed the use of SVMs in prediction of chemical data across multiple fields and found it's performance better compared to other methods. Ji *et al.* (Arumugasamy *et al.* 2021) compared the performance of SVM, BPNN, GRNN and multilinear regression (MLR) in predicting DO concentration in hypoxic river systems and found SVM has the best performance among the methods compared. Arumugasamy *et al.* (2021) compared the performance of ANN and SVM techniques for prediction of bio-polymer molecular weight and the results from both training and testing samples indicated SVM as a proper solution with respect to the characteristic of a polymerization problem.

Other than the traditional machine learning methods, there are also state-of-the art approaches to optimization algorithms and hybrid methods. Latest research work on optimization algorithms have been presented in Cai *et al.* (2021) and Deng *et al.* (2020). Li *et al.* (2020) tested hybrid methods: a wavelet BPNN hybrid model (WNN) and wavelet NARX hybrid model (WNARX). It was found that hybrid models produce predictions with best accuracy, with the exact better model of the two depending on the toxic metal being predicted. Chou *et al.* (2018) studied ensemble models of the traditional methods using voting, bagging, stacking and tiering techniques, in addition to the MetaFA-LSSVR hybrid model. MetaFA-LSSVR utilizes the firefly algorithm (FA) for optimizing parameters for least squares support vector machines for regression (LSSVR). The experimental results showed that the ensemble method of ANN using the tiering technique produces predictions with the lowest root-mean-square error (RMSE) and mean absolute error (MAE), while the MetaFA-LSSVR method produces the highest linear correlation coefficient (R). Nieto *et al.* (2019) used the artificial bee colony (ABC) algorithm for optimizing

parameters of SVMs using different kernels. Wang *et al.* (2008) proposed an online SVM method where the data was inputted sequentially, and the successive optimal model was calculated using the previous optimal model. The online SVM method was found to be superior to the traditional SVM method compared in the study. Yeganeh *et al.* (2012) combined partial least squares (PLS) for feature selection with SVM for the hybrid SVM-PLS method. The SVM-PLS method produced results which were more accurate than traditional SVM and at reduced computation time. Yuchi *et al.* (2019) studied the use of blended methods where the results of different methods were averaged. They also proposed the use of predictions from one model as an input feature for prediction in another method and found that the blended method using RFR predictions in MLR gave the best performance compared to standalone MLR, standalone RFR and the average of both.

These works reflected that decision trees, artificial neural networks and support vector machines can be effective to predict environmental pollution indicators. This project therefore uses the three different machine learning methods of random forest (RF), multilayer perceptron (MLP) and support vector regression (SVR) in training the prediction models. RF is an ensemble of decision trees where multiple different decision trees were used and their predictions combined to provide better performance as compared to a single tree. MLP is a basic version of a feedforward artificial neural network. SVR is the variant of SVM that is used in regression tasks instead of classification.

2. MATERIAL AND METHODS

2.1. Area of the research work

The dataset used for this project was obtained from the European Environment Agency. Waterbase (European Environmental Agency 2020), a database of water quality measurements in water bodies throughout the European Economic Area, was used for this research. The part of the dataset used is the Water Quality ICM Aggregated Data. The raw dataset consists of 3,510,775 samples taken from 1931 to 2018. Each sample contains 31 features such as the monitoring site, water body category, type of determinant measured, procedure of measurement, units of measurement, year of measurement, indicators of data sample quality and statistics of each data sample such as mean, minimum, maximum, median and standard deviation. The rest of the features are related metadata of each data sample. This research uses the mean values as the indicative values of each aggregated data sample due to them being the most complete statistics of each sample as opposed to other statistics such as median.

Meanwhile a field study was carried out to explore the potential of the learning framework on a different dataset. The aim was to study the viability of applying the same learning framework to a dataset that can have different relations between the features and target value due to a different water body category from which the samples were taken. The field study also observed the effects on performance of using a dataset that is much smaller in terms of number of data samples and number of features.

2.2. System flow

Figure 1 presents the flow diagram of the proposed system for BOD5 prediction. It consists of four layers, namely: (1) data preprocessing layer, (2) machine learning model training layer, (3) optimization layer and (4) decision-making layer. Layer 1 involves cleaning and reformatting the dataset while Layer 2 works to train machine learning models. Three machine learning methods were implemented in Layer 2: RF, MLP and SVR. Layer 3 meanwhile works to optimize the models via the use of GA for hyperparameter tuning and sequential feature selection (SFS) to select the best subset of features. Performance metrics and feature importance are calculated in this layer. Finally, Layer 4 selects the best model out of all models trained using the three different machine learning methods. The criteria for best model selection was the model with the highest correlation coefficient (R^2) value.

2.3. Data preprocessing

As shown in Figure 2, the dataset was preprocessed to remove unusable or low-quality data samples. Since the mean values were used as the indicator values, all data samples where the mean value was missing were removed from the dataset. Subsequently, all data samples where the readings obtained did not meet the required standard and where values for key features were missing were removed. This was done by checking the features of each data sample that indicates the quality of data, such as the Level of Quality (LOQ) features. Data samples that contained any of the LOQ features indicating that the sample is not up to quality standards, were removed. Thereafter, only the features relevant to the project were retained in the dataset. These features include the monitoring site, water body category, type of measurement, year of measurement and the mean measurement values of each data sample. Duplicate or conflicting data with the same monitoring site, type of measurement and year of measurement were removed. To fit the model training algorithm, the dataset was reformatted so that each data sample consists of data for the monitoring site, year of measurement, water body category, and values for all types of

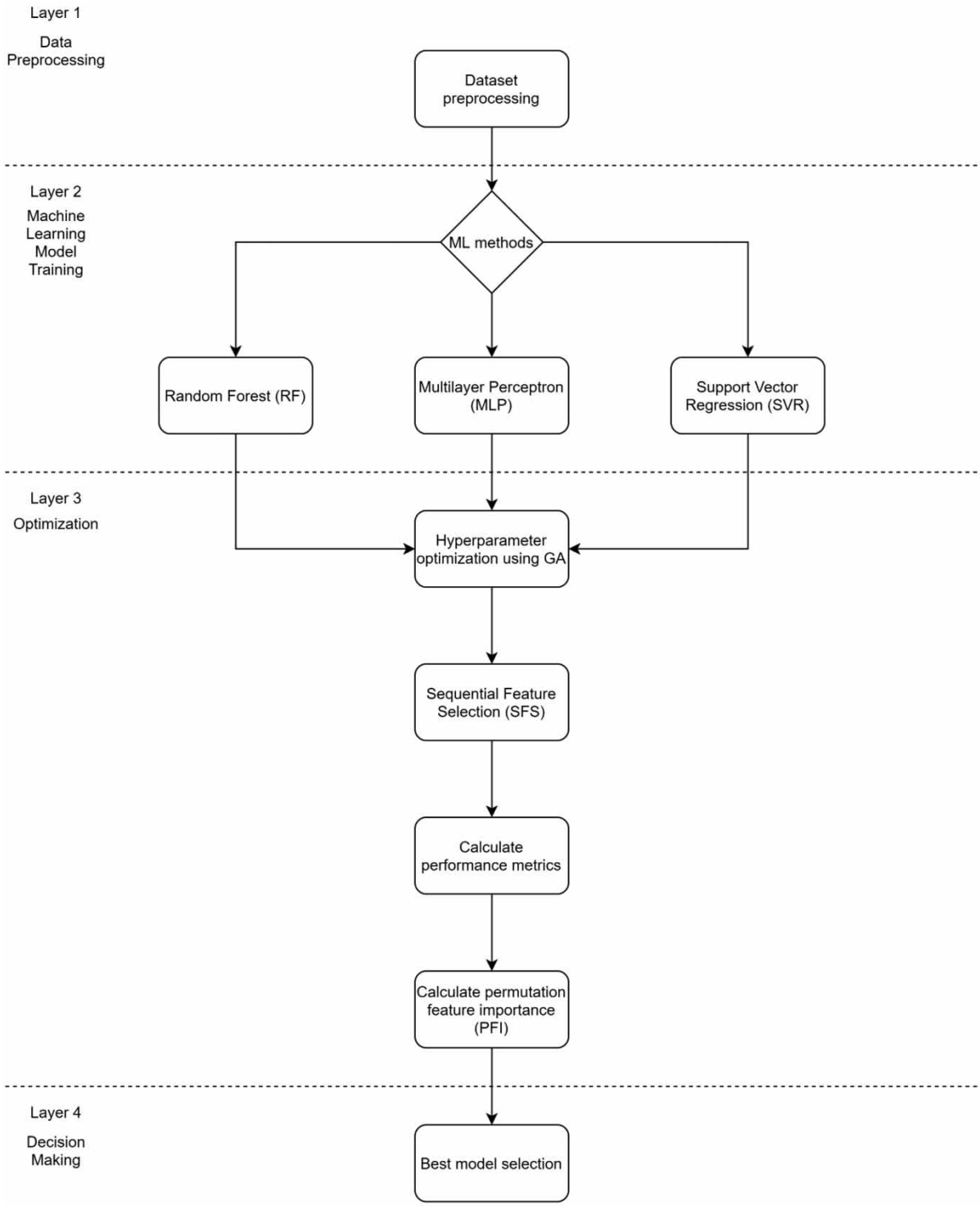

Figure 1 | Flow diagram of the proposed system.

measurements that are present. All data samples with missing value for BOD5 were removed since it is the target feature for prediction. Rare features with values missing in more than 90% of all data samples were removed from the dataset. After that, sparse data points that have more than half of the feature values missing were also removed from the dataset. The feature set was then rechecked for compliance with the above requirement. The missing values for the remaining dataset were filled with median values of each determinant. All values in the dataset were then standardized by centering to mean and scaling to unit

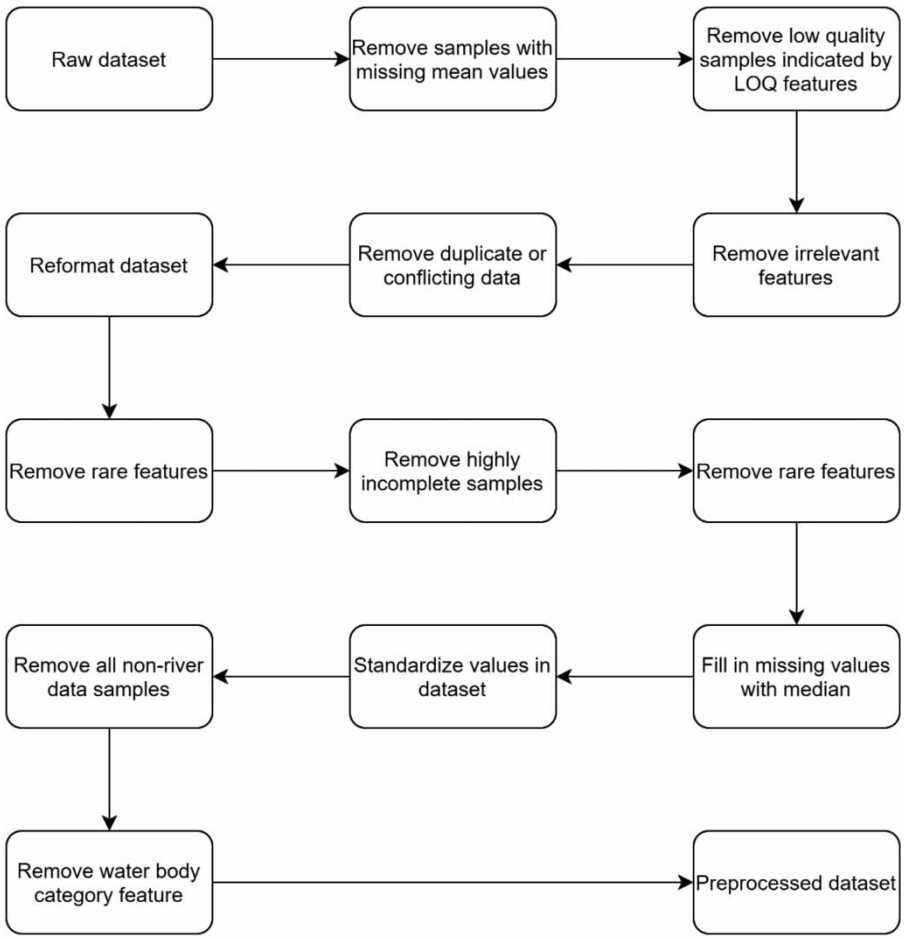

Figure 2 | Data preprocessing process.

variance of each feature due to requirements of some machine learning methods such as SVR. After that, all non-river water sample data were removed due to huge variations in conditions between types of water bodies that would negatively affect the model. River water samples were chosen because this is the category that is most prevalent in the dataset. The resulting preprocessed dataset consists of 1,436 data samples and 48 features including the target feature – BOD5.

2.4. Model training

2.4.1. Random forest (RF)

Random forest (RF) (Breiman 2001) is a machine learning technique which utilizes an ensemble of decision trees for classification or regression tasks. Each decision tree in random forest was trained with a random sample with replacement of the training data, and the results of all trees were combined using majority vote for classification and averaging for regression. The use of random samples of the training data for multiple decision trees reduces overfitting compared to using the entire training set with a single decision tree. RF has an advantage over other machine learning methods which create a 'black-box' model as a model created with RF can be easily interpreted by humans.

2.4.2. Multilayer perceptron (MLP)

Multilayer perceptron (MLP) (Hastie *et al.* 2009) is a variant of a feedforward artificial neural network. An artificial neural network consists of connected nodes resembling the neurons in a biological brain. MLP consists of a minimum of 3 layers of nodes: the input layer, hidden layer and output layer. Other than the input layer nodes, each node receives inputs from the other nodes, and the output of each node is calculated using a nonlinear activation function. The connections between nodes have weights which determine their relative importance. The learning process for MLP involves continually adjusting the weights in the network to minimize the error rate using backpropagation. Backpropagation computes the gradient of the

weight space with respect to error calculated by a loss function and updates the weights in the network using methods such as stochastic gradient descent.

2.4.3. Support vector regression (SVR)

Support Vector Regression (SVR) (Drucker *et al*. 1996) is a variant of SVM but used for regression tasks. SVM maps the original input space into a high-dimensional input space and performs linear regression in the high-dimensional space by constructing a maximum margin separator which minimizes expected generalization error instead of training error. The original inputs were mapped using kernel functions, which take as input the dot products of pairs of input points. The use of dot products allows the SVM to map the inputs efficiently compared to calculating the corresponding points of each input in a high-dimensional space. To prevent overfitting, SVM allows a soft margin which allows for misclassifications but tries to minimize the cost calculated by a cost function for each misclassification. Compared to SVM where the cost function only considers data points within the margin, SVR does not consider data points close to the model prediction.

2.5. Optimization of training parameters and feature selection

The training parameters for each method were optimized using GA searching. GA is a search algorithm that utilizes sets of chromosomes across multiple generations. Each chromosome contains a candidate solution for the search. During the first generation, a specified number of chromosomes are initialized and evaluated. Afterwards, a new set of child chromosomes are generated by using the parent chromosomes from a previous generation and then undergoing a mutation process. These child chromosomes were then evaluated, and their performance compared to the parent chromosomes. The next generation of parent chromosomes were then chosen from this pool of previous generation chromosomes and child chromosomes. New generations were then generated repeatedly until a terminating condition is met.

For hyperparameter optimization, each chromosome represents a set of hyperparameter values for the machine learning method. The first generation of chromosomes is initialized randomly. The algorithm tests the performance of the models trained with the sets of parameters from each chromosome. The child chromosomes are created using two randomly selected parents and then undergo mutation by randomly offsetting each value in the child chromosome. The next generation of parent chromosomes are chosen by taking the best chromosomes from the pool of parent and child chromosomes. The process repeats until the number of generations reach a set limit and the algorithm returns the chromosome containing the best set of parameters for each model.

Feature selection for each method was done using the model trained using the optimized parameters. SFS was used as the feature selection method. SFS divides the entire feature set into two different subsets, which are the chosen features subset and the remaining features subset. SFS starts with an empty chosen feature subset and then greedily chooses the best feature from the remaining features subset to add into the subset at each step. Each feature in the remaining features subset was tested by training the model with a feature set consisting of the entire chosen features subset and the feature from the remaining features subset that is currently being tested. The best feature at each step is the feature that provides the biggest increase in performance to the model at each step. The best feature is then moved from the remaining features subset into the chosen features subset. The algorithm repeats until adding any of the feature from the remaining features subset will not improve performance of the model further. SFS then returns the chosen features subset which gives the best performance for the model.

2.6. Model evaluation

The models were trained and tested using 10-fold cross validation where the dataset was split into 10 parts, with 9 parts used for training and 1 part used for testing. The process was repeated 10 times with a different part being used for testing each time. The average performance across all 10 runs was recorded as the performance for each model. The performance of the trained models was evaluated on 5 metrics: correlation coefficient (R^2), MSE, relative MSE, MAE and relative MAE.

Permutation feature importance (PFI) was calculated for the models trained using each method. PFI is a method for calculating feature importance that randomly shuffles the values of each feature in the model and determines the effect on the performance of the model. The formula for the calculation can be found in Figure 3 (Scikit-learn n.d.).

2.7. Best model selection

The best model is selected from the three models each trained using a different machine learning method. The criteria for selection as the best model are the correlation coefficients (R^2) of each model obtained during evaluation after optimization of hyperparameters and feature selection. The model with the highest R^2 score would be chosen as the best model.

- Input: fitted predictive model m, tabular dataset (training or validation) D.
- Computer the reference score s of the model m on dataset D (for instance the accuracy for a classifier or the R^2 for a regressor).
- For each feature j (column of D):
    - For each repetition K in $1, \ldots, K$:
        - Randomly shuffle column j of dataset D to generate a corrupted version of the data named $\widetilde{D}_{k,j}$.
        - Compute the score $s_{k,j}$ of model m on corrected data $\widetilde{D}_{k,j}$
    - Compute importance i_j for feature f_j defined as:

$$i_j = s - \frac{1}{K}\sum_{k=1}^{K} s_{k,j}$$

Figure 3 | Formula for calculation of permutation feature importance (PFI).

2.8. Software

Dataset preprocessing was performed using the pandas library (The pandas development team 2021) (McKinney 2010). Training and evaluation of the RF and SVR methods were implemented using the scikit-learn library (Pedregosa *et al.* 2011). MLP was trained and evaluated using the PyTorch library (Paszke *et al.* 2019).

3. RESULTS AND DISCUSSION

3.1. Experimental setting

Samples where the mean value is missing were first removed from the dataset. Subsequently, all low-quality data which are indicated by the various LOQ features were removed. All unnecessary features were then removed. Duplicate or conflicting data which have the same values for water sampling location, type of measurement and year of measurement were removed. The dataset was then reformatted such that each data sample consisted of water sampling location, year of measurement, type of water body and all the different types of measurement values in the dataset. All data samples having missing values for the BOD5 feature were removed. Features that only had values in less than 10% of the data samples were removed by counting the number of missing values for each feature. Data samples that had missing values in more than half of all types of measurements were also removed by counting number of missing values for each data sample. All features were then subjected to the minimum values present in 10% data samples check as above, again to remove those that no longer meet the requirement after removal of samples. Using the sklearn library, the missing values in the dataset were then filled in and all the values were then standardized. By checking for the type of water body, all data samples that do not belong to a river water sample were removed. The feature indicating type of water body was then removed.

Table 1 | Default parameter values of each machine learning method

Method	Parameters	
RF	n_estimators	100
	max_depth	None
MLP	hidden_layer_sizes	(100, 100, 100)
	solver	adam
	learning_rate_init	0.001
SVR	kernel	rbf
	gamma	scale
	C	1
	coef0	0
	degree	3

Table 2 | Range of values for parameters of each machine learning method

Method	Parameters	
RF	n_estimators	20–200
	max_depth	5–50
MLP	hidden_layer_sizes	(1–100, 1–100, 1–100)
	solver	sgd, adam
	learning_rate_init	0–0.01
SVR	kernel	poly, rbf, sigmoid
	gamma	scale, auto
	C	0.1–10
	coef0	0–10
	degree	1–3

Table 3 | Best parameter values obtained from a genetic algorithm (GA) search for each machine learning method

Method	Parameters	
RF	n_estimators	151
	max_depth	38
MLP	hidden_layer_sizes	(29, 69, 75)
	solver	adam
	learning_rate_init	0.008227074264420868
SVR	kernel	rbf
	gamma	auto
	C	4.439991951383791
	coef0	5.723662221079469
	degree	2

Random seeding was performed to ensure consistent and reproducible results for each model. Evaluation of each trained model was done by training with 10-fold cross validation. The models were first trained with each method using the default training parameters as shown in Table 1.

GA was used to search for the best parameter set of each method. The GA algorithm uses 10 chromosomes per generation across 10 generations. Child chromosomes were generated by randomly selecting two parents and taking the average of their chromosome values. During each generation only the 10 best chromosomes between the parent and child chromosomes were

Table 4 | List of chosen features using manual feature selection

Features
Water temperature
Total suspended solids
Nitrate
Nitrite
Phosphate
CODCr
Ammonium
Total phosphorus
Dissolved oxygen
Chlorophyll a
CODMn

kept and used as parent chromosomes for next generation. At the end, the algorithm outputs the optimal set of parameters for each method. Table 2 shows the possible ranges for each parameter of each method. For RF, n_estimators is the number of trees used and has uniform distribution in integers from 20 to 200. max_depth is the maximum depth of each tree and has uniform distribution in integers from 5 to 50. For MLP, hidden_layer_sizes define the number of nodes for all three hidden layers of the network. The number of nodes for each layer has uniform distribution in integers from 1 to 100. solver is the optimizer used when training the network and has equal probability for both. learning_rate_init is the learning rate during training and has uniform distribution. For SVR, kernel is the kernel used for the SVR algorithm and has equal probability for each. gamma is the kernel coefficient for when rbf, poly or sigmoid kernels are used and has equal probability for each. If gamma = scale, 1/(n_features * X.var()) is used as the value. If gamma = auto, 1/n_features is used as the value. C is the regularization parameter and is calculated by taking the power of 100 to a uniformly distributed number between 0 and 1, then dividing by 10. coef0 is the independent term in the kernel function and has uniform distribution. degree is the degree of the poly kernel and has equal probability for 1, 2 or 3.

The best sets of parameters found by GA for each method are as shown in Table 3. Feature selection was first conducted manually with the advice from a domain expert. The features chosen were assumed to be more correlated to BOD5 values in the water samples. The chosen list of features is listed in Table 4 and used consistently across all three machine learning methods.

Afterwards, feature selection was done using SFS with the optimal parameter set for each method. SFS starts with an empty chosen feature list and at each step finds the best feature to add into the list. The process stops when adding any of the remaining feature would not increase the model performance further. The best set of features found by SFS for each method are shown in Table 5.

Table 5 | List of chosen features for each machine learning method using sequential feature selection

Method	Features
RF	Total suspended solids Nitrate Nitrite CODCr Ammonium Chlorophyll a Copper and its compounds Iron and its compounds Aluminium and its compounds Benzo(g,h,i)perylene Vanadium and its compounds CODMn AOX EDTA
MLP	CODCr Nitrite Total suspended solids Ammonium Chlorophyll a Manganese and its compounds AOX Aluminium and its compounds Iron and its compounds Acid neutralizing capacity
SVR	Total suspended solids Nitrite CODCr Ammonium Iron and its compounds Aluminium and its compounds Total dissolved solids EDTA

Table 6 | Performance metrics using default parameter values

Method	R^2	MSE	Relative MSE	MAE	Relative MAE	Model Training Time
RF	0.6897242343363986	0.6069281621790829	0.202064956806941	0.5097239333488733	0.1697026946415008	0.365s
MLP	0.6002781376741629	0.7184918967199982	0.23920793781525912	0.5283550803663198	0.17590557358491232	2.765s
SVR	0.6014518673809747	0.7692460179979574	0.2561055378326711	0.5330737564070847	0.17747656522738245	0.088s
OLSLR	0.5172842754813942	0.8718954876604146	0.2902806872920887	0.5847165424826792	0.19467002894850743	0.005s
DTR	0.38333576045506673	1.1471815725883916	0.381931848544531	0.7107092701048952	0.236616863271644	0.032s

3.2. Results

3.2.1 Performance

Table 6 shows the results for each method trained using default parameters.

The ordinary least squares linear regression (OLSLR) statistical method and decision tree regression (DTR) a single decision tree method, were also tested against the three proposed machine learning methods. Comparison of the results shows that OLSLR is inferior to all three proposed methods, which is consistent with previous findings that statistical methods perform worse than machine learning methods. DTR also produced worse results than RF, showing that the ensemble of multiple decision trees in RF has the advantage over single decision tree methods such as DTR.

In terms of model training time, OLSLR was the fastest at 0.005 s while MLP was the slowest at 2.765 s which is well within acceptable range. The best performing MLP model was able to predict BOD5 values at a very fast rate of 7.64 microseconds per sample.

Optimization of training parameters was performed using GA, and the results are shown in Table 7.

The performance for each method improved over their results in Table 7. This shows that the GA algorithm is successful in finding better hyperparameters for each method. Feature selection was then performed by training the models with the manually chosen set of features. Table 8 shows the results for each method.

Feature selection using SFS was then performed on each method instead and the results are as shown in Table 9.

The performance of models trained with a manually chosen set of features are similar to those with no feature selection at all. However, the models using SFS for feature selection outperform the manual feature selection models. For SFS, the best performing model is the MLP model. The MLP model has the highest R^2 value and the lowest MSE, relative MSE, MAE and relative MAE. The relative MSE and relative MAE values indicate that the predictions made by the model have on average around 15% of variation from the real values. Although the RF and SVR models have worse performance, the difference is not very big and the range of error in their predictions were just slightly bigger. This shows that the learning framework proposed in this project can train good models regardless of the ML method used.

3.2.2 Feature importance

Due to the higher performance of models trained using features selected from SFS, only the SFS models have their feature importance values calculated. The calculated PFI for each method are shown in Table 10.

Table 7 | Performance metrics using best parameter values obtained from genetic algorithm (GA) search

Method	R^2	MSE	Relative MSE	MAE	Relative MAE
RF	0.6954848762138784	0.5957618502060089	0.19834734986902758	0.5046604714760955	0.16801691324532536
MLP	0.722905352804904	0.5287630607917452	0.17604140275247915	0.48636654549522185	0.16192630550373913
SVR	0.6252840832045239	0.702609864530186	0.23392032331914037	0.5317462106019659	0.17703458460679725

Table 8 | Performance metrics using manual feature selection

Method	R^2	MSE	Relative MSE	MAE	Relative MAE
RF	0.6919636842006298	0.6002600256314178	0.19984493010946647	0.5082347766753114	0.16920690881762057
MLP	0.7295849974514856	0.5201503620448693	0.17317397179651509	0.4864683507089941	0.161960199574579
SVR	0.6328097363595326	0.6938239290089379	0.23099521654005215	0.5326070493699635	0.17732118417378495

Table 9 | Performance metrics using sequential feature selection

Method	R^2	MSE	Relative MSE	MAE	Relative MAE
RF	0.7285857116585056	0.5285300978302614	0.17596384225408207	0.49243842120958925	0.1639478187246049
MLP	0.7672791942775417	0.4326263301609101	0.14403454340237976	0.4611935091550118	0.15354543142710791
SVR	0.6961395292290989	0.5753817456143058	0.19156218943215148	0.5041845730061698	0.1678584720824952

Table 10 | Permutation feature importance (PFI) values for each machine learning method

Method	PFI
RF	['CODCr', 0.6945141372017399] ['Ammonium', 0.17846086262625951] ['Nitrite', 0.12572423698897228] ['Chlorophyll a', 0.08226019535914775] ['Total suspended solids', 0.0643038225783798] ['Aluminium and its compounds', 0.02668710048518753] ['Nitrate', 0.021305861897961232] ['Iron and its compounds', 0.02053630943988869] ['EDTA', 0.020335977148789642] ['Copper and its compounds', 0.018294481840751064] ['AOX', 0.017574975859233666] ['CODMn', 0.015718546117567667] ['Vanadium and its compounds', 0.014344399088356729] ['Benzo(g,h,i)perylene', 0.006265465364715306]
MLP	['CODCr', 0.36663533276211024] ['Nitrite', 0.27800461387879294] ['Ammonium', 0.19501086089403452] ['Chlorophyll a', 0.1161780562545599] ['Manganese and its compounds', 0.09679991530698473] ['Acid neutralizing capacity', 0.0898337615257988] ['Aluminium and its compounds', 0.07254567114086552] ['Total suspended solids', 0.06691121196590821] ['Iron and its compounds', 0.04799456384765577] ['AOX', 0.039585462115436565]
SVR	['CODCr', 0.569827248640375] ['Ammonium', 0.19634061530442914] ['Nitrite', 0.17398903278814423] ['Total suspended solids', 0.09762556630098815] ['Total dissolved solids', 0.08824618596319636] ['Iron and its compounds', 0.08607595403860682] ['Aluminium and its compounds', 0.05841903402958719] ['EDTA', 0.034719252921629064]

From the calculated PFI values, CODCr is found to be the most important feature across all models. Ammonium and nitrite were also present in the top 3 of every model together with CODCr, indicating that these 3 features are the most important in terms of correlation to the BOD5 values in water samples.

3.3. Field study: China lake dataset

3.3.1 Data preprocessing and model training

The dataset used here is a small private dataset consisting of water samples collected from two different lakes in Jiangsu Province of China over 2 years. The cleaned dataset consists of 120 data samples and 24 features including the target BOD5.

3.3.2 Optimization of training parameters and feature selection

The MLP method was used here as it gave the best performance in our Waterbase dataset. Parameter optimization was performed, and the best set of parameters are shown in Table 11.

Table 11 | Best parameter values obtained from genetic algorithm (GA) search

Method	Parameters	
MLP	hidden_layer_sizes	(12, 58, 72)
	solver	adam
	learning_rate_init	0.009507712972773643

Table 12 | List of chosen features using sequential feature selection (SFS)

Method	Features
MLP	Arsenic Dissolved Oxygen Mercury Anionic Surfactant Zinc Permanganate Index

After obtaining the optimal parameters, feature selection was performed using SFS. The list of chosen features are shown in Table 12.

The features selected through SFS here show a big difference to the features selected in the Waterbase dataset. The most important feature of CODCr in Waterbase is not selected at all when using the lake dataset.

3.3.3. Performance

The performance of the trained model is shown in Table 13.

The R^2 score for the China Lake dataset was lower compared to the Waterbase dataset. This implies that the model did not fit the data very well. The poor R^2 score is attributed to the smaller number of data samples in the China lake dataset.

Despite the low R^2 value, the relative MAE was similar to the models trained on the Waterbase dataset, with an average prediction error of around 15%. In addition, the relative MSE value was found to be approximately 9% which is an improvement from the Waterbase models. This shows that the learning framework detailed in this project can be used on a different water sample dataset that has much smaller number of samples and from a different type of water body.

3.3.4. Feature importance

The calculated feature importance is shown in Table 14. The calculated PFI values show that the most important features for the lake dataset are Permanganate Index, Dissolved Oxygen and Arsenic. This shows that there is a difference in relation between the different features and the BOD5 value of the water samples when compared to using the Waterbase dataset, possibly due to the different type of water body from which the samples were taken. The feature importance values may also be inaccurate due to the poorer fit of the model indicated by the R^2 score of 0.47013405024708144.

3.4. Discussion

Most of the other applications of machine learning to prediction of pollution indicators often focus on other more specific indicators in their prediction, such as concentrations of nitrate in Ransom *et al.* (2017) or Chl-a in Nieto *et al.* (2019) and Park *et al.* (2014). BOD5 prediction was made by Dogan *et al.* (2009) but only using ANN and the water samples were limited to only one

Table 13 | Performance metrics

Method	R^2	MSE	Relative MSE	MAE	Relative MAE	Model Training Time
MLP	0.47013405024708144	0.2157973944859386	0.0891724770603052	0.3491713086764017	0.14428566474231477	1.012s[a]

[a]Built with optimized hyperparameters and best set of features from feature selection.

Table 14 | Permutation feature importance (PFI) values

Method	PFI
MLP	['Permanganate Index', 0.8669850776151206] ['Dissolved oxygen', 0.6263821118251902] ['Arsenic', 0.6052708377547389] ['Zinc', 0.4632491490914945] ['Mercury', 0.41420375018514] ['Anionic Surfactant', 0.28063279274667663]

location. In contrast, this project explores the use of multiple different ML methods in the BOD5 prediction, and the dataset used in the project consists of water samples from a large number of different locations. The geographical diversity of the data and use of different ML methods enable the system to be more robust and accurate compared to more limited systems.

For further improvements to the system, a wider range of training parameters can be tested for each method to further improve the performance. For example, the current best model trained using MLP uses three hidden layers each with maximum 100 nodes, but a wider and deeper network may be able to give more accurate predictions. The parameters of each method can also be allowed to vary by a larger range. This however would require more computational resources for training each model and for the higher number of models needed to be trained to cover the wider search space.

Alternative methods of feature selection can also be performed. The current sequential feature selection (SFS) method chooses the best feature to include at each step, but another method such as testing randomly chosen subsets of all features may be able to find a set of features that gives better performance. However, this also requires much greater computational resources as the current SFS system has $O(N^2)$ time complexity for a dataset with N features, while testing every subset of all features will require $O(2^N)$ time.

4. CONCLUSIONS

This study employed multiple machine learning methods such as random forest (RF), support vector regression (SVR) and multilayer perceptron (MLP) to train a best model that can accurately predict the BOD5 values in water samples based on other physical and chemical properties of the water. A public dataset (Waterbase) was pre-processed prior to training on models with different ML methods. The training parameters were optimized using genetic algorithm (GA) and feature selection was done using a technique called sequential feature selection (SFS). MLP yielded the best performance with an R^2 score of 0.7672791942775417 together with relative MSE and relative MAE of around 15%. Feature importance calculations indicated that CODCr, ammonium and nitrate are features that highly correlate to BOD5. The proposed machine learning framework was also tested on a small private dataset consisting of water samples collected from two different lakes in Jiangsu Province of China. Compared to the model trained using the Waterbase dataset, the trained model was found to have a similar range of prediction error (around 15%), similar relative MAE (around 14%) and achieved about 6% better relative MSE. The experimental results show that the model is capable of predicting the BOD5 values with decent accuracy and is useful as an alternative to traditional ways of measurement. The system can be used in warning systems that allow water management agencies to respond quickly to any pollution events so that the damage to the environment can be reduced.

The proposed learning framework was tested on a publicly available dataset with relatively small sample size. In the future, we could consider obtaining a bigger and more complete dataset by implementing a standard set of physical and chemical properties alongside BOD5 measurement in water samples and consolidating results across multiple water management agencies. Such a dataset would allow the prediction models to be trained to a higher accuracy, and the performance can keep improving by training on any new measurement results as they come into the dataset. Moving forward, a wider range of training parameters can be tested to achieve better prediction performance.

DATA AVAILABILITY STATEMENT

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

REFERENCES

Arumugasamy, S., Chen, Z., Le, D. & Pakalapati, H. 2021 Comparison between artificial neural networks and support vector machine modeling for polycaprolactone synthesis via enzyme catalyzed polymerization. *Process Integration and Optimization for Sustainability* **24**, 5. (1).

Breiman, L. 2001 Random forests. *Machine Learning* **45**, 5–32.

Cai, X., Zhao, H., Shang, S., Zhou, Y., Deng, W., Chen, H. & Deng, W. 2021 An improved quantum-inspired cooperative co-evolution algorithm with muli-strategy and its application. *Expert Systems with Applications* **171**, 114629.

Chan, W., Le, D.-K., Chen, Z., Tan, J. & Chew, I. 2021 Resource allocation in multiple energy-integrated biorefinery using neuroevolution and mathematical optimization. *Process Integration and Optimization for Sustainability* (2021). https://doi.org/10.1007/s41660-020-00151-6.

Chislock, M. F., Doster, E., Zitomer, R. A. & Wilson, A. E. 2013 *Eutrophication: Causes, Consequences, and Controls in Aquatic Ecosystems*. (Nature Education) Retrieved April 24, 2021, from: https://www.nature.com/scitable/knowledge/library/eutrophication-causes-consequences-and-controls-in-aquatic-102364466/

Chou, J.-S., Ho, C.-C. & Hoang, H.-S. 2018 Determining quality of water in reservoir using machine learning. *Ecological Informatics* **44**, 57–75.

Cipullo, S., Snapir, B., Prpich, G., Campo, P. & Coulon, F. 2019 Prediction of bioavailability and toxicity of complex chemical mixtures through machine learning models. *Chemosphere* **215**, 388–395.

Delzer, G. & McKenzie, S. 2003 *Five-day Biochemical Oxygen Demand. In USGS TWRI Book 9*. USGS, Reston, Virginia, pp. BOD-1–BOD-21.

Deng, W., Xu, J., Gao, X.-Z. & Zhao, H. 2020 An enhanced MSIQDE algorithm with novel multiple strategies for global optimization problems. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 1–10. DOI: 10.1109/TSMC.2020.3030792.

Dogan, E., Sengorur, B. & Koklu, R. 2009 Modeling biological oxygen demand of the Melen River in Turkey using an artificial neural network technique. *Journal of Environmental Management* **90** (2), 1229–1235.

Drucker, H., Burges, C., Kaufman, L., Smola, A. & Vapnik, V. 1996 Support Vector Regression Machines. In: *Proceedings of the 9th International Conference on Neural Information Processing Systems*. MIT Press, Cambridge, MA, pp. 155–161.

European Environmental Agency 2020 *Waterbase – Water Quality ICM*. Retrieved From European Environment Agency, Copenhagen. See: https://www.eea.europa.eu/data-and-maps/data/waterbase-water-quality-icm.

Hastie, T., Tibshirani, R. & Friedman, J. 2009 *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, NY.

Huang, K., Xiao, Q., Meng, X., Geng, G., Wang, Y., Lyapustin, A., Gu, D. & Liu, Y. 2018 Predicting monthly high-resolution PM2.5 concentrations with random forest model in the North China Plain. *Environmental Pollution* **242** (A), 675–683.

Li, P., Hua, P., Gui, D., Niu, J., Pei, P., Zhang, J. & Krebs, P. 2020 A comparative analysis of artificial neural networks and wavelet hybrid approaches to long-term toxic heavy metal prediction. *Scientific Reports* **10**, 13439.

McKinney, W. 2010 Data structures for statistical computing in Python. In: *Proceedings of the 9th Python in Science Conference*, pp. 56–61.

Nieto, P. G., García-Gonzalo, E., Fernández, J. A. & Muñiz, C. D. 2019 Water eutrophication assessment relied on various machine learning techniques: a case study in the Englishmen Lake (Northern Spain). *Ecological Modelling* **404**, 91–102.

Park, Y., Cho, K. H., Park, J., Cha, S. M. & Kim, J. H. 2014 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* **502**, 31–41.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J. & Chintala, S. 2019 Pytorch: an imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems* **32**, 8024–8035.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, É. 2011 Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* **12** (85), 2825–2830.

Ransom, K. M., Nolan, B. T., Traum, J. A., Faunt, C. C., Bell, A. M., Gronberg, J. M., Wheeler, D. C., Rosecrans, C. Z., Jurgens, B., Schwarz, G. E., Belitz, K., Eberts, S. M., Kourakos, G. & Harter, T. 2017 A hybrid machine learning model to predict and visualize nitrate concentration throughout the Central Valley aquifer, California, USA. *Science of The Total Environment* **601–602**, 1160–1172.

Reshetilov, A., Arlyapov, V., Alferov, V. & Reshetilova, T. 2013 BOD Biosensors: Application of Novel Technologies and Prospects for the Development. In: Rinken, T. (ed.), *State of the Art in Biosensors – Environmental and Medical Applications*. IntechOpen.

Sapankevych, N. I. & Sankar, R. 2009 Time series prediction using support vector machines: a survey. *IEEE Computational Intelligence Magazine* **4** (2), 24–38.

Scikit-learn n.d. *Permutation feature importance*. Retrieved from scikit-learn. Available from: https://scikit-learn.org/stable/modules/permutation_importance.html

Shiklomanov, I. 1993 World fresh water resources. In: Gleick, P. H. (ed.), *Water in Crisis: A Guide to the World's Fresh Water Resources*. Oxford University Press, New York.

The pandas development team 2021 pandas-dev/pandas: Pandas 1.2.3. Zenodo. See: https://zenodo.org/record/4572994#.YPg49egzaUk.

United Nations, n.d. *World population prospects 2019*. Department of Economic and Social Affairs, Population Dynamics, United Nations, New York. See: https://population.un.org/wpp2019/.

United States Geological Survey (USGS) n.d. *Biological Oxygen Demand (BOD) and Water*. USGS, US Dept of the Interior, Washington DC. See: https://www.usgs.gov/special-topic/water-scienceschool/science/biological-oxygen-demand-bod-and-water.

Wang, W., Men, C. & Lu, W. 2008 Online prediction model based on support vector machine. *Neurocomputing* **71** (4–6), 550–558.

Yeganeh, B., Shafie Pour Motlagh, M., Rashidi, Y. & Kamalan, H. 2012 Prediction of CO concentrations based on a hybrid Partial Least Square and Support Vector Machine model. *Atmospheric Environment* **55**, 357–365.

Yuchi, W., Gombojav, E., Boldbaatar, B., Galsuren, J., Enkhmaa, S., Beejin, B., Naidan, G., Ochir, C., Legtseg, B., Byambaa, T., Barn, P., Henderson, S. B., Janes, C. R., Lanphear, B. P., McCandless, L. C., Takaro, T. K., Venners, S. A., Webster, G. M. & Allen, R. W. 2019 Evaluation of random forest regression and multiple linear regression for predicting indoor fine particulate matter concentrations in a highly polluted city. *Environmental Pollution* **245**, 746–753.

Zinatloo-Ajabshir, S., Sadat Morassaei, M., Amiri, O. & Salavati-Niasari, M. 2020 Green synthesis of dysprosium stannate nanoparticles using Ficus carica extract as photocatalyst for the degradation of organic pollutants under visible irradiation. *Ceramics International* **46** (5), 6095–6107.

First received 6 May 2021; accepted in revised form 18 June 2021. Available online 29 June 2021