


Prediction of biochemical oxygen demand with genetic algorithm-based support vector regression

Yan Zheng Liu^a and Zhiyuan Chen ^{b,*}

^a School of Computer Science, University of Nottingham, Nottingham NG8 1BB, UK

^b School of Computer Science, University of Nottingham Malaysia, Jln Broga, Semenyih, Selangor 43500, Malaysia

*Corresponding author. E-mail: zhiyuan.chen@nottingham.edu.my

 ZC, 0000-0002-4915-1593

ABSTRACT

Five-day biochemical oxygen demand (BOD5) is a vital wastewater contamination strength indicator. The process of measuring BOD5 is to measure the mass of molecular oxygen consumed in 1 L of water at 20 °C over a 5-day incubation period. It is a time-consuming process and often too late for water management agencies to make a timely reaction if the result of measurement shows a water body is seriously polluted. Biosensors can simplify the process of BOD5 measurement; however, the measurement results often deviate significantly from the measured BOD5 values. The main aim of this research is to identify a machine learning model, which could predict BOD5 value from historical data and make it easier to detect water pollution in advance and timely adopt treatment measures. Three machine learning techniques, linear regression, support vector regression (SVR) and multi-layer perceptron (MLP) and two optimization processes have been studied in this research. Four main steps, preprocessing (one-time only), model training, model evaluation (testing) and analysis have been implemented in the experiments. With three feature selection strategies, the results of the experiment showed that SVR with genetic algorithm (GA) optimizer achieved the best performance with R^2 of 0.694 and the lowest MAE of 0.109.

Key words: biochemical oxygen demand prediction, machine learning model, multi-layer perceptron, support vector regression, wastewater quality indicator

HIGHLIGHTS

- Genetic algorithm-based support vector regression has been proposed to predict the biochemical oxygen demand values from simple variables that are easily measured.
- Comparison experiments have been conducted among three popular machine learning techniques with two optimization processes.
- The best model SVR with genetic algorithm optimizer achieved the best performance with R^2 of 0.694 and the lowest MAE of 0.109.

ABBREVIATIONS

Cu	cuprum
Zn	zinc
Se	selenium
As	arsenic
Hg	mercury
Cd	cadmium
Cr	chromium
Pb	plumbum
NH ₄ -N	ammonia-nitrogen
P	total phosphorus
N	total nitrogen
pH	acidity/alkalinity
DO	dissolved oxygen
PMI	permanganate index
COD	chemical oxygen demand

This is an Open Access article distributed under the terms of the Creative Commons Attribution Licence (CC BY 4.0), which permits copying, adaptation and redistribution, provided the original work is properly cited (<http://creativecommons.org/licenses/by/4.0/>).

PFCs	perfluorochemical
CN	cyanide
VP	volatile phenol
AS	anionic surfactant
SOx	sulfide
BOD5	5-day biochemical oxygen demand

INTRODUCTION

Water is an essential resource for all creatures and the management of water quality is of great importance. Among the properties of water bodies, 5-day biochemical oxygen demand (BOD5) and chemical oxygen demand (COD) are highly valued in environmental projects (Verma & Singh 2013; Najafzadeh & Ghaemi 2019). However, the measuring process of BOD5 and COD is cumbersome and time-consuming. The standard method of measuring BOD5 is to measure the mass of molecular oxygen consumed per liter of water in 5 days at a temperature of 20 °C (Great Britain, Parliament and House of Commons 1912). Measuring BOD5 in a big natural water system is more complicated since the test needs to be applied to a mess of water samples (Delzer & McKenzie 2003). The accuracy of the BOD5 measure is mainly based on the bacterial growth in the sample water; therefore, it has high uncertainty. According to Basant *et al.* (2010), laboratory test results are flawed because they do not account for oxygen taken up by organisms. Moreover, the existence of various inorganic interfering radicals sometimes distorts the results of the experiment on COD (Verma & Singh 2013). To avoid such difficulties and errors, attempts have been made to predict the BOD5 and COD values from simple variables that are easily measured and with high accuracy, such as temperature, pH, DO, TSS, etc. Linear regression (LR), artificial neural networks (ANNs), random forest (RF) and support vector regression (SVR) were applied in recent research work (Verma & Singh 2013; Najafzadeh & Ghaemi 2019; Ooi *et al.* 2021) and showed good performance.

Early in the 1790s, scientists and researchers started to predict BOD5 using mathematical methods (Padgett & Papadopoulos 1979). That was, in fact, a process of guessing the random phenomenon with mathematical equations. In recent years, machine learning methods have been proposed to predict BOD5. In 2013, Verma & Singh (2013) presented a machine learning method, an ANN to predict BOD and COD. The result showed a training error rate of 0.0611. In 2016, echo state network (ESN) combined with particle swarm optimization (PSO) was proposed (Qiao *et al.* 2016). Compared to the standard ESN, the proposed PSO-ESN eliminated the poor solution caused by measurement errors. It also significantly outperformed another machine learning method, AMGA-RBF, and achieved a training error rate of 0.0339 and a testing error rate of 0.0483 separately. Then in 2019, multivariate adaptive regression spline (MARS) and least square-support vector machine (LS-SVM) were used to predict the indices of BOD5 and COD (Najafzadeh & Ghaemi 2019). The results showed that both methods performed better than ANNs in the prediction of BOD5. In Cipullo *et al.*'s (2019) research, several machine learning methods were compared, including RF, SVR and multi-layer perceptron (MLP). The results showed that MLP got the best performance in predicting BOD5, with an R^2 score of 0.7672, and relative MSE and relative MAE were both approximately 15%. The machine learning methods were enhanced with the genetic algorithm (GA) and the sequential feature selection (SFS) method.

Machine learning methods have also been applied to other aspects of water quality prediction. LR, support vector machines, ANNs, classification and regression trees were compared with the baseline and some integration scenarios in research work on determining the quality of water in reservoirs (Chou *et al.* 2018). The in-depth analysis and detailed comparison demonstrated that the ensemble ANNs model with a tiering method was more accurate than other models. In Cipullo *et al.*'s (2019) research, ANN and RF were used to predict temporal changes in the bio-availability of complex chemical mixtures in contaminated soils. In Li *et al.*'s (2020) research, a back-propagation neural network (BPNN) and a nonlinear autoregressive network with exogenous inputs (NARX) were used for the long-term prediction of toxic metals. These two methods were integrated with the wavelet transform method and the results showed that the wavelet transform can improve long-term concentration predictions.

In most of the studies, the numbers of input features for the prediction of BOD5 were within 4–9 (Verma & Singh 2013; Najafzadeh & Ghaemi 2019). Field studies conducted on different rivers indicated that there was no general relationship between the BOD value and an acceptable level of accuracy (Najafzadeh & Ghaemi 2019). In Abdalrahman and Refaie's research (Alsulaili & Refaie 2020), the correlation coefficient between each pair of single water parameters were small and the performance of prediction BOD using a combination of less than four input features was not very good (with an R^2 score of 0.3440). In recent years, researchers started considering the importance of different features for certain water

quality indices. Particularly, the SFS feature selection method (Ooi *et al.* 2021) and permutation feature importance (PFI) method for feature importance ranking have demonstrated good performance.

In this research, 23 water indexes are investigated while previous research works normally use less than 9 input features to predict BOD5. These input features have also been split into different chemical categories to find a relationship between different water index types and BOD5. Furthermore, the SFS feature selection process has been integrated with different machine learning models to find the most suitable features set that can predict BOD5 more accurately.

MATERIAL AND METHODS

Back-propagation artificial neural network (BPNN), PSO-ESN, MLR, LS-SVM and MARS algorithms have been studied in BOD5 prediction. Among these machine learning methods, LR, MLP and SVR are widely used due to their good performance. In this research, we study these three algorithms with two optimization methods, GA and grid search optimization. Manual feature selection based on chemical categories and sequential feature selector (SFS) have been implemented to find the features that are most related to BOD5.

LR method

LR is a well-known and well-understood algorithm in statistics and machine learning. It can be used in prediction by establishing the relationship between independent and dependent variables (Chou *et al.* 2016). LR has been proven to have a good performance in the prediction of water quality. In Chou *et al.* (2018), ordinary least squares LR was applied, which was a version of LR that fitted a linear model with coefficients $w = (w_1, \dots, w_p)$ to minimize the residual sum of squares between the observed and the predicted values of the target column (BOD5).

The mathematical expression of LR:

The simple regression problem with just one input value and one output value can be shown in the following equation.

$$y = B_0 + B_1 * x \quad (1)$$

B_0 is the scale factor and B_1 is the bias coefficient which gives the line an additional degree of freedom. Problems with multiple input values can be represented by Equation (2). \vec{x} represents an n -dimensional input vector, and B_1 represents the n -scale factors for the n -dimensional input vector.

$$y = B_0 + B_1 * \vec{x} \quad (2)$$

Learning techniques for LR include:

- Ordinary least squares. This is the sum of all the squared values of the distances from each data point to the regression line. The learning process aims to minimize this quantity. It is not easy to implement the ordinary least squares, because it needs a large memory to fit the data and compare all the ordinary least squares.
- Gradient descent. In this process, the coefficient of B_1 is updated in the direction towards minimizing the sum of the squared errors step by step until the minimum sum squared error is achieved or no further improvement is possible. It is useful and easy to implement for problems with large datasets.
- Regularization. The regularization method is the upgrade of the ordinary least squares. It not only minimizes the sum of the squared errors but also reduces the complexity of the model. This includes lasso regression, which minimizes the absolute sum of the coefficient; and ridge regression, which minimizes the squared absolute sum of the coefficient (Bhattacharyya 2018).

Multi-layer perceptron

MLP is a powerful neural network that works in both prediction and classification tasks. It has been analyzed in several studies and performs well on a wide variety of problems. MLP consists of an input layer, multiple hidden layers and an output layer. Each layer is composed of nodes. The structure of three-layer MLP neural networks is shown in Figure 1. The computation takes place in each node in hidden layers and an output layer, which consists of two parts as shown in Figure 2. The first part is the sum of all nodes from the upper layer with weights. The second part is called the activation function $f(a)$. The activation function is usually tanh or the logistic sigmoid function. Fully connected three-layer MLPs

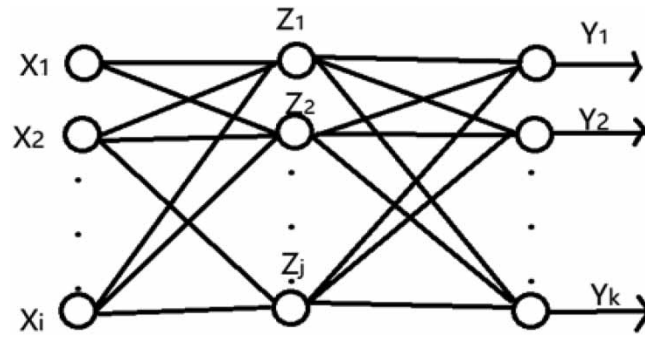


Figure 1 | MLP neural network.

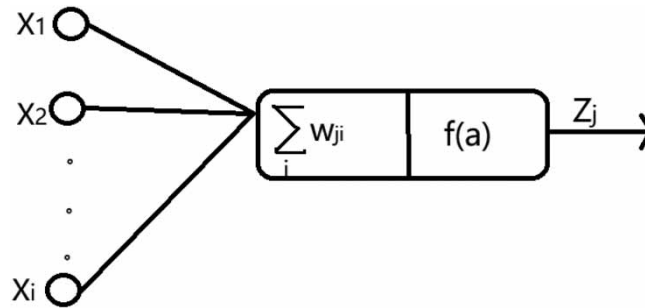


Figure 2 | Computation of each node in MLP.

are considered in this paper since these networks have been shown to approximate any continuous function (Cybenko 1989; Hornik *et al.* 1989, 1990).

Support vector machine

SVM is a well-known machine learning method. SVM for regression is called SVR. The difference between SVR and simple regression methods is that we try to fit the error with a certain threshold in SVR and try to minimize the error rate with simple regression.

The theory of SVMs for nonlinear classification problems is shown in Figure 3. The target of the original SVM for regression problems is to find a hyperplane that can classify points into two classes (red crosses and green circles) with the max margin.

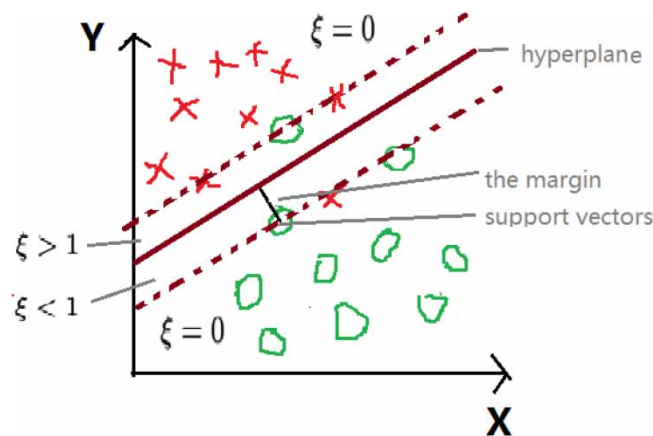


Figure 3 | SVMs for binary classification problem.

The data points in Figure 3 can be expressed as Equation (3) and the function of the hyperplane can be represented as Equation (4).

$$D = (x_1, y_1), (x_2, y_2), \dots, (x_m, y_m) \quad (3)$$

$$H(x) = w * x - b \quad (4)$$

The perpendicular distance between the hyperplane and the closest data points is the margin, and the closest data points are support vectors. The equation of margin is $\text{margin} = 1/w^2$. The values of $H(x)$ are 1 and -1 at the margin. For data points with positive values, the classification results are positive. Data points with negative values will be classified as negative. If the data points are linearly separable, the data points should all be in the area where $F(x) > 1$ or $F(x) < -1$. However, data points are not always separable, such as the unassigned green circle and the two red crosses. To deal with such cases, the soft margin SVM allows classification errors while still maximizing the margin by introducing slack variables ξ_i . ξ_i measures the degree of misclassification. Then the objective function for maximizing the margin is to minimize Equation (5). For nonlinearly separable datasets, sometimes hyperplanes cannot be found. To address this issue, a kernel trick is used to transform data points into a higher-dimensional space where a hyperplane exists. The theory and the kernel tricks of SVR are almost the same as SVM for regression. The difference is that the slack variable ξ is decomposed into two variables, ξ and ξ' in SVR. In this study, three different kernel tricks are applied.

$$Q(w, b, \xi_i) = \frac{1}{2} + C \sum \xi_i \quad (5)$$

Optimization methods

Hyper-parameters are model configuration arguments that are set manually to guide the learning process. The effect of a hyper-parameter is known generally, but the best hyper-parameters for a given dataset are not clear. So the step to find the best hyper-parameters is an important procedure in the learning process. This procedure is called hyper-parameter optimization or hyper-parameter search.

• Grid search

One of the simplest optimization methods is grid search. It finds the best fitting solution by searching a defined search space. The search space is a collection of n -dimensional vectors, each dimension represents a hyper-parameter and the dimension scale is the possible value of the hyper-parameter, such as real value, integral value or categorical value. Each vector in the grid search is a fitting solution for the learning process, which is a combination of hyper-parameters.

• Evolutionary optimization

Evolutionary optimization is a type of advanced optimization method. The genetic algorithm (GA) is an evolutionary optimization method. It is inspired by genetics. Chromosomes in genetic algorithms represent a set of variables. Each chromosome represents a solution to the given problem. The variables can be Boolean, Integer, Floating or String or any combination of the above. Each generation includes a set of different chromosomes.

The workflow of GA optimization is shown in Figure 4. The main steps are:

1. Generate initial population. The first generation is generated randomly.
2. Calculate the evaluation metrics, such as R^2 , RMSE, and MAE.
3. Check for termination. The algorithm is possible to stop according to the value of the cost function, the maximum number of iterations or by a stall generation. The maximum number of iterations is the most commonly used and it guarantees that the algorithm will produce results within a certain time, regardless of whether it has reached the extremum or not.
4. Selection. Among all individuals in the current population, those with better performance can be selected. The elite genes of these individuals are passed on to the next generation. The offspring is created by recombining the genes of the chosen individuals from the last generation. The selection process guarantees the offspring individuals could inherit the best

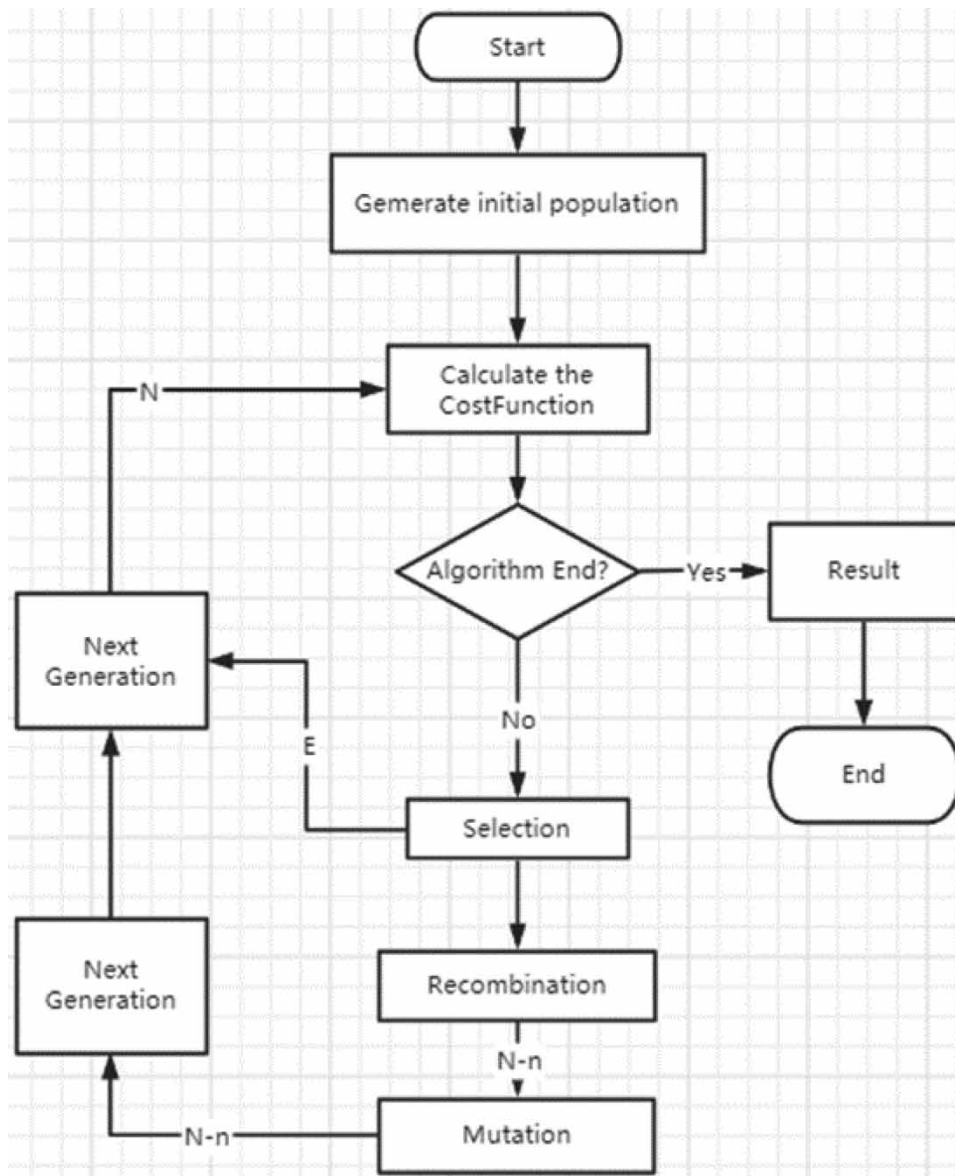


Figure 4 | GA optimization workflow.

possible combination of genes from their parents' generation and that the best value of the cost function (maximum or minimum) cannot get worse.

5. Mutation. Genes from the selection process are recombined with random changes. These random changes are from those that passed the crossover and mutation. Mutation is important and has impacts on experiment results (De Falco *et al.* 2002).

Sequential feature selection

An SFS algorithm is a group of greedy search algorithms. The motivation behind the feature selection algorithm is to automatically select the most relevant feature subset to the problem. The difference between SFS and feature reduction methods (such as PCA) is whether the importance of the original features can be shown in the result of the function. Therefore, to find the most related features to BOD5, sequence feature selection is a better algorithm. In this study, two sequential feature selectors have been applied: are sequential forward selection (SFS) and sequential forward floating selection (SFFS).

• Sequential forward selection

The input of SFS is the whole-dimensional feature. The output is a subset of that feature set. The main steps of SFS are:

1. Initialize an empty feature set $X_k = \Phi$.
2. Find a feature x^+ that maximized the evaluation function.
3. Add x^+ to X_k .
4. Repeat steps 2,3 until the desired number of features is reached.

Sequential forward floating selection

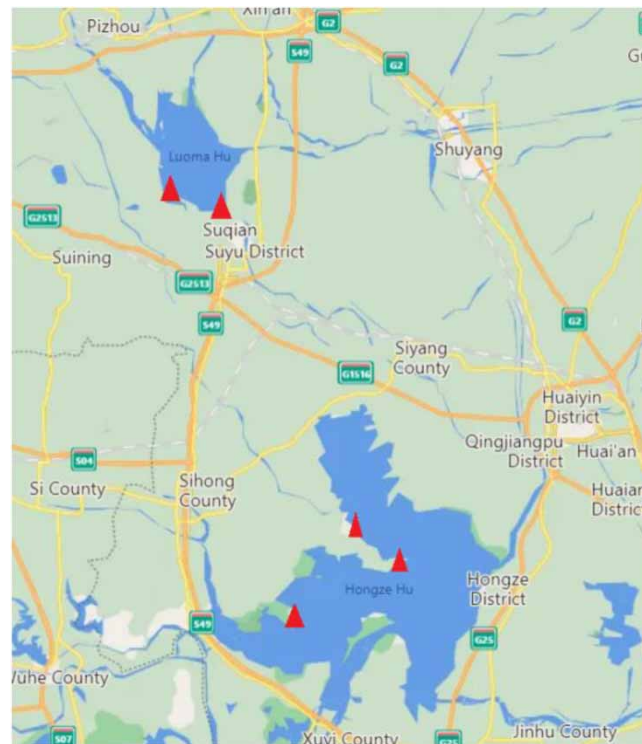
The input of SFFS is the whole-dimensional feature set. The output is a subset of that feature set. SFFS is an enhanced method of SFS. The main steps of SFS are:

1. Initialize an empty feature set $X_k = \Phi$.
2. Find a feature x^+ that maximized the evaluation function.
3. Add x^+ to X_k .
4. Find a feature y^+ that maximized the evaluation function. If y^+ cannot be found, remove x^+ and repeat step 2.
5. Repeat steps 2,3,4 until the desired number of features is reached.

Data preparation

The dataset used in this study was collected from two inland lakes in Suqian, Jiangsu Province, China, called Hongze Hu and Luoma Hu. The data were recorded monthly from 2018 to 2019. The geographical locations of the five test sites are shown in Figure 5. The parameters of water quality in the dataset include both simple chemical variables and complex variables, which are difficult to measure. The simple variables include pH, mercury concentrations, lead concentrations, etc. The complex variables are BOD5, and COD. Details of all variables could be found in Appendix 1.

The features are used as inputs to the prediction model in this study, which include all the water quality indexes we can access. In the first set of experiments, all features are selected to train the three machine learning models. And manually



Test Sites ▲

Figure 5 | Geographical locations of test sites.

selected three types of features, which are heavy metal pollution indexes, mineral content and composition indexes have been fed into the three machine learning models in the second set of the experiment. Finally, the feature selection algorithm SFS is applied. By comparing the prediction results of machine learning models trained on all features, manually selected features and features selected by SFS, we identify an optimal model, which is a combination of one machine learning model and a suitable feature selection method.

Experiment design

All experiments include four main steps: preprocessing (one-time only), model training, model evaluation (testing) and analysis. The overall experiment design and workflow are shown in Figure 6. Data preprocessing includes removing repeatability, filling in missing data, changing error format and deleting or replacing abnormal data points. The model training process begins with feature selection. Three sets of experiments are taken on different feature selection strategies. The first set of experiments is trained on all features except for the first three features, which are TestSite, Year and Month. The second set of experiments is trained on different chemical categories. The third set of experiments is with the SFS strategy. For each experiment, the dataset is split into a training dataset and a testing dataset. Then the training dataset is split into k subsets, and the model is trained and fitted k times on the k subsets in each experiment to find the best hyper-parameters. This is the k -fold cross-validation method, which can help to avoid over-fitting problems because the fitting results are tested k times on new data. Then the model with the best hyper-parameters is applied to the test dataset to evaluate the performance. The evaluation operation follows each training process and results are recorded. The data analysis runs through the entire experimental process. After each experiment, basic analysis is required to adjust the next step of the experiments. The overall analysis is done after all the experiments are completed.

Experiment settings

The default values and search space (value range for hyper-parameter optimization) could be found in Table 1. The model with the best R^2 score is selected in the training process. The trained model is tested and the predicted values for the target property (BOD5 values) are obtained. The predicted values can then be used to calculate the evaluation metrics.

The evaluation metrics in this research include:

R^2 score: Regression score function. It represents the goodness of fitting. The best possible value is 1.0, and it can also be negative (which means the models are arbitrarily worse).

RMSE: Root mean square error is widely used for evaluating the performance of a model in predicting quantitative data. It involves giving more weight to the larger error value using the quadratic formula. But if the data scale is smaller than 1, this does not work.

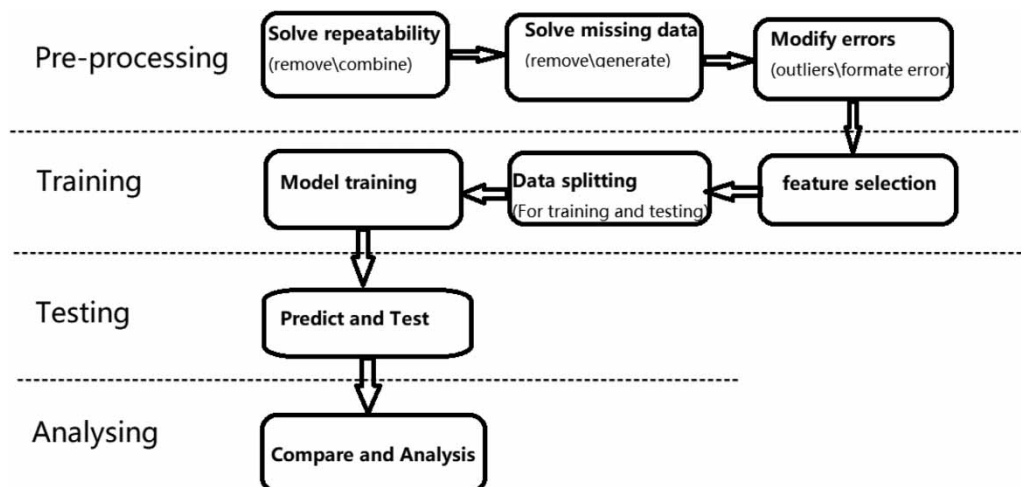


Figure 6 | Experiment process.

Table 1 | Experiment parameters

Method	Parameter	Default	Search space
LR	normalize	True	True, False
	positive	False True	True, False True, False
	fit_intercept	rbf	'linear', 'poly', 'rbf', 'sigmoid'
MLP	kernel	3	37
	degree epsilon	0.1 1.0	0.1 1 0.1 2
	C	100	20 80
	hidden_layer_sizes	relu	'identity', 'logistic', 'tanh', 'relu'
SVR	activation	'adam' 0.0001	'sgd', 'adam'
	solver alpha	'constant'	0.0001 1
	learning_rate	0.001	constant', 'invscaling', 'adaptive'
	learning_rate_init	Default	0.001 0.08
	Parameter	True	Search Space
	normalize	False True	True, False
	positive	rbf	True, False True, False

MAE: Mean absolute error is the average of all absolute error values. The difference to RMSE is that it does not give more weight to larger error values. For data on a small scale, it is a better metric than RMSE. In our experiments, data are scaled to the range $(-1,1)$. Hence, MAE is of higher reference value than RMSE.

MAPE: Mean absolute percentage error is the sum of the absolute ratio of the error values and the true values. Both MAPE and MAE are widely used metrics for evaluating errors in problems that have small-scale datasets and work best when the dataset has no extremes and no zeros.

RESULTS AND DISCUSSION

Results on all features

The results of the experiments with all features are shown in [Table 2](#). The performances of SVR and MLP are very similar and are much better than LR. As discussed in the evaluation metrics section, an R^2 score and MAE have been calculated in this study due to higher reference values. MLP has a slightly higher R^2 score of 0.637 and a lower MAE of 0.129 than SVR (0.626 and 0.133). Therefore, in this set of experiments, the MLP model achieves the best performance.

The SVR and MLP models perform much worse than LR before optimization. After implementing the grid search optimization, the experimental results have changed slightly. And with the implementation of GA optimization, the performance of SVR and MLP improved significantly. For R^2 values, SVR increased from 0.198 to 0.626, and MLP increases from 0.143 to 0.637. The results show that GA optimization has remarkable advantages over grid search. The best parameters for MLP and SVR with the GA search are shown in [Table 3](#).

Results on different feature types

All features have been grouped into three categories, including heavy metal, mineral and a composite index.

Table 2 | Experiments on all features

Methods		R^2 score	RMSE	MAE	MAPE
LR	default params	0.288	0.568	0.323	0.226
	grid search	0.258	0.580	0.337	0.226
SVR	default params	0.198	0.583	0.339	0.323
	grid search	0.361	0.570	0.325	0.298
	GA search	0.626	0.365	0.133	0.130
MLP	default params	0.143	0.552	0.304	0.198
	grid search	0.556	0.397	0.158	0.150
	GA search	0.637	0.359	0.129	0.141

Table 3 | Best parameters of MLP and SVR model (GA search)

Methods	Parameters	Value
SVR	kernel	rbf
	degree	5
	epsilon	1.192
	C	1.51
MLP	hidden_layer_sizes	47
	activation	logistic
	solver	'adam'
	alpha	0.339
	learning_rate	Adaptive
	learning_rate_init	0.001

- Heavy metal: Cu, Zn, Se, As, Hg, Cd, Cr, Pb.
- Mineral: NH₄-N (ammonia-nitrogen), P, N.
- Composite index: pH, DO (dissolved oxygen), PMI (permanganate index), COD, PFCs (perfluorochemical), CN (cyanide), VP (volatile phenol), AS (anionic surfactant), Petro, SO_x (sulfide), chlorophyll a, alpha

Evaluation results of different types of features are shown in Table 4. Performances for mineral input are worst. For R^2 score, all methods give negative values, which means the models fit poorly. Performance for the composite index is the best. The highest R^2 score for the composite index is 0.656, which is almost the same level as the set of experiments with all parameters. That proves that minerals in water have fewer effects on BOD₅, composite has a greater effect and heavy metal is in the middle.

Results on sequential feature selector

Sequential feature selector (SFS) is applied to select features that are most related to BOD₅. Figure 7 shows the fitness of models when the number of selected features increases. For LR, performance increases quickly when the number of selected features are smaller than 12. This means the improvement of performance is contributed by the first 12 selected features. For SVR and MLP, the k_{features} values are 13 and 17 on which points performance stops increasing. So, the selected features are used as input features to train LR, SVR and MLP models. The features selected are shown in Table 5. Features selected by SFS based on the three models are DO, PMI, COD, Cu, Zn, Se, VP, Petro and Alpha.

The evaluation results of models based on SFS are shown in Table 6. Compared to data in Table 2, only the performance of SVR has been improved from 0.626 to 0.694 in R^2 score and from 0.133 to 0.109 in MAE. The performances of LR and MLP become worse than all features. That shows the SFS based on MLP and LR models do not work very well.

CONCLUSION

In this research, various machine learning strategies with different optimization processes for the prediction of BOD₅ value from historical data have been studied. The results show that GA-based SVR with sequential feature selector achieves the best

Table 4 | Experiment results for different feature types

Feature	Method	R^2	Score	RMSE	MAE
Heavy metal	LR	0.205	0.751	0.564	0.366
	MLP	-0.604	0.755	0.570	0.256
	SVR	0.214	0.528	0.279	0.181
Mineral	LR	-0.058	0.773	0.598	0.416
	MLP	-0.080	0.619	0.384	0.226
	SVR	-0.143	0.637	0.406	0.221
Composite index	LR	0.272	0.540	0.291	0.193
	MLP	0.656	0.350	0.122	0.123
	SVR	-0.450	0.730	0.532	0.252

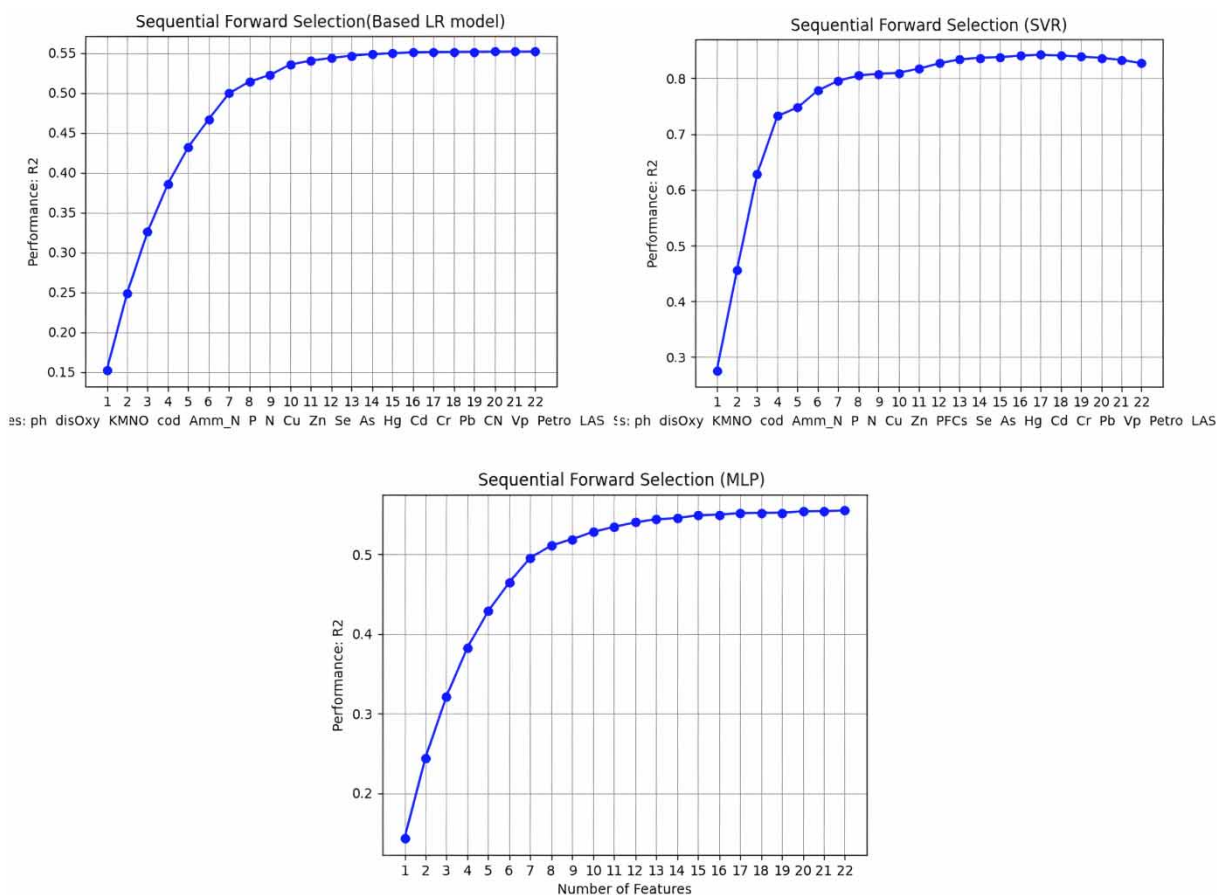


Figure 7 | Performance of feature selection.

Table 5 | Selected features

Methods	Features
LR	DO, PMI, COD, Cu, Zn, Se, Cd, CN, VP, Petro, AS, Alpha
SVR	DO, PMI, COD, P, Cu, Zn, PFCs, Se, Hg, VP, Petro, SOx, Alpha
MLP	pH, DO, PMI, COD, NH_N, P, Cu, Zn, PFCs, Se, As, Cd, CN, Vp, Petro, LAS, Alpha

Table 6 | Experiments on sequential feature selector

Methods	R ² score	RMSE	MAE	MAPE
LR	0.534	0.407	0.165	0.153
SVR	0.694	0.329	0.109	0.114
MLP	0.409	0.458	0.210	0.159

performance with R^2 of 0.694 and the lowest MAE of 0.109. For experiments with all features, MLP is a little bit better than SVR, but when combined with SFS, SVR shows better performance. Grid search and the GA algorithm have been applied to optimize the machine learning models and from the experimental results, GA optimization presents remarkable advantages over grid search.

It is interesting to find MLP combined with SFS did not work very well. But MLP without SFS performs at a similar level to SVR. Hence, one of the future research directions could be in identifying proper feature selection methods for MLP. Meanwhile, it might be worth considering other optimization methods, such as tiering method and the least squares method. Lastly, the data studied in this research are in a relatively small dataset with only 2 years of data record. Testing the optimal model with a larger dataset might need to be conducted in the future.

ACKNOWLEDGEMENT

We would like to thank Prof. Jian Cui from Province and Chinese Academy of Sciences, Institute of Botany, Nanjing, China for providing the research dataset.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

- Alsulaili, A. & Refaie, A. 2020 Artificial neural net-work modeling approach for the prediction of five-day biological oxygen demand and wastewater treatment plant performance. *Water Supply* **21** (5), 186121877. doi:10.2166/ws.2020.199.
- Basant, N., Gupta, S., Malik, A. & Singh, K. P. 2010 Linear and nonlinear modeling for simultaneous prediction of dissolved oxygen and biochemical oxygen demand of the surface water – a case study. *Chemometrics and Intelligent Laboratory Systems* **104** (2), 1722180.
- Bhattacharyya, S. 2018 Ridge and Lasso Regression: L1 and L2 Regularization. Available from: <https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcfb0b> (accessed 26 September 2018).
- Chou, J.-S., Chong, W. K. & Bui, D.-K. 2016 Nature-inspired metaheuristic regression system: programming and implementation for civil engineering applications. *Journal of Computing in Civil Engineering* **30**, 1–17.
- Chou, J.-S., Ho, C.-C. & Hoang, H.-S. 2018 Determining quality of water in reservoir using machine learning. *Ecological Informatics* **44**, 57275. <https://doi.org/10.1016/j.ecoinf.2018.01.005>.
- 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**, 3882395. <https://doi.org/10.1016/j.chemosphere.2018.10.056>.
- Cybenko, G. 1989 Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems* **2** (4), 3032314. issn:1435-568X. doi:10.1007/BF02551274.
- De Falco, I., Della Cioppa, A. & Tarantino, E. 2002 Mutation-based genetic algorithm: performance evaluation. *Applied Soft Computing* **1** (4), 2852299. [https://doi.org/10.1016/S1568-4946\(02\)00021-2](https://doi.org/10.1016/S1568-4946(02)00021-2).
- Delzer, G. C. & McKenzie, S. W. 2003 Five-day biochemical oxygen demand. *US Geological Survey Techniques of Water-Resources Investigations, book 9* (G.C. Delzer and S.W. McKenzie, eds.). USGS, Reston, VA.
- Hornik, K., Stinchcombe, M. & White, H. 1989 Multilayer feedforward networks are universal approximators. *Neural Networks* **2** (5), 3592366.
- Hornik, K., Stinchcombe, M. & White, H. 1990 Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks* **3** (5), 5512560. [https://doi.org/10.1016/0893206080\(90\)90005-6](https://doi.org/10.1016/0893206080(90)90005-6).
- 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** (1), 13439. doi:10.1038/s41598-020-70438-8.
- Najafzadeh, M. & Ghaemi, A. 2019 Prediction of the five-day biochemical oxygen demand and chemical oxygen demand in natural streams using machine learning methods. *Environmental Monitoring and Assessment* **191** (6), 380. doi:10.1007/s10661-019-7446-8.
- Ooi, K. S., Chen, Z., Poh, P. E. & Cui, J. 2021 BOD5 prediction using machine learning methods. *Water Supply*, ws2021202. doi:10.2166/ws.2021.202.
- Padgett, W. J. & Papadopoulos, A. S. 1979 Stochastic models for prediction of BOD and DO in streams. *Ecological Modelling* **6** (4), 2892303.
- Qiao, J. F., Li, R. X., Chai, W. & Han, H. G. 2016 Prediction of BOD based on PSO-ESN neural network. *Control Engineering of China* **23** (4), 4632467.
- Great Britain, Parliament and House of Commons. 1912 *Royal Commission on Sewage Disposal. Eighth Report of the Commissioners*. House of Commons, London.
- Verma, A. K. & Singh, T. N. 2013 Prediction of water quality from simple field parameters. *Environmental Earth Sciences* **69** (3), 8212829.

First received 7 July 2022; accepted in revised form 13 April 2023. Available online 24 April 2023